A learning theory for quantum photonic processors and beyond

Matteo Rosati

Dipartimento di Ingegneria Civile, Informatica e delle Tecnologie Aeronautiche, Università Roma Tre, Via Vito Volterra 62, Rome, I-00146, Italy

We consider the tasks of learning quantum states, measurements and channels generated by continuous-variable (CV) quantum circuits. This family of circuits is suited to describe optical quantum technologies and in particular it includes state-of-the-art photonic processors capable of showing quantum advantage. We define classes of functions that map classical variables, encoded into the CV circuit parameters, to outcome probabilities evaluated on those circuits. We then establish efficient learnability guarantees for such classes, by computing bounds on their pseudo-dimension or covering numbers, showing that CV quantum circuits can be learned with a sample complexity that scales polynomially with the circuit's size, i.e., the number of modes. Our results show that CV circuits can be trained efficiently using a number of training samples that, unlike their finite-dimensional counterpart, does not scale with the circuit depth.

Contents

1 Introduction

The last years have seen an incredible advancement in hardware solutions for quantum technologies. In particular, the recent demonstration of a quantum computational advantage via photonic circuits [1, 2] finally paves the way for the realization of full-fledged quantum information processing with light, a solution that bears intrinsic advantages with respect to other platforms, in terms of scalability, robustness and deployability [3, 4, 5]. At the same time, the increased control of infinite-dimensional quantum states in several other platforms, such as cavity [6, 7] or mechanical resonators [8], is pushing the boundaries of continuous-variable (CV) quantum information processing beyond photonics. Finally, the increased interplay between qubit and CV platforms [9, 10] spurs the interest into the development of quantum error correction codes [11, 12, 13] and provides an alternative to more standard approaches for quantum technologies. A combination of the aforementioned events thus marks a renewed surge of interest into CV information processing.

From a theoretical perspective, the characterization of the information-processing capabilities of quantum devices has been recently subject to a paradigm shift, thanks to the introduction of statistical learning techniques [14, 15, 16, 17, 18], which underly the success of classical machine learning [19, 20, 21]. In this approach, one recognizes that a successful use of quantum devices often requires two ingredients: (i) the estimation of quantities of interest about the quantum states or processes running in the device; (ii) the optimization of the device's parameter setup based on the estimated data, in order to maximize the device's performance in a specific task. Therefore, in this setting one can characterize the information-processing capabilities of a quantum device by quantifying its sample complexity (SC), i.e., the number of repetitions of an experiment that are required to estimate specific quantities and/or find a good device setup for carrying out a certain task.

To the best of our knowledge, research in this direction has focused almost exclusively on finite-dimensional systems so far, showing that, surprisingly, one can learn many properties of a d-dimensional quantum state using a number of samples that scales only as $\log d$, i.e., *linearly in the system size*. Unfortunately, most of these results cannot make any prediction for the case $d = \infty$, leaving open a quite looming question: can CV quantum systems be learned efficiently?

In particular, Aaronson [14] was the first to prove that it is possible to *learn efficiently* an approximation of an unknown quantum state by estimating the expected values of binary observables, with a SC linear in the number of qubits; instead, the SC turns out to be exponential in the number of qubits for the dual problem of learning an unknown quantum measurement [15], and a similar scaling is believed to hold for learning general quantum processes [14, 17]. By restricting to the class of polynomial-depth quantum circuits, Caro and Datta [17] and Popescu [22] showed that efficient process learnability

can be recovered. More recently, Caro et al. [18] computed the SC of circuits when employed as encoders of classical information into the quantum device.

The only known estimates for CV systems were introduced by [23], which provide energy-dependent SC bounds for *online learning* of Gaussian states, a task which has in general larger SC than the ones treated here. Furthermore, after this article was published on a preprint server, three more works appeared providing methods to perform shadow tomography of CV states [24, 25]. This is a task related with the one we consider here, though more difficult, in the sense that it requires the estimation of M observables with high precision, whereas here we want to obtain estimates which are precise with high probability, on average with respect to a distribution over observables (see [26] for a detailed description of the differences between these two settings). Hence one can expect that the sample complexity obtained in these further works is larger than the one given in the present manuscript. Observe also that each of these works quantifies non-Gaussianity in a different way, suitable in each case to the specific methods employed to bound the sample complexity. A similar reasoning applies to the CV learning no-free-lunch theorem in [27], where the authors optimize case-dependent cost functions that can be computed by applying a circuit to the target unknown state or gate. In our case, instead, one is able to find an approximation of the measurement statistics obtained by applying random circuits to the target unknown state or gate, which is a more expensive task in principle.

In this article, we introduce a statistical learning theory for infinite-dimensional quantum systems, studying the learnability of CV quantum states and processes, e.g., those that can be implemented in a photonic processor. We show that the simple class of Gaussian states and operations can be learned efficiently, with a SC scaling quadratically in the number of bosonic modes n (independent degrees of freedom, analogous to the number of qubits) of the system, i.e., also linearly in the system size and without any dependence on cutoffs. Unfortunately, it is well-known that non-Gaussian operations are required to realize a universal CV quantum computer [28]. Therefore, we consider next a wider class of operations including Gaussian, photocounting and generalized Gaussian (GG) operations, the latter being a non-Gaussian set introduced by Bourassa et al. [29], which can approximate several key non-Gaussian resources, e.g., Gottesman-Kitaev-Preskill (GKP), cat and Fock states. We show that also this class is learnable with a SC quadratic in the number of modes, hence quadratically in the system size, provided that its parameters satisfy certain constraints, scaling with the systems' energy, thus establishing efficient theoretical learnability for a wide range of non-Gaussian CV quantum devices. Interestingly, at variance with finite-dimensional systems, in the CV case it appears that the SC does not directly depend on the circuit's depth, but rather on the degree of non-Gaussianity of the circuit.

The article is structured as follows: in Sec. 2 we describe in detail our problem setting, defining two different learning tasks for which our results apply, namely the standard learning an approximation of an unknown state or process, and the more general learning of a near-optimal circuit for tasks with linear loss function in the outcome probabilities, e.g., state discrimination or synthesis (2.1). We then do a brief recap of CV circuits, including Gaussian components, photodetection and a recently introduced class of non-Gaussian components (2.2). Finally, we provide a compact presentation of our main results, stating SC bounds for both learning tasks with several possible CV circuit classes that arise from combining states, measurements and channels of different kinds (2.3). In Sec. 3 we present our methods in detail and prove our results: we start with a statement of key theorems in classical statistical learning of probability-valued functions (3.1), then proceed by computing complexity measures for three CV circuit classes (3.2) and finally

apply these results to the computation of SC bounds for both tasks previously introduced (3.3, 3.4). In Sec. 5 we discuss the significance of our results and elaborate on future research directions.

2 Problem setting and results summary

2.1 Learning quantum circuits from measurement outcomes

2.1.1 Learning quantum states, measurements and channels

Consider a quantum circuit comprising preparation, evolution and measurement. The circuit is described by an input quantum state ρ, with $\mathrm{Tr}\left[\rho\right] = 1$ and $\rho \geq 0$, a noisy quantum channel Φ, i.e., a completely positive and trace-preserving map, and a measurement operator $\mathbb{1} \geq M \geq 0$ [1]. When running the circuit with a given setup (ρ, Φ, M), the outcome is a bit $b \in \{0, 1\}$, taking the value $b = 1$ when M is "accepted", which happens with probability $P(M|\Phi, \rho) = \mathrm{Tr}\left[M \cdot \Phi(\rho)\right]$, as prescribed by the Born rule.

Suppose now that ρ is unknown and we want to approximate it with a state $\sigma \in \mathcal{C}$, where $\mathcal{C}$ is a class of hypotheses. We measure the quality of our approximation by how well the hypothesis σ is able to reproduce the statistics obtained from ρ by applying channels and measurements from a class $\mathcal{S}$, sampled according to a probability distribution $Q(\Phi, M)$, via the true loss function

$$L_{\sigma,\rho}(\Phi, M) := \left|P(M|\Phi, \sigma) - P(M|\Phi, \rho)\right|. \tag{1}$$

Naturally, the quality of the hypotheses will depend on the employed hypothesis class: in statistical learning theory, one typically distinguishes between the *realizable* setting, where the unknown state $\rho \in \mathcal{C}$, and the more general *agnostic* setting, where $\rho \notin \mathcal{C}$. Our results are applicable to both these learning settings; for example, we show that it is possible to find a Gaussian approximation to a non-Gaussian state efficiently. Other constraints on the hypothesis class arise naturally from the theory of CV systems, see below Sec. 2.2 and 2.3.

In practice, as shown in Fig. 1, we will rely on a finite training set of binary outcomes $b_i \in \{0, 1\}$ obtained by running the circuit several times, each with a different sample channel Φ_i and measurement operator M_i, for $i = 1, \cdots, T$. Therefore, the best we can do is trying to find a candidate $\sigma \in \mathcal{C}$ with small loss on each training instance (Φ_i, M_i, b_i), as measured by the empirical loss function

$$\hat{L}_{\sigma,\rho}(\Phi_i, M_i, b_i) := \left|P(M_i|\Phi_i, \sigma) - b_i\right|. \tag{2}$$

A learning theorem then guarantees that the chosen hypothesis σ can approximate the statistics of the unknown ρ in terms of the true loss function, even on unseen instances, a property called generalization. The minimum number of samples T_0 that guarantees generalization will depend on the specific precision and error thresholds requested and

[1] Note that, as typical in the learning setting that we consider, one is guaranteed to find an algorithm that approximates the expected value of M, i.e., the statistics of a binary-outcome measurement $\{M, \mathbb{1} - M\}$, *on average* over M sampled in a certain sample class $\mathcal{S}$. Similarly, the statistics of multi-outcome measurements $\{M(m)\}_{m\in\mathcal{M}}$ can be approximated using the same algorithm and by including all measurements operators $M(m)$ in the sample class $\mathcal{S}$. If one however is interested in obtaining good approximations of all $P(M(m)|\Phi, \rho)$ at once, then the SC will scale at least linearly in $|\mathcal{M}|$, as pointed out in [14] via a union-bound argument. An improvement over this scaling is conceivable only by allowing the algorithm to perform joint measurements directly on the quantum samples or by changing the requirements, e.g., as in shadow tomography [26].

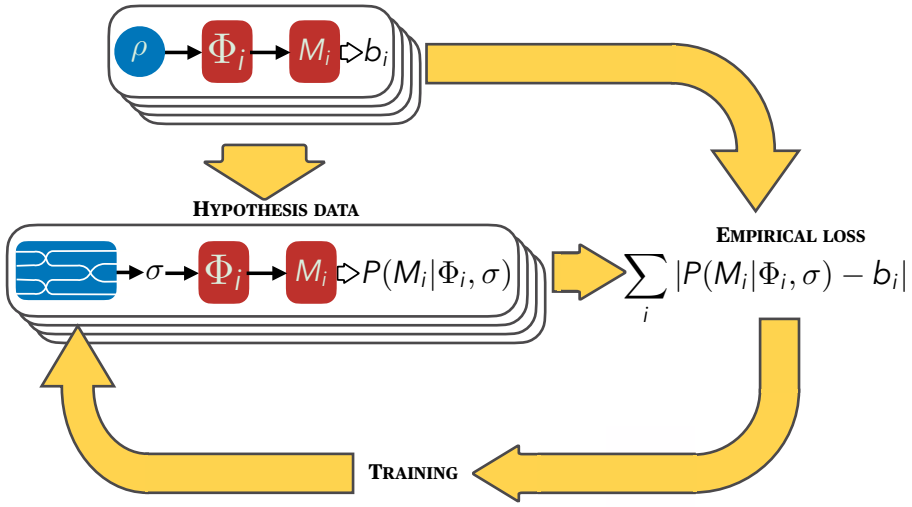

Figure 1: Depiction of the simplest learning problem studied for CV architectures. We aim to reproduce the statistics of an unknown state ρ with respect to random couples of CV channels and measurements (Φ_i, M_i). For each i, the state is passed through channel Φ_i and measured with a binary measurement M_i obtaining outcome b_i. Then the learner produces a hypothesis σ by tuning the parameters of a CV circuit, e.g., a photonic chip, and obtains its output statistics on the channels and measurements of the training set. A suitable empirical loss evaluates how well σ approximates ρ on the training set. This can be used to optimize σ iteratively.

on the chosen hypothesis class. We thus arrive at the following operational learnability statement:

Definition 1. *(State learnability) Let ρ be an unknown quantum state, $\mathcal{C}$ a set of hypothesis quantum states (not necessarily containing ρ) and $\mathcal{S}$ a set of channel-measurement couples. Suppose there is an algorithm that takes as input $T \geq T_0$ training samples $\{(\Phi_t, M_t, b_t)\}_{t=1}^T$ according to the unknown distribution $Q(\Phi, M) \cdot P(M|\Phi, \rho)$, and outputs a hypothesis $\sigma \in \mathcal{C}$ that is η-good on each training instance, i.e.,*

$$\hat{L}_{\sigma,\rho}(\Phi_i, M_i, b_i) < \eta \; \forall i = 1, \cdots, T. \tag{3}$$

We say that an approximation to ρ in $\mathcal{C}$ with respect to sample channels and measurements in $\mathcal{S}$ can be learned if, with probability at least $1-\delta$ with respect to the training sample, i.e., $\{(\Phi_t, M_t, b_t)\}_{t=1}^T \sim Q^T$, it holds

$$\Pr_{(\Phi,M)\sim Q}(L_{\sigma,\rho}(\Phi, M) < \eta + \gamma) \geq 1 - \epsilon, \tag{4}$$

for all $\eta, \gamma, \epsilon, \delta > 0$ and distributions Q on $\mathcal{S}$. The sample complexity (SC) is the minimum number of samples $T_0(\delta, \epsilon, \gamma, \mathcal{C}, \mathcal{S})$ sufficient for learnability (4).

This statement is a generalization of [14] to the agnostic case, where $\rho \in \mathcal{C}$ and η cannot be decreased arbitrarily. In the simpler realizable case, one can bring η as close to zero as desired, thus converging to the true quantum state $\rho \in \mathcal{C}$.

Note that the only role of the quantum circuit in this setting is to provide binary outcomes b_i according to the Born rule. Hence, one can easily interchange the role of states, channels and measurements as hypotheses and samples, obtaining similar definitions of

(i) *Measurement learnability* [15], where the objective is to approximate an unknown measurement, $\mathcal{C} = \{M\}$ is a class of measurement operators and $\mathcal{S} = \{(\rho, \Phi)\}$ is a class of states and channels;

(ii) *Channel learnability* [14, 17], where the objective is to approximate an unknown channel, $\mathcal{C} = \{\Phi\}$ is a class of channels and $\mathcal{S} = \{(\rho, M)\}$ is a class of states and measurement operators.

In Sec. 3 we provide SC bounds for these learning problems in the case where $\mathcal{C}$ and $\mathcal{S}$ are classes of CV states, channels and measurements.

2.1.2 Learning circuits for discrimination and synthesis

So far we have considered the problem of finding a quantum state (or, more generally, circuit) that approximates the measurement statistics of an unknown target state (resp. circuit), based on training samples received from the latter. On the other hand, many problems in quantum information theory have the objective of finding a family of quantum circuits that carries out a certain task with optimal performance, on average with respect to different possible realizations, for example:

(i) *Discrimination*: there is a set of states $\{\rho_x\}_{x\in\mathcal{X}}$ and a probability distribution $Q(x)$ over $\mathcal{X}$. We receive an unknown sample ρ_x, with $x \sim Q$, and the task is to find a multi-outcome measurement $\{M_x\}_{x\in\mathcal{X}}$, with operators in a desired measurement class $\mathcal{C}$, that guesses x with maximum average success probability, i.e.,

$$P_{\text{succ}}(\mathcal{C}) := \max_{\{M_x\}_{x\in\mathcal{X}}: M_x\in\mathcal{C}\forall x} \mathbb{E}_{x\sim Q}\text{Tr}\left[M_x\rho_x\right]. \tag{5}$$

In particular, within the CV setting considered in this paper, the discrimination of Gaussian states with non-Gaussian resources plays a major role in quantum optical communication [30, 31, 32]. Here we will focus on a variational approach to the problem, where the measurement class can be equivalently described as an information-processing channel, represented by a noisy quantum circuit Φ in a given channel class $\mathcal{C}$, followed by a rank-1 projective measurement in a fixed basis $\{|x\rangle\langle x|\}_{x\in X}$, whose outcome x corresponds to guessing for the state ρ_x. The optimization then yields an optimal circuit within the chosen class, with average success probability

$$P_{\text{succ}}(\mathcal{C}) := \max_{\Phi\in\mathcal{C}} \mathbb{E}_{x\sim Q} \langle x| \Phi(\rho_x) |x\rangle. \tag{6}$$

Note that, by a Naimark-dilation argument, this setting can also include POVM measurements realizable by embedding the states in a larger Hilbert space. Furthermore, one can also consider higher-rank projections or coarse-grainings of the measurement outcomes by convex combination of the rank-1 measurement operators, i.e., $M_x = \sum_{y\in\mathcal{X}} p(y|x)|y\rangle\langle y|$ with fixed combination coefficients.

This approach is extremely flexible, e.g., it can accomodate also channel discrimination problems. In this case, the randomness can be attributed to the channel itself, i.e., the training set is constructed by sampling from a set of channels $\{\Phi_x\}_{x\in\mathcal{X}}$ that we want to discriminate, labelled by $x \sim Q$. We can optimize the input state and output measurements in a certain class $\mathcal{C} = \{(\rho, \{M_x\}_{x\in\mathcal{X}})\}$, with the aim of maximizing the average success probability:

$$P_{\text{succ}}(\mathcal{C}) = \max_{(\rho,\{M_x\})\in\mathcal{C}} \mathbb{E}_{x\sim Q}\text{Tr}\left[M_x\Phi_x(\rho)\right]. \tag{7}$$

(ii) *Pure-state synthesis and embedding problems*: starting from a noisy input state sampled from a set $\{\rho_x\}_{x\in X}$ via a classical random variable $x \sim Q$, we want to produce

a corresponding pure target state $\{|\psi_x\rangle\langle\psi_x|\}_{x\in\mathcal{X}}$ for each x. We can do so via a variational noisy quantum circuit, represented by a quantum channel Φ from a desired class $\mathcal{C}$. The circuit is chosen to maximize the average fidelity within the class, i.e.,

$$F(\mathcal{C}) := \max_{\Phi\in\mathcal{C}} \mathbb{E}_{x\sim Q} \langle\psi_x| \Phi(\rho_x) |\psi_x\rangle . \tag{8}$$

Note that, by properly choosing the input and target states, as well as the variational circuit class, one can adapt this setting to several quantum information processing tasks, e.g., distillation and error correction problems. For example, in a distillation problem one starts with a mixed state $\rho_x = \sigma_x^{\otimes k}$, where σ_x is not maximally entangled of which k copies are available, and wants to convert it, via an LOCC channel Φ, into a pure maximally entangled target state $|\psi_x\rangle$ (similarly, one can think of distillation in other resource theories, e.g., that of coherence [33]). The random variable x represents different preparations of the mixed state and different target states. The figure of merit (8) quantifies the average distillation fidelity. Instead, in error correction, the states ρ_x might represent the result of a noisy quantum computation, depending on a classical input x, and the objective is to find a channel Φ that can correct the errors, mapping the states to the ideal output of the computation $|\psi\rangle_x$. The average error-correction fidelity is then (8) too. Finally, by setting the input state $\rho_x = |x\rangle\langle x|$ as a computational-basis encoding of the random variable x, we can optimize Φ as an embedding map that tries to encodes the classical variable x into a target quantum $|\psi_x\rangle$ via a complex circuit, with average encoding fidelity given again by (8).

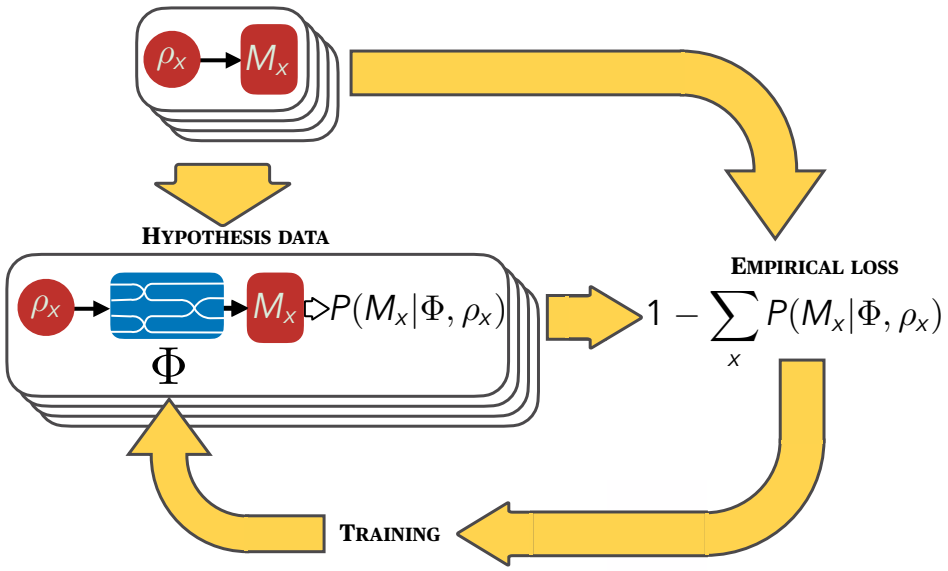

Figure 2: Depiction of the discrimination task learning problem studied for CV architectures. We receive random samples (ρ_x, M_x, b_x) from a class of input states $\{\rho_x\}$ to be discriminated with a fixed measurement device $\{M_x\}$. Then the learner produces a hypothesis channel Φ that processes the input ρ_x before measurement, and obtains its output statistics. A suitable empirical loss evaluates how well Φ is able to discriminate the states in the training set. This can be used to optimize Φ iteratively.

Traditional methods in the literature tackle these problems from a standard optimization perspective, both analytical and numerical [34, 35, 36, 37, 38, 39, 40]. We show instead that it is possible to formulate a learning problem where a quantum circuit is trained to

maximize the average acceptance probability based on samples $x \in \mathcal{X}$, as shown in Fig. 2. The learner receives training samples of the form (ρ_x, M_x), where $x \sim Q$, and tries to classify each input by assigning a bit value $b = 1$ with probability $P(M_x|\Phi, \rho_x)$, where Φ represents a noisy quantum circuit in a given class $\mathcal{C}$, or $b = 0$ otherwise. The target decision rule is trivial, in the sense that each couple (ρ_x, M_x) is always accepted, and the loss is measured by

$$L_\Phi(\rho, M) := (1 - P(M|\Phi, \rho)). \tag{9}$$

A learning guarantee then provides conditions for the identification of a hypothesis Φ whose average loss is close to the minimum in the class, given by

$$\min_{\Phi \in \mathcal{C}} \mathbb{E}_{x\sim Q} L_\Phi(\rho_x, M_x) = 1 - \max_{\Phi \in \mathcal{C}} \mathbb{E}_{x\sim Q} P(M_x|\Phi, \rho_x), \tag{10}$$

which is equivalent to finding a hypothesis with close-to-optimal success probability or fidelity in the above-described discrimination and synthesis problems. Clearly, in most applications the learning problem will be intrinsically agnostic, since an ideal circuit that classifies each state correctly, i.e., such that $P(M_x|\Phi, \rho_x) = 1$ for all $x \in \mathcal{X}$ might not exist at all. This brings us to the following operational task learnability statement:

Definition 2. *(Optimal task learnability) Let $\mathcal{S} = \{(\rho_x, M_x)\}_{x\in\mathcal{X}}$ be a set of (not necessarily known) state-measurement couples associated with a quantum-information-processing task, Q a distribution on $\mathcal{X}$ and $\mathcal{C}$ a set of hypothesis quantum channels. Suppose there is an algorithm that takes as input a training sample $\{(\rho_{x_t}, M_{x_t})\}_{t=1}^T$ of size $T \geq T_0$ according to the unknown distribution $Q(x)$, and outputs a hypothesis $\Phi \in \mathcal{C}$ that minimizes the cumulative empirical loss on the sample, i.e.,*

$$\sum_{t=1}^{T} L_\Phi(\rho_{x_t}, M_{x_t}). \tag{11}$$

We say that a near-optimal circuit in $\mathcal{C}$ for the task represented by $\{(\rho_x, M_x)\}_{x\in X}$ can be learned if, with probability at least $1 - \delta$ with respect to the training sample, i.e., $\{(\rho_{x_t}, M_{x_t})\}_{t=1}^T \sim Q^T$, it holds

$$\mathbb{E}_{x\sim Q} L_\Phi(\rho_x, M_x) - \inf_{\Gamma\in\mathcal{C}} \mathbb{E}_{x\sim Q} L_\Gamma(\rho_x, M_x) \leq \epsilon \tag{12}$$

for all $\epsilon, \delta > 0$ and distributions Q on $\mathcal{X}$. The sample complexity (SC) is the minimum number of samples $T_0(\delta, \epsilon, \mathcal{C}, \mathcal{S})$ sufficient for learnability (12).

Thanks to the similarity of the induced hypothesis classes employed in Definitions 1 and 2, in Sec. 3 we are able to derive SC bounds also for the latter learning task, in the case of CV quantum circuits. Previous combinatorial-dimension bounds can be combined with our results to obtain SC bounds for information-processing tasks also in finite-dimensions [14, 15, 17].

2.2 Continuous-variable quantum information

A CV bosonic quantum system can be described as a quantum harmonic oscillator with position and momentum operators ξ_1, ξ_2 (also called quadratures), satisfying the canonical commutation relations $[\xi_1, \xi_2] = i\hbar$, and Hamiltonian $H := \frac{h\nu}{2}(\xi_1^2 + \xi_2^2)$, where ν is the oscillator's natural frequency, h is Planck's constant and $\hbar = \frac{h}{2\pi}$ (for simplicity, we henceforth set $h\nu = 1 = \hbar$ [2]). The oscillator can model, for example, the state of a mechanical

[2] This amounts to measuring energy in units of $h\nu$ and time in units of $1/(2\pi\nu)$.

resonator or a travelling electromagnetic pulse with fixed polarization, wavevector and frequency. A single CV quantum system is called mode and it plays an analogous role to that of a qubit for two-dimensional quantum systems. A system of n modes has a vector of quadrature operators $\boldsymbol{\xi} := (\xi_1, \xi_2, \cdots, \xi_{2n-1}, \xi_{2n})^T$, were T is the transpose [3].

A compact way to describe CV systems is via their Wigner function, a quasi-probability distribution defined for every operator O of the Hilbert space as

$$W_O(\mathbf{r}) := \int_{\mathbb{R}^{2n}} \frac{d^{2n}\mathbf{s}}{(2\pi)^n} e^{-\mathrm{i}\mathbf{s}^T\Omega\mathbf{r}} \operatorname{Tr}\left[O \cdot D(\mathbf{s})^\dagger\right], \tag{13}$$

where $D(\mathbf{s}) := \exp(\mathrm{i}\mathbf{s}^T\Omega\boldsymbol{\xi})$ is the displacement operator and $\Omega := \bigoplus_{i=1}^n \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is the symplectic form. Note that, if O is Hermitian, its Wigner function is real-valued; if additionally $O = \rho$ is a normalized quantum state, then its Wigner function is normalized as well, i.e., $\int_{\mathbb{R}^{2n}} d^{2n}\mathbf{r} W_\rho(\mathbf{r}) = 1$.

As observed in Sec. 2.1, our central object of interest will be the outcome probability of a CV quantum circuit, which can be easily determined via integration of the Wigner functions of the final circuit state ρ and the measurement operator M:

$$P(M|\rho) = \operatorname{Tr}[M\rho] = \int_{\mathbb{R}^{2n}} d^{2n}\mathbf{r} W_M(\mathbf{r}) W_\rho(\mathbf{r}). \tag{14}$$

Finally, key quantities in the analysis of CV systems are the first and second moments of the quadratures, i.e., the mean-values vector $\mathbf{m} \in \mathbb{R}^{2n}$ and covariance matrix $V \in \mathrm{Sym}_{2n}(\mathbb{R})$[4] with components

$$m_i := \operatorname{Tr}[\xi_i \rho], \quad V_{i,j} := \frac{1}{2}\operatorname{Tr}[(\xi_i\xi_j + \xi_j\xi_i)\rho]. \tag{15}$$

Note that the covariance matrix of a quantum state must respect the uncertainty principle $V + \mathrm{i}\Omega \geq 0$, implying $V \geq 0$.

2.2.1 Gaussian circuits

The simplest class of CV circuits comprises only Gaussian states, channels and measurements. Indeed, any element of this class can be experimentally realized using lasers, optical instrumentation and photodetectors; furthermore, a programmable integrated-photonic chip including all such elements is well within reach of current technology [2, 5, 4]. Gaussian states $\rho_{\mathbf{m},V}$ on n modes are completely determined by their mean $\mathbf{m} \in \mathbb{R}^{2n}$ and covariance matrix $V \in \mathrm{Sp}_{2n}(\mathbb{R})$: their Wigner function is Gaussian with the same mean and covariance matrix V, i.e.,

$$W_{\rho_{\mathbf{m},V}}(\mathbf{r}) = \mathcal{G}_{\mathbf{m},V}(\mathbf{r}) := \frac{e^{-\frac{1}{2}(\mathbf{r}-\mathbf{m})^T V^{-1}(\mathbf{r}-\mathbf{m})}}{\sqrt{\operatorname{Det}(2\pi V)}}. \tag{16}$$

Similarly, general-dyne Gaussian measurement operators $M_{\mathbf{m},V}$ have Wigner function $W_{M_{\mathbf{m},V}}(\mathbf{r}) = \mathcal{G}_{\mathbf{m},V}(\mathbf{r})$, the most common ones being homo- and hetero-dyne measurements (the latter corresponding to $V = \frac{\mathbb{1}}{2}$ and arbitrary $\mathbf{m}$). Finally, a noisy Gaussian

[3] Given the limited use of quadrature operators in the article, we use greek letters to represent operator-valued quantities and latin letters to represent complex-valued quantities.

[4] In the article we use the following notation for sets of square matrices of order $2n$ with real coefficients (and similarly for complex coefficients): $\mathrm{Mat}_{2n}(\mathbb{R})$ is the set of all such matrices, $\mathrm{Sym}_{2n}(\mathbb{R})$ is the subset of symmetric matrices, $\mathrm{Sp}_{2n}(\mathbb{R})$ is the subset of symplectic matrices.

channel is determined by a displacement vector $\mathbf{d} \in \mathbb{R}^{2n}$ and two matrices $X \in \mathrm{Mat}_{2n}(\mathbb{R})$, $Y \in \mathrm{Sym}_{2n}(\mathbb{R})$ satisfying the constraint $Y + \mathrm{i}\Omega \geq X\Omega X^T$, which amounts to preserving the uncertainty principle. The action of a Gaussian channel $\Phi_{\mathbf{d},X,Y}$ preserves Gaussian states, transforming their mean and covariance as follows:

$$\mathbf{m} \mapsto \mathbf{m}_{\mathrm{out}} := X\mathbf{m} + \mathbf{d}, \quad V \mapsto V_{\mathrm{out}} := XVX^T + Y. \tag{17}$$

Therefore, the outcome probability of a Gaussian circuit is

$$P_{\mathrm{g}}(M_{\mathbf{m}',V'}|\Phi_{\mathbf{d},X,Y}, \rho_{\mathbf{m},V}) = \mathcal{G}_{\mathbf{m}_{\mathrm{out}},V_{\mathrm{out}}+V'}(\mathbf{m}'). \tag{18}$$

2.2.2 Photodetection

Non-Gaussian resources are needed to realize universal quantum computation with CV systems. One such resource is photodetection, whose single-mode measurement operators $M_k := |k\rangle\langle k|$ correspond to projections onto Fock states $|k\rangle$, i.e., states with a fixed number k of photons (or excitations) such that $H|k\rangle = k + \frac{1}{2}$. If we allow the detection of up to K photons in a single mode, an n-mode photodetection measurement is hence described by a vector of outcomes $\mathbf{k} \in \{0, \cdots, K\}^n$ as $M_{\mathbf{k}} := |\mathbf{k}\rangle\langle\mathbf{k}| = \bigotimes_{i=1}^n |k_i\rangle\langle k_i|$.

The outcome probability of a photodetection measurement on a Gaussian circuit is

$$P_{\mathrm{gp}}(M_{\mathbf{k}}|\Phi_{\mathbf{d},X,Y}, \rho_{\mathbf{m},V}) = (2\pi)^n \mathcal{G}_{\mathbf{m}_{\mathrm{out}},V_{\mathrm{out}}+\frac{1}{2}}(0) H_{\mathbf{k}}^{(\tilde{V})}(\tilde{\mathbf{m}}), \tag{19}$$

with $H_{\mathbf{k}}$ a multivariate Hermite polynomial [41] defined as

$$H_{\mathbf{k}}^{(V)}(\mathbf{m}) := \mathcal{G}_{0,V}(\mathbf{m})^{-1} \prod_{i=1}^n \left(-\frac{\partial}{\partial_{2i-1}}\right)^{\mathbf{k}_i} \left(-\frac{\partial}{\partial_{2i}}\right)^{\mathbf{k}_i} \mathcal{G}_{0,V}(\mathbf{m}), \tag{20}$$

while $\tilde{V} \in \mathrm{Mat}_{2n}(\mathbb{R})$ and $\tilde{\mathbf{m}} \in \mathbb{R}^{2n}$ are quadratic functions of V_{out} and $\mathbf{m}_{\mathrm{out}}$ [42]. In the following, we refer to Gaussian plus photodetection (GP) circuits as those obtained by combining Gaussian channels with photodetection measurements, possibly coarse-grained. The latter term refers to a post-processing of outcomes that lumps together different photocounting events $\mathbf{k}$ with probabilities $q_{\mathbf{k}} \geq 0$ such that $\sum_{\mathbf{k}} q_{\mathbf{k}} \leq 1$, resulting in a measurement operator $M_q := \sum_{\mathbf{k}} q_{\mathbf{k}} |\mathbf{k}\rangle\langle\mathbf{k}|$ and clearly

$$P_{\mathrm{gp}}(M_q|\Phi_{\mathbf{d},X,Y}, \rho_{\mathbf{m},V}) = \sum_{\mathbf{k}} q_{\mathbf{k}} P_{\mathrm{gp}}(M_{\mathbf{k}}|\Phi_{\mathbf{d},X,Y}, \rho_{\mathbf{m},V}). \tag{21}$$

2.2.3 Generalized Gaussian circuits

More generally, Bourassa et al. [29] introduced for simulation purposes a class of non-Gaussian circuits that we call generalized Gaussian (GG). An arbitrary GG state is defined by a possibly infinite set of complex-valued coefficients, means and covariances $\{c_i \in \mathbb{C}, \mathbf{m}_i \in \mathbb{C}^{2n}, V_i \in \mathrm{Mat}_{2n}(\mathbb{C})\}_{i\in\mathcal{I}}$. The Wigner function of a GG state $\rho_{\{c,\mathbf{m},V\}_{\mathcal{I}}}$ is then given by a complex combination of complex-valued Gaussians:

$$W_{\rho_{\{c,\mathbf{m},V\}_{\mathcal{I}}}}(\mathbf{r}) := \sum_{i\in\mathcal{I}} c_i \mathcal{G}_{\mathbf{m}_i,V_i}(\mathbf{r}), \tag{22}$$

where we omitted the explicit dependence on i of all the parameters on the left-hand side, for simplicity. Notice that, while the total Wigner function must be real, since ρ is Hermitian, its components instead are in general complex-valued and do not correspond to physically valid quantum states. Nevertheless we require that $\mathrm{Re}V_i \geq 0$ for all $i \in \mathcal{I}$,

so that all Gaussian components are bounded, and $\sum_{i\in\mathcal{I}} V_i + \mathrm{i}\Omega \geq 0$, so that the full covariance matrix respects the uncertainty principle. Furthermore, the overall Wigner function must be normalized, implying $\sum_{i\in\mathcal{I}} c_i = 1$. Furthermore, we stress that this class is far from a simple mathematical abstraction. For example, it includes typical non-Gaussian states, e.g., cat states, or arbitrary-precision approximations theoreof, e.g., Fock or GKP states, which can all be written as complex superpositions of pure Gaussian states. We refer the reader to [29] for a detailed discussion.

Similarly, GG measurement operators $M_{\{c,\mathbf{m},V\}_\mathcal{I}}$ have a Wigner function like (22), with the only difference that in general it is sub-normalized, i.e., $\sum_{i\in\mathcal{I}} c_i \leq 1$.As for the case of GG states, this class of measurements includes projections on the most relevant non-Gaussian states or approximations thereof.

Finally, we can perform a GG channel on an arbitrary input state ρ by introducing ancillary modes in a non-Gaussian state, performing a Gaussian channel on the input plus ancillary modes, and finally performing a non-Gaussian measurement on the ancillary modes, possibly followed by another Gaussian channel conditioned on the measurement outcome. The overall action of a GG channel on a Gaussian state is determined by a set of complex-valued coefficients, displacement vectors and matrices $\{c_i \in \mathbb{C}, \mathbf{d}_i \in \mathbb{C}^{2n}, X_i, Y_i \in \mathrm{Mat}_{2n}(\mathbb{C})\}_{i\in\mathcal{I}}$ as follows:

$$\Phi_{\{c,\mathbf{d},X,Y\}_\mathcal{I}} : \rho_{\mathbf{m},V} \mapsto \rho_{\{c,\mathbf{m}_{\mathrm{out}},V_{\mathrm{out}}\}_\mathcal{I}}, \tag{23}$$

with $\sum_i c_i = 1$ in order to guarantee the trace-preserving property.

Therefore, the outcome probability of a GG circuit is

$$\begin{aligned} &P_{\mathrm{gg}}(M_{\{c',\mathbf{m}',V'\}_{\mathcal{I}'}} | \Phi_{\{c'',\mathbf{d},X,Y\}_{\mathcal{I}''}}, \rho_{\{c,\mathbf{m},V\}_\mathcal{I}}) \\ &= \sum_{i\in\mathcal{I},j\in\mathcal{I}',k\in\mathcal{I}''} c_i c'_j c''_k \, \mathcal{G}_{\mathbf{m}^{(ijk)}_{\mathrm{out}}, V^{(ik)}_{\mathrm{out}}+V'_j}(\mathbf{m}'_j), \end{aligned} \tag{24}$$

where $\mathbf{m}^{(ijk)}_{\mathrm{out}} := X_k\mathbf{m}_i + \mathbf{d}_k$, $V^{(ik)}_{\mathrm{out}} := X_k V_i X_k^T + Y_k$, and the sum runs over all indices defining the GG state, channel and measurement, which will be left implicit in the rest of the article.

2.3 Results summary

We are now ready to state our main result: we establish that quantum states generated by n-mode CV circuits can be learned efficiently in n, for all the practically relevant circuit classes that can be constructed with components from Sec. 2.2:

Theorem 1. *Let ρ be an unknown quantum state, $\mathcal{C}_{\mathrm{g,gg}}$ the hypothesis classes of Gaussian and GG states, and $\mathcal{S}_{\mathrm{g,gp,gg}}$ the hypothesis classes of Gaussian, GP and GG measurements and channels on an n-mode CV architecture. Then it is possible to learn an ϵ-approximation of ρ inside $\mathcal{C}$ using samples from $\mathcal{S}$ (in the sense of Def. 1) with SC polynomial in n and ϵ^{-1}. Explicit SC upper bounds for each of the above classes are given in Table 1, hiding the dependence on ν, δ and up to sub-leading* log *factors.*

Here K is the maximum photon-number cutoff for photodetection measurements, while B is a constant restricting the GG class (see 3.2.3). Note that the same SC bounds of Table 1 hold for channel and measurement learning, i.e., swapping the roles of $\mathcal{S}$ and $\mathcal{C}$, except for one notable difference: photodetection measurement learning with Gaussian states and channels has SC scaling exponentially with the number of modes, $O(K^n)$ (see 3.2.2).

Similarly, we establish learnability of CV circuits for discrimination and synthesis tasks with polynomial encoding functions:

		hypothesis states	
		$\mathcal{C}_{\rm g}$	$\mathcal{C}_{\rm gg}$
Sample channels and measurements	$\mathcal{S}_{\rm g}$	$O(n^2\epsilon^{-2}\log n)$	$O(n^2\epsilon^{-4}B^2)$
	$\mathcal{S}_{\rm gp}$	$O(n^2\epsilon^{-2}\log(nK))$	unknown
	$\mathcal{S}_{\rm gg}$	$O(n^2\epsilon^{-4}B^2)$	$O(n^2\epsilon^{-4}B^2)$

Table 1: Sample complexity of learning Gaussian or GG states with Gaussian, GP or GG channels and measurements. Here n is the number of modes of the CV circuit, ϵ the error, K the photon-number cutoff and B a constant restricting the GG class.

Theorem 2. *Let $\mathcal{C}$ be a class of n-mode CV quantum channels and $(\rho_{\boldsymbol{\gamma}(x)}, M_{\boldsymbol{\gamma}(x)})_{x\in\mathcal{X}}$ the states and measurements describing a discrimination or synthesis task, where $\boldsymbol{\gamma}(x)$ is a vector of encoding functions that are polynomial in x of order at most ℓ. In particular, let $\mathcal{C}_{\rm g,gp,gg}$ be the classes of Gaussian, GP and GG circuit families.*

Then it is possible to learn an ϵ-optimal channel in $\mathcal{C}$ for the task described by $(\rho_{\boldsymbol{\gamma}(x)}, M_{\boldsymbol{\gamma}(x)})_{x\in\mathcal{X}}$ (in the sense of Def. 2) with SC polynomial in n, ϵ^{-1}, ℓ. Explicit SC upper bounds are given in Table 2.

		Sample complexity
Circuit class	$\mathcal{C}_{\rm g}$	$O(\ell n^2\epsilon^{-2}\log n)$
	$\mathcal{C}_{\rm gp}$	$O(\ell n^2\epsilon^{-2}\log(nK))$
	$\mathcal{C}_{\rm gg}$	$O(\ell n^2\epsilon^{-4}B^2)$

Table 2: Sample complexity of learning Gaussian, Gaussian plus photodetection, or GG circuits for discrimination and synthesis tasks. Here n is the number of modes of the CV circuit, ϵ the error, K the photon-number cutoff, B a constant restricting the GG class and ℓ the order of the encoding polynomials.

3 Sample complexity of CV circuits

3.1 Bounding the SC of probability-valued functions

Our results build on the study of learnability for probability-valued function classes, pioneered in the works [43, 44, 45] and more recently employed in the quantum setting by [14]. The following theorems bound the number of samples required to approximate a probabilistic relation between two random variables, in terms of a suitable effective dimension of the function class employed for approximation. Thus, by computing such dimension for the various classes induced by processing and measuring CV circuits, as done in the next sections, we can bound their sample complexity.

We start by introducing two quantities that measure the effective dimension of a function class:

Definition 3. *(Fat-shattering- and pseudo-dimension) Let $\mathcal{F} := \{f : x \in \mathcal{X} \mapsto f(x) \in \mathbb{R}\}$ be a class of real-valued functions on a set $\mathcal{X}$, take $\gamma > 0$ and a sequence of k inputs $\mathbf{x} \in \mathcal{X}^k$. We say that $\mathbf{x}$ is γ-shattered by $\mathcal{F}$ if there exists a sequence of thresholds $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that, for each sequence of k bits $(b_1, \cdots, b_k) \in \{0,1\}^k$ determining a binary classification pattern of the inputs $\mathbf{x}$ there exists a concept $f \in \mathcal{F}$ that reproduces such pattern with thresholds $\boldsymbol{\alpha}$ and safety margin γ, i.e.,*

$$f(x_i) \geq \alpha_i + \gamma \;\forall i : b_i = 1 \text{ and } f(x_i) \leq \alpha_i - \gamma \;\forall i : b_i = 0. \tag{25}$$

The γ-fat-shattering dimension, $\mathrm{fat}_{\mathcal{F}}(\gamma)$*, is the largest k such that there exists a sequence $\mathbf{x} \in \mathcal{X}^k$ that is γ-shattered by $\mathcal{F}$. If there is no such k, then* $\mathrm{fat}_{\mathcal{F}}(\gamma) = \infty$.

The pseudo-dimension is

$$\mathrm{Pdim}(\mathcal{F}) := \lim_{\gamma' \to 0} \mathrm{fat}_{\mathcal{F}}(\gamma') \geq \mathrm{fat}_{\mathcal{F}}(\gamma) \tag{26}$$

for all $\gamma > 0$.

In order to prove our results we rely on two theorems from the classical statistical learning theory of real-valued functions. In particular, a line of works originating from Kearns and Schapire [43] (see also [44, 45, 46]) focused on obtaining learning guarantees for so-called probabilistic concepts (p-concepts); these are functions $f : x \mapsto [0, 1]$ that can be assimilated to the probability distribution of a random bit b conditioned on a random input x, i.e., $b = 1$ with probability $f(x)$ or $b = 0$ otherwise. In this setting, Kearns and Schapire identified two distinct learning problems, given a hypothesis class $\mathcal{F}$ of p-concepts: (i) *learning a model of probability*, where one is interested in finding a hypothesis $h \in \mathcal{F}$ that is close to the true p-concept f on each input x; and (ii) the simpler problem of *learning a decision rule*, where one is interested in finding a hypothesis $h \in \mathcal{F}$ that minimizes the mis-classification error on average with respect to the bit b and the input x.

Aaronson [14] discussed both these learning frameworks, linking them with p-concepts originating from quantum states and measurements via the Born rule. Here we observe that problem (i), i.e., learning a model of probability, is a proxy for state learnability (see Definition 1), and it requires an extension of Aaronson's results to the agnostic case:

Theorem 3. *(Agnostic p-concept learning) Let $\mathcal{X}$ be a sample space, Q a distribution on $\mathcal{X}$ and $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$ a hypothesis class. Consider an unknown p-concept $f : \mathcal{X} \mapsto [0, 1]$ (not necessarily in $\mathcal{F}$) and a training set $\{(x_t, b_t)\}_{t=1}^{T}$, where $x_t \sim Q$ and $b_t = 1$ with probability $f(x_t)$ or $b_t = 0$ otherwise. Suppose that we choose a hypothesis $h \in \mathcal{F}$ with small quadratic loss on each element of the training set, i.e., $(h(x_t) - b_t)^2 < \eta$ for all $t = 1, \cdots, T$. Then with probability at least $1 - \delta$ over the sample it holds*

$$\Pr_{x \sim Q}((h(x) - f(x))^2 < \eta + \gamma) \geq 1 - \epsilon \tag{27}$$

for all $\eta, \gamma, \epsilon, \delta > 0$, provided that

$$T \geq T_0(\epsilon, \gamma, \delta, \mathcal{F}) = O\left(\frac{1}{\epsilon}\left(d \log^2 \frac{d}{\gamma\epsilon} + \log \frac{1}{\delta}\right)\right), \tag{28}$$

where $d = \mathrm{fat}_{\mathcal{F}}\left(\frac{\gamma}{8}\right)$,

The proof of this Theorem is provided in Appendix A. We can also easily switch to the linear loss, more commonly used to measure the distance between quantum circuit outputs, via the following corollary:

Corollary 1. *Under the same assumptions of Theorem 3, suppose that we choose a hypothesis $h \in \mathcal{F}$ with small linear loss on each element of the training set, i.e., $|h(x_i) - b_i| < \eta$ for all $i = 1, \cdots, T$. Then with probability at least $1 - \delta$ over the sample it holds*

$$\Pr_{x \sim Q}(|h(x) - f(x)| < \eta + \gamma) \geq 1 - \epsilon \tag{29}$$

for all $\eta, \gamma, \epsilon, \delta > 0$ with the same sample complexity as (28)*.*

Proof. We apply Theorem 3 with $\eta \mapsto \eta^2$ and $\gamma \mapsto \gamma^2$. Then the sample complexity remains of the same order of magnitude in $\log \frac{1}{\gamma}$ and the learning guarantee (27) implies that

$$|h(x) - f(x)| < \sqrt{\eta^2 + \gamma^2} \leq \sqrt{(\eta + \gamma)^2} = \eta + \gamma \tag{30}$$

holds with probability at least $1 - \epsilon$ with respect to $x \sim Q$ and at least $1 - \delta$ with respect to the sample. □

On the other hand, problem (ii) above, i.e., learning a decision rule, is a proxy for task learnability (see Definition 2) and it can be solved by applying directly a Theorem by Aaronson:

Theorem 4. *(Prediction [14, Suppl. Mat. Thm. 8]) Under the same assumptions of Theorem 3, given T samples $\{(x_t, b_t)\}_{t=1}^T$ such that $x_t \sim Q$ and $b_t \sim f(x_t)$, i.e., $b_t = 1$ with probability $f(x_t)$ or $b_t = 0$ otherwise, suppose there exists a learning algorithm that outputs a hypothesis $h \in \mathcal{F}$ minimizing the empirical total-variation loss $\sum_{t=1}^T |h(x_t) - b_t|$. Define the mis-classification error of hypothesis h as*

$$\Delta_{h,f}(x) := \mathbb{E}_{b \sim f(x)} |h(x) - b| = f(x)(1 - h(x)) + (1 - f(x))h(x). \tag{31}$$

Then with probability at least $1 - \delta$ over the sample it holds

$$\mathbb{E}_{x \sim Q} \left[\Delta_{h,f}(x)\right] - \inf_{c \in \mathcal{F}} \mathbb{E}_{x \sim Q} \left[\Delta_{c,f}(x)\right] \leq \epsilon, \tag{32}$$

for all $\epsilon, \delta > 0$, provided that the training set has size at least

$$T = O\left(\frac{1}{\epsilon^2}\left(d \log^2 \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right), \tag{33}$$

with $d = \mathrm{fat}_{\mathcal{F}}\left(\frac{\epsilon}{10}\right)$.

In particular, by choosing a trivial target concept $f(x) = 1$ one can recover the task optimization of Definition 2 .

These Theorems guarantee that it is possible to identify "good" approximations to an unknown probabilistic process within a certain function class, using a number of training samples that scales linearly with the fat-shattering dimension of the class. In Secs. 3.2.1,3.2.2 we compute the pseudo-dimension of function classes comprising output probabilities of GG and GP circuits. This immediately gives SC bounds using the Theorems above, since the pseudo-dimension is the largest among fat-shattering dimensions.

In general, however, the pseudo-dimension can be infinite while the SC is still finite. This is the case for the function class of outcome probabilities of GG circuits (Sec. 3.2.3), hence we need to consider a tighter measure of complexity, known as covering numbers:

Definition 4. *(Covering number) Let $\mathcal{Y}$ be a subset of a normed vector space with distance measure $d(y_1, y_2)$ for all $y_1, y_2 \in \mathcal{Y}$. Then $\mathcal{W} \subseteq \mathcal{Y}$ is an internal ϵ-cover of $\mathcal{Y}$ if, for each $y \in \mathcal{Y}$ there exists a $w \in \mathcal{W}$ that is ϵ-close to it, i.e., $d(y, w) \leq \epsilon$.*

Now let $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be a class of $\mathcal{Y}$-valued functions and consider a subset of the input set $\Xi \subseteq \mathcal{X}$. The restriction of $\mathcal{F}$ to the input subset Ξ is defined as

$$\mathcal{F}|_{\Xi} := \{F \in \mathcal{Y}^{\Xi} : \exists f \in \mathcal{F} \text{ s.t. } F(x) = f(x)\ \forall x \in \Xi\}, \tag{34}$$

i.e., the set of distinct functions on Ξ whose value coincides with that of a function from $\mathcal{F}$ on all inputs in the subset.

We define the uniform ϵ-covering number of $\mathcal{F}$ at scale k as the size of the largest ϵ-cover of $\mathcal{F}|_\Xi$, optimized over all input subsets Ξ of size k, i.e.,

$$\mathcal{N}_d(\epsilon, \mathcal{F}, k) := \max\{|\mathcal{W}| : \exists \Xi \subseteq \mathcal{X}, |\Xi| = k \text{ s.t. } \mathcal{W} \text{ } \epsilon\text{-cover of } \mathcal{F}|_\Xi\}, \tag{35}$$

where d is the distance measure on $\mathcal{Y}$ used to construct the covers.

Furthermore, the covering number can be related with the pseudo-dimension as follows:

Lemma 1. *([21, Theorem 12.2]) Let $\mathcal{F}$ be a function class with real, bounded output, i.e., $f \in [-B, B]$ for all $f \in \mathcal{F}$, and d any p-norm-distance. Then for all $\epsilon > 0$ and $k \in \mathbb{N}$ it holds*

$$\log \mathcal{N}_d(\epsilon, \mathcal{F}, k) \leq \mathrm{Pdim}(\mathcal{F}) \log \frac{2eBk}{\mathrm{Pdim}(\mathcal{F})\epsilon}. \tag{36}$$

Finally, we state some useful Lemmas on the properties of the pseudo-dimension of composed classes.

The following result can be obtained by a simple generalization of the same result obtained for real function classes:

Lemma 2. *([47, Lemma 5]) Let $\mathcal{F} := \{f : \mathcal{X} \to \mathcal{Y}\}$ be a function class and $g : \mathcal{Y} \to \mathcal{Z}$ a strictly monotonous function. Then* $\mathrm{Pdim}(g(\mathcal{F})) = \mathrm{Pdim}(\mathcal{F})$.

The following result is stated by [48] and it can be obtained by combining the same result for Vapnik-Chervonenkis (VC) dimension with the characterization of pseudo-dimension in terms of VC dimension.

Lemma 3. *([49]) Let $\mathcal{F}_3 := \mathcal{F}_1 + \mathcal{F}_2$, then* $\mathrm{Pdim}(\mathcal{F}_3) = O(\mathrm{Pdim}(\mathcal{F}_1) + \mathrm{Pdim}(\mathcal{F}_2))$.

The following are standard results in classical statistical learning theory.

Lemma 4. *([21, Theorem 8.3]) Let $\mathcal{F} := \{f_a(x) : \mathcal{X} \mapsto \mathbb{R} \mid a \in \mathbb{R}^d, f_a(x) = \mathrm{poly}_\ell(a) \forall x \in \mathcal{X}\}$ be a class of functions that, for each fixed input $x \in \mathcal{X}$, are polynomials of order ℓ in their d real parameters. Then* $\mathrm{Pdim}(\mathcal{F}) \leq 2d \log(12\ell)$.

Lemma 5. *([21, Lemma 14.13]) Let $\mathcal{F}$ be a real-valued function class and $g : \mathbb{R} \to \mathbb{R}$ be $L - Lipschitz$. Then $\mathcal{N}_2(\epsilon, g(\mathcal{F}), k) \leq \mathcal{N}_2(\frac{\epsilon}{L}, \mathcal{F}, k)$.*

The following result follows from a general theorem in statistical learning, by restricting to proper convex combinations, i.e., with coefficients summing to less than one:

Lemma 6. *([21, Theorem 14.14]) Let $\mathcal{F}$ be a function class satisfying three properties: (i) $\mathcal{F} = -\mathcal{F}$, (ii) $0 \in \mathcal{F}$ and (iii) $f \in [-B, B]$ for all $f \in \mathcal{F}$. Then*

$$\log \mathcal{N}_2(\epsilon, k, \mathfrak{C}(\mathcal{F})) \leq \left\lceil \left(\frac{B}{\epsilon}\right)^2 \right\rceil \log \mathcal{N}_2(\epsilon, \mathcal{F}, k). \tag{37}$$

The following result regards the covering number of a product of classes.

Lemma 7. *Let $\mathcal{F}_i \subseteq [-B_i, B_i]^X$, $i = 1, 2$, be two bounded real-valued function classes on a space X and define their product*

$$\mathcal{F} := \mathcal{F}_1 \cdot \mathcal{F}_2 = \{f = f_1 \cdot f_2 : f_i \in \mathcal{F}_i, i = 1, 2\}. \tag{38}$$

Then

$$\mathcal{N}(B_1\epsilon_1 + B_2\epsilon_2, \mathcal{F}, m) \leq \mathcal{N}(\epsilon_1, \mathcal{F}_1, m)\mathcal{N}(\epsilon_2, \mathcal{F}_2, m). \tag{39}$$

Proof. Suppose that $\mathcal{W}_i$ are ϵ_i-covers for the two classes and consider the class $\mathcal{W} = \mathcal{W}_1 \cdot \mathcal{W}_2$. Then for any $f = f_1 \cdot f_2 \in \mathcal{F}$ and $i = 1, 2$ there exist $w_i \in \mathcal{W}_i$ such that $d(w_i, f_i) \leq \epsilon_i$ by construction, where d is the distance used to define the covers. By the triangle inequality it follows

$$\begin{aligned} d(w_1 w_2, f_1 f_2) &\leq d(w_1 w_2, f_1 w_2) + d(f_1 w_2, f_1 f_2) \\ &\leq B_1 d(w_1, f_1) + B_2 d(w_2, f_2) \leq \epsilon \end{aligned} \tag{40}$$

for $\epsilon = B_1\epsilon_1 + B_2\epsilon_2$, which implies that $\mathcal{W}$ is an ϵ-cover of $\mathcal{F}$ of size $|\mathcal{W}_1| \cdot |\mathcal{W}_2|$. □

3.2 Pseudo-dimension and covering numbers of CV circuits

In this Section we bound the pseudo-dimension or covering number of function classes representing probability distributions that arise from manipulating and measuring CV classes. These bounds, together with the learning theorems for classical probability distributions of Sec. 3.1, immediately provide the sample complexity bounds given in Tables 1, 2.

3.2.1 Gaussian circuits

Consider an unknown n-mode CV state ρ and suppose that we can probe it by applying channels and measurement operators in the Gaussian class

$$\mathcal{S}_{\mathrm{g}}(n) := \{(\Phi_{\mathbf{d},X,Y}, M_{\mathbf{m}',V'})\}, \tag{41}$$

and obtaining measurement outcomes. We want to "imitate" these outcomes by finding an approximation of ρ in the class of Gaussian states

$$\mathcal{C}_{\mathrm{g}}(n) := \{\sigma_{\mathbf{m},V}\}. \tag{42}$$

We can tackle this state learning problem by defining a class of functions that take as input Gaussian channels and measurements and output their corresponding "acceptance" probability on a Gaussian state:

$$\begin{aligned} \mathcal{F}_{\mathrm{g}}(n) := \{f_{\mathbf{m},V} : (\mathbf{d}, X, Y, \mathbf{m}', V') \\ \mapsto P_{\mathrm{g}}(M_{\mathbf{m}',V'} | \Phi_{\mathbf{d},X,Y}, \sigma_{\mathbf{m},V})\}, \end{aligned} \tag{43}$$

where different functions are effectively parametrized by different Gaussian states. The key ingredient in bounding the SC for this learning problem is a calculation of the pseudo-dimension of $\mathcal{F}_{\mathrm{g}}(n)$:

Theorem 5. $\mathrm{Pdim}(\mathcal{F}_{\mathrm{g}}(n)) = O\left(n^2 \log n\right)$.

Proof. Define the function classes

$$\begin{aligned} \mathcal{F}_{\mathrm{e}} := \{f_{\mathbf{m},V} : (\mathbf{d}, X, Y, \mathbf{m}', V') \\ \mapsto (\mathbf{m}_{\mathrm{out}} - \mathbf{m}')^T (V_{\mathrm{out}} + V')^{-1} (\mathbf{m}_{\mathrm{out}} - \mathbf{m}')\}, \end{aligned} \tag{44}$$

and

$$\mathcal{F}_{\mathrm{d}} := \{f_V : (X, Y, V') \mapsto \mathrm{Det}\left(2\pi(V_{\mathrm{out}} + V')\right)\}, \tag{45}$$

and note that

$$\begin{aligned} \log \mathcal{F}_{\mathrm{g}} \subset \mathcal{F}_{\mathrm{e}} + \frac{1}{2} \log \mathcal{F}_{\mathrm{d}} := \{f_1 + f_2 \\ \mid f_1 \in \mathcal{F}_{\mathrm{e}}, f_2 \in \frac{1}{2} \log \mathcal{F}_{\mathrm{d}}\}, \end{aligned} \tag{46}$$

where the sum of two function classes contains the sum of all couples of functions from each class, while $g(\mathcal{F})$ is the function class obtained by applying the function $g(\cdot)$ to every element of $\mathcal{F}$. Hence we have

$$\mathrm{Pdim}(\mathcal{F}_{\mathrm{g}}) = \mathrm{Pdim}(\log \mathcal{F}_{\mathrm{g}}) < \mathrm{Pdim}\left(\mathcal{F}_{\mathrm{e}} + \frac{1}{2}\log \mathcal{F}_{\mathrm{d}}\right) \leq O(\mathrm{Pdim}(\mathcal{F}_{\mathrm{e}}) + \mathrm{Pdim}(\mathcal{F}_{\mathrm{d}})), \quad (47)$$

where we have used the properties of pseudo-dimension under composition with monotonous functions (Lemma 2) and sum (Lemma 3).

We are then left to evaluate the dimension of the two function classes (44,45), which can be done observing that they comprise functions that are polynomial in the parameters and applying Lemma 4.

The function class $\mathcal{F}_{\mathrm{d}}$ is characterized by the $n(2n+1)$ parameters $\{V_{i,j} : i \geq j\}$. A function $f(X, Y, V') \in \mathcal{F}_{\mathrm{d}}$ is related to these parameters via the determinant of the matrix $V_{\mathrm{out}} + V' \in \mathrm{Sym}_{2n}(\mathbb{R})$; here V' is the covariance matrix of the Gaussian measurement, therefore an input to the function itself, while V_{out} is the output covariance matrix after the Gaussian channel, therefore it depends both on the inputs X, Y and on the parameters as given in (17): $V_{\mathrm{out}} = XVX^T + Y$. The determinant implies that f is a polynomial of order $2n$ in the entries of V_{out}, which are in linear relation with the entries of V, i.e., the function parameters. We conclude that any $f \in \mathcal{F}_{\mathrm{d}}$ is a polynomial function of order $2n$ in the real parameters $\{V_{i,j} : i \geq j\}$, and by applying Lemma 4 with $\ell = 2n$ and $d = n(2n+1)$ we obtain

$$\mathrm{Pdim}(\mathcal{F}_{\mathrm{d}}) \leq 2(2n^2 + n)\log(24n). \quad (48)$$

The function class $\mathcal{F}_{\mathrm{e}}$ instead is characterized by the same parameters of $\mathcal{F}_{\mathrm{d}}$ plus the $2n$ parameters $\mathbf{m}$, for a total of $2n^2+3n$ parameters. The dependence on these parameters is via the function $(\mathbf{m}_{\mathrm{out}} - \mathbf{m}')^T(V_{\mathrm{out}} + V')^{-1}(\mathbf{m}_{\mathrm{out}} - \mathbf{m}')$. First, we focus on the matrix inverse $(V_{\mathrm{out}} + V')^{-1}$, whose elements are given by

$$((V_{\mathrm{out}} + V')^{-1})_{i,j} = \frac{M_{i,j}}{\mathrm{Det}\,(V_{\mathrm{out}} + V')}, \quad (49)$$

where $M_{i,j}$ are the minors of $V_{\mathrm{out}} + V'$, i.e., polynomials of order $2n-1$ in the matrix elements. We can therefore further decompose this class as

$$\log \mathcal{F}_{\mathrm{e}} = \log \mathcal{F}_{\mathrm{quad}} + \log \mathcal{F}_{\mathrm{d}}, \quad (50)$$

with

$$\mathcal{F}_{\mathrm{quad}} := \{f_{\mathbf{m},V} : (\mathbf{d}, X, Y, \mathbf{m}', V') \mapsto (\mathbf{m}_{\mathrm{out}} - \mathbf{m}')^T (M_{i,j})_{i,j=1,\cdots,2n} (\mathbf{m}_{\mathrm{out}} - \mathbf{m}')\}. \quad (51)$$

Functions in $\mathcal{F}_{\mathrm{quad}}$ depend on the parameters $\{V_{i,j} : i \geq j\} \cup \{m_i\}$ via polynomials of order $2n-1$, i.e., the minors, plus $2n$, i.e., the vector of mean values $\mathbf{m}_{out} = X\mathbf{m}$ (as per (17)), for a total degree of $4n-1$. Hence, applying Lemma 4 with $\ell = 4n-1$ and $d = 2n^2+3n$ we conclude that $\mathrm{Pdim}(\mathcal{F}_{\mathrm{quad}}) \leq 2(2n^2+3n)\log(12(4n-1))$. Putting this together with (47,48,50) and applying again Lemmas 2 and 3 we obtain the theorem statement:

$$\mathrm{Pdim}(\mathcal{F}_{\mathrm{g}}) \leq O(\mathrm{Pdim}(\mathcal{F}_{\mathrm{quad}}) + 2\mathrm{Pdim}(\mathcal{F}_{\mathrm{d}})) \leq O(n^2 \log n). \quad (52)$$

□

If instead we are interested in learning a Gaussian measurement operator or a Gaussian channel, the corresponding function classes comprise the same functions of $\mathcal{F}_{\mathrm{g}}$, though with exchanged roles of variables and parameters. For example, for measurement learning we have to consider functions

$$f_{\mathbf{m}',V'} : (\mathbf{m}, V, \mathbf{d}, X, Y) \mapsto P_{\mathrm{g}}(M_{\mathbf{m}',V'}|\Phi_{\mathbf{d},X,Y}, \sigma_{\mathbf{m},V}), \tag{53}$$

while for channel learning

$$f_{\mathbf{d},X,Y} : (\mathbf{m}, V, \mathbf{m}', V') \mapsto P_{\mathrm{g}}(M_{\mathbf{m}',V'}|\Phi_{\mathbf{d},X,Y}, \sigma_{\mathbf{m},V}). \tag{54}$$

It is straightforward to see that the pseudo-dimension of these classes is still bounded by $O(n^2 \log n)$. Indeed, functions parametrized by measurements coincide with those parametrized by states, while functions parametrized by channels still have a number of parameters $O(n^2)$, i.e., the matrix elements of X, Y, and depend on polynomials of order $O(n)$ in these variables, e.g., the dependence of V_{out} on X is quadratic, implying a polynomials of order $2n$ in the matrix elements of X.

3.2.2 Gaussian circuits plus photodetection

Consider now the sample class of n-mode Gaussian channels plus photodetection measurements

$$\mathcal{S}_{\mathrm{gp}}(n, K) := \{(\Phi_{\mathbf{d},X,Y}, M_q)\}, \tag{55}$$

where M_q is a coarse-grained photon-number measurement operator and K is the maximum photon-number cutoff (see Sec. 2.2.2). Define then the function class of outcome probabilities obtained by applying a channel and measurement from $\mathcal{S}_{\mathrm{gp}}(n, K)$ to a state in $\mathcal{C}_{\mathrm{g}}(n)$:

$$\begin{aligned}\mathcal{F}_{\mathrm{gp}}(n, K) := \{f_{\mathbf{m},V} : (\mathbf{d}, X, Y, q) \\ \mapsto P_{\mathrm{gp}}(M_q|\Phi_{\mathbf{d},X,Y}, \sigma_{\mathbf{m},V})\}.\end{aligned} \tag{56}$$

We can compute the pseudo-dimension of this class as well:

Theorem 6. $\mathrm{Pdim}(\mathcal{F}_{\mathrm{gp}}(n, K)) = O(n^2 \log(nK))$.

Proof. Following the proof above, we observe that

$$\mathrm{Pdim}(\mathcal{F}_{\mathrm{gp}}) \leq O(\mathrm{Pdim}(\mathcal{F}_{\mathrm{g}}) + \mathrm{Pdim}(\mathcal{F}_{\mathrm{p}})), \tag{57}$$

with $\mathcal{F}_{\mathrm{e,d}}$ as before and

$$\mathcal{F}_{\mathrm{p}} := \left\{ f_{\mathbf{m},V} : (\mathbf{d}, X, Y, q) \mapsto \sum_{\mathbf{k}} q_{\mathbf{k}} H_{\mathbf{k}}^{(\tilde{V})}(\tilde{\mathbf{m}}) \right\}. \tag{58}$$

Hence we only need to compute the pseudo-dimension of the latter class. Note that $\sum_{\mathbf{k}} q_{\mathbf{k}} H_{\mathbf{k}}^{(\tilde{V})}(\tilde{\mathbf{m}})$ is a polynomial of order nK in the vector $\tilde{V}\tilde{\mathbf{m}}$, by definition. In turn, both the matrix $\tilde{V}$ and the vector $\tilde{m}$ depend quadratically on V_{out} and m_{out} (see 2.2.2 and [42]), which are linear functions of the parameters $\{V_{i,j} : i \geq j\} \cup \{m_i\}$, as discussed before. Hence $\tilde{V}\tilde{\mathbf{m}}$ is a polynomial of order 4 in the parameters. We conclude that a function $f_{\mathbf{m},V} \in \mathcal{F}_{\mathrm{p}}$ is a polynomial of order $4nK$ in its $2n^2 + 3n$ parameters and we conclude

$$\mathrm{Pdim}(\mathcal{F}_{\mathrm{p}}) \leq 2(2n^2 + 3n) \log(28nK). \tag{59}$$

Combining this pseudo-dimension with Theorem 5 and (57) we obtain the theorem statement. □

If we consider an analogous function class for channel learning, comprising functions

$$f_{\mathbf{d},X,Y} : (\mathbf{m}, V, q) \mapsto P_{\text{gp}}(M_q|\Phi_{\mathbf{d},X,Y}, \sigma_{\mathbf{m},V}), \tag{60}$$

the pseudo-dimension stays the same. For measurement learning instead we have functions

$$f_q : (\mathbf{m}, V, \mathbf{d}, X, Y) \mapsto P_{\text{gp}}(M_q|\Phi_{\mathbf{d},X,Y}, \sigma_{\mathbf{m},V}) \tag{61}$$

which are linear in the parameters $q_{\mathbf{k}}$. However the number of such parameters is $(K+1)^n$ and therefore the pseudo-dimension is exponential in the number of modes.

3.2.3 Generalized Gaussian circuits

Consider now the hypothesis class of n-mode GG states

$$\mathcal{C}_{\text{gg}}(n) := \{\rho_{\{c,\mathbf{m},V\}_{\mathcal{I}}}\} \tag{62}$$

and the sample class of n-mode GG channels and measurements

$$\mathcal{S}_{\text{gg}}(n) := \{(\Phi_{\{c'',X,Y\}_{\mathcal{I}''}}, M_{\{c',\mathbf{m}',V'\}_{\mathcal{I}'}})\}. \tag{63}$$

Define then the function class of outcome probabilities obtained by applying a channel and measurement from $\mathcal{S}_{\text{gg}}(n)$ to a state in $\mathcal{C}_{\text{gg}}(n)$:

$$\begin{aligned}
\mathcal{F}_{\text{gg}}(n) := \Bigg\{ & f_{\{c,\mathbf{m},V\}_{\mathcal{I}}} : (\{c'', X, Y\}_{\mathcal{I}''}, \{c', \mathbf{m}', V'\}_{\mathcal{I}'}) \\
& \mapsto P_{\text{gg}}(M_{\{c',\mathbf{m}',V'\}_{\mathcal{I}'}}|\Phi_{\{c'',x,Y\}_{\mathcal{I}''}}, \sigma_{\{c,\mathbf{m},V\}_{\mathcal{I}}}) \Big| \ \forall i,j,k \\
& B_1 \geq \left|(\mathbf{m}_{\text{out}}^{(ijk)} - \mathbf{m}'_j)^T (V_{\text{out}}^{(ik)} + V'_j)^{-1} (\mathbf{m}_{\text{out}}^{(ijk)} - \mathbf{m}'_j)\right|, \\
& B_2 \geq \sum_{i,j,k} \left|c_i c'_j c''_k\right|, \ B_3 \leq \left|\text{Det}\left(2\pi\left(V_{\text{out}}^{(ik)} + V'_j\right)\right)\right|^{\frac{1}{2}} \Bigg\},
\end{aligned} \tag{64}$$

where $B_{1,2,3}$ can be interpreted as mild constraints on the absolute values of the complex parameters and variables determining GG states, channels and measurements that concur to the creation of the class. Note that these constraints are connected with the degree of non-Gaussianity of the GG state or measurement, e.g., B_2 controls the number of components in the sum, while $B_{1,3}$ control the strength of the various terms in the sum; see Appendix 4.1 for a more explicit interpretation of the first constraint and Appendix 4.2 for an evaluation of these constraints for relevant classes of GG states, i.e., GKP, cat and Fock states.

It is easy to show that the pseudo-dimension of $\mathcal{F}_{\text{gg}}$ is infinite, due to the potentially infinite number of terms in the complex-combination. Yet, we can show that its covering number is still finite, provided that $B_{1,2,3}$ are finite and that the coefficients of the GG measurements and channels we are learning are fixed. This further constraint is also mild, in the sense that it does not trivialize the learning problem. It amounts to fixing some properties of the setup of the device, i.e., evolution and measurement, that is employed to test the unknown state.

For example, for the most relevant cases in [29], fixing the GG measurement coefficients amounts to *fixing a specific projection state* $|\psi\rangle\langle\psi|$ in the GG class, e.g., a well-defined

GKP, cat or Fock state; an admissible class of measurement operators will thus have the form

$$M(U) = U^\dagger|\psi\rangle\langle\psi|U, \tag{65}$$

where U are Gaussian unitaries labelling the measurement outcome, e.g., displacement operators $U = D(\mathbf{r})$. Generalizing this construction, another example are measurement operators of the form $M(U) = U^\dagger M_0 U$, where now M_0 is a fixed GG operator, not necessarily projective nor rank-1.

Instead, fixing GG channel coefficients amounts to consider conditional dynamics based on partial measurements with *fixed non-Gaussian components*, e.g.,

$$\Phi(\rho) = \sum_i \Phi_A^{(i)}(\mathrm{Tr}_S[M_A^{(i)} V_{SA}(\rho_S \otimes \sigma_A) V_{SA}^\dagger]), \tag{66}$$

where σ and $\{M^{(i)}\}_i$ are respectively a GG state and measurement on an ancilla register A, coupled to the input system S via a Gaussian unitary V, and $\Phi^{(i)}$ are Gaussian quantum channels acting on the system conditional on the measurement outcome i. An admissible class of GG channels of this form is obtained by fixing $\{M^{(i)}\}_i$ and σ, while varying V and $\{\Phi^{(i)}\}_i$. This is fine for most practical applications, where non-Gaussian evolution is based on a well-defined measurement and ancilla, e.g., Fock damping, squeezed ancilla-assisted gates and the GKP T-gate mentioned in Ref. [29].

Theorem 7. *Let $\{c'_j\}_{j\in\mathcal{I}'}$, $\{c''_k\}_{k\in\mathcal{I}''}$ be fixed and $\mathcal{N}_2(\epsilon, \mathcal{F}_{\mathrm{gg}}, k)$ be the covering number with respect to the Euclidean distance. Then*

$$\log \mathcal{N}_2(\epsilon, \mathcal{F}_{\mathrm{gg}}, k) \leq \left\lceil \left(\frac{2B}{\epsilon}\right)^2 \right\rceil d \log \frac{2ek\tilde{B}}{d\epsilon}, \tag{67}$$

with $d = O(n^2 \log n)$, $B = \frac{B_2}{B_3}$ and $\tilde{B} = O(\log(B_1 B))$.

Proof. Since the outcome probability (24) is a real function we can rewrite it as $P_{\mathrm{gg}} = \mathrm{Re}P_{\mathrm{gg}}$, obtaining

$$P_{\mathrm{gg}} = B_2 \sum_{ijk} \mathrm{Re}\left[p_{ijk} \frac{e^{-R_{ijk}}}{M_{ijk}} e^{-\mathrm{i}(I_{ijk} + A_{ijk} - \theta_{ijk})} \right] \tag{68}$$

$$= B_2 \sum_{ijk} p_{ijk} \frac{e^{-R_{ijk}}}{M_{ijk}} \cos(I_{ijk} + A_{ijk} - \theta_{ijk}) \tag{69}$$

where in the first equality we have defined the real and imaginary parts of the Gaussian's exponent,

$$R_{ijk} := \frac{1}{2}\mathrm{Re}\left[(\mathbf{m}_{\mathrm{out}}^{(ijk)} - \mathbf{m}'_j)^T (V_{\mathrm{out}}^{(ik)} + V'_j)^{-1} (\mathbf{m}_{\mathrm{out}}^{(ijk)} - \mathbf{m}'_j) \right], \tag{70}$$

$$I_{ijk} := \frac{1}{2}\mathrm{Im}\left[(\mathbf{m}_{\mathrm{out}}^{(ijk)} - \mathbf{m}'_j)^T (V_{\mathrm{out}}^{(ik)} + V'_j)^{-1} (\mathbf{m}_{\mathrm{out}}^{(ijk)} - \mathbf{m}'_j) \right], \tag{71}$$

the radial and angular parts of the Gaussian's normalization,

$$M_{ijk} := \left| \mathrm{Det}\left(2\pi \left(V_{\mathrm{out}}^{(ik)} + V'_j \right) \right) \right|^{\frac{1}{2}}, \tag{72}$$

$$A_{ijk} := \frac{1}{2} \arccos \frac{\mathrm{Re}\,\mathrm{Det}\left(2\pi \left(V_{\mathrm{out}}^{(ik)} + V'_j \right) \right)}{M_{ijk}}, \tag{73}$$

and the radial and angular parts of the product of the complex combination coefficients,

$$p_{ijk} := \frac{|c_i c'_j c''_k|}{B_2}, \quad \theta_{ijk} := \arccos \frac{\mathrm{Re}(c_i c'_j c''_k)}{B_2\, p_{ijk}}, \tag{74}$$

the former rescaled so that $\sum_{i,j,k} p_{ijk} \leq 1$. Let us then define the function classes

$$\mathcal{F}_{\mathrm{exp}} := \left\{ f_{(\mathbf{m},V)_i} : ((X,Y)_k, (\mathbf{m}', V')_j) \mapsto \frac{e^{-R_{ijk}}}{M_{ijk}} \right\}, \tag{75}$$

$$\mathcal{F}_{\mathrm{im}} := \left\{ f_{(\mathbf{m},V)_i} : ((X,Y)_k, (\mathbf{m}', V')_j) \mapsto I_{ijk} \right\}, \tag{76}$$

$$\mathcal{F}_{\mathrm{ang}} := \left\{ f_{(\mathbf{m},V)_i} : ((X,Y)_k, (\mathbf{m}', V')_j) \mapsto A_{ijk} \right\}, \tag{77}$$

$$\mathcal{F}_{\mathrm{const}} := \{\theta_{ijk}\}, \quad \mathcal{F}_{\mathrm{trig}} := \cos(\mathcal{F}_{\mathrm{im}} + \mathcal{F}_{\mathrm{ang}} + \mathcal{F}_{\mathrm{const}}), \tag{78}$$

where the latter makes use of the sum and composition operations defined before. In particular, note that each value of θ_{ijk} determines a different function of $\mathcal{F}_{\mathrm{trig}}$ and that we can assume without loss of generality $\theta_{ijk} \in [0, 2\pi)$.

Putting all these definitions together, we conclude that $\mathcal{F}_{\mathrm{gg}} \subseteq B_2 \mathfrak{C}(\mathcal{F}_{\mathrm{exp}} \cdot \mathcal{F}_{\mathrm{trig}})$, where $\mathfrak{C}$ represents the convex-hull operation, and the product of two function classes contains all the products of two functions, one from each class. We stress that the convex-hull coefficients p_{ijk} can depend exclusively on the parameters of the function class and not on the input variables, which obliges us to fix the channel and measurement coefficients in the assumptions of the Theorem. Then we can apply Lemmas 6,7 to prove that

$$\begin{aligned} \log \mathcal{N}_2(\epsilon, \mathcal{F}_{\mathrm{gg}}, k) &\leq \left\lceil \left(\frac{2B_2}{\epsilon B_3} \right)^2 \right\rceil \log \mathcal{N}_2 \left(\frac{\epsilon}{2B_2}, \mathcal{F}_{\mathrm{exp}} \cdot \mathcal{F}_{\mathrm{trig}}, k \right) \\ &\leq \left\lceil \left(\frac{2B_2}{\epsilon B_3} \right)^2 \right\rceil \log(\mathcal{N}_2\left(\tilde{\epsilon}, \mathcal{F}_{\mathrm{exp}}, k\right) \mathcal{N}_2\left(\tilde{\epsilon}, \mathcal{F}_{\mathrm{trig}}, k\right)). \end{aligned} \tag{79}$$

where $\tilde{\epsilon} = \frac{\epsilon B_3}{2B_2(1+B_3)}$, and we have used the fact that functions in $\mathcal{F}_{\mathrm{exp}}$ are bounded by $1/B_3$, while those in $\mathcal{F}_{\mathrm{trig}}$ by 1. Note furthermore that, by Lemma 5 and since $\cos(x)$ is 1-Lipschitz, it holds $\mathcal{N}_2\left(\tilde{\epsilon}, \mathcal{F}_{\mathrm{trig}}, k\right) \leq \mathcal{N}_2\left(\tilde{\epsilon}, \mathcal{F}_{\mathrm{im}} + \mathcal{F}_{\mathrm{ang}} + \mathcal{F}_{\mathrm{const}}, k\right)$. We then proceed to evaluate the pseudo-dimensions of $\mathcal{F}_{\mathrm{exp}}$, $\mathcal{F}_{\mathrm{im}}$, $\mathcal{F}_{\mathrm{ang}}$ and $\mathcal{F}_{\mathrm{const}}$, which in turn will allow to bound their covering numbers via Eq. (36).

For the first class we can follow the same steps of Theorem 5, observing that $\log \mathcal{F}_{\mathrm{exp}} \subset \mathrm{Re}\tilde{\mathcal{F}}_{\mathrm{e}} + \frac{1}{4} \log |\tilde{\mathcal{F}}_{\mathrm{d}}|^2$, with $\tilde{\mathcal{F}}_{\mathrm{e,d}}$ generalizing the classes of Eqs. (44,45) to complex-valued parameters and variables, without symmetry constraints. This implies that the functions in $\tilde{\mathcal{F}}_{\mathrm{d}}$ and $\tilde{\mathcal{F}}_{\mathrm{e}}$ are polynomials in, respectively, $8n^2$ and $2(4n^2 + 2n)$ real parameters, of the same order as $\mathcal{F}_{\mathrm{d}}$ and $\mathcal{F}_{\mathrm{e}}$. Furthermore, note that the degree of a complex polynomial does not change by taking its real part, while it doubles by taking its radial component squared. Therefore we conclude that the functions in $\mathrm{Re}\tilde{\mathcal{F}}_{\mathrm{e}}$ are polynomials of degree $4n + 1$, implying

$$\mathrm{Pdim}(\mathrm{Re}\tilde{\mathcal{F}}_{\mathrm{e}}) \leq 4(4n^2 + 2n) \log(12(4n + 1)), \tag{80}$$

while the functions in $|\tilde{\mathcal{F}}_{\mathrm{d}}|^2$ are polynomials of degree $4n$, implying

$$\mathrm{Pdim}(|\tilde{\mathcal{F}}_{\mathrm{d}}|^2) \leq 16n^2 \log(48n), \tag{81}$$

so we conclude that $\mathrm{Pdim}(\mathcal{F}_{\mathrm{exp}}) = O(n^2 \log n) =: d$. We can now apply Eq. (36), observing that the functions in $\mathcal{F}_{\mathrm{exp}}$ have range $[0, \frac{1}{B_3}]$, obtaining

$$\log \mathcal{N}_2\left(\tilde{\epsilon}, \mathcal{F}_{\mathrm{exp}}, k\right) \leq d \log \frac{2ek}{d\tilde{\epsilon} B_3}. \tag{82}$$

We proceed similarly to bound the covering number of $\mathcal{F}_{\text{im}} + \mathcal{F}_{\text{ang}} + \mathcal{F}_{\text{const}}$. Firstly, it is well-known that $\text{Pdim}(\mathcal{F}_{\text{const}}) = 1$; secondly, observe that $\mathcal{F}_{\text{im}} = \text{Im}\tilde{\mathcal{F}}_{\text{e}}$, hence we conclude immediately $\text{Pdim}(\mathcal{F}_{\text{im}}) = \text{Pdim}(\text{Re}\tilde{\mathcal{F}}_{\text{e}})$. The class $\mathcal{F}_{\text{ang}}$ instead requires a bit more work: using the fact that arccos (restricted to the interval $[-1, 1]$) and log are strictly monotonous functions we have, applying again Lemma 2, that

$$\begin{aligned}
\text{Pdim}(\mathcal{F}_{\text{ang}}) &\leq \text{Pdim}(\text{Re}\tilde{\mathcal{F}}_{\text{d}} \cdot |\tilde{\mathcal{F}}_{\text{d}}|^{-\frac{1}{2}}) && (83)\\
&\leq \text{Pdim}(\log \text{Re}\tilde{\mathcal{F}}_{\text{d}} - \frac{1}{4}\log |\tilde{\mathcal{F}}_{\text{d}}|^2) && (84)\\
&\leq O(\text{Pdim}(\text{Re}\tilde{\mathcal{F}}_{\text{d}}) + \text{Pdim}(|\tilde{\mathcal{F}}_{\text{d}}|^2)). && (85)
\end{aligned}$$

It is then easy to show that $\text{Pdim}(\text{Re}\tilde{\mathcal{F}}_{\text{d}}) = 16n^2 \log(24n)$ and conclude that $\text{Pdim}(\mathcal{F}_{\text{ang}}) = O(n^2 \log n)$ as well. Finally, we apply Eq. (36), observing that for any function $f \in (\mathcal{F}_{\text{im}} + \mathcal{F}_{\text{ang}} + \mathcal{F}_{\text{const}})$ it holds $|f| \leq |I_{ijk} + A_{ijk} + \theta_{ijk}| \leq B_1 + 3\pi$, so that

$$\log \mathcal{N}_2 (\tilde{\epsilon}, \mathcal{F}_{\text{trig}}, k) \leq d \log \frac{2e(B_1 + 9)k}{d\tilde{\epsilon}}. \tag{86}$$

By combining Eqs. (79,82,86) we obtain the claim. □

3.3 SC bounds for CV state, measurement and channel learning

The quantities computed in Sec. 3.2 can be straightforwardly employed to obtain SC bounds for the state, measurement and channel learning problems described in Sec. 2.1.1. The simplest case is that where only Gaussian or GP circuit components are involved:

Proof. (Theorem 1 for Gaussian or GP circuits) The function classes $\mathcal{F}_{\text{g}}$, Eq. (43), and $\mathcal{F}_{\text{gp}}$, Eq. (56), arise when trying to learn Gaussian states with Gaussian or GP channels and measurements. The SC of these tasks can then be evaluated by applying Theorem 1 to the classes $\mathcal{F}_{\text{g,gp}}$, concluding that the following number of samples is sufficient for learning

$$\begin{aligned}
T_{\text{g}} &= O\left(\frac{1}{\nu^4\epsilon^2}\left(n^2 \log(n) \log^2 \frac{1}{\nu\epsilon} + \log \frac{1}{\delta}\right)\right), && (87)\\
T_{\text{gp}} &= O\left(\frac{1}{\nu^4\epsilon^2}\left(n^2 \log(nK) \log^2 \frac{1}{\nu\epsilon} + \log \frac{1}{\delta}\right)\right), && (88)
\end{aligned}$$

where we employed $\text{fat}_{\mathcal{F}}(\cdot) \leq \text{Pdim}(\mathcal{F})$ and the pseudo-dimension bounds of Theorems 5,6. □

An analogous proof technique can be employed for channel or measurement learning. Let us note that these scalings are perhaps not particularly surprising, since tomography of Gaussian states and processes also requires a number of measurements scaling polynomially with the number of modes. Indeed, for an n-mode Gaussian state $\rho_{\mathbf{m},V}$, the mean $\mathbf{m}$ can be estimated with error ϵ via $O(n\epsilon^{-2})$ single-mode homodyne measurements, while V requires additional $O(n^2\epsilon^{-2})$ homodyne measurements. Therefore the SC complexity for learning a Gaussian state using Gaussian or photodetection measurements seems comparable with that of performing full tomography.

On the other hand, the surprising result is that a similar scaling is obtained when the state is not Gaussian, but rather GG, as proved below:

Proof. (Theorem 1 for GG circuits) In this case we would need to apply Theorem 1 to the function class $\mathcal{F}_{\text{gg}}$ (64). However, we only have a bound on its covering number and not on its fat-shattering or pseudo-dimension.

We then need to repeat the proof of [45, Theorem 12], upon which the SC of Theorems 1,4 relies; indeed, we can apply directly our covering number bound, right after the following statement is proved (Theorem 12 thereof):

$$\begin{aligned}\Pr_{\{(x_t,b_t)\}\sim(Q\cdot P)^{\times T}}&\left[\left|\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(x_t,b_t)-\mathbb{E}\mathcal{L}(x,b)\right|\geq\epsilon'\right]\\&\leq 4\mathcal{N}_{\tilde{1}}\left(\frac{\epsilon'}{2}-\alpha,\mathcal{F},2T\right)e^{-\alpha^2\frac{T}{2}},\end{aligned}\tag{89}$$

where $\alpha=\lceil 1/(\epsilon'\kappa)\rceil^{-1}$, $0<\kappa\leq 1/4$. Note that we can take $\mathcal{L}(x,b)=(f(x)-b)^2$ for all $f\in\mathcal{F}_{\text{gg}}$ to recover our learning problem (as shown in [14, Theorem 8.1]).

Importantly, the covering number $\mathcal{N}_{\tilde{1}}$ is computed with respect to a rescaled 1-norm distance, $d_{\tilde{1}}(\mathbf{x},\mathbf{y}):=\frac{1}{T}\sum_{t=1}^{T}|x_t-y_t|$, for $\mathbf{x},\mathbf{y}\in\mathcal{X}^{\times T}$, rather then the Euclidean distance $d_2(\mathbf{x},\mathbf{y})$ used in Eq. (67). We can bound this covering number as follows:

$$\begin{aligned}\mathcal{N}_{\tilde{1}}\Big(\frac{\epsilon'}{2}-\alpha,\mathcal{F},2T\Big)&\leq\mathcal{N}_{\tilde{1}}\left(\frac{\epsilon'}{4},\mathcal{F},2T\right)\leq\mathcal{N}_2\left(\frac{\epsilon'}{4},\mathcal{F},2T\right)\\&\leq\mathcal{N}_2\left(\frac{\epsilon'}{4},\mathcal{F}_{gg},2T\right)^2\qquad(90)\\&\leq\exp\left(2\left\lceil\left(\frac{8B}{\epsilon'}\right)^2\right\rceil d\log\frac{16eT\tilde{B}}{d\epsilon'}\right).\qquad(91)\end{aligned}$$

The first inequality above follows from the fact that the covering number is a decreasing function of the covering's size and that $\epsilon'/2-\alpha\geq\epsilon'/4$; the second one instead follows from the fact that covering numbers increase when computed with respect to a larger distance and clearly $d_{\tilde{1}}\leq d_2$; the third inequality follows from $\mathcal{F}=(\mathcal{F}_{gg})^2$ and Lemma 7; finally, the last one follows from Theorem 7 above, with $d=O(n^2\log n)$, $B=\frac{B_2}{B_3}$ and $\tilde{B}=O(\log(B_1B))$.

Using this bound and [45, Lemma 18], with $y_1=4$, $y_2=2d\left\lceil\left(\frac{8B_2}{\epsilon' B_3}\right)^2\right\rceil$, $y_3=\frac{16eB}{d\epsilon'}$ and $y_4=\frac{\alpha^2}{2}$, we can now obtain a sufficient T for learnability to hold, i.e., such that the right-hand side of Eq. (89) is bounded by δ:

$$T_{\text{gg}}=O\left(\frac{dB^2}{\epsilon'^4}\log\frac{B\tilde{B}}{\epsilon' d}+\frac{1}{\epsilon'^2}\log\frac{1}{\delta}\right),\tag{92}$$

which implies the Theorem's statement for $\epsilon'=\nu^2\epsilon$. □

Also in this case, the proof can be easily extended to the problems of channel and measurement learning. Importantly, note that a crucial condition for applying Theorem 7 is that the GG combination coefficients are fixed for the sample class. Hence, for example, we can guarantee efficient learnability of arbitrary GG states only with respect to GG channels and measurements with fixed coefficients. This covers interesting cases, e.g.: (i) Gaussian channels and measurements, (ii) channels generated by a Gaussian interaction with a fixed GG ancillary state and (iii) measurements generated by projecting on a fixed GG state after applying a Gaussian unitary.

The resulting SC is $O(n^2\epsilon^{-4})$, obtained by inserting in (92) $\epsilon'=O(\epsilon)$ and $d=O(n^2)$ and hiding logarithmic terms. Thus, the SC of learning this non-Gaussian class of states is only ϵ^{-2}-larger than the SC of learning Gaussian states or performing Gaussian state tomography. Furthermore, it is clear that Gaussian state tomography does not work at

all in this case, since it allows a correct reconstruction of just the mean and covariance matrix of the unknown GG state, which are not sufficient to reconstruct its measurement statistics.

3.4 SC bounds for CV circuit compilation

The techniques employed in Sec. 3.3 can also be applied to bound the SC of learning the optimal family of quantum circuits for a discrimination or synthesis task, as described in Sec. 2.1.2. However, there are two major differences: (i) we have to consider a prediction version of the problem, i.e., applying Theorem 4; (ii) we have to consider function classes of the form $\mathcal{F} \circ \mathcal{E}$, obtained via composition of an encoding class $\mathcal{E} := \{x \mapsto \boldsymbol{\gamma}(x)\}$, that maps input variables into circuit parameters, and a probability-valued function class, that maps circuits into outcome probabilities as before, i.e., $\mathcal{F}_{\text{g,gp,gg}}$. Indeed, in the case of task learnability, differently form the learning setting of Sec. 2.1.1, one can allow the freedom to optimize also a classical function that maps the input random variable x to a vector $\boldsymbol{\gamma}(x)$, which determines the Gaussian and non-Gaussian parameters of the hypotheses. For example, suppose we want to learn an optimal Gaussian measurement $\{M_x\}$ for a discrimination problem. Then, each M_x will be parametrized in terms of a mean-value vector and covariance matrix $(\mathbf{m}_x, V_x)$, but the specific dependence of these on the classical input x still requires to be specified. One can do so via the classical of functions $\boldsymbol{\gamma}(x)$, whose components give the components of $\mathbf{m}_x$ and matrix elements of V_x. In this way, one can effectively allow a compilation not only of the quantum circuit performing the measurement, but also of the classical post-processing of the emasurement outcomes.

Given the function classes $\mathcal{F}$ and $\mathcal{E}$, if the resulting function class $\mathcal{F} \circ \mathcal{E}$ also has bounded pseudo-dimension or covering number, then we obtain learnability also in this case. Unfortunately, the properties of such complexity measures under composition of classes are not simple to apply in our case, and hence we cannot provide a formula for general $\mathcal{E}$. Still, we can prove learnability for the specific class of polynomial encoding functions:

Proof. (Theorem 2) The SC of learning Gaussian or GP circuits for discrimination and synthesis can be computed by applying Theorem 4 to the function classes $\mathcal{F}_{\text{g,gp}} \circ \mathcal{E}_\ell$, where

$$\mathcal{E}_\ell := \{x \mapsto \boldsymbol{\gamma}(x) | \gamma_i(x) = \text{poly}_\ell(x) \forall i\} \tag{93}$$

is a class of encoding functions that are polynomials of order ℓ in the input x.

For their pseudo-dimension we can proceed as in Theorems 5,6, up until the point where we need to evaluate the pseudo-dimension of $\mathcal{F}_{\text{e,d,p}} \circ \mathcal{E}_\ell$, which are still classes of polynomials in their defining parameters. In particular, note that the encoding $\mathcal{E}_\ell$ contributes a total of ℓ additional parameters for each input of $\mathcal{F}$, which has always $O(n^2)$ inputs, and that each of these parameters appears with a linear dependence in the encoding function. The only thing we have to be careful about is counting the order of the original functions in the inputs. For example consider the class $\mathcal{F}_{\text{d}}$, which is a polynomial of order $2n$ in the $2n(2n+1)$ parameters $\{V_{i,j} : i \geq j\}$, as before, but also a polynomial of the same order in the variables $\{V'_{i,j} : i \geq j\}$ and a polynomial of order $4n$ in the variables $\{X, Y\}$. However, when we compose it with the encoding class, $\mathcal{F}_{\text{d}} \circ \mathcal{E}_\ell$, each of these variables becomes a parametrized function from $\mathcal{E}_\ell$, in the form of a polynomial of order 1 in ℓ additional parameters. Thus we conclude that the functions in $\mathcal{F}_{\text{d}} \circ \mathcal{E}_\ell$ are polynomials of order $O(n)$ in $O(\ell n^2)$ parameters. Similar considerations hold for the other two classes. Therefore, the overall modification to the pseudo-dimensions is the presence of $O(\ell n^2)$ parameters

and we obtain

$$T_{\rm g} = O\left(\frac{1}{\nu^4\epsilon^2}\left(\ell n^2 \log(n)\log^2\frac{1}{\nu\epsilon} + \log\frac{1}{\delta}\right)\right), \tag{94}$$

$$T_{\rm gp} = O\left(\frac{1}{\nu^4\epsilon^2}\left(\ell n^2 \log(nK)\log^2\frac{1}{\nu\epsilon} + \log\frac{1}{\delta}\right)\right). \tag{95}$$

In the case of GG circuits, the proof is analogous but we further need to restrict to encodings that map the input x to GG circuit parameters with fixed combination coefficients. □

Finally, note that the class of polynomial encodings covers several cases of interest, since polynomials of arbitrary degree can be used to approximate any desired function. We believe that similar results can be obtained for larger classes of encoding functions, e.g., neural networks, but leave this proof open for future works.

4 Discussion of learnability for GG states

4.1 Sufficient conditions for learnability of GG states

Here we manipulate further the B_1-constraint for GG learnability, clarifying its physical significance. For a generic complex-valued vector $\mathbf{x}$ and matrix A we have

$$|\mathbf{x}^T A\mathbf{x}| = \langle \mathbf{x}^*, A\mathbf{x}\rangle \leq ||\mathbf{x}^*|| \cdot ||A\mathbf{x}|| \tag{96}$$

$$\leq ||\mathbf{x}||^2 \cdot ||A||_\infty = ||\mathbf{x}||^2 \cdot \lambda_{\max}(|A|), \tag{97}$$

where $\langle \mathbf{s}, \mathbf{r}\rangle = \mathbf{s}^\dagger \mathbf{r}$ is the standard Euclidean inner product over the complex numbers and $||\mathbf{s}|| = \sqrt{\langle \mathbf{s}, \mathbf{s}\rangle}$ the induced norm. Futhermore, the first inequality is obtained via the Cauchy-Schwartz inequality, while the second one from the definition of the operator norm $||A||_\infty := \sup_{\mathbf{x}} ||A\mathbf{x}||/||\mathbf{x}||$. Applying this result with $\mathbf{x} = \mathbf{m}_{\rm out}^{(ijk)} - \mathbf{m}'_j$ and $A = (V_{\rm out}^{(ik)} + V'_j)^{-1}$ we obtain

$$I_{ijk} \leq ||\mathbf{m}_{\rm out}^{(ijk)} - \mathbf{m}'_j||^2 \cdot \lambda_{\min}(|V_{\rm out}^{(ik)} + V'_j|)^{-1}. \tag{98}$$

If the mean values $||\mathbf{m}_{\rm out}^{(ijk)}||, ||\mathbf{m}'_j||$ are bounded for each i, j, then the first term in the product above is also bounded; this condition implies that the SC increases when trying to learn a GG state with larger mean. As for the second term, it is upper-bounded whenever the minimum eigenvalue of $|V_{\rm out}^{(ik)} + V'_j|$ is lower-bounded. For example, using Weyl's inequality we obtain the bound

$$\lambda_{\min}(|V_{\rm out}^{(ik)} + V'_j|)^2 \geq \lambda_{\min}(V'_j V'^\dagger_j) \tag{99}$$

$$+ \lambda_{\min}(V_{\rm out}^{(ik)} V_{\rm out}^{(ik)\dagger}) + \lambda_{\min}(V'_j V_{\rm out}^{(ik)\dagger} + V_{\rm out}^{(ik)} V'^\dagger_j).$$

This condition is related to how much the single components of the GG state violate the uncertainty principle, since we know that the latter bounds the eigenvalues of the covariance matrix of a quantum state away from zero. Hence, the more unphysical the single GG state components are, the larger the SC required to learn it. Finally, note that a lower bound on $\lambda_{\min}(|A|)$ implies also a lower bound on $|\mathrm{Det}\,(A)\,|$ and therefore one could absorb the third constraint B_3 into the first one.

4.2 Learnability constraints for useful classes of GG states

Here we analyze in more detail the significance of the GG learnability constraints on specific kinds of non-Gaussian states that can be approximated by the GG class. We employ the analysis of [29, Section IV] and estimate the explicit dependence of $B_{1,2,3}$ in Theorem 7 on the states' parameters. Specifically, we need to obtain a lower bound on $\lambda_{\min}(|V_i|)$ and upper bounds on $||\mathbf{m}_i||$, $\sum_i |c_i|$, since then the constants have to scale at least as $B_1 = O(||\mathbf{m}_i|| \lambda_{\min}(|V_i|)^{-1})$, $B_2 = O(\sum_i |c_i|)$ and $B_3 = \Omega(\lambda_{\min}(|V_i|)^{\frac{1}{2}})$ with the parameters. Clearly, there is an extra dependence on the channel and measurement parameters that we don't analyze here, though we assume that these have bounded parameters and non-zero displacements.

Finite-energy GKP states the relevant parameters are the damping factor ϵ and the lattice size L, i.e., the maximum position or momentum allowed when truncating the state; the total number of terms in the sum is L^2. Then, from [29, Eq. 43] we obtain $\lambda_{\min}(V_i) \gtrsim \epsilon$, $||\mathbf{m}_i|| \lesssim (1-2\epsilon)2L$ at first order in ϵ, i.e., when the states approximate GKP states well, and $\sum_i |c_i| \leq L^2 ||\mathbf{m}_i|| \lesssim (1-2\epsilon)2L^3$. Therefore

$$B_1 = O(L\epsilon^{-1}),\ B_2 = O(L^3),\ B_3 = \Omega(\epsilon^{\frac{1}{2}}). \tag{100}$$

Cat states these can be expressed exactly as a GG state with four components [29, Eq. 55], with $V_i = \mathbb{1}$, $||\mathbf{m}_i|| \lesssim |\alpha|$ and $\sum_i |c_i| = 1$ for the "+" cat state, while $\sum_i |c_i| = (1+e^{-2|\alpha|^2})(1-e^{-2|\alpha|^2})^{-1}$ for the "-" cat state. Here α is the amplitude of the coherent states and a larger $|\alpha|$ identifies a "more macroscopic" superposition; hence the region of interest is that of large $|\alpha|$, where $\sum_i |c_i| \sim 1$ also for the "-" cat state. In conclusion, for large-energy cat states, i.e., $|\alpha|^2 \gg 1 + O(e^{-2|\alpha|^2})$, we have

$$B_1 = O(|\alpha|),\ B_2 = O(1),\ B_3 = \Omega(1). \tag{101}$$

Fock states for a cutoff K on the maximum photon number, we can approximate Fock states using squeezed GG states [29, Eq. 61], governed by a parameter $r < \frac{1}{\sqrt{K}}$, which approaches exact Fock states in the limit $r \to 0$. Then we have $\lambda_{\min}(|V_i|) \lesssim 1 + O(Kr^2)$, $||\mathbf{m}_i|| = 0$ and $\sum_i |c_i| \lesssim K e^{O(r^{-2}) + O(K \log r^{-2})}$. Therefore $B_1 = B_3 = O(1)$ while

$$B_2 = e^{O(r^{-2}) + O(K \log r^{-2}) + O(\log K)}, \tag{102}$$

which means that photon-number states can be learned with exponential difficulty in the photon-number K and the approximation precision r^{-2}, consistently with the predictions of Eq.(61).

5 Conclusions

The interpretation of our results relies on a good understanding of two key terms: (i) *learnability* means that we can find a quantum circuit that approximates an unknown process using only samples of random measurements performed on the latter; (ii) *efficiency* means that the number of samples needed to reach a good approximation is polynomial in the circuit's characteristics, e.g. its size.

In this article, we established efficient learnability for CV quantum circuits comprising Gaussian and certain non-Gaussian components. These circuits describe, for example,

photonic processors at the state-of-the-art, as well as more futuristic ones based on approximate cat or GKP states. Moreover, our results are platform-independent and hence also apply to other infinite-dimensional quantum circuits based, for example, on mechanical or cavity-coupled resonators.

Our learnability result for CV circuits applies to a variety of information-processing tasks of general interest: (i) approximating the measurement statistics of an unknown state or process and (ii) approximating the optimal circuit for state or process discrimination and for pure-state synthesis.

Furthermore, we believe that our methods can be extended to prove efficient learnability for CV circuits comprising CV states *with bounded P-function* and Gaussian or GG channels and measurements. Indeed, every CV state can be written as an affine combination of Gaussian coherent states via the P-representation [50]; hence the Gaussian or GG measurement statistics on this state is itself an affine combination of probabilities of the form Eqs. (18,24) and we can bound the covering number of these classes using the same methods of Theorem 7.

Finally, our results are a starting point for a systematic study of statistical learning theory in CV quantum systems, leaving several open problems of interest, including: learnability of general CV circuits (see comment above); sample-complexity lower bounds; explicit learning algorithms for CV circuits and their application to specific tasks (e.g., the discovery of optimal decoders for optical communication [51, 52, 53, 31] and state discrimination [54, 39] and the approximation of quantum processes with restricted resources [33]); analysis of more challenging learning settings, where one is allowed to choose the quantum measurements to be applied on the samples and has to guarantee the approximation of multiple measurement operators at once [26, 55, 56].

6 Acknowledgements

M.R. acknowledges support from the Einstein Foundation. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 845255 and the project PNRR - Finanziato dall'Unione europea - MISSIONE 4 COMPONENTE 2 INVESTIMENTO 1.2 - "Finanziamento di progetti presentati da giovani ricercatori" - Id MSCA_0000011-SQUID - CUP F83C22002390007 (Young Researchers) - Finanziato dall'Unione europea - NextGenerationEU.

A Agnostic learning of probabilistic concepts with point-wise guarantees

Here we prove a learning theorem for probabilistic concepts analogous to [14, Supp. Mat. Theorem 1], but dropping the assumption that the unknown concept f is contained in the hypothesis class $\mathcal{F}$. We use a quadratic loss to measure the distance of a hypothesis $h \in \mathcal{F}$ from the true concept on input x. In this agnostic setting $f \notin \mathcal{F}$, hence the aim is to ensure that, if one finds a hypothesis $h \in \mathcal{F}$ with small loss on the training set, then h generalizes well, i.e., its loss is small also on unseen inputs. For the proof we need the following theorem on function learning from interpolation by Anthony and Bartlett [46], which applies to the agnostic setting:

Theorem 8. *([46, Corollary 3.3]) Let $\mathcal{X}$ be a sample space, Q a distribution on $\mathcal{X}$ and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ a hypothesis class of real-valued functions. Consider an unknown concept $f : \mathcal{X} \mapsto \mathbb{R}$ (not necessarily in $\mathcal{F}$) and a training set $\{(x_i, f(x_i)\}_{i=1}^{T}$, where $x_i \sim Q$. Suppose*

that we choose a hypothesis $h \in \mathcal{F}$ with small loss on each element of the training set, i.e., $|h(x_i) - f(x_i)| < \eta$ for all $i = 1, \cdots, T$. Then with probability at least $1 - \delta$ over the sample it holds

$$\Pr_{x\sim Q}(|h(x) - f(x)| < \eta + \gamma) \geq 1 - \epsilon, \tag{103}$$

i.e., the hypothesis has small loss also on unseen instances of the random variable x, provided that

$$T \geq T_0(\epsilon, \gamma, \delta) = O\left(\frac{1}{\epsilon}\left(d \log^2 \frac{d}{\gamma\epsilon} + \log \frac{1}{\delta}\right)\right), \tag{104}$$

where $d = \mathrm{fat}_{\mathcal{F}}\left(\frac{\gamma}{8}\right)$, for all $\eta, \epsilon, \gamma, \delta > 0$.

The key difference of this Theorem with respect to [44, 45] is that it provides convergence guarantees for the point-wise loss, i.e., for all x, rather than for its expectation. This requirement is necessary for quantum state (measurement, and channel) learning à la Aaronson, where one wants to identify an operator that reproduces the observed statistics with small error on each input. Indeed, this problem corresponds to that of *learning a model of probability* introduced by Kearns and Schapire [43]. In contrast, the circuit learning tasks introduced in Sec. 2.1.2 only request a convergence guarantee for the expected loss, thus corresponding to the problem of *learning a decision rule* also introduced by Kearns and Schapire [43].

We are now ready to state and prove our theorem, relying on a useful trick by [14] to learn from measurement outcomes rather than probability values.

Theorem 9. *(Theorem 3 repeated) Let $\mathcal{X}$ be a sample space, Q a distribution on $\mathcal{X}$ and $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ a hypothesis class. Consider an unknown p-concept $f : \mathcal{X} \mapsto [0,1]$ (not necessarily in $\mathcal{F}$) and a training set $\{(x_i, b_i)\}_{i=1}^T$, where $x_i \sim Q$ and $b_i = 1$ with probability $f(x_i)$ or $b_i = 0$ otherwise. Suppose that we choose a hypothesis $h \in \mathcal{F}$ with small quadratic loss on each element of the training set, i.e., $(h(x_i) - b_i)^2 < \eta$ for all $i = 1, \cdots, T$. Then with probability at least $1 - \delta$ over the sample it holds*

$$\Pr_{x\sim Q}((h(x) - f(x))^2 < \eta + \gamma) \geq 1 - \epsilon \tag{105}$$

for all $\eta, \gamma, \epsilon, \delta > 0$, provided that

$$T \geq T_0(\epsilon, \gamma, \delta, \mathcal{F}) = O\left(\frac{1}{\epsilon}\left(d \log^2 \frac{d}{\gamma\epsilon} + \log \frac{1}{\delta}\right)\right), \tag{106}$$

where $d = \mathrm{fat}_{\mathcal{F}}\left(\frac{\gamma}{8}\right)$.

Proof. Following [14, Suppl. Mat. Theorem 1], starting from $\mathcal{F}$ we construct a suitable class of functions that take as input both random variables $x \in \mathcal{X}$ and $b \in [0,1]$ as follows:

$$\mathcal{F}^* = \{h^* : (x, b) \mapsto (h(x) - b)^2\}. \tag{107}$$

We can now apply Theorem 8 to the sample space $\mathcal{X} \times \{0,1\}$, with probability distribution $Q \cdot f(x)$ and true concept $f^*(x, b) = 0$. Suppose that a minimization algorithm finds a concept $h^* \in \mathcal{F}$ with η-small loss on each element of the training set, i.e., for all $i = 1, \cdots, T$ it holds

$$\eta > |h^*(x_i, b_i) - f^*(x_i, b_i)| = (h(x_i) - b_i)^2. \tag{108}$$

Then, if

$$T \geq T_0(\epsilon, \gamma, \delta, \mathcal{F}^*) = O\left(\frac{1}{\epsilon}\left(d \log^2 \frac{d}{\gamma\epsilon} + \log \frac{1}{\delta}\right)\right), \tag{109}$$

with $d = \mathrm{fat}_{\mathcal{F}^*}\left(\frac{\gamma}{8}\right)$, with probability at least $1-\delta$ over the sample it holds

$$1-\epsilon \leq \Pr_{x\sim Q, b\sim f(x)}(|h^*(x,b) - f^*(x,b)| < \eta+\gamma) \tag{110}$$
$$= \Pr_{x\sim Q}((h(x)-B)^2 < \eta+\gamma, B=b) \tag{111}$$
$$= \Pr_{x\sim Q}(\cup_{b\in\{0,1\}}(h(x)-b)^2 < \eta+\gamma) \tag{112}$$
$$\leq \Pr_{x\sim Q}(\mathbb{E}_{b\sim f(x)}(h(x)-b)^2 < \eta+\gamma) \tag{113}$$
$$\leq \Pr_{x\sim Q}((h(x)-f(x))^2 < \eta+\gamma), \tag{114}$$

where the first inequality follows from Theorem 8, the first equality from conditioning on the binary random variable B with distribution $f(x)$, and the second equality from the fact that the events $B=0$ and $B=1$ are disjoint. Furthermore, the second and third inequalities follow from the fact that $(E_1 \Rightarrow E_2)$ implies $\Pr(E_1) \leq \Pr(E_2)$ for any two events E_1, E_2. In particular, for the second inequality we observed that

$$(h(x)-b)^2 < \eta+\gamma \;\forall b\in\{0,1\} \Rightarrow \mathbb{E}_{b\sim f(x)}(h(x)-b)^2 < \eta+\gamma, \tag{115}$$

while for the third one we have

$$\eta+\gamma > \mathbb{E}_{b\sim f(x)}(h(x)-b)^2 = f(x)(1-h(x))^2 + (1-f(x))h(x)^2 \tag{116}$$
$$= (h(x)-f(x))^2 + f(x) - f(x)^2 \geq (h(x)-f(x))^2, \tag{117}$$

since $f(x)\in[0,1]$. Finally, we observe that $d = \mathrm{fat}_{\mathcal{F}^*}(\alpha) \leq 2\mathrm{fat}_{\mathcal{F}}(\alpha)$, obtaining the desired sample complexity. □

References

[1] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, Peng Hu, Xiao-Yan Yang, Wei-Jun Zhang, Hao Li, Yuxuan Li, Xiao Jiang, Lin Gan, Guangwen Yang, Lixing You, Zhen Wang, Li Li, Nai-Le Liu, Chao-Yang Lu, and Jian-Wei Pan. "Quantum computational advantage using photons". Science (80-.). **370**, 1460 LP – 1463 (2020).

[2] Lars S. Madsen, Fabian Laudenbach, Mohsen Falamarzi. Askarani, Fabien Rortais, Trevor Vincent, Jacob F. F. Bulmer, Filippo M. Miatto, Leonhard Neuhaus, Lukas G. Helt, Matthew J. Collins, Adriana E. Lita, Thomas Gerrits, Sae Woo Nam, Varun D. Vaidya, Matteo Menotti, Ish Dhand, Zachary Vernon, Nicolás Quesada, and Jonathan Lavoie. "Quantum computational advantage with a programmable photonic processor". Nature **606**, 75–81 (2022).

[3] Francesco Hoch, Simone Piacentini, Taira Giordani, Zhen-Nan Tian, Mariagrazia Iuliano, Chiara Esposito, Anita Camillini, Gonzalo Carvacho, Francesco Ceccarelli, Nicolò Spagnolo, Andrea Crespi, Fabio Sciarrino, and Roberto Osellame. "Reconfigurable continuously-coupled 3D photonic circuit for Boson Sampling experiments". npj Quantum Inf. **8**, 55 (2022). arXiv:2106.08260.

[4] J. Eli Bourassa, Rafael N Alexander, Michael Vasmer, Ashlesha Patil, Ilan Tzitrin, Takaya Matsuura, Daiqin Su, Ben Q Baragiola, Saikat Guha, Guillaume Dauphinais, Krishna K Sabapathy, Nicolas C Menicucci, and Ish Dhand. "Blueprint for a scalable photonic fault-tolerant quantum computer". Quantum **5**, 1–38 (2021). arXiv:2010.02905.

[5] Jianwei Wang, Fabio Sciarrino, Anthony Laing, and Mark G. Thompson. "Integrated photonic quantum technologies". Nat. Photonics **14**, 273–284 (2020).
[6] Stefan Krastanov, Victor V. Albert, Chao Shen, Chang-Ling Zou, Reinier W. Heeres, Brian Vlastakis, Robert J. Schoelkopf, and Liang Jiang. "Universal Control of an Oscillator with Dispersive Coupling to a Qubit" (2015). arXiv:1502.08015.
[7] Alec Eickbusch, Volodymyr Sivak, Andy Z. Ding, Salvatore S. Elder, Shantanu R. Jha, Jayameenakshi Venkatraman, Baptiste Royer, S. M. Girvin, Robert J. Schoelkopf, and Michel H. Devoret. "Fast Universal Control of an Oscillator with Weak Dispersive Coupling to a Qubit" (2021). arXiv:2111.06414.
[8] Patricio Arrangoiz-Arriola, E. Alex Wollack, Zhaoyou Wang, Marek Pechal, Wentao Jiang, Timothy P. McKenna, Jeremy D. Witmer, and Amir H. Safavi-Naeini. "Resolving the energy levels of a nanomechanical oscillator" (2019). arXiv:1902.04681.
[9] Ulrik L. Andersen, Jonas S. Neergaard-Nielsen, Peter Van Loock, and Akira Furusawa. "Hybrid discrete- and continuous-variable quantum information". Nat. Phys. **11**, 713–719 (2015). arXiv:1409.3719.
[10] Jacob Hastrup, Kimin Park, Jonatan Bohr Brask, Radim Filip, and Ulrik Lund Andersen. "Measurement-free preparation of grid states". npj Quantum Inf. **7**, 17 (2021).
[11] Marios H. Michael, Matti Silveri, R. T. Brierley, Victor V. Albert, Juha Salmilehto, Liang Jiang, and S. M. Girvin. "New class of quantum error-correcting codes for a bosonic mode". Phys. Rev. X**6** (2016). arXiv:1602.00008.
[12] Kyungjoo Noh, S M Girvin, and Liang Jiang. "Encoding an Oscillator into Many Oscillators". Phys. Rev. Lett.**125** (2020). arXiv:1903.12615.
[13] Allan D. C. Tosta, Thiago O. Maciel, and Leandro Aolita. "Grand Unification of continuous-variable codes" (2022). arXiv:2206.01751.
[14] Scott Aaronson. "The Learnability of Quantum States". Proc. R. Soc. A Math. Phys. Eng. Sci. **463**, 3089–3114 (2006). arXiv:0608142.
[15] Hao-Chung Cheng, Min-Hsiu Hsieh, and Ping-Cheng Yeh. "The Learnability of Unknown Quantum Measurements". Quantum Inf. Comput. **16**, 0615–0656 (2016). arXiv:1501.00559.
[16] Ryan Sweke, Jean Pierre Seifert, Dominik Hangleiter, and Jens Eisert. "On the quantum versus classical learnability of discrete distributions". Quantum**5** (2021). arXiv:2007.14451.
[17] Matthias C. Caro and Ishaun Datta. "Pseudo-dimension of quantum circuits". Quantum Mach. Intell. **2**, 14 (2020). arXiv:2002.01490.
[18] Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. "Encoding-dependent generalization bounds for parametrized quantum circuits". Quantum **5**, 582 (2021).
[19] V.N. Vapnik. "An overview of statistical learning theory". IEEE Trans. Neural Networks **10**, 988–999 (1999).
[20] Vladimir N. Vapnik. "The Nature of Statistical Learning Theory". Volume 13. Springer New York. New York, NY (2000).
[21] Martin Anthony and Peter L. Bartlett. "Neural Network Learning". Cambridge University Press. (1999).
[22] Claudiu Marius Popescu. "Learning bounds for quantum circuits in the agnostic setting". Quantum Inf. Process. **20**, 286 (2021).
[23] Srinivasan Arunachalam, Yihui Quek, and John Smolin. "Private learning implies quantum stability" (2021). arXiv:2102.07171.
[24] Srilekha Gandhari, Victor V. Albert, Thomas Gerrits, Jacob M. Taylor,

and Michael J. Gullans. "Continuous-Variable Shadow Tomography" (2022). arXiv:2211.05149.

[25] Simon Becker, Nilanjana Datta, Ludovico Lami, and Cambyse Rouzé. "Classical shadow tomography for continuous variables quantum systems" (2022). arXiv:2211.07578.

[26] Scott Aaronson. "Shadow tomography of quantum states". In Proc. 50th Annu. ACM SIGACT Symp. Theory Comput. Pages 325–338. New York, NY, USA (2018). ACM. arXiv:1711.01053.

[27] Tyler Volkoff, Zoë Holmes, and Andrew Sornborger. "Universal Compiling and (No-)Free-Lunch Theorems for Continuous-Variable Quantum Learning". PRX Quantum **2**, 040327 (2021).

[28] Seth Lloyd and Samuel L. Braunstein. "Quantum Computation over Continuous Variables". Phys. Rev. Lett. **82**, 1784–1787 (1999).

[29] J. Eli Bourassa, Nicolás Quesada, Ilan Tzitrin, Antal Száva, Theodor Isacsson, Josh Izaac, Krishna Kumar Sabapathy, Guillaume Dauphinais, and Ish Dhand. "Fast Simulation of Bosonic Qubits via Gaussian Functions in Phase Space". PRX Quantum **2**, 040315 (2021). arXiv:2103.05530.

[30] Matteo Rosati, Andrea Mari, and Vittorio Giovannetti. "Capacity of coherent-state adaptive decoders with interferometry and single-mode detectors". Phys. Rev. A **96**, 012317 (2017). arXiv:1703.05701.

[31] Marco Fanizza, Matteo Rosati, Michalis Skotiniotis, John Calsamiglia, and Vittorio Giovannetti. "Squeezing-enhanced communication without a phase reference". Quantum **5**, 608 (2021). arXiv:2006.06522.

[32] M. Bilkis, Matteo Rosati, and John Calsamiglia. "Reinforcement-learning calibration of coherent-state receivers on variable-loss optical channels". In 2021 IEEE Inf. Theory Work. Pages 1–6. IEEE (2021).

[33] Maria Garcia Diaz, Benjamin Desef, Matteo Rosati, Dario Egloff, John Calsamiglia, Andrea Smirne, Michalis Skotiniotis, and Susana F. Huelga. "Accessible coherence in open quantum system dynamics". Quantum**4** (2020). arXiv:1910.05089.

[34] Kenji Nakahira. "Quantum process discrimination with restricted strategies". Phys. Rev. A **104**, 062609 (2021). arXiv:2104.09038.

[35] Kenji Nakahira and Kentaro Kato. "Generalized quantum process discrimination problems". Phys. Rev. A **103**, 062606 (2021). arXiv:2104.09759.

[36] Jasminder S. Sidhu, Michael S. Bullock, Saikat Guha, and Cosmo Lupo. "Unambiguous discrimination of coherent states" (2021). arXiv:2109.00008.

[37] F E Becerra, J Fan, G Baumgartner, J Goldhar, J T Kosloski, and a Migdall. "Experimental demonstration of a receiver beating the standard quantum limit for multiple nonorthogonal state discrimination". Nat. Photonics **7**, 147–152 (2013).

[38] M. T. DiMario and F. E. Becerra. "Demonstration of optimal non-projective measurement of binary coherent states with photon counting". npj Quantum Inf. **8**, 84 (2022).

[39] Matteo Rosati, Giacomo De Palma, Andrea Mari, and Vittorio Giovannetti. "Optimal quantum state discrimination via nested binary measurements". Phys. Rev. A - At. Mol. Opt. Phys. **95**, 1–10 (2017). arXiv:1701.02233.

[40] Antonio Assalini, Nicola Dalla Pozza, and Gianfranco Pierobon. "Revisiting the Dolinar receiver through multiple-copy state discrimination theory". Phys. Rev. A - At. Mol. Opt. Phys. **84**, 1–6 (2011). arXiv:1107.5452.

[41] Sharif Rahman. "Wiener-Hermite Polynomial Expansion for Multivariate Gaussian Probability Measures" (2017). arXiv:1704.07912.

[42] V. V. Dodonov, O. V. Manko, and V. I. Manko. "Multidimensional Hermite polynomials and photon distribution for polymode mixed light". Phys. Rev. A **50**, 813–817 (1994).
[43] Michael J. Kearns and Robert E. Schapire. "Efficient distribution-free learning of probabilistic concepts". J. Comput. Syst. Sci. **48**, 464–497 (1994).
[44] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. "Scale-sensitive dimensions, uniform convergence, and learnability". J. ACM **44**, 615–631 (1997).
[45] Peter L. Bartlett and Philip M. Long. "Prediction, Learning, Uniform Convergence, and Scale-Sensitive Dimensions". J. Comput. Syst. Sci. **56**, 174–190 (1998).
[46] Martin Anthony and Peter L. Bartlett. "Function Learning from Interpolation". Comb. Probab. Comput. **9**, 213–225 (2000).
[47] David Haussler. "Decision theoretic generalizations of the PAC model for neural net and other learning applications". Inf. Comput. **100**, 78–150 (1992).
[48] Ohad Asor, Hubert Haoyang Duan, and Aryeh Kontorovich. "On the additive properties of the fat-shattering dimension". IEEE Trans. Neural Networks Learn. Syst. **25**, 2309–2312 (2014).
[49] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". J. Comput. Syst. Sci. **55**, 119–139 (1997).
[50] Alessio Serafini. "Quantum Continuous Variables". CRC Press. Boca Raton, FL : CRC Press, Taylor & Francis Group, [2017] — (2017).
[51] Janis Nötzel and Matteo Rosati. "Operating Fiber Networks in the Quantum Limit" (2022). arXiv:2201.12397.
[52] Matteo Rosati and Vittorio Giovannetti. "Achieving the Holevo bound via a bisection decoding protocol". J. Math. Phys.**57** (2016). arXiv:1506.04999.
[53] Matteo Rosati, Andrea Mari, and Vittorio Giovannetti. "Capacity of coherent-state adaptive decoders with interferometry and single-mode detectors". Phys. Rev. A **96**, 012317 (2017).
[54] M. Bilkis, M. Rosati, R. Morral Yepes, and J. Calsamiglia. "Real-time calibration of coherent-state receivers: Learning by trial and error". Phys. Rev. Res. **2**, 033295 (2020). arXiv:2001.10283.
[55] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". Nat. Phys. **16**, 1050–1057 (2020).
[56] Marco Fanizza, Yihui Quek, and Matteo Rosati. "Learning quantum processes without input control" (2022). arXiv:2211.05005.