Automated design of quantum-optical experiments for device-independent quantum key distribution

X. Valcarce,[1,*] P. Sekatski,[2] E. Gouzien,[1] A. Melnikov,[3] and N. Sangouard[1]
[1]*Université Paris-Saclay, CEA, CNRS, Institut de Physique Théorique, 91191 Gif-sur-Yvette, France*
[2]*Department of Applied Physics, University of Geneva, 1205 Geneva, Switzerland*
[3]*Terra Quantum AG, 9000 St. Gallen, Switzerland*

(Received 5 October 2022; accepted 31 March 2023; published 7 June 2023)

Device-independent quantum key distribution (DIQKD) reduces the vulnerability to side-channel attacks of standard quantum key distribution protocols by removing the need for characterized quantum devices. The higher security guarantees come, however, at the price of a challenging implementation. Here, we tackle the question of the conception of an experiment for implementing DIQKD with photonic devices. We introduce a technique combining reinforcement learning, an optimization algorithm, and a custom efficient simulation of quantum optics experiments to automate the design of photonic setups maximizing a given function of the measurement statistics. Applying the algorithm to DIQKD, we get unexpected experimental configurations leading to high key rates and to a high resistance to loss and noise. These configurations might be helpful to facilitate a first implementation of DIQKD with photonic devices and for future developments targeting improved performances.

DOI: 10.1103/PhysRevA.107.062607

I. INTRODUCTION

In quantum key distribution (QKD) [1,2], two separated parties connected by a public quantum channel, Alice and Bob, aim at expanding a string of random bits, i.e., a key. The secrecy and correctness of this key generally rely on the assumptions that (i) the devices used to generate the key behave according to quantum theory, (ii) the parties' locations are isolated to prevent unwanted information leakage, (iii) Alice and Bob get access to random numbers, (iv) they can process classical information on trusted computers, and (v) their quantum devices are trusted and perfectly calibrated to carry out precisely the state and measurements foreseen by the protocol [3–6].

When these assumptions are not met due to imperfections or simplifications in the QKD implementation, hacking becomes possible and compromises the security of the key [7]. In order to reduce the vulnerability to these attacks, new QKD protocols relying on fewer assumptions are desirable. In device-independent QKD (DIQKD) especially, assumption (v) is removed. The state structure produced by the source, the underlying Hilbert space dimension, and the operators describing the measurement apparatus are unknown and their choice is even given to the eavesdropper, Eve.

The higher security level of DIQKD comes at the price of a challenging implementation. DIQKD is entanglement based [2]; i.e., the key is distilled from the results of measurements on entangled states. It requires high-quality entangled states where the quality is quantified by means of a nonlocal game. A high winning probability of the Bell-Clauser-Horne-Shimony-Holt (CHSH) game [8], for example, ensures that Alice and Bob's state is closed to a two-qubit maximally entangled state [9], and that their measurement outcomes are unpredictable to Eve [10]. Many entangled pairs are also required because the postprocessing steps needed to distill an actual key from the measurement outcomes are bit consuming.

Significant progress on the preparation of high-quality entanglement using single trapped ions was recently reported experimentally [11], and culminated with the first distribution of a device-independent key [12] (see also Ref. [13] for recent results aiming to extend DIQKD over hundreds of meters with single atoms). A next step aims at implementing DIQKD with a purely photonic platform, which is plausibly closer to what is expected for a commercial device. On the positive side, the Bell-CHSH game has already been properly implemented using a photon pair source producing polarization entanglement and photon counting techniques for polarization measurements [14–18]. The reported winning probabilities are, however, very close to what can be obtained with classical strategies and are thus not sufficient to realize DIQKD. The main issue of these demonstrations is the photon states which are different from ideal two-qubit states [19]. Another problem is loss—a fraction of photons are lost on the way from the source to the detectors [20]. While the most advanced realization of DIQKD with photonic devices uses polarization entanglement and measurements [21], a natural question is how to combine currently available photonic resources to facilitate the realization of DIQKD in this setting.

By combining Gaussian and non-Gaussian operations, researchers have imagined optical circuits that are capable of winning Bell games with a probability higher than classical strategies [22–24]. Since the number of possible arrangements of optical elements grows exponentially with the number of operations considered, all possible combinations of these operations have likely not been considered, and simple configurations might have been missed. Recent developments on integrated circuits also invite us to explore complex solutions with large numbers of modes and operations [25,26]. Furthermore, substantial theoretical efforts are devoted to the

*xavier.valcarce@ipht.fr

development of security proofs using the distribution of measurement results directly instead of the winning probability of a Bell game obtained from this distribution [27–33]. The need to find optical circuits facilitating the implementation of DIQKD, the possibility of implementing complex optical circuits, and the search for an optical setup producing exactly a given probability distribution of results push us to provide automated solutions to design optical experiments.

Machine learning [34–37] is becoming more and more useful in automation of problem solving in quantum physics research [38–43]. Inspired by Ref. [44], we introduce a technique combining reinforcement learning [45], an optimization algorithm [46], and a custom efficient simulation of quantum optics experiments to design photonic setups maximizing a given function of the measurement statistics. Applying the algorithm to DIQKD, it discovered new, unexpected experimental configurations leading to high key rates in both ideal and lossy cases. The relative simplicity of one of these settings together with its resistance to detector inefficiencies and noise, or the high key rate of a more advanced setting, could be helpful for a first implementation of DIQKD and for future developments.

II. DIQKD PROTOCOL

The protocol is divided in rounds. Each round starts with the creation of entangled systems, half sent to Alice, the other half to Bob, and finishes with randomly chosen measurements. Alice can choose one out of two measurement settings labeled $\hat{A}_x$ with $x \in \{0, 1\}$ and Bob has the choice between three measurement settings called $\hat{B}_y$ with $y \in \{0, 1, 2\}$. For each measurement input, one out of two possible outcomes is obtained that we label A_x for Alice and B_y for Bob, with $\{A_x, B_y\} \in \{0, 1\}$. The settings $x, y \in \{0, 1\}$ are used in a CHSH game in which a round is won if the outcomes of Alice and Bob are the same for the pair of settings $\{\hat{A}_0, \hat{B}_0\}$, $\{\hat{A}_0, \hat{B}_1\}$, and $\{\hat{A}_1, \hat{B}_0\}$ and different when the settings choice is $\{\hat{A}_1, \hat{B}_1\}$. The winning probability ω of the CHSH game is given by $(4 + S)/8$ where the CHSH score S is defined as

$$S = \langle \hat{A}_0 \hat{B}_0 \rangle + \langle \hat{A}_0 \hat{B}_1 \rangle + \langle \hat{A}_1 \hat{B}_0 \rangle - \langle \hat{A}_1 \hat{B}_1 \rangle, \tag{1}$$

with $\langle \hat{A}_x \hat{B}_y \rangle = p(A_x = B_y | \hat{A}_x, \hat{B}_y) - p(A_x \neq B_y | \hat{A}_x, \hat{B}_y)$. The setting $\hat{B}_2$ is ideally chosen to produce outcomes correlated with the results of $\hat{A}_0$.

After many rounds, Alice (Bob) forms a raw key $\mathbf{A}$ ($\mathbf{B}$) from the results of her (his) measurements. Bob then uses error correction to reconstruct a copy of Alice's string and estimate the Bell value S. A randomness extractor is finally applied to obtain the final secret key. In the asymptotic limit of a large number of rounds, the key generation rate when optimal one-way error correction and privacy amplification is used is bounded by [47] $r = H(\mathbf{A}|E) - H(\mathbf{A}|\mathbf{B})$, with H the von Neumann entropy. The first term, which quantifies Eve's uncertainty about the reference key $\mathbf{A}$, can be lower bounded by a function of the CHSH score [48]. When the protocol further includes a step where artificial noise is added to the measurement outcomes, Alice is instructed to generate a new raw key $\mathbf{A}'$ by flipping each of the bits of her initial raw key $\mathbf{A}$ independently with probability p before the postprocessing steps. Eve's uncertainty can increase depending on the value of p. In this case, the key rate is given by [49]

$$r \leqslant 1 - I_p(S) - H(\mathbf{A}'|\mathbf{B}), \tag{2}$$

with

$$I_p(S) = h\left(\frac{1 + \sqrt{(S/2)^2 - 1}}{2}\right) - h\left(\frac{1 + \sqrt{1 - p(1-p)(8 - S^2)}}{2}\right),$$

h being the binary entropy.

III. PHOTONIC CIRCUITS

A. Photonic circuits under consideration

We consider an experiment involving n bosonic modes initially in the vacuum state. Their state is then manipulated by applying single-mode and two-mode operations on any mode or pair of modes in any order. $n - m$ of these modes are measured with non-photon-number-resolving detectors. The state preparation is finalized if the desired combination of measurement outcomes (click or no click) is obtained on the measured modes. The remaining m modes are split between Alice and Bob, to which they apply a sequence of operations chosen from the same set to define the measurement settings. All the modes are finally detected by means of non-photon-number-resolving detectors, yielding one of the 2^m possible results. In the examples below, we explore circuits up to $\{m, n\} = \{2, 4\}$ to keep a reasonable implementation complexity.

The set of possible operations we consider is a fair representation of operations that are routinely used in quantum optics experiments. It includes single-mode squeezers, phase shifters and displacements for the single-mode operations, two-mode squeezers, and beam splitters for the two-mode operations. The use of a photon detector is motivated by the need for non-Gaussian operations to obtain statistics that cannot be reproduced by locally causal models and hence for producing a key device independently.

B. Reference photonic circuit

The most commonly used optical setup for realizing Bell tests [14–18,21] uses a combination of two-mode squeezing operations on the vacuum for producing polarization-entangled photon pairs and standard polarization measurements (see Fig. 1). By optimizing the squeezing parameters, the setting choice, and the amount of noise in the noisy preprocessing and by binning the measurement results appropriately, Eq. (2) yields a key rate of $\sim$0.2522 in the ideal case, i.e., without noise and loss [49]. When considering detectors with nonunit efficiencies, the key rate decreases as shown in Fig. 3 (see blue dashed line). The critical detection efficiency, that is, the minimum detection efficiency needed to generate a positive key rate, is 82.6% [49]. This serves as a reference to benchmark the performance of alternative circuits enabling DIQKD.

C. Simulating quantum-optical circuits

To quantify the performance of a set of photonic circuits, we need to efficiently compute their measurement statistics.

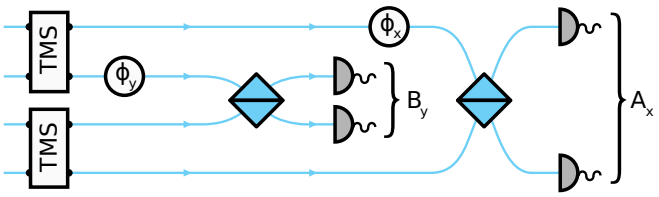

FIG. 1. Most commonly used photonic experiment for realizing Bell tests which can naturally be envisioned for implementing DIQKD. Alice and Bob receive each two modes, one from each of two two-mode squeezers (TMS rectangle) operating on the vacuum. The settings choice $\hat{A}_x$ ($\hat{B}_y$) is obtained by choosing the relative phase ϕ_x (ϕ_y) of the two modes and the transmission of the beam splitter (square) combining them. The measurements are finalized by placing two non-photon-resolving detectors (grey half circle) at the output of the beam splitters. The outcomes A_x and B_y are obtained by binning the click and no-click events produced by the photon detectors.

This is achieved here by describing states using first and second moments of quadrature operators. Formally, if a_i, $a_i^\dagger$ are the bosonic operators for the mode $i = \{1, \ldots, n\}$, the corresponding dimensionless quadrature operators are given by $\hat{x}_i = \frac{a_i^\dagger + a_i}{2}$ and $\hat{p}_i = \mathrm{i}\frac{a_i^\dagger - a_i}{2}$ with $[\hat{x}_i, \hat{p}_i] = \frac{\mathrm{i}}{2}$. We collect these $2n$ operators in a vector $\mathbf{q} = (\hat{x}_1, \hat{p}_1, \ldots, \hat{x}_n, \hat{p}_n)$ and label the ith component of this vector q_i. The displacement vector $\boldsymbol{\mu}$ and the covariance matrix Σ associated to a given state have elements given by $\mu_i = \langle q_i \rangle$ and $\Sigma_{ij} = \frac{1}{2}\langle q_i q_j + q_j q_i \rangle - \mu_i \mu_j$. $\boldsymbol{\mu}$ and Σ give a faithful representation of any n-mode Gaussian states in terms of $2n^2 + 3n$ real parameters (see Refs. [50–52] and Appendix A). As long as Gaussian operations only are applied on the initial vacuum, the state remains Gaussian and can be represented efficiently by the displacement vector and the covariance matrix.

The only exception that we consider to produce non-Gaussianity is the photon detection. Interestingly, as we show in Appendix A, if one starts with an n-mode Gaussian state and measures one mode with a single photon detector, the state of the remaining $n - 1$ modes is Gaussian when it is conditioned on a no-click event or a difference between two Gaussian states when the conditioning is on a click. Hence, the conditional state of the $n - 1$ modes can be fully described by at most two pairs of displacement vectors and covariance matrices. For each additional heralding operation, the number of parameters required to describe the state has to be doubled. Nevertheless, if the number of modes used for heralding remains low, we obtain a memory-efficient exact representation of the state associated to a given circuit. This is precisely our regime of interest, since we want a reasonable heralding rate to end up with feasible proposals.

IV. AUTOMATED DESIGN OF QUANTUM OPTICS EXPERIMENTS

The automated design of quantum optics experiments is based on reinforcement learning—a machine learning paradigm in which an agent is interacting with an environment and learns a task by trial and error. The agent is a routine which specifies the order with which operations are placed on the different modes. The environment efficiently models the series of operations proposed by the agent in order to deduce the measurement statistics and set the parameters of chosen operations to optimize the key rate. The maximal key rate computed by the environment is fed back to the agent as a reward.

The task of the agent is to invent photonic circuits suitable for DIQKD. It learns to do so by repeatedly interacting with a virtual optical circuit inside an episode until a stop condition is met. Specifically, at the beginning of each episode e, the agent perceives a state $s_1(e)$ which is a representation of the (empty) optical circuit at the first step $k = 1$. After a deliberation phase, the agent places one or several optical elements corresponding to an action $a_1(e)$ on the bosonic modes. This produces a new circuit which is analyzed by the environment. The agent then receives back the new state of the circuit $s_2(e)$ together with the associated reward $r_2(e)$ which can be adapted depending on the property of circuits that is desired. An interaction step starts again and the end of the episode is reached when a given circuit depth is achieved. The agent learns from past experiences $\{s_i(k), r_i(k)\}$ by updating the policy behind the deliberation process.

The task of the environment is to simulate the optical circuits proposed by the agent and optimize its parameters in order to compute the reward associated to the circuit. To simulate the circuit, we developed a package, QuantumOpticalCircuits.jl [53], written in JULIA [54], using the efficient state representation described in the previous section. To find the parameters of circuits leading to the highest key rates according to the bound given in Eq. (2), we used the Nelder-Mead algorithm, a suitable algorithm for the optimization of multidimensional nonlinear objective functions which does not require an analytical or numerical gradient to be supplied [46]. The optimization is performed in the ideal case, i.e., with unit detector efficiency. To get access to the critical detection efficiency, the optimal parameters are computed by first considering detectors with unit efficiencies. The efficiency is then decreased and a new parameter optimization is performed, starting from the ones resulting in the best key rate at the previous step. The process starts again with a smaller detection efficiency until the key rate drops below a certain threshold.

A. Choice of the policy

Transition in the environment, i.e., the change in state and reward from an action, is the result of a heuristic numerical optimization. In such a case, it is natural to use model-free reinforcement learning, learning purely from trial and error, not trying to construct a model of that transition [45]. Furthermore, infinitely many states can be observed and learning which action performs best for each state is computationally impossible. Instead, we use a policy gradients method, which aims at learning directly a stochastic policy mapping states to actions [55]. Finally, the sampling cost, i.e., the cost to simulate an episode, is high from the numerical optimization. In this scope, we used the proximal policy optimization (PPO) algorithm [56], a model-free policy gradient algorithm known to be sample efficient compared to similar reinforcement learning algorithms, e.g., trust region policy optimization (TRPO) [57]. The details of this algorithm are available in Appendix B.

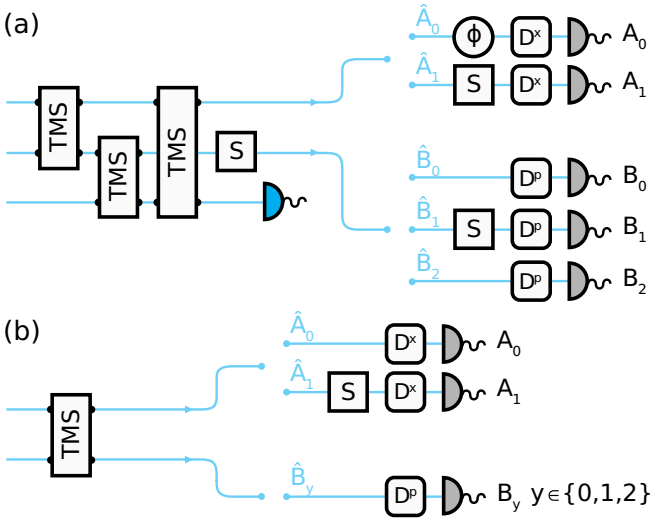

FIG. 2. Photonic setups resulting from an automated search when (a) the highest key rate is favored in the absence of imperfection and (b) the highest tolerance to detector inefficiency is favored. In setup (a), the state preparation uses three two-mode squeezed (TMS) states operating on the pair of modes {1, 2}, {2, 3}, and {1, 3}, respectively, followed by a single mode squeezers (S) on mode 2. Alice and Bob receive the state of modes 1,2 conditioned on a click at the photon detector on mode 3. For each of his measurement settings $y = \{0, 1, 2\}$, Bob performs a displacement operation in the p direction (D^p). In the case $y = 1$, a single-mode squeezer is applied before the displacement operation. On Alice's side, the first measurement setting is a phase shifter (Φ) followed by a displacement operation along x (D^x). For the second setting, a single-mode squeezer replaces the phase shifter. In setup (b), the preparation step is made with a single two-mode squeezer (TMS) operating in vacuum. Alice's settings are set by either a displacement for $\hat{A}_0$ or a single-mode squeezer followed by a displacement for $\hat{A}_1$. Bob's settings correspond to displacement operations with different amplitudes for each input.

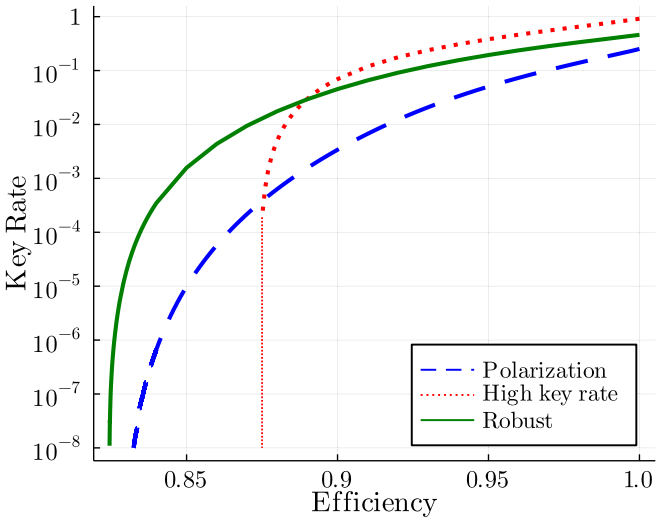

FIG. 3. Key rate as a function of the detector efficiency. The blue dashed line is associated with the reference circuit shown in Fig. 1. The orange dotted line corresponds to the circuit found when high key rates are favored for unit efficiency detections [Fig. 2(a)]. The green solid line is associated to the circuit found when favoring the resistance to nonunit detection efficiency [Fig. 2(b)].

B. Results

In the first step, we define a reward which favors a high key rate in the absence of loss and noise. The setup found by the agent after a few training steps is depicted in Fig. 2(a). It involves three modes, one mode serving as a heralding after a series of two-mode and single-mode squeezed operations. This setup yields a key rate of ∼0.914, much higher than the reference circuit (key rate of ∼0.252). The resistance to loss of this unexpected setup is characterized by optimizing the key rate as a function of the detection efficiency η (the detectors of Alice and Bob and the one used for the heralding are assumed to have the same efficiency). From the result given in Fig. 3 (orange dashed line), we see that the setup of Fig. 2(a) provides a higher key rate than the reference circuit for $\eta \gtrsim 87.5\%$.

In the second step, we adapt the reward to favor loss-tolerant circuits. Concretely, we look for circuits with a minimum detection efficiency for achieving a key rate of at least 10^{-4}. The most interesting setup was found for $n = 2$ and uses a single two-mode squeezed operation on the vacuum for state preparation [see Fig. 2(b)]. As shown in Fig. 3 (green solid line), the circuit leads to a key rate of $\sim 10^{-8}$ for $\eta = 82.45\%$, while the same rate is obtained for the reference circuit for $\eta = 83.25\%$. Moreover, depending on the detection efficiency, the key rate can be up to two orders of magnitude larger. Finally, when including dark counts with a probability of 10^{-3}, a key rate of 10^{-4} is obtained with the reference circuit for $\eta \sim 91.1\%$ while the proposed circuit takes $\eta \sim 86.1\%$. Note that the parameters of circuits presented in Fig. 2 are given in Appendix C.

V. CONCLUSION

We presented an algorithm combining an efficient modeling of Gaussian and heralded processes, an optimization and reinforcement learning to automate the search of optical circuits producing desired statistics. We showed its usefulness in DIQKD, for both mid- and long-term goals aiming respectively to facilitate a first photonic realization and to deliver high key rates. We do expect the proposed algorithm to remain useful in case new protocols and security proofs are proposed as the formula of the key rate that was used to guide the agent can be updated to include future developments.

ACKNOWLEDGMENTS

The authors would like to thank Jean-Daniel Bancal, Jean Etesse, Anthony Martin, Ernest Tan, Märta Tschudin, Ramona Wolf, and Julian Zivy for fruitful discussions and Dowinson Nguyen for the illustrations. We acknowledge funding by the Institut de Physique Théorique (IPhT), Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), by a French national quantum initiative managed by Agence Nationale de la Recherche in the framework of France 2030 with the Reference No. ANR-22-PETQ-0009 and by the European Union's Horizon Europe research and innovation programme under the project "Quantum Security Networks Partnership" (QSNP, Grant Agreement No. 101114043).

TABLE I. Single-mode Gaussian operations. When the operation acts on the ith mode, the nonzero components of $\boldsymbol{d}$ and the nontrivial block of the matrix M appear at positions $2i-1$ and $2i$. For the displacement we also consider the particular cases denoted $\mathrm{D}^{\hat{x}}$ for $\mathrm{Im}\{\alpha\}=0$ and $\mathrm{D}^{\hat{p}}$ for $\mathrm{Re}\{\alpha\}=0$.

Operation	U	$\boldsymbol{d}$	M	Parameters
Displacement (D)	$e^{\alpha a_i^\dagger - \alpha^* a_i}$	$(\dots, 0, \mathrm{Re}\alpha, \mathrm{Im}\alpha, 0, \dots)^T$	$\mathbb{1}_{2n}$	$\alpha \in \mathbb{C}$
Phase shifter (Φ)	$e^{-\mathrm{i}\theta a_i^\dagger a_i}$	$\vec{0}$	$\begin{pmatrix} \mathrm{c}_\theta & \mathrm{s}_\theta \\ -\mathrm{s}_\theta & \mathrm{c}_\theta \end{pmatrix} \oplus \mathbb{1}_{2n-2}$	$\theta \in \mathbb{R}$
Single-mode squeezer (S)	$e^{\frac{1}{2}(z^* a_i^2 - z a_i^{\dagger 2})}$	$\vec{0}$	$\begin{pmatrix} \mathrm{C}_r - \mathrm{S}_r\mathrm{c}_\theta & -\mathrm{S}_r\mathrm{s}_\theta \\ -\mathrm{S}_r\mathrm{s}_\theta & \mathrm{C}_r + \mathrm{S}_r\mathrm{c}_\theta \end{pmatrix} \oplus \mathbb{1}_{2n-2}$	$z = re^{\mathrm{i}\theta} \in \mathbb{C}$

APPENDIX A: MODELING PHOTONIC CIRCUITS

In this Appendix we present a concise and self-contained derivation of how we model the photonic circuits discussed in the main text with a low number of parameters. We start by giving a short summary containing all the formulas required to use this representation. The derivations are presented afterwards.

1. Summary

Consider n bosonic modes and the associated quadrature operators $\hat{x}_i = \frac{a_i^\dagger + a_i}{2}$ and $\hat{p}_i = \mathrm{i}\frac{a_i^\dagger - a_i}{2}$ that we collected in a vector $\mathbf{q} = (\hat{x}_1, \hat{p}_1, \dots, \hat{x}_n, \hat{p}_n) = (q_1, q_2, \dots, q_{2n-1}, q_{2n})$. Any Gaussian n-mode state ρ can be faithfully represented by the first two moments $(\boldsymbol{\mu}, \Sigma)$ of the quadrature operators on the state $\mu_i = \langle q_i \rangle = \mathrm{tr}\rho\, q_i$ and $\Sigma_{ij} = \frac{1}{2}\langle q_i q_j + q_j q_i \rangle - \mu_i \mu_j$. Here $\boldsymbol{\mu}$ is the displacement (column) vector ($2n$ real parameters) and $\Sigma = \Sigma^T$ is the covariance matrix ($2n^2 + n$ real parameters). In particular, for the vacuum state one finds

$$\text{vacuum state } |0\rangle\langle 0|^{\otimes n}: \quad (\boldsymbol{\mu}, \Sigma) = \left(\vec{0}, \frac{1}{4}\mathbb{1}\right). \tag{A1}$$

A Gaussian operation T maps Gaussian states to Gaussian states. It can be represented by means of a pair $(\boldsymbol{d}, M)$ with a $2n$ real column vector $\boldsymbol{d}$ and a symplectic $2n \times 2n$ matrix M. When acting on a state the Gaussian operation transforms the displacement vector and the covariance matrix as

$$\mathrm{T}: (\boldsymbol{\mu}, \Sigma) \mapsto (M\boldsymbol{\mu} + \boldsymbol{d}, M\Sigma M^T). \tag{A2}$$

In Tables I and II we give the representation of all single-mode and two-mode Gaussian transformations in terms of $(\boldsymbol{d}, M)$ as well as the corresponding unitary representation $\mathrm{T}: |\psi\rangle \mapsto U|\psi\rangle$. We use the notation $\mathrm{c}_\theta = \cos(\theta)$, $\mathrm{s}_\theta = \sin(\theta)$, $\mathrm{C}_r = \cosh(r)$, and $\mathrm{S}_r = \sinh(r)$. Next we consider the transformations of the state resulting from measuring out the mode i with a single photon detector of efficiency η. Both of the states $\rho_{\neg i}$ and $\rho_{\circ i}$ of the remaining $(n-1)$ modes resulting respectively from discarding the measurement outcome (or tracing out the mode i, $\rho_{\neg i}$) and conditioning on the no-click outcome ($\rho_{\circ i}$) are Gaussian. The displacement vectors and quadrature moments of the resulting states are given in Table III. To express the state we use the transformation TR_i which simply drops the components of a vector or rows at columns of a matrix at positions $2i-1, 2i$,

$$\mathrm{TR}_i[\boldsymbol{\mu}] \equiv (\dots, \mu_{2i-2}, \mu_{2i+1}, \dots),$$
$$\mathrm{TR}_i[\Sigma] \equiv \left(\begin{array}{ccc|cc} \ddots & \vdots & & \vdots & \\ \cdots & \Sigma_{2i-2,2i-2} & & \Sigma_{2i-2,2i+1} & \cdots \\ \hline \cdots & \Sigma_{2i+1,2i-2} & & \Sigma_{2i+1,2i+1} & \cdots \\ & \vdots & & \vdots & \ddots \end{array}\right), \tag{A3}$$

and the matrix F given by

$$F = \begin{pmatrix} \frac{4\eta}{2-\eta} & 0 \\ 0 & \frac{4\eta}{2-\eta} \end{pmatrix}_{2i-1,2i} \oplus 0_{2n-2}. \tag{A4}$$

The probabilities to observe the no-click ($p_{\circ i}$) and the click ($p_{\bullet i}$) outcomes are given by

$$p_{\circ i} = \frac{2}{2-\eta}\sqrt{\frac{(\det\Sigma)^{-1}}{\det(\Sigma^{-1}+F)}} \times e^{-\frac{1}{2}\boldsymbol{\mu}^T\left(\Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1}+F)^{-1}\Sigma^{-1}\right)\boldsymbol{\mu}}$$
$$\text{and} \quad p_{\bullet i} = 1 - p_{\circ i}. \tag{A5}$$

TABLE II. Two-mode Gaussian operations. When the operation acts on modes i and j the nontrivial block of the matrix M appears at positions $2i-1$, $2i$, $2j-1$, and $2j$.

Operation	U	$\boldsymbol{d}$	M	Parameters
Beam splitter (BS)	$e^{\theta(a_i^\dagger a_j - a_i a_j^\dagger)}$	$\vec{0}$	$\begin{pmatrix} \mathrm{c}_\theta & 0 & \mathrm{s}_\theta & 0 \\ 0 & \mathrm{c}_\theta & 0 & \mathrm{s}_\theta \\ -\mathrm{s}_\theta & 0 & \mathrm{c}_\theta & 0 \\ 0 & -\mathrm{s}_\theta & 0 & \mathrm{c}_\theta \end{pmatrix} \oplus \mathbb{1}_{2n-4}$	$\theta \in \mathbb{R}$
Two-mode squeezer (TMS)	$e^{z^* a_i a_j - z a_i^\dagger a_j^\dagger}$	$\vec{0}$	$\begin{pmatrix} \mathrm{C}_r & 0 & -\mathrm{S}_r\mathrm{c}_\theta & -\mathrm{S}_r\mathrm{s}_\theta \\ 0 & \mathrm{C}_r & -\mathrm{S}_r\mathrm{s}_\theta & \mathrm{S}_r\mathrm{c}_\theta \\ -\mathrm{S}_r\mathrm{c}_\theta & -\mathrm{S}_r\mathrm{s}_\theta & \mathrm{C}_r & 0 \\ -\mathrm{S}_r\mathrm{s}_\theta & \mathrm{S}_r\mathrm{c}_\theta & 0 & \mathrm{C}_r \end{pmatrix} \oplus \mathbb{1}_{2n-4}$	$z = re^{\mathrm{i}\theta} \in \mathbb{C}$

TABLE III. The possible states resulting from an n-mode Gaussian state $\rho \simeq (\boldsymbol{\mu}, \Sigma)$ after measuring out the mode i with a single photon detector with efficiency η. The transformation TR_i defined in Eq. (A3) simply removes the components of a vector or rows and columns of a matrix at positions $2i-1$ and $2i$. The matrix F defined in Eq. (A4) depends on the detector efficiency. The state $\rho_{\bullet i}$ is not Gaussian but can be decomposed as a difference of two Gaussian states [see Eq. (A6)].

Transformation	Density matrix	Displacement vector	Covariance matrix
Tracing out the mode i (discarding the outcome)	$\rho_{\neg i}$	$\boldsymbol{\mu}_{\neg i} \equiv \mathrm{TR}_i(\boldsymbol{\mu})$	$\Sigma_{\neg i} \equiv \mathrm{TR}_i(\Sigma)$
Conditioning to no-click outcome on the mode i	$\rho_{\circ i}$	$\boldsymbol{\mu}_{\circ i} \equiv \mathrm{TR}_i[(\Sigma^{-1}+F)^{-1}\Sigma^{-1}\boldsymbol{\mu}]$	$\Sigma_{\circ i} \equiv \mathrm{TR}_i[(\Sigma^{-1}+F)^{-1}]$
Conditioning to click outcome on the mode i	$\rho_{\bullet i}$	$\times$	$\times$

Importantly, the state $\rho_{\bullet i}$ conditional on the click outcome is non-Gaussian, but can be expressed as a difference of two Gaussian states:

$$\rho_{\bullet i} = \frac{\rho_{\neg i} - p_{\circ i}\rho_{\circ i}}{1-p_{\circ i}} \simeq \left\{ \frac{1}{1-p_{\circ i}}, (\boldsymbol{\mu}_{\neg i}, \Sigma_{\neg i}) \right\}. \left\{ -\frac{p_{\circ i}}{1-p_{\circ i}}, (\boldsymbol{\mu}_{\circ i}, \Sigma_{\circ i}) \right\}. \tag{A6}$$

For any subsequent heralding, the number of terms in the sum is doubled. Generally, we are thus interested in states $\rho = \sum_k w_k \rho_k$ that can be represented as a quasimixture ($\rho \simeq \{w_k, (\boldsymbol{\mu}_k, \Sigma_k)\}$) of Gaussian state ρ_k, where the weights w_k can be negative. In total such a representation requires $2^{n-m}(2m^2+3m+1)$ real parameters, where the number of modes used for heralding $n-m$ is kept low for the setups of interest.

Finally, we give a compact formula that computes the measurement statistics. When m modes are measured with single photon detectors, the outcomes are labeled by a bitstring $\boldsymbol{k}$ of length m, where each bit specifies if the detector on the mode i clicks ($k_i = 1$) or not ($k_i = 0$). For a Gaussian state described by $(\boldsymbol{\mu}, \Sigma)$ the probability to observe the outcome $\boldsymbol{k}$ is given by

$$\mathrm{Prob}(\boldsymbol{k}) = \sum_{\boldsymbol{\ell} \in \mathbb{S}_{\boldsymbol{k}}} (-1)^{\boldsymbol{k}\cdot\boldsymbol{\ell}} \langle \hat{o}_{\boldsymbol{\ell}} \rangle \quad \text{with} \tag{A7}$$

$$\langle \hat{o}_{\boldsymbol{\ell}} \rangle = \left(\frac{2}{2-\eta}\right)^{|\boldsymbol{\ell}|} \times \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(\Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1}+O_{\boldsymbol{\ell}})^{-1}\Sigma^{-1})\boldsymbol{\mu}\right)}{\sqrt{\det(\Sigma)\det(\Sigma^{-1}+O_{\boldsymbol{\ell}})}},$$

$$O_{\boldsymbol{\ell}} = \frac{4\eta}{2-\eta} \bigoplus_{i=1}^{m} \begin{pmatrix} \ell_i & 0 \\ 0 & \ell_i \end{pmatrix}. \tag{A8}$$

Here $\mathbb{S}_{\boldsymbol{k}}$ is the set of all bitstrings $\boldsymbol{\ell}$ of length m whose components ℓ_i are fixed to 1 for all i such that $k_i = 0$ (a set of size $2^{|\boldsymbol{k}|}$), and $|\boldsymbol{\ell}| = \sum_{i=1}^{m} \ell_i$ is the Hamming weight of a bitstring.

2. Wigner representation

Let us first consider a single bosonic mode associated with creation and annihilation operators $a^\dagger$ and a. The Dirac delta operator, defined as

$$\delta(a-\alpha) = \frac{1}{\pi^2} \int d^2\beta \, e^{(a^\dagger - \alpha^*)\beta - (a-\alpha)\beta^*}, \tag{A9}$$

where α and β are complex numbers, possesses various properties and in particular

$$\pi \int d^2\alpha \, \delta(a-\alpha)\mathrm{tr}\rho\, \delta(a-\alpha) = \rho \tag{A10}$$

for any density operator ρ (see Ref. [58], Sec. 4). This suggests that $\delta(a-\alpha)$ can be used for representing density operators with the quasiprobability distribution

$$W_\rho(\alpha) = \mathrm{tr}\, \rho\, \delta(a-\alpha) \tag{A11}$$

satisfying $-2\pi^{-1} \leqslant W_\rho(\alpha) \leqslant 2\pi^{-1}$ [58]. This representation—the Wigner representation—can be extended to the multimode case. Consider n bosonic modes with the operator $a_i, a_i^\dagger$ associated to the mode $i = \{1, n\}$. The Wigner representation of an n-mode state ρ is defined by the following extension of the monomode case:

$$W_\rho(\boldsymbol{\alpha}) = \mathrm{tr}\, \rho \bigotimes_{i=1}^{n} \delta(a_i - \alpha_i), \tag{A12}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}^T \in \mathbb{C}^n$. As before, the link between the state ρ of the n modes and the Wigner function is given by

$$\rho = \pi^n \int d^2\boldsymbol{\alpha}\, W_\rho(\boldsymbol{\alpha}) \bigotimes_{i=1}^{n} \delta(a_i - \alpha_i). \tag{A13}$$

3. Wigner representation of the vacuum state

An n-mode state ρ is called Gaussian if its Wigner function is Gaussian, i.e., equal to probability density function $N(\boldsymbol{\alpha}; \boldsymbol{\mu}, \Sigma)$ of a multivariate normal distribution [50,51]

$$W_\rho(\boldsymbol{\alpha}) = N(\boldsymbol{\alpha}; \boldsymbol{\mu}, \Sigma) = \frac{\exp\left[-\frac{1}{2}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\mu})^T \Sigma^{-1}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\mu})\right]}{\sqrt{\det(2\pi\,\Sigma)}}. \tag{A14}$$

$\tilde{\boldsymbol{\alpha}}$ is the $\mathbb{R}^{2n}$ vector constructed from $\boldsymbol{\alpha}$ in the following way:

$$\tilde{\boldsymbol{\alpha}} = \{\mathrm{Re}(\alpha_1), \mathrm{Im}(\alpha_1), \ldots, \mathrm{Re}(\alpha_n), \mathrm{Im}(\alpha_n)\}^T. \tag{A15}$$

$\boldsymbol{\mu}$ is the displacement (column) vector, and Σ is the covariance matrix. Together they are given by $2n^2 + 3n$ real parameters with

$$\mu_i = \langle q_i \rangle, \tag{A16}$$

$$\Sigma_{ij} = \tfrac{1}{2}\langle q_i q_j + q_j q_i \rangle - \mu_i \mu_j, \tag{A17}$$

with q_i the ith component of the vector $\mathbf{q} = (\hat{x}_1, \hat{p}_1, \ldots, \hat{x}_n, \hat{p}_n)$ composed of the dimensionless quadrature operators $\hat{x}_i = \frac{a_i^\dagger + a_i}{2}$ and $\hat{p}_i = \mathrm{i}\frac{a_i^\dagger - a_i}{2}$ which satisfy $[\hat{x}_i, \hat{p}_i] = \frac{\mathrm{i}}{2}$.

A single-mode vacuum state, $|0\rangle$, is an example of a Gaussian state. Its Wigner function is characterized by a zero displacement vector $\boldsymbol{\mu} = \vec{0}_2$ (as $\langle 0|\hat{x}|0\rangle = \langle 0|\hat{p}|0\rangle = 0$) and a covariance matrix proportional to identity $\Sigma = \frac{1}{4}\mathbb{1}_2$ (as $\langle 0|\hat{x}^2|0\rangle = \langle 0|\hat{p}^2|0\rangle = \frac{1}{4}$ and $\langle 0|\{\hat{x}, \hat{p}\}|0\rangle = \langle 0|\frac{\mathrm{i}}{4}(aa^\dagger -$

$aa^\dagger)|0\rangle = 0$). Extending this to n modes, we get that the Wigner function of the n-mode vacuum state is given by Eq. (A14) with

$$\boldsymbol{\mu} = \vec{0}_{2n}, \tag{A18}$$

$$\Sigma = \tfrac{1}{4}\mathbb{1}_{2n}, \tag{A19}$$

where the subscript specifies the size of the objects.

4. Gaussian operations

Among possible operations on bosonic systems, Gaussian operations are those mapping Gaussian states to Gaussian states. They are thus fully characterized by their effect on the displacement vector and covariance matrix [51]

$$\boldsymbol{\mu} \mapsto M\boldsymbol{\mu} + \boldsymbol{d} \quad \text{and} \quad \Sigma \mapsto M\Sigma M^T, \tag{A20}$$

where, for an n-mode system, $\boldsymbol{d} \in \mathbb{R}^{2n}$ and M is a symplectic matrix, i.e., a $2n \times 2n$ matrix satisfying

$$M^T \Omega M = \Omega \quad \text{with} \quad \Omega = \mathbb{1}_n \otimes \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{A21}$$

Note that a single-mode Gaussian operation characterized by M and $\boldsymbol{d}$ acting on the mode i of an n-mode state has a vector $\boldsymbol{d} = (\dots, 0, d_{2i-1}, d_{2i}, 0 \dots)^T$ with all elements that are zero except at positions $2i-1$ and $2i$, and a symplectic matrix $M = \mathbb{1}_{2(i-1)} \oplus M' \oplus \mathbb{1}_{2(n-i)}$ with the only nontrivial 2×2 block M' appearing at positions $2i-1$ and $2i$. Similarly a two-mode Gaussian operation acting on modes i and j will only have nontrivial elements appearing at positions $2i-1$, $2i$, $2j-1$, and $2j$.

Below we list the single- and two-mode Gaussian operations. We first define each operation by its unitary representation U with its action on the state given by $|\Psi\rangle \mapsto U|\Psi\rangle$ (the Schrödinger picture). Then we compute its action $o \mapsto U^\dagger o U$ on the ladder and quadrature operator (Heisenberg picture). Finally we obtain the representation of the operation in terms of the pair $\boldsymbol{d}$ and M. We only specify the nontrivial block of M and the nonzero elements of $\boldsymbol{d}$.

a. Phase shifter. The phase shifter (Φ) is a single-mode operator given by the unitary operator $U_\Phi(\theta) = \exp(-i\theta a^\dagger a)$. When applied on the bosonic operators it gives

$$a \mapsto e^{-i\theta}a, \quad a^\dagger \mapsto e^{i\theta}a^\dagger, \tag{A22}$$

$$\hat{x} \mapsto \cos(\theta)\hat{x} + \sin(\theta)\hat{p}, \quad \hat{p} \mapsto \cos(\theta)\hat{p} - \sin(\theta)\hat{x}. \tag{A23}$$

We thus find that the corresponding symplectic transformation is characterized by

$$\boldsymbol{d} = \vec{0}, \quad M = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{A24}$$

b. Displacement. The unitary operator associated to a displacement (D) with amplitude $\alpha \in \mathbb{C}$ is given by $U_{\mathrm{D}}(\alpha) = \exp(\alpha a^\dagger - \alpha^* a)$. This operation shifts the ladder and quadrature operators according to

$$a \mapsto a + \alpha, \quad a^\dagger \mapsto a^\dagger + \alpha^*, \tag{A25}$$

$$\hat{x} \mapsto \hat{x} + \mathrm{Re}\{\alpha\}, \quad \hat{p} \mapsto \hat{p} + \mathrm{Im}\{\alpha\}. \tag{A26}$$

The symplectic transformation of a displacement is hence characterized by

$$\boldsymbol{d} = \begin{pmatrix} \mathrm{Re}\{\alpha\} \\ \mathrm{Im}\{\alpha\} \end{pmatrix}, \quad M = \mathbb{1}_2. \tag{A27}$$

We denote D^x and D^p the displacement operations in the directions $\hat{x}$ and $\hat{p}$, respectively. A displacement in $\hat{x}$ can be seen as a displacement with α real. Similarly, a displacement in $\hat{p}$ is a displacement with α imaginary.

c. Single-mode squeezer. The single-mode squeezing operation (S) is given by $U_{\mathrm{S}}(z) = \exp[\frac{1}{2}(z^* a^2 - z(a^\dagger)^2)]$, where the complex parameter z can be written as $z = re^{i\theta}$. Single-mode squeezing transforms the ladder and quadrature operators as

$$a \mapsto \cosh(r)a - e^{i\theta}\sinh(r)a^\dagger \quad a^\dagger \mapsto \cosh(r)a^\dagger - e^{-i\theta}\sinh(r)a \tag{A28}$$

$$\hat{x} \mapsto [\cosh(r) - \sinh(r)\cos(\theta)]\hat{x} - \sinh(r)\sin(\theta)\hat{p} \quad \hat{p} \mapsto [\cosh(r) + \sinh(r)\cos(\theta)]\hat{p} - \sinh(r)\sin(\theta)\hat{x}. \tag{A29}$$

Hence, single-mode squeezing can be expressed by the symplectic transformation characterized by

$$\boldsymbol{d} = \vec{0}, \quad M = \cosh(r)\mathbb{1}_2 - \sinh(r)\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix}. \tag{A30}$$

d. Two-mode squeezer. A two-mode squeezer (TMS) acting on modes i, j is defined by the unitary $U_{\mathrm{TMS}}(z) = \exp(z^* a_i a_j - z a_i^\dagger a_j^\dagger)$, with $z = re^{i\theta}$. This changes the bosonic operators as follows:

$$\begin{aligned} a_i &\mapsto \cosh(r)a_i - e^{i\theta}\sinh(r)a_j^\dagger, \quad a_i^\dagger \mapsto \cosh(r)a_i^\dagger - e^{-i\theta}\sinh(r)a_j, \\ a_j &\mapsto \cosh(r)a_j - e^{i\theta}\sinh(r)a_i^\dagger, \quad a_j^\dagger \mapsto \cosh(r)a_j^\dagger - e^{-i\theta}\sinh(r)a_i, \end{aligned} \tag{A31}$$

$$\begin{aligned} \hat{x}_i &\mapsto \cosh(r)\hat{x}_i - \sinh(r)[\cos(\theta)\hat{x}_j + \sin(\theta)\hat{p}_j], \quad \hat{p}_i \mapsto \cosh(r)\hat{p}_i + \sinh(r)[\cos(\theta)\hat{p}_j - \sin(\theta)\hat{x}_j], \\ \hat{x}_j &\mapsto \cosh(r)\hat{x}_j - \sinh(r)[\cos(\theta)\hat{x}_i + \sin(\theta)\hat{p}_i], \quad \hat{p}_j \mapsto \cosh(r)\hat{p}_j + \sinh(r)[\cos(\theta)\hat{p}_i - \sin(\theta)\hat{x}_i]. \end{aligned} \tag{A32}$$

The corresponding symplectic transformation is given by

$$\boldsymbol{d} = \vec{0}, \quad M = \begin{pmatrix} \cosh(r) & 0 & -\sinh(r)\cos(\theta) & -\sinh(r)\sin(\theta) \\ 0 & \cosh(r) & -\sinh(r)\sin(\theta) & \sinh(r)\cos(\theta) \\ -\sinh(r)\cos(\theta) & -\sinh(r)\sin(\theta) & \cosh(r) & 0 \\ -\sinh(r)\sin(\theta) & \sinh(r)\cos(\theta) & 0 & \cosh(r) \end{pmatrix}. \tag{A33}$$

e. Beam splitter. A beam splitter (BS) on modes i, j is given by the unitary $U_{\rm BS}(\theta) = \exp[\theta(a_i^\dagger a_j - a_i a_j^\dagger)]$, where the transmittivity is given by $\cos^2(\theta)$ and the reflectivity is $\sin^2(\theta)$. This Gaussian operation maps the operators to

$$\begin{aligned} a_i &\mapsto \cos(\theta)a_i + \sin(\theta)a_j, \quad a_i^\dagger \mapsto \cos(\theta)a_i^\dagger + \sin(\theta)a_j^\dagger, \\ a_j &\mapsto \cos(\theta)a_j - \sin(\theta)a_i, \quad a_j^\dagger \mapsto \cos(\theta)a_j^\dagger - \sin(\theta)a_i^\dagger, \end{aligned} \tag{A34}$$

$$\begin{aligned} \hat{x}_i &\mapsto \cos(\theta)\hat{x}_i + \sin(\theta)\hat{x}_j, \quad \hat{p}_i \mapsto \cos(\theta)\hat{p}_i + \sin(\theta)\hat{p}_j, \\ \hat{x}_j &\mapsto \cos(\theta)\hat{x}_j - \sin(\theta)\hat{x}_i, \quad \hat{p}_j \mapsto \cos(\theta)\hat{p}_j - \sin(\theta)\hat{p}_i. \end{aligned} \tag{A35}$$

The corresponding symplectic transformation is characterized by

$$\boldsymbol{d} = \vec{0}, \quad M = \begin{pmatrix} \cos(\theta) & 0 & \sin(\theta) & 0 \\ 0 & \cos(\theta) & 0 & \sin(\theta) \\ -\sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & -\sin(\theta) & 0 & \cos(\theta) \end{pmatrix}. \tag{A36}$$

We note that the transformations D, S, and TMS with arbitrary complex parameters α and $z = re^{i\theta}$ can be decomposed as the same transformations with real parameters ($\alpha = \alpha^*$ and $z = r$) combined with two phase shifters.

5. Measuring a Gaussian state with single photon detectors

We here consider the measurement of one or several modes in a multimode state with non-photon-number-resolving (NPNR) detectors. The positive operator-valued measure (POVM) associated to the "no-click" event of a NPNR detection with efficiency η operating on a mode with bosonic operators a and $a^\dagger$ is given by $R^{a^\dagger a}$ where $R = (1-\eta)$ while the "click" event is obviously related to the POVM $\mathbb{1} - R^{a^\dagger a}$. Although such a measurement is not a Gaussian operation, we show that the multimode state conditioned on the outcome of such a measurement on one or several modes can be written as a mixture of Gaussian states and, hence, its Wigner function can be written as a difference between two densities of a multivariate normal distribution.

a. Tracing out a mode. Let $W_\rho(\boldsymbol{\alpha})$ the Wigner function of an n-mode state ρ. When the mode i is traced out, the resulting state $\rho_{\neg 1}$ is given by

$$\begin{aligned} \rho_{\neg i} = {\rm tr}_i \rho &= {\rm tr}_i\, \pi^n \int d^2\boldsymbol{\alpha} W(\boldsymbol{\alpha}) \bigotimes_{j=1}^{n} \delta(a_j - \alpha_j) \\ &= \pi^{n-1} \int d^2\boldsymbol{\alpha} W(\boldsymbol{\alpha}) \left({\rm tr}_i \pi\, \delta(a_i - \alpha_i)\right) \bigotimes_{\substack{j=1 \\ j\neq i}}^{n} \delta(a_j - \alpha_j) \\ &= \pi^{n-1} \int d^2\boldsymbol{\alpha} W(\boldsymbol{\alpha}) \bigotimes_{\substack{j=1 \\ j\neq i}}^{n} \delta(a_j - \alpha_j). \end{aligned} \tag{A37}$$

The second equality is obtained from the definition given in Eq. (A12) while the third inequality uses ${\rm tr}\delta(a-\alpha) = \pi$ (see Ref. [58]). This shows that when the mode i is traced out, the Wigner function of the remaining modes is simply given by

$$W_{\neg i}(\overline{\boldsymbol{\alpha}}) = \int d^2\alpha_i W(\boldsymbol{\alpha}), \tag{A38}$$

that is, $\overline{\boldsymbol{\alpha}}$ is obtained from the vector $\boldsymbol{\alpha}$ by removing the components $2i$ and $2i+1$. Since the marginals of a multivariate normal distribution are also normal, Gaussianity is preserved by the trace; that is, $\rho_{\neg i}$ remains Gaussian if ρ is Gaussian.

Concretely, the displacement vector $\mu_{\neg i}$ of $\rho_{\neg i}$ can be obtained by removing the components $2i$ and $2i+1$ of the displacement vector of ρ. Its covariance matrix $\Sigma_{\neg i}$ is obtained by removing the rows and columns $2i$ and $2i+1$ of the covariance matrix of ρ.

b. Outcome probabilities. Let us consider a NPNR detector operating on a single mode with bosonic operators a and $a^\dagger$ and state ρ. The probability of having a "no click" is given by

$$p_{\text{no-click}} = {\rm tr}\,\rho\, R^{a^\dagger a} = \int d^2\alpha W_\rho(\alpha) \underbrace{{\rm tr}\pi\,\delta(a-\alpha)R^{a^\dagger a}}_{f_R(\alpha)}. \tag{A39}$$

As $\delta(a-\alpha)$ can be written as $\delta(a-\alpha) = \frac{1}{\pi^2}\int d^2\beta\, e^{\alpha\beta^* - \alpha^*\beta} {\rm D}(\beta)$ [58] with ${\rm D}(\beta)$ the displacement operator with amplitude β, we have

$$\begin{aligned} f_R(\alpha) &= \frac{1}{\pi}\int d^2\beta\, e^{\alpha\beta^* - \alpha^*\beta} {\rm tr} R^{a^\dagger a} {\rm D}(\beta) \\ &= \frac{1}{\pi}\int d^2\beta\, e^{\alpha\beta^* - \alpha^*\beta} e^{|\beta|^2/2} {\rm tr} R^{a^\dagger a} e^{-\beta^* a} e^{\beta a^\dagger} \\ &= \frac{1}{\pi}\int d^2\beta\, e^{\alpha\beta^* - \alpha^*\beta} e^{|\beta|^2/2} {\rm tr} e^{\beta a^\dagger} R^{a^\dagger a} e^{-\beta^* a} \\ &= \frac{1}{\pi^2}\int d^2\beta\, e^{\alpha\beta^* - \alpha^*\beta} e^{|\beta|^2/2} \int d^2\gamma\, e^{\beta\gamma^*} e^{(R-1)|\gamma|^2} e^{-\beta^*\gamma} \\ &= \frac{1}{\pi(1-R)}\int d^2\beta e^{\alpha\beta^* - \alpha^*\beta} e^{-\frac{1}{2}\frac{1+R}{1-R}|\beta|^2} \\ &= \frac{2}{(1+R)} e^{-2|\alpha|^2\frac{1-R}{1+R}}, \end{aligned} \tag{A40}$$

where we used a writing of $\mathrm{D}(\beta)$ as $e^{|\beta|^2/2}e^{-\beta^* a}e^{\beta a^\dagger}$ in the second equality, the cyclic property of the trace in the third equality, and the writing of $R^{a^\dagger a}$ in the normal order $:e^{(R-1)a^\dagger a}:$ in the fourth equality. We deduce that the probability for having a "no click" for any monomode state ρ can be computed from its Wigner function as

$$P_{\text{no-click}} = \int d^2\alpha W_\rho(\alpha) f_R(\alpha)$$
$$\text{with} \quad f_R(\alpha) = \frac{2}{(1+R)} e^{-2|\alpha|^2 \frac{1-R}{1+R}}. \tag{A41}$$

c. Heralding on photon detections. We now consider the case with n modes with a photon detection on mode i. The subnormalized state resulting from a "no-click event" on mode i is given by

$$\begin{aligned}\tilde{\rho}_{\circ i} &= \mathrm{tr}_i R^{a_i^\dagger a_i}\rho \\ &= \mathrm{tr}_i R^{a_i^\dagger a_i}\pi^n \int d^2\boldsymbol{\alpha} W(\boldsymbol{\alpha}) \bigotimes_{j=1}^{n} \delta(a_j - \alpha_j) \\ &= \pi^{n-1}\int d^2\boldsymbol{\alpha} W(\boldsymbol{\alpha})(\mathrm{tr}_i \pi R^{a_i^\dagger a_i}\delta(a_i-\alpha_i)) \bigotimes_{\substack{j=1\\ j\neq i}}^{n} \delta(a_j-\alpha_j) \\ &= \pi^{n-1}\int d^2\boldsymbol{\alpha} W(\boldsymbol{\alpha}) f_R(\alpha_i) \bigotimes_{\substack{j=1\\ j\neq i}}^{n} \delta(a_j-\alpha_j),\end{aligned} \tag{A42}$$

where f_R is defined in Eq. (A40). From Eq. (A12), the corresponding subnormalized Wigner function is

$$\tilde{W}_{\circ i}(\overline{\boldsymbol{\alpha}}) = \int d^2\alpha_i W(\boldsymbol{\alpha}) f_R(\alpha_i), \tag{A43}$$

with $\overline{\boldsymbol{\alpha}}$ the vector with $(n-1)$ elements constructed by dropping the ith element of the vector $\boldsymbol{\alpha}$. Let us compute the normalization for Gaussian states. We have

$$\tilde{W}_{\circ i}(\overline{\boldsymbol{\alpha}}) = \int d\tilde{\boldsymbol{\alpha}}_i \frac{\exp\left(-\frac{1}{2}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu})^T\Sigma^{-1}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu})\right)}{\sqrt{\det(2\pi\,\Sigma)}} f_R(\tilde{\boldsymbol{\alpha}}_i), \tag{A44}$$

where we denoted $\tilde{\boldsymbol{\alpha}}_i = \binom{\mathrm{Re}\alpha_i}{\mathrm{Im}\alpha_i}$. From Eq. (A40), we have

$$\begin{aligned} f_R(\tilde{\boldsymbol{\alpha}}_i) &= \frac{2}{(1+R)}\exp\left(-|\tilde{\boldsymbol{\alpha}}_i|^2\frac{2(1-R)}{1+R}\right) \\ &= \frac{2}{(1+R)}\exp\left(-\frac{2(1-R)}{1+R}\tilde{\boldsymbol{\alpha}}_i^T\tilde{\boldsymbol{\alpha}}_i\right) \\ &= \frac{2}{(1+R)}\exp\left(-\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T F\tilde{\boldsymbol{\alpha}}\right),\end{aligned} \tag{A45}$$

where F is a matrix composed of $n-1$ blocks F_j, each block being a 2×2 block, such that

$$F_{j\neq i} = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}, \quad F_{j=i} = \frac{4(1-R)}{1+R}\mathbb{1}_2. \tag{A46}$$

Equation (A44) can therefore be written as

$$\begin{aligned}\tilde{W}_{\circ i}(\overline{\boldsymbol{\alpha}}) = \frac{2}{(1+R)\sqrt{\det(2\pi\,\Sigma)}}\int d\tilde{\boldsymbol{\alpha}}_i \\ \times \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu})^T\Sigma^{-1}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu}) - \frac{1}{2}\tilde{\boldsymbol{\alpha}}^T F\tilde{\boldsymbol{\alpha}}\right).\end{aligned} \tag{A47}$$

The integral can be rewritten as an integral of a multivariate distribution with a constant factor. To do so, we start by expressing the term in the exponent as

$$\begin{aligned}&\tfrac{1}{2}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu})^T\Sigma^{-1}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu}) + \tfrac{1}{2}\tilde{\boldsymbol{\alpha}}^T F\tilde{\boldsymbol{\alpha}} \\ &\quad = \tfrac{1}{2}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{w})^T(\Sigma^{-1}+F)(\tilde{\boldsymbol{\alpha}}-\boldsymbol{w}) \\ &\qquad + \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{w}^T(\Sigma^{-1}+F)\boldsymbol{w}\end{aligned} \tag{A48}$$

with $\boldsymbol{w} = (\Sigma^{-1}+F)^{-1}\Sigma^{-1}\boldsymbol{\mu}$. This leads to

$$\begin{aligned}&\exp\left(-\frac{1}{2}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu})^T\Sigma^{-1}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\mu}) - \frac{1}{2}\tilde{\boldsymbol{\alpha}}^T F\tilde{\boldsymbol{\alpha}}\right) \\ &\quad = \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{w})^T(\Sigma^{-1}+F)(\tilde{\boldsymbol{\alpha}}-\boldsymbol{w})\right)\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{w}^T(\Sigma^{-1}+F)\boldsymbol{w}\right) \\ &\quad = N(\boldsymbol{\alpha};\boldsymbol{w},(\Sigma^{-1}+F)^{-1})\frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{w}^T(\Sigma^{-1}+F)\boldsymbol{w}\right)}{\sqrt{\det((\Sigma^{-1}+F)/2\pi)}},\end{aligned} \tag{A49}$$

where we used the identity $\det(2\pi X) = 1/\det(X^{-1}/2\pi)$. Equation (A47) can now be rewritten as

$$\tilde{W}_{\circ i}(\overline{\boldsymbol{\alpha}}) = \frac{2}{1+R}\sqrt{\frac{1}{\det(\Sigma)\det(\Sigma^{-1}+F)}}e^{-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{w}^T(\Sigma^{-1}+F)\boldsymbol{w}}\int d^2\alpha_i N(\boldsymbol{\alpha};\boldsymbol{w},(\Sigma^{-1}+F)^{-1}). \tag{A50}$$

Note that the marginal of the multivariate normal distribution is normalized. We deduce that the properly normalized Wigner function $W_{\circ i}(\boldsymbol{\alpha})$, which can be written as $\tilde{W}_{\circ i}(\boldsymbol{\alpha}) = p_{\circ i}W_{\circ i}(\boldsymbol{\alpha})$, is given by

$$W_{\circ i}(\overline{\boldsymbol{\alpha}}) = N(\overline{\boldsymbol{\alpha}};\boldsymbol{\mu}',\Sigma'), \tag{A51}$$

where the displacement vector $\boldsymbol{\mu}'$ is obtained by removing the elements $2i$ and $2i+1$ of $(\Sigma^{-1}+F)^{-1}\Sigma^{-1}\boldsymbol{\mu}$ and the covariance matrix Σ' is obtained by removing the rows and columns $2i$ and $2i+1$ of $(\Sigma^{-1}+F)^{-1}$. $p_{\circ i}$, the probability of a no-click outcome when applying a NPNR detector on mode i, is given by

$$p_{\circ i} = \frac{2}{1+R}\sqrt{\frac{1}{\det(\Sigma)\det(\Sigma^{-1}+F)}} \times e^{-\frac{1}{2}\boldsymbol{\mu}^T(\Sigma^{-1}-\Sigma^{-1}(\Sigma^{-1}+F)^{-1}\Sigma^{-1})\boldsymbol{\mu}}. \tag{A52}$$

We can finally express the state $\rho_{\bullet i}$ conditioned on a "click" on mode i by considering its connection with the subnormalized state

$$\tilde{\rho}_{\bullet i} = \mathrm{tr}_i\rho(\mathbb{1} - R^{a_i^\dagger a_i}) = \rho_{\neg i} - p_{\circ i}\rho_{\circ i} = p_{\bullet i}\rho_{\bullet i}, \tag{A53}$$

where $p_{\bullet i} = 1 - p\circ i$ is the probability of having a click on mode i. We deduce

$$\rho_{\bullet i} = \frac{\rho_{\neg i} - p_{\circ i}\rho_{\circ i}}{1 - p_{\circ i}} \tag{A54}$$

and the Wigner function of this normalized state is given by

$$W_{\bullet i}(\overline{\boldsymbol{\alpha}}) = \frac{W_{\neg i}(\overline{\boldsymbol{\alpha}}) - p_{\circ i}W_{\circ i}(\overline{\boldsymbol{\alpha}})}{1 - p_{\circ i}} = \frac{W_{\neg i}(\overline{\boldsymbol{\alpha}}) - \tilde{W}_{\circ i}(\overline{\boldsymbol{\alpha}})}{1 - p_{\circ i}}. \tag{A55}$$

This shows that the Wigner function of the conditional state $\rho_{\bullet i}$ can be written as a weighted sum of Gaussian Wigner functions. Note that Gaussian operations acting on the conditional state can be accounted for by first considering their actions on individual Gaussian Wigner functions and by then recombining the two branches according to the weighted sum of initial Wigner functions.

d. Statistics of NPNR detections on multiple modes. We finally consider the detection of m modes with NPNR detectors and show the way to compute the probability of outcomes. We represent an arbitrary outcome by a vector $\mathbf{k}$ where the ith component K_i equals 0 for a no-click event and 1 for a click. The probability of getting such an outcome is given by

$$\langle E_{\mathbf{k}}\rangle = \left\langle \bigotimes_{i|k_i=0} R^{a_i^\dagger a_i} \bigotimes_{j|k_j=1} (\mathbb{1} - R^{a_j^\dagger a_j}) \right\rangle. \tag{A56}$$

According to Eq. (A39), we have

$$\begin{aligned}\langle E_{\mathbf{k}}\rangle &= \int d^2\boldsymbol{\alpha}\, W_\rho(\boldsymbol{\alpha}) \sum_{\boldsymbol{\ell}\in S_{\boldsymbol{k}}} (-1)^{k\cdot\ell} \prod_{i=0}^{n} (f_R(\alpha_i))^{l_i} \\ &= \sum_{\boldsymbol{\ell}\in S_{\boldsymbol{k}}} (-1)^{\boldsymbol{k}\cdot\boldsymbol{\ell}} \langle \hat{o}_{\boldsymbol{\ell}}\rangle,\end{aligned} \tag{A57}$$

where $S_{\boldsymbol{k}}$ is the set containing all strings $\boldsymbol{\ell}$ with n bits where the components ℓ_i are fixed to 1 for all i such that $k_i = 0$. This set contains $2^{|\boldsymbol{k}|}$ terms with $|\boldsymbol{k}| = \sum_i k_i$. From the value of $f_R(\alpha)$ given in Eq. (A40), we have

$$\begin{aligned}\langle \hat{o}_{\boldsymbol{\ell}}\rangle &= \left(\frac{2}{1+R}\right)^{|\boldsymbol{\ell}|} \int d^2\boldsymbol{\alpha}\, W(\boldsymbol{\alpha}) \exp\left(-\frac{2(1-R)}{1+R}\sum_i \ell_i|\alpha_i|^2\right) \\ &= \left(\frac{2}{1+R}\right)^{|\boldsymbol{\ell}|} \int d^2\boldsymbol{\alpha}\, W(\boldsymbol{\alpha}) \exp\left(\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T O_{\boldsymbol{\ell}}\tilde{\boldsymbol{\alpha}}\right) \\ &= \left(\frac{2}{1+R}\right)^{|\boldsymbol{\ell}|} \int d^2\boldsymbol{\alpha}\, N(\boldsymbol{\alpha};\boldsymbol{\mu},\Sigma)\, e^{-\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T O_{\boldsymbol{\ell}}\tilde{\boldsymbol{\alpha}}} \\ &= \left(\frac{2}{1+R}\right)^{|\boldsymbol{\ell}|} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(\Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1}+O_{\boldsymbol{\ell}})^{-1}\Sigma^{-1})\boldsymbol{\mu}\right)}{\sqrt{\det(\Sigma)\det(\Sigma^{-1}+O_{\boldsymbol{\ell}})}}.\end{aligned} \tag{A58}$$

where $|\boldsymbol{\ell}| = \sum_i \ell_i$ and $O_{\boldsymbol{\ell}}$ is the block-diagonal matrix

$$O_{\boldsymbol{\ell}} = \frac{4(1-R)}{1+R}\bigoplus_{i=1}^{m} \ell_i \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{A59}$$

The last equality is obtained by noting that

$$(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\mu})^T\Sigma^{-1}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\mu}) + \tilde{\boldsymbol{\alpha}}^T O_{\boldsymbol{\ell}}\tilde{\boldsymbol{\alpha}} = (\tilde{\boldsymbol{\alpha}} - \boldsymbol{w})^T(\Sigma^{-1} + O_\ell)(\tilde{\boldsymbol{\alpha}} - \boldsymbol{w}) + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{w}^T(\Sigma^{-1} + O_\ell)\boldsymbol{w} \tag{A60}$$

for $\boldsymbol{w} = (\Sigma^{-1} + O_\ell)^{-1}\Sigma^{-1}\,\boldsymbol{\mu}$ and

$$\int d^2\boldsymbol{\alpha} N(\boldsymbol{\alpha};\boldsymbol{w},(\Sigma^{-1}+O_\ell)^{-1}) = 1. \tag{A61}$$

The expected statistics is obtained by combining Eqs. (A57) and (A58).

APPENDIX B: AUTOMATED DESIGN IMPLEMENTATION

We here provide a description of how we automatize the design of quantum photonics experiments for DIQKD. This automatization is based on reinforcement learning, a subfield of machine learning. Reinforcement learning (RL) algorithms aim at finding out what action an *agent* should take when

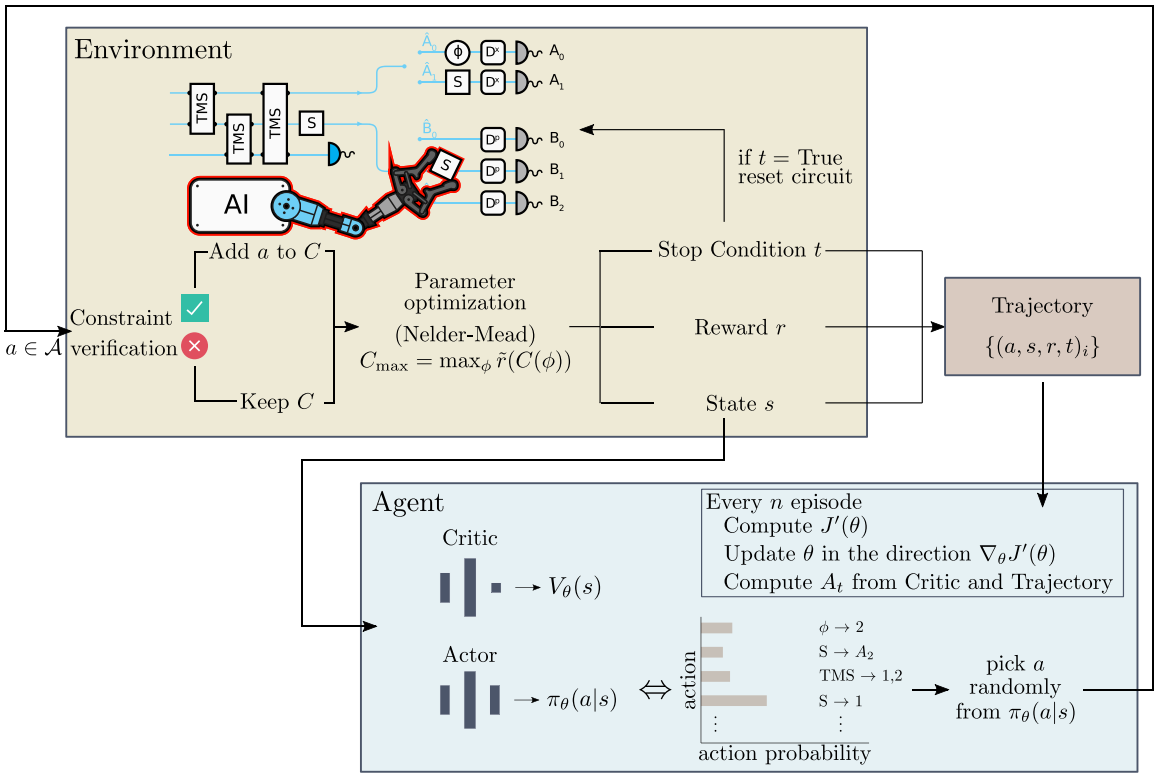

FIG. 4. Reinforcement learning for automated circuit design. An action a is sent to the environment. If it verifies some constraints (avoid redundancy, avoid beam splitter on the vacuum, etc.) it is added to the optical circuit C; otherwise C is unchanged. The circuit's parameters are then optimized. A resulting reward r, state s, and stop condition t are passed both to the agent and to a memory or trajectory. If the stop condition is *true* the circuit is reset to empty modes and the episode ends. Otherwise, for s and the policy π_θ the agent picks a new action and sends it to the environment. Every n episode, the agent uses the trajectory to update the parameters θ to maximize future rewards.

interacting with its *environment* in order to maximize the cumulative reward (see Fig. 4). Here, we give a quick overview of policy gradients and the proximal policy optimization algorithm—the type of agent we used. Then we dive into the details of our implementation before concluding with its convergence.

We emphasize that our algorithm is inspired by the one presented in Ref. [44]. There, it has shown how to automate the search of optical experiments for winning nonlocal games using projective simulation, simulated annihilating, and a numerical framework based on Fock representation. Our approach differs in all these aspects—we used policy-gradient-based reinforcement learning, Nelder-Mead optimization, and a faster and more reliable custom-made numerical framework based on Gaussian representation of states to simulate optical circuits.

1. Policy gradient overview

The behavior of an agent during its interaction with the environment is captured by its (stochastic) *policy* $\pi_\theta(a|s)$: the probability distribution of taking action $a \in \mathcal{A}$ when perceiving a state s. Policy gradient methods are a type of RL algorithms that aim at learning the parameters θ of the policy $\pi_\theta(a|s)$ in order to find the strategy maximizing an estimate of the sum of total future rewards. This learning occurs by sampling trajectories—a sequence of state s, action a, and reward r—for a given policy, followed by an update of the policy's parameters that favors the trajectories with the highest cumulative reward.

For a given reward function $J_\theta(r, s)$, gradient ascent is used to update the parameters θ by a certain amount given by the learning step. Note that this learning step is a hyperparameter crucial to training stability; e.g., a too small step results in a stagnation of the policy whereas a too large step changes may hinders the convergence of the policy.

For a concise overview of different policy gradient methods, consult Ref. [55]. Here, we note that policy gradient algorithms can also be used in model-free RL. Model-free algorithms optimize the policy without any knowledge of the internal functioning of the environment, i.e., the reward function and the transition function (the probability to obtain an output state given an input state and an action). This approach is very versatile and well suited for environments with a complex transition model, e.g., if the state transition depends on an internal complex stochastic process such as a nonlinear optimization as in our case (see below).

2. Proximal policy optimization

For our implementation we use a model-free policy gradient method known as proximal policy optimization (PPO). This algorithm is known to be sample efficient and has seen numerous successful applications while remaining fairly simple to implement and use.

PPO is an *on-policy* algorithm; i.e., policy updates are computed based on the latest policy π_{old} (parametrized by some parameter value $\tilde{\theta}$) and a batch of trajectories $\{(\boldsymbol{a}_t, \boldsymbol{s}_t, \boldsymbol{r}_t)\}$ sampled from this policy. The *clipped surrogate* reward function used by PPO reads [56]

$$J_\theta^{\text{CLIP}} = \mathbb{E}_t[\min(R_t(\theta)A_t(\theta),\ \text{clip}(R_t(\theta), 1-\epsilon, 1+\epsilon)A_t(\theta))]. \tag{B1}$$

Here, $R_t(\theta) = \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)/\pi_{\text{old}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ is the probability ratio between a policy π_θ and the old policy π_{old}, $A_t(\theta)$ is the advantage function, ϵ is a hyperparameter, and the expected value $\mathbb{E}_t$ is taken over the batch of trajectories. The *clip* function is used to constrain the amplitude of policy changes, quantified by $R_t(\theta)$, to be within the interval $[1-\epsilon, 1+\epsilon]$. Specific to PPO, this limitation increases the robustness to suboptimal learning step choices by preventing large policy updates. The advantage function $A_t(\theta)$ quantifies how good was the choice of the action sequence $\boldsymbol{a}_t$ in reaction to the perceived states $\boldsymbol{s}_t$ when compared to a baseline. The baseline reward expected from a state s_k is estimated with the *value* function $V_\theta(s_k)$ (see below). For each step k of a given trajectory the advantage function thus compares the returned reward r_k plus the reward predicted for the next state, $V_\theta(s_{k+1})$ (resulting from the action a_k), to the predicted total baseline reward $V_\theta(s_k)$ for the observed state s_k. Formally, for a trajectory segment containing T steps, the advantage is given by

$$\begin{aligned} A_t &= \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \quad \text{with} \\ \delta_k &= r_k + \gamma V_\theta(s_{k+1}) - V_\theta(s_k), \end{aligned} \tag{B2}$$

where λ and γ are hyperparameters that can be tuned to discount rewards that are delayed in time from the action (see Ref. [56] for details). In summary, the product $R_t(\theta)A_t(\theta)$ appearing in the reward function J_θ^{CLIP} is large if the choice of the action sequence $\boldsymbol{a}_t$ is advantageous when compared to the baseline ($A_t > 0$) and the policy π_θ is more likely to choose these actions than π_{old} ($R_t > 1$).

The value function $V_\theta(s)$ estimates the predicted cumulative reward obtained when following the policy π_θ starting from the state s. It shares the parameters θ with the policy, and is learned by minimizing the error term

$$L_\theta^V = (V_\theta(s) - V^{\text{target}})^2, \tag{B3}$$

where V^{target} is the computed value from the trajectory batch. Since θ is common to both the policy and the value function, this error term can simply be added to the reward function given in Eq. (B1) [56].

Finally, to enhance the exploration of the search space, an entropic penalty term $H(\pi_\theta)$ is added to the reward function. It favors policies that are not deterministic, but explore different actions.

To summarize, the total PPO reward function used to optimize the policy for higher cumulative reward is

$$J_\theta = J_\theta^{\text{CLIP}} + \mathbb{E}_t\big[-c_1 L_\theta^V + c_2 H(\pi_\theta)\big], \tag{B4}$$

where weights c_1, c_2 are hyperparameters.

PPO implementation. We used an implementation of PPO available in the ReinforcementLearning.jl package [59]. This implementation uses two neural networks—an actor network and a critic network—sharing parameters θ.

The actor network acts as the policy. It takes as an input the state received from the environment. The output layer has one neuron for each of the possible actions. The higher is the value of the activation function for one of these neurons the higher is the probability to choose the corresponding action.

The critic network acts as the value function V_θ. It also takes the observed state as the input. The output layer is composed of a single neuron whose activation value is directly proportional to the outcome of V_θ.

We tested multiple hyperparameter configurations. We settled on the choice of a single hidden layer of 256 neurons for both neural networks. For the hyperparameters of the advantage function, we set the discount factor to 0.99 and the smoothing parameter λ to 0.95. We consider learning from trajectory of size $T = 32$. The reward function is parametrized by a clip range of $\epsilon = 0.1$ and weights $c_1 = 0.5$ and $c_2 = 10^{-3}$.

3. Interaction with the environment

The PPO agent interacts with an environment that plays the role of a virtual laboratory where the n-mode photonic setup is implemented and the DIQKD protocol is executed. As described in the main text, the photonic circuit can be decomposed in two "phases": the state preparation phase, ending with the heralding measurements of all but the two first modes, which is followed by the measurement phase, where Alice and Bob can each perform local operations on their mode before measuring it with a NPNR detector. Actions that are taken by the agent correspond to placing optical elements on specific locations in the circuit.

A *step* starts with the agent taking an action, then the environment updates the photonic circuit, optimizes the circuit parameters, and returns a corresponding reward, the updated state, and a stop condition. This stop condition is a Boolean variable that becomes true when the total number of actions taken by the agent reaches a threshold we fixed at 15. When the stop condition is true the circuit and the action counter are reset. An *episode* is defined as the series of actions taken until the stop condition occurs.

The optical elements we consider are displacement (D), phase shifter (Φ), single-mode squeezer (S), two-mode squeezer (TMS), and beam-splitter (BS) operations. These are all Gaussian operations that are detailed in the previous section. An action a is composed of an optical element and a location on the circuit, i.e., on which mode it acts on. In addition, we denote $a(\phi)$ the action a with its optical element parametrized by ϕ which is either a real number, if a is a phase shifter or a beam splitter, or a complex number otherwise.

In the state preparation phase, we allow for phase shifters and squeezers to be applied on all of the n modes. This results in the following set of actions:

$$\mathcal{A}_{\text{prep}} = \left\{\Phi^{(i)}, \text{S}^{(i)}, \text{TMS}^{(ij)}, \text{BS}^{(ij)}\right\} \tag{B5}$$

with $i, j \in \{1, \ldots, n\}$ specifying the modes on which they act. In the measurement phase, we consider Alice (Bob) to always perform a displacement along the x (p) direction before the NPNR detector. In addition, Alice and Bob can perform

actions from the sets

$$\mathcal{A}_{\text{meas}}^{\text{Alice}} = \{\text{D}_0^p, \text{S}_0, \text{PS}_0, \text{D}_1^p, \text{S}_1, \text{PS}_1\},$$
$$\mathcal{A}_{\text{meas}}^{\text{Bob}} = \{\text{D}_0^x, \text{S}_0, \text{PS}_0, \text{D}_1^x, \text{S}_1, \text{PS}_1, \text{D}_2^x, \text{S}_2, \text{PS}_2\}. \quad \text{(B6)}$$

Here, the subscripts denote the choice of the measurement setting following which the operation is performed (Alice and Bob only receive a single mode). Combining the three sets, we obtain the total set of all possible actions that can be taken in our environment,

$$\mathcal{A} = \mathcal{A}_{\text{prep}} \cup \mathcal{A}_{\text{meas}}^{\text{Alice}} \cup \mathcal{A}_{\text{meas}}^{\text{Bob}}. \quad \text{(B7)}$$

4. Circuit parameter optimization

A succession of N actions $a_1, \ldots, a_N$ from $\mathcal{A}$ with the corresponding parameter values $\phi_1, \ldots, \phi_N$ defines a photonic circuit

$$C(\boldsymbol{\phi}) = \{a_1(\phi_1), \ldots, a_n(\phi_N)\}, \quad \forall a_i \in \mathcal{A}. \quad \text{(B8)}$$

In order to avoid trivial actions and redundancy, we add some constraints on such circuits. If a phase shifter or a beam splitter is added on an empty mode, i.e., acting on the vacuum, we can simply discard the corresponding action from C. If two identical actions a are either consecutive or separated by actions that commute with a, the latter occurrence of a is discarded from C. Finally, we constrained the actions in the measurement set A_{meas} to be unique; i.e., for each $a \in A_{\text{meas}} = \mathcal{A}_{\text{meas}}^{\text{Alice}} \cup \mathcal{A}_{\text{meas}}^{\text{Bob}}$ we discard any other occurrence of a in C.

The circuit $C_{\max}$ is the circuit with parameters $\boldsymbol{\phi}_{\max}$ and noisy preprocessing probability p optimized to maximize the key rate obtained from Eq. (2). Such an optimization can be hard to perform since the key rate defined in Eq. (2) is only valid for CHSH score $S > 2$. To help the optimization to converge, we define the *extended key rate* as the continuation of the key rate formula

$$\tilde{r} = 1 - \tilde{I}_p - H(\mathbf{A}'|\mathbf{B}) \quad \text{(B9)}$$

with

$$\tilde{I}_p = \begin{cases} h\left(\frac{1+\sqrt{(|S|/2)^2-1}}{2}\right) - h\left(\frac{1+\sqrt{1-p(1-p)(8-|S|^2)}}{2}\right) & \text{if } |S| > 2, \\ 1 + h\left(\frac{1+\sqrt{|S|/2}}{2}\right) - h\left(\frac{1+\sqrt{1-p(1-p)|S|^2}}{2}\right) & \text{otherwise} \end{cases}$$

(see Fig. 5) to all possible values of the CHSH score, as plotted in Fig. 5. Note that $\tilde{r}$ is negative for local values of CHSH $|S| \leqslant 2$. The extended key rate function is what we used to optimize the parameters of the circuit C, i.e., to numerically solve

$$\boldsymbol{\phi}_{\max} = \text{argmax}_{\boldsymbol{\phi}}\, \tilde{r}(C(\boldsymbol{\phi})). \quad \text{(B10)}$$

Concretely, this optimization was done using the Nelder-Mead method [46]. For a new circuit, we use multiple random starting parameters. However, for a circuit C' constructed by adding a new action to a previously optimized circuit $C_{\max}$, we optimize the parameters of C' starting from $(\boldsymbol{\phi}_{\max}, 0)$. To avoid local optima, two additional optimizations with different starting points are performed, one from a random point and one from $(\boldsymbol{\phi}_{\max} + \varepsilon \boldsymbol{u}, 0)$, where $\boldsymbol{u}$ is a uniformly distributed real vector with value in [0,1] and ε is a scalar we fixed to 0.2.

In the case where $n \geqslant 3$ some modes are used for heralding in the state preparation phase. In this case, there exist circuits that never produce the heralding event, e.g., when heralding on an empty (vacuum) mode. Similarly, some circuit parameter values also render heralding impossible. In both of these cases, we fix the key rate to a dummy value of -1. This allows the minimization to run over the entire space of parameters and avoids the implementation of complicated parameter constraints.

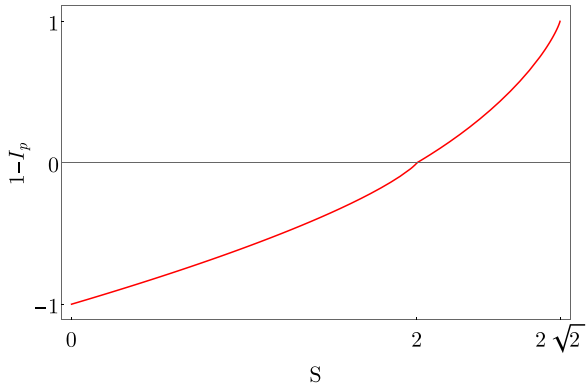

FIG. 5. Evolution of the extended version of $1 - I_p$ with the CHSH score S in the absence of noisy preprocessing ($p = 0$).

We tested our circuit optimization strategy on the reference photonic implementation of DIQKD shown in Fig. 1. We were able to recover similar key rates as the ones derived analytically in Ref. [49]. In particular, we found the same efficiency threshold for key rate higher than 10^{-9}. Below this order of magnitude, the optimization becomes too unstable and the numerical quantum optics simulation starts to induce non-negligible error due to matrix inverse operations.

5. Reward function

For a given circuit C, the environment computes a reward evaluating the performance of the circuit for a fixed task. We investigated two tasks for (i) finding circuits maximizing the key rate in the absence of loss, and (ii) finding a loss-tolerant circuit, i.e., maximizing loss while maintaining a key rate greater that 10^{-4}.

In the first case, a naive approach would be to use the extended key rate $\tilde{r}$ in Eq. (B9) as the reward:

$$r = \tilde{r}(C_{\max}). \quad \text{(B11)}$$

To help the PPO algorithm to converge, we reshape this reward using a secondary reward based on the CHSH value S attained by $C_{\max}$,

$$r = \frac{1}{1+\varepsilon}[\tilde{r}(C_{\max}) + \varepsilon\, |S(C_{\max})|], \quad \text{(B12)}$$

where we fixed the weight $\varepsilon = 10^{-2}$.

In the second case, the reward is the minimum efficiency so that the circuit can output a key rate greater than the threshold set to 10^{-4}. Denote C^η the circuit with NPNR detectors of efficiency η. To evaluate this reward, we start by optimizing the circuit C^η in the perfect case, i.e., with efficiency $\eta = 1$. We then lower η by a step s depending on the order of magnitude of the key rate $\tilde{r}(C_{\max})$ following

$$s = \max\left(2 \times 10^{\lfloor \log_{10}(\tilde{r}(C_{\max}))-1 \rfloor}, 10^{-3}\right). \quad \text{(B13)}$$

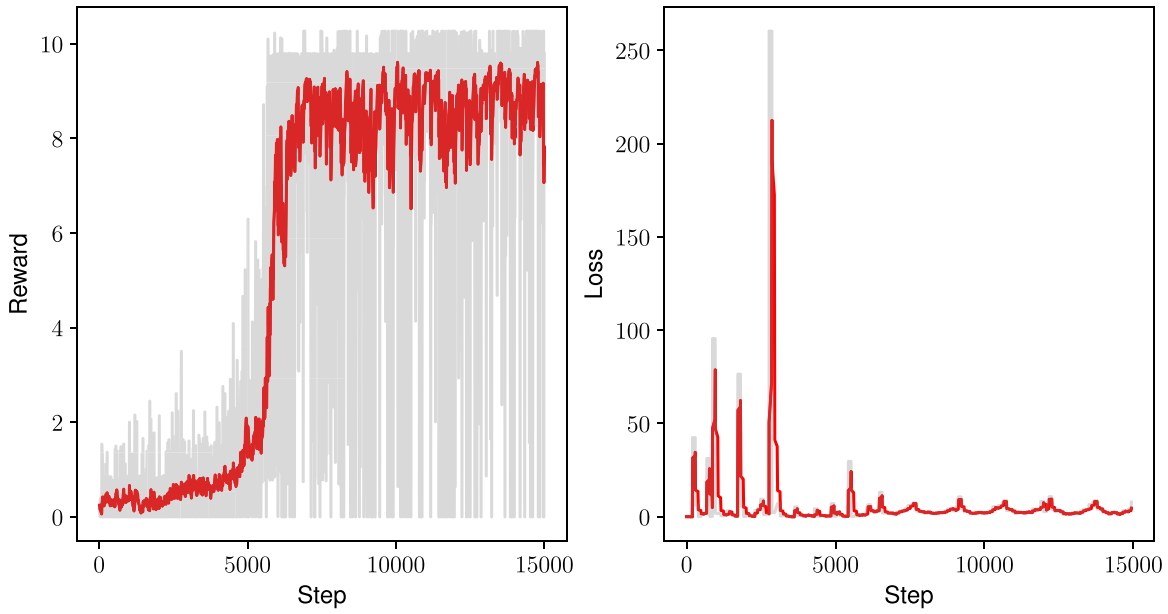

FIG. 6. Evolution of the reward r and loss J_θ with learning steps, in the case of PPO aiming to design quantum-optical circuits maximizing the key rate in a lossless scenario. The grey curves represent the raw data obtained during learning. The red curves are the smoothed evolution obtained using Eq. (B19) on the raw data and with weight $\omega = 0.9$.

For the newly obtained efficiency, we optimize the circuit again, starting from the optimal parameters found at the previous step. Eventually, η is too small and the optimal key rate found by the optimization does not exceed the threshold. We label $\eta_{\min}$ the smallest efficiency before this happens. The reward is then given by

$$r = \eta_{\min}. \tag{B14}$$

Note that in both tasks, the PPO algorithm learns from the total reward gathered per episode. The reward attributed at step i is thus defined as

$$\begin{cases} r_i - r_{i-1} & \text{if } r_i - r_{i-1} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{B15}$$

so that only the highest reward obtained during an episode matters.

6. The state perceived by the agent

For each circuit C, the state s returned by the environment to the PPO algorithm is taken from its optimized version $C_{\max}$ at the measurement phase, just before the NPNR detectors.

For each setting pair $\{x, y\}$ we extract the quantum state, i.e., the covariance matrix Σ and the displacement vector $\boldsymbol{\mu}$ for each Gaussian state composing the conditional state as in Eq. (A55). Note that, thanks to symmetries of the Hermitian nature of the covariance matrix on m modes, a single covariance matrix is parametrized by $(2m^2 + m)$ real parameters. Denote $\vec{\Sigma}$ the vector containing these real parameters. For a Gaussian state, all parameters are contained in the vector

$$\vec{p} = \begin{pmatrix} \vec{\Sigma} \\ \boldsymbol{\mu} \end{pmatrix}. \tag{B16}$$

Trivially, a heralded state represented as a quasimixture of 2^{n-m} Gaussian states with weight w_i can be represented by the vector

$$\boldsymbol{p} = \begin{pmatrix} w_1 \\ \vec{p}_1 \\ \vdots \\ w_{2^{n-m}} \\ \vec{p}_{2^{n-m}} \end{pmatrix} = \begin{pmatrix} w_1 \\ \vec{\Sigma}_1 \\ \boldsymbol{\mu}_1 \\ \vdots \\ w_{2^{n-m}} \\ \vec{\Sigma}_{2^{n-m}} \\ \boldsymbol{\mu}_{2^{n-m}} \end{pmatrix}. \tag{B17}$$

Labeling $\boldsymbol{p}_{xy}$ the vector containing the information of the quantum state for the setting choice x, y, the state returned by the environment is the vector

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{p}_{00} \\ \boldsymbol{p}_{01} \\ \vdots \\ \boldsymbol{p}_{12} \end{pmatrix}. \tag{B18}$$

7. PPO convergence

To grasp the convergence of our algorithm, we investigated the evolution of different factors with learning steps. The reward function, also called *loss*, used to optimize the policy, Eq. (B4), should converge to zero, i.e., showing a convergence of the learning algorithm and a successful descent of the gradient of this function. When the policy converges to a more optimal one, the reward received from the environment should, on average, increase with learning steps.

The evolution of these quantities is depicted in Fig. 6. As an example, we choose a scenario where the PPO is aiming for quantum-optical circuits maximizing key rates in a lossless scenario. In this case, the reward is simply the key rate

TABLE IV. Key rate and optical device parameters for the circuit depicted in Fig. 2(a). for different loss values η considered to be the same for all NPNR photon detectors. Parameters are for (a) state generation parameters and (b) measurement settings. Single- and two-mode squeezers have two parameters r, θ which are given in separate columns with names containing r and θ as a superscript. p denotes the noisy preprocessing probability.

(a)

η	Key rate	TMS^r_{12}	TMS^θ_{12}	TMS^r_{23}	TMS^θ_{23}	TMS^r_{13}	TMS^θ_{13}	SMS^r_2	SMS^θ_2
0.0	9.1372×10^{-1}	4.9723×10^{-2}	1.5727	1.8764×10^{-5}	9.8943×10^{-1}	1.8567×10^{-5}	2.5610	-5.1173×10^{-2}	1.1440×10^{-2}
0.01	7.6769×10^{-1}	5.3058×10^{-2}	1.5728	3.7045×10^{-4}	9.4867×10^{-1}	3.6731×10^{-4}	2.5192	-5.2103×10^{-2}	4.6623×10^{-3}
0.02	6.5271×10^{-1}	5.8176×10^{-2}	1.5701	3.9093×10^{-4}	9.7914×10^{-1}	3.8866×10^{-4}	2.5493	-5.2998×10^{-2}	-3.0361×10^{-4}
0.03	5.5266×10^{-1}	6.3723×10^{-2}	1.5689	4.2081×10^{-4}	1.0056	4.1874×10^{-4}	2.5762	-5.3323×10^{-2}	-7.4331×10^{-3}
0.04	4.6313×10^{-1}	7.0504×10^{-2}	1.5731	3.9112×10^{-4}	9.8700×10^{-1}	3.8951×10^{-4}	2.5583	-5.4637×10^{-2}	7.3763×10^{-3}
0.05	3.8185×10^{-1}	7.7611×10^{-2}	1.5706	4.2287×10^{-4}	1.0027	4.2152×10^{-4}	2.5735	-5.3403×10^{-2}	-3.8425×10^{-4}
0.06	3.0744×10^{-1}	8.5503×10^{-2}	1.5709	4.4170×10^{-4}	1.0528	4.4066×10^{-4}	2.6236	-5.1200×10^{-2}	7.4612×10^{-4}
0.07	2.3905×10^{-1}	9.4187×10^{-2}	1.5710	5.3904×10^{-4}	1.0958	5.3820×10^{-4}	2.6667	-4.7661×10^{-2}	1.7861×10^{-4}
0.08	1.7625×10^{-1}	1.0426×10^{-1}	1.5708	7.9601×10^{-4}	1.1090	7.9513×10^{-4}	2.6798	-4.3630×10^{-2}	-2.3382×10^{-4}
0.09	1.1926×10^{-1}	1.1654×10^{-1}	1.5709	5.8393×10^{-4}	1.1251	5.8299×10^{-4}	2.6960	-4.0096×10^{-2}	2.9306×10^{-4}
0.1	6.9422×10^{-2}	1.3162×10^{-1}	1.5708	6.2310×10^{-4}	1.1442	6.2116×10^{-4}	2.7150	-3.6519×10^{-2}	-1.9758×10^{-4}
0.102	6.0558×10^{-2}	1.3495×10^{-1}	1.5708	5.2328×10^{-4}	1.1048	5.2148×10^{-4}	2.6756	-3.5594×10^{-2}	1.1047×10^{-4}
0.104	5.2132×10^{-2}	1.3832×10^{-1}	1.5708	5.3724×10^{-4}	1.1237	5.3521×10^{-4}	2.6946	-3.4552×10^{-2}	-1.5648×10^{-4}
0.106	4.4179×10^{-2}	1.4174×10^{-1}	1.5708	5.5538×10^{-4}	1.1332	5.5312×10^{-4}	2.7040	-3.3381×10^{-2}	1.1994×10^{-5}
0.108	3.6734×10^{-2}	1.4516×10^{-1}	1.5710	4.8560×10^{-4}	1.1484	4.8350×10^{-4}	2.7193	-3.2064×10^{-2}	3.9609×10^{-4}
0.11	2.9836×10^{-2}	1.4859×10^{-1}	1.5708	2.9157×10^{-4}	1.1589	2.9024×10^{-4}	2.7298	-3.0676×10^{-2}	1.1961×10^{-3}
0.112	2.3521×10^{-2}	1.5205×10^{-1}	1.5709	3.5600×10^{-4}	1.1596	3.5436×10^{-4}	2.7305	-2.8905×10^{-2}	2.2124×10^{-3}
0.114	1.7832×10^{-2}	1.5542×10^{-1}	1.5708	5.4197×10^{-4}	1.1669	5.3943×10^{-4}	2.7378	-2.7330×10^{-2}	-7.1217×10^{-4}
0.116	1.2817×10^{-2}	1.5881×10^{-1}	1.5708	2.6547×10^{-4}	1.1815	2.6423×10^{-4}	2.7522	-2.5443×10^{-2}	-4.7823×10^{-4}
0.118	8.5244×10^{-3}	1.6211×10^{-1}	1.5708	2.7929×10^{-4}	1.1937	2.7800×10^{-4}	2.7645	-2.3568×10^{-2}	-7.5283×10^{-4}
0.12	5.0165×10^{-3}	1.6536×10^{-1}	1.5710	2.9523×10^{-4}	1.2059	2.9391×10^{-4}	2.7768	-2.1515×10^{-2}	6.6004×10^{-4}
0.121	3.5785×10^{-3}	1.6694×10^{-1}	1.5709	3.0890×10^{-4}	1.2246	3.0756×10^{-4}	2.7954	-2.0588×10^{-2}	-8.9291×10^{-4}
0.122	2.3646×10^{-3}	1.6856×10^{-1}	1.5711	3.2344×10^{-4}	1.2425	3.2206×10^{-4}	2.8133	-1.9541×10^{-2}	2.6670×10^{-3}
0.123	1.3864×10^{-3}	1.7007×10^{-1}	1.5711	3.3647×10^{-4}	1.2583	3.3506×10^{-4}	2.8293	-1.8604×10^{-2}	2.1192×10^{-3}
0.124	6.5655×10^{-4}	1.7161×10^{-1}	1.5710	3.5080×10^{-4}	1.2693	3.4937×10^{-4}	2.8401	-1.7989×10^{-2}	7.9132×10^{-3}
0.125	1.8979×10^{-4}	1.7308×10^{-1}	1.5711	3.6626×10^{-4}	1.2815	3.6475×10^{-4}	2.8523	-1.7036×10^{-2}	1.4553×10^{-2}
0.126	2.8585×10^{-6}	1.7437×10^{-1}	1.5708	1.0760×10^{-4}	1.2278	1.0716×10^{-4}	2.7979	-1.5724×10^{-2}	7.6604×10^{-2}
0.1261	2.4878×10^{-7}	1.7426×10^{-1}	1.5731	1.0177×10^{-4}	1.2164	1.0133×10^{-4}	2.7870	-1.7422×10^{-2}	1.0454×10^{-1}

(b)

η	Key rate	A_0: PS	A_0: D	A_1: SMS	A_1: D	B_0: D	B_1: SMS	B_1: D	B_2: D	p
0.0	9.1372×10^{-1}	1.5707×10^{-3}	2.2232×10^{-1}	-1.9493×10^{-1}	-6.3047×10^{-1}	-1.6790×10^{-1}	2.7628×10^{-1}	6.8478×10^{-1}	2.2199×10^{-1}	1.0000×10^{-9}
0.01	7.6769×10^{-1}	2.3369×10^{-4}	2.3975×10^{-1}	-1.9173×10^{-1}	-6.2839×10^{-1}	-1.6243×10^{-1}	2.8520×10^{-1}	7.0228×10^{-1}	2.4115×10^{-1}	-9.2873×10^{-10}
0.02	6.5271×10^{-1}	-8.0489×10^{-4}	2.6220×10^{-1}	-1.8428×10^{-1}	-6.2299×10^{-1}	-1.5334×10^{-1}	2.9663×10^{-1}	7.2669×10^{-1}	2.6435×10^{-1}	2.3234×10^{-9}
0.03	5.5266×10^{-1}	-2.3541×10^{-4}	2.8403×10^{-1}	-1.7780×10^{-1}	-6.1976×10^{-1}	-1.4537×10^{-1}	3.0614×10^{-1}	7.4890×10^{-1}	2.8736×10^{-1}	1.1074×10^{-8}
0.04	4.6313×10^{-1}	8.9333×10^{-4}	3.0692×10^{-1}	-1.7054×10^{-1}	-6.1750×10^{-1}	-1.3720×10^{-1}	3.1607×10^{-1}	7.7165×10^{-1}	3.1127×10^{-1}	2.0419×10^{-7}
0.05	3.8185×10^{-1}	5.4205×10^{-5}	3.3159×10^{-1}	-1.6266×10^{-1}	-6.1522×10^{-1}	-1.2794×10^{-1}	3.2359×10^{-1}	7.9474×10^{-1}	3.3707×10^{-1}	-6.0571×10^{-6}
0.06	3.0744×10^{-1}	6.9204×10^{-5}	3.5722×10^{-1}	-1.5410×10^{-1}	-6.1378×10^{-1}	-1.1854×10^{-1}	3.2982×10^{-1}	8.1813×10^{-1}	3.6390×10^{-1}	5.8982×10^{-5}
0.07	2.3905×10^{-1}	1.2173×10^{-4}	3.8416×10^{-1}	-1.4479×10^{-1}	-6.1308×10^{-1}	-1.0865×10^{-1}	3.3433×10^{-1}	8.4167×10^{-1}	3.9229×10^{-1}	-3.2192×10^{-4}
0.08	1.7625×10^{-1}	-3.4803×10^{-5}	4.1305×10^{-1}	-1.3407×10^{-1}	-6.1313×10^{-1}	-9.7854×10^{-2}	3.3794×10^{-1}	8.6579×10^{-1}	4.2301×10^{-1}	-1.3200×10^{-3}
0.09	1.1926×10^{-1}	3.9122×10^{-5}	4.4497×10^{-1}	-1.2094×10^{-1}	-6.1373×10^{-1}	-8.5505×10^{-2}	3.4164×10^{-1}	8.9149×10^{-1}	4.5766×10^{-1}	-4.8523×10^{-3}
0.1	6.9422×10^{-2}	1.9935×10^{-5}	4.8152×10^{-1}	-1.0426×10^{-1}	-6.1445×10^{-1}	-7.0290×10^{-2}	3.4499×10^{-1}	9.1957×10^{-1}	4.9826×10^{-1}	1.6936×10^{-2}
0.102	6.0558×10^{-2}	-9.6271×10^{-6}	4.8940×10^{-1}	-1.0046×10^{-1}	-6.1457×10^{-1}	-6.6840×10^{-2}	3.4542×10^{-1}	9.2542×10^{-1}	5.0711×10^{-1}	2.1503×10^{-2}
0.104	5.2132×10^{-2}	5.2224×10^{-5}	4.9736×10^{-1}	-9.6748×10^{-2}	-6.1486×10^{-1}	-6.3434×10^{-2}	3.4573×10^{-1}	9.3125×10^{-1}	5.1609×10^{-1}	2.7155×10^{-2}
0.106	4.4179×10^{-2}	-2.6837×10^{-5}	5.0539×10^{-1}	-9.2925×10^{-2}	-6.1519×10^{-1}	-5.9919×10^{-2}	3.4585×10^{-1}	9.3706×10^{-1}	5.2517×10^{-1}	3.4114×10^{-2}
0.108	3.6734×10^{-2}	1.1644×10^{-4}	5.1341×10^{-1}	-8.9279×10^{-2}	-6.1573×10^{-1}	-5.6551×10^{-2}	3.4572×10^{-1}	9.4272×10^{-1}	5.3429×10^{-1}	4.2646×10^{-2}
0.11	2.9836×10^{-2}	1.0767×10^{-4}	5.2146×10^{-1}	-8.5561×10^{-2}	-6.1627×10^{-1}	-5.3036×10^{-2}	3.4558×10^{-1}	9.4834×10^{-1}	5.4341×10^{-1}	5.3124×10^{-2}
0.112	2.3521×10^{-2}	-1.9377×10^{-6}	5.2963×10^{-1}	-8.1750×10^{-2}	-6.1683×10^{-1}	-4.9322×10^{-2}	3.4498×10^{-1}	9.5394×10^{-1}	5.5270×10^{-1}	6.6013×10^{-2}
0.114	1.7832×10^{-2}	5.9934×10^{-5}	5.3735×10^{-1}	-7.8393×10^{-2}	-6.1787×10^{-1}	-4.6363×10^{-2}	3.4448×10^{-1}	9.5919×10^{-1}	5.6155×10^{-1}	8.1947×10^{-2}
0.116	1.2817×10^{-2}	-3.7851×10^{-5}	5.4524×10^{-1}	-7.4857×10^{-2}	-6.1881×10^{-1}	-4.3045×10^{-2}	3.4364×10^{-1}	9.6446×10^{-1}	5.7058×10^{-1}	1.0191×10^{-1}
0.118	8.5244×10^{-3}	8.7705×10^{-6}	5.5282×10^{-1}	-7.1775×10^{-2}	-6.2006×10^{-1}	-4.0042×10^{-2}	3.4271×10^{-1}	9.6938×10^{-1}	5.7929×10^{-1}	1.2741×10^{-1}
0.12	5.0165×10^{-3}	1.0671×10^{-4}	5.6027×10^{-1}	-6.8865×10^{-2}	-6.2153×10^{-1}	-3.7223×10^{-2}	3.4148×10^{-1}	9.7409×10^{-1}	5.8794×10^{-1}	1.6088×10^{-1}
0.121	3.5785×10^{-3}	1.0696×10^{-4}	5.6387×10^{-1}	-6.7483×10^{-2}	-6.2235×10^{-1}	-3.5895×10^{-2}	3.4098×10^{-1}	9.7640×10^{-1}	5.9204×10^{-1}	1.8198×10^{-1}
0.122	2.3646×10^{-3}	5.6720×10^{-5}	5.6751×10^{-1}	-6.6126×10^{-2}	-6.2326×10^{-1}	-3.4557×10^{-2}	3.4023×10^{-1}	9.7863×10^{-1}	5.9631×10^{-1}	2.0721×10^{-1}

TABLE IV. (*Continued.*)

η	Key rate	A_0: PS	A_0: D	A_1: SMS	A_1: D	B_0: D	B_1: SMS	B_1: D	B_2: D	p
0.123	1.3864×10^{-3}	2.7193×10^{-4}	5.7096×10^{-1}	-6.4867×10^{-2}	-6.2408×10^{-1}	-3.3274×10^{-2}	3.3967×10^{-1}	9.8075×10^{-1}	6.0029×10^{-1}	2.3804×10^{-1}
0.124	6.5655×10^{-4}	9.9099×10^{-5}	5.7424×10^{-1}	-6.3973×10^{-2}	-6.2521×10^{-1}	-3.2334×10^{-2}	3.3942×10^{-1}	9.8289×10^{-1}	6.0407×10^{-1}	2.7753×10^{-1}
0.125	1.8979×10^{-4}	-1.9151×10^{-4}	5.7762×10^{-1}	-6.3854×10^{-2}	-6.2669×10^{-1}	-3.1611×10^{-2}	3.3897×10^{-1}	9.8491×10^{-1}	6.0821×10^{-1}	3.3330×10^{-1}
0.126	2.8585×10^{-6}	-7.7997×10^{-4}	5.8036×10^{-1}	-6.2803×10^{-2}	-6.2774×10^{-1}	-3.0501×10^{-2}	3.3751×10^{-1}	9.8633×10^{-1}	6.1170×10^{-1}	4.4008×10^{-1}
0.1261	2.4878×10^{-7}	4.8657×10^{-4}	5.7970×10^{-1}	-6.2676×10^{-2}	-6.2699×10^{-1}	-2.9760×10^{-2}	3.3945×10^{-1}	9.8663×10^{-1}	6.1103×10^{-1}	4.7140×10^{-1}

obtained for perfect detector efficiency. Furthermore, since the policy is stochastic, in order to get more relevant statistics, we trained the PPO on ten environments simultaneously. That is, a single agent interacts with ten optical setups in parallel. The reward in Fig. 6 is the cumulative reward received from these environments. Because of the stochastic nature of the learning process, it is relevant to study a smoothed evolution of the reward and loss with learning steps. We choose to define the smoothed evolution s of a quantity x at step t as

$$s(x,t) = \begin{cases} x(0) & \text{if } t = 0 \\ s(t-1)\omega + x(t)(1-\omega) & \text{otherwise,} \end{cases} \tag{B19}$$

for some weight $\omega \in [0, 1]$.

In Fig. 6, we see three phases occurring during the policy optimization. First is an exploration phase, where the policy is close to random and where the rewards obtained are relatively low. The loss is contained in a [0,100] interval. Then, a drastic increase in the reward occurs. This is first triggered by a high loss of around 250. Finally, both the reward and the loss plateau, with loss getting close to zero and the average reward over the ten environments oscillating around 0.84. This can be interpreted as a convergence of the policy (low loss) in a relevant region (high reward). Note that the spikes in the loss evolution are due to gradient descent on batches—some batch of data will randomly contain episodes with better rewards than others.

TABLE V. Key rate and optical device parameters for the circuit depicted in Fig. 2(b) for different loss values η considered to be the same for all NPNR photon detectors.

η	Key rate	TMS^r	TMS$^\theta$	A_0: D^x	A_1: SMS	A_1: D^x	B_0: D^p	B_1: D^p	B_2: D^p	p
0.0	4.6009×10^{-1}	7.4170×10^{-1}	1.5708	2.3779×10^{-2}	-3.1225×10^{-1}	-7.9996×10^{-1}	2.9255×10^{-1}	-3.8842×10^{-1}	-2.0662×10^{-2}	3.6645×10^{-9}
0.01	3.9246×10^{-1}	7.2670×10^{-1}	1.5708	3.3502×10^{-2}	-3.1168×10^{-1}	-7.9141×10^{-1}	2.8421×10^{-1}	-3.9874×10^{-1}	-2.9351×10^{-2}	2.1021×10^{-5}
0.02	3.3557×10^{-1}	7.1115×10^{-1}	1.5708	3.4711×10^{-2}	-3.1335×10^{-1}	-7.9398×10^{-1}	2.8304×10^{-1}	-4.0083×10^{-1}	-2.9978×10^{-2}	1.6367×10^{-4}
0.03	2.8411×10^{-1}	6.9311×10^{-1}	1.5708	3.4806×10^{-2}	-3.1546×10^{-1}	-7.9759×10^{-1}	2.8216×10^{-1}	-4.0152×10^{-1}	-2.9529×10^{-2}	-5.8623×10^{-4}
0.04	2.3708×10^{-1}	6.7195×10^{-1}	1.5708	3.4325×10^{-2}	-3.1768×10^{-1}	-8.0118×10^{-1}	2.8088×10^{-1}	-4.0115×10^{-1}	-2.8496×10^{-2}	1.4921×10^{-3}
0.05	1.9419×10^{-1}	6.4693×10^{-1}	1.5708	3.3433×10^{-2}	-3.2005×10^{-1}	-8.0434×10^{-1}	2.7863×10^{-1}	-3.9975×10^{-1}	-2.6997×10^{-2}	-3.1229×10^{-3}
0.06	1.5546×10^{-1}	6.1770×10^{-1}	1.5708	3.2269×10^{-2}	-3.2236×10^{-1}	-8.0644×10^{-1}	2.7495×10^{-1}	-3.9727×10^{-1}	-2.5186×10^{-2}	5.7512×10^{-3}
0.07	1.2103×10^{-1}	5.8384×10^{-1}	1.5708	3.0851×10^{-2}	-3.2452×10^{-1}	-8.0720×10^{-1}	2.6936×10^{-1}	-3.9328×10^{-1}	-2.3043×10^{-2}	9.6353×10^{-3}
0.08	9.1120×10^{-2}	5.4533×10^{-1}	1.5708	2.9309×10^{-2}	-3.2626×10^{-1}	-8.0579×10^{-1}	2.6113×10^{-1}	-3.8767×10^{-1}	-2.0729×10^{-2}	1.5054×10^{-2}
0.09	6.5879×10^{-2}	5.0244×10^{-1}	1.5708	2.7573×10^{-2}	-3.2714×10^{-1}	-8.0176×10^{-1}	2.4984×10^{-1}	-3.7990×10^{-1}	-1.8174×10^{-2}	2.2277×10^{-2}
0.1	4.5338×10^{-2}	4.5579×10^{-1}	1.5708	2.5684×10^{-2}	-3.2674×10^{-1}	-7.9437×10^{-1}	2.3491×10^{-1}	-3.6944×10^{-1}	-1.5525×10^{-2}	3.1541×10^{-2}
0.11	2.9350×10^{-2}	4.0610×10^{-1}	1.5708	2.3559×10^{-2}	-3.2433×10^{-1}	-7.8290×10^{-1}	2.1601×10^{-1}	-3.5568×10^{-1}	-1.2804×10^{-2}	4.3136×10^{-2}
0.12	1.7566×10^{-2}	3.5416×10^{-1}	1.5709	2.1184×10^{-2}	-3.1912×10^{-1}	-7.6619×10^{-1}	1.9295×10^{-1}	-3.3789×10^{-1}	-1.0095×10^{-2}	5.7584×10^{-2}
0.13	9.4578×10^{-3}	3.0070×10^{-1}	1.5707	1.8415×10^{-2}	-3.1026×10^{-1}	-7.4333×10^{-1}	1.6590×10^{-1}	-3.1523×10^{-1}	-7.4617×10^{-3}	7.5736×10^{-2}
0.14	4.3676×10^{-3}	2.4609×10^{-1}	1.5708	1.5200×10^{-2}	-2.9614×10^{-1}	-7.1216×10^{-1}	1.3524×10^{-1}	-2.8647×10^{-1}	-5.0579×10^{-3}	9.9171×10^{-2}
0.15	1.5731×10^{-3}	1.9043×10^{-1}	1.5707	1.1507×10^{-2}	-2.7524×10^{-1}	-6.7037×10^{-1}	1.0180×10^{-1}	-2.4964×10^{-1}	-2.9720×10^{-3}	1.3115×10^{-1}
0.16	3.4921×10^{-4}	1.3291×10^{-1}	1.5708	7.2657×10^{-3}	-2.4498×10^{-1}	-6.1388×10^{-1}	6.6762×10^{-2}	-2.0079×10^{-1}	-1.3032×10^{-3}	1.8008×10^{-1}
0.161	2.8648×10^{-4}	1.2691×10^{-1}	1.5708	6.8140×10^{-3}	-2.4118×10^{-1}	-6.0683×10^{-1}	6.3112×10^{-2}	-1.9485×10^{-1}	-1.1657×10^{-3}	1.8631×10^{-1}
0.162	2.3197×10^{-4}	1.2105×10^{-1}	1.5704	6.3903×10^{-3}	-2.3769×10^{-1}	-6.0040×10^{-1}	5.9713×10^{-2}	-1.8875×10^{-1}	-1.0167×10^{-3}	1.9330×10^{-1}
0.163	1.8506×10^{-4}	1.1517×10^{-1}	1.5707	5.9594×10^{-3}	-2.3351×10^{-1}	-5.9303×10^{-1}	5.6217×10^{-2}	-1.8270×10^{-1}	-9.0929×10^{-4}	2.0090×10^{-1}
0.164	1.4511×10^{-4}	1.0904×10^{-1}	1.5707	5.5199×10^{-3}	-2.2939×10^{-1}	-5.8546×10^{-1}	5.2683×10^{-2}	-1.7614×10^{-1}	-8.1228×10^{-4}	2.0850×10^{-1}
0.165	1.1153×10^{-4}	1.0276×10^{-1}	1.5711	5.0270×10^{-3}	-2.2507×10^{-1}	-5.7759×10^{-1}	4.9036×10^{-2}	-1.6903×10^{-1}	-6.9031×10^{-4}	2.1761×10^{-1}
0.166	8.3706×10^{-5}	9.6287×10^{-2}	1.5707	4.5631×10^{-3}	-2.2056×10^{-1}	-5.6962×10^{-1}	4.5414×10^{-2}	-1.6135×10^{-1}	-5.9316×10^{-4}	2.2587×10^{-1}
0.167	6.1064×10^{-5}	9.0015×10^{-2}	1.5710	4.1256×10^{-3}	-2.1595×10^{-1}	-5.6068×10^{-1}	4.1838×10^{-2}	-1.5403×10^{-1}	-5.0248×10^{-4}	2.3633×10^{-1}
0.168	4.3020×10^{-5}	8.3875×10^{-2}	1.5708	3.7193×10^{-3}	-2.1110×10^{-1}	-5.5223×10^{-1}	3.8439×10^{-2}	-1.4620×10^{-1}	-4.4855×10^{-4}	2.4747×10^{-1}
0.169	2.9019×10^{-5}	7.6837×10^{-2}	1.5708	3.1944×10^{-3}	-2.0594×10^{-1}	-5.4199×10^{-1}	3.4600×10^{-2}	-1.3722×10^{-1}	-3.3272×10^{-4}	2.5950×10^{-1}
0.17	1.8507×10^{-5}	7.0047×10^{-2}	1.5699	2.8113×10^{-3}	-1.9983×10^{-1}	-5.3160×10^{-1}	3.0950×10^{-2}	-1.2808×10^{-1}	-3.0096×10^{-4}	2.7186×10^{-1}
0.171	1.0952×10^{-5}	6.2750×10^{-2}	1.5715	2.3506×10^{-3}	-1.9393×10^{-1}	-5.2119×10^{-1}	2.7193×10^{-2}	-1.1761×10^{-1}	-1.6298×10^{-4}	2.8703×10^{-1}
0.172	5.8329×10^{-6}	5.5781×10^{-2}	1.5715	1.8728×10^{-3}	-1.8825×10^{-1}	-5.1090×10^{-1}	2.3722×10^{-2}	-1.0697×10^{-1}	-1.2892×10^{-4}	3.0609×10^{-1}
0.173	2.6476×10^{-6}	4.7932×10^{-2}	1.5706	1.5120×10^{-3}	-1.7837×10^{-1}	-4.9469×10^{-1}	1.9849×10^{-2}	-9.5389×10^{-2}	-9.7213×10^{-5}	3.2488×10^{-1}
0.174	9.1917×10^{-7}	3.8504×10^{-2}	1.5711	1.0041×10^{-3}	-1.7120×10^{-1}	-4.8087×10^{-1}	1.5460×10^{-2}	-7.9199×10^{-2}	-6.4233×10^{-5}	3.5135×10^{-1}
0.175	1.6981×10^{-7}	3.2052×10^{-2}	1.5722	8.0323×10^{-4}	-1.7038×10^{-1}	-4.7369×10^{-1}	1.2678×10^{-2}	-6.6971×10^{-2}	1.1025×10^{-5}	3.9505×10^{-1}
0.1756	1.0871×10^{-8}	3.1382×10^{-2}	1.5694	6.6202×10^{-4}	-1.6206×10^{-1}	-4.5882×10^{-1}	1.2088×10^{-2}	-6.8129×10^{-2}	-1.9528×10^{-4}	4.3388×10^{-1}

APPENDIX C: RESULTS

We here present the results of our automated approach of photonic-based DIQKD experimental design. Quantum-optical circuits maximizing our designed rewards are presented, as well as their corresponding parameters for different values of efficiency.

1. Discovered quantum-optical experiments for DIQKD

When using the reward given in Eq. (B12) to maximize the key rate in a noiseless scenario, the PPO algorithm converges to the circuit depicted in Fig. 2(a). Parameters optimizing the key rate for different efficiencies are given in Table IV. Similar circuits with extra gates that can freely be discarded (e.g., a phase shifter on the third mode just before the heralding operation) were also found by the agent. The occurrence of these similar circuits is simply explained by the reward not punishing for greater circuit depths. Furthermore, adding a "noncontributing" gate can nevertheless increase the reward due to a better parameter optimization.

In the case of the reward given in Eq. (B14) targeting circuits with a high loss tolerance, the PPO algorithm settles on the circuit drawn in Fig. 2(b). Table V lists the best parameter choices for different values of efficiency.

[1] C. H. Bennett and G. Brassard, Theor. Comput. Sci. **560**, 7 (2014).
[2] A. K. Ekert, Phys. Rev. Lett. **67**, 661 (1991).
[3] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Rev. Mod. Phys. **74**, 145 (2002).
[4] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, Rev. Mod. Phys. **81**, 1301 (2009).
[5] H.-K. Lo, M. Curty, and K. Tamaki, Nat. Photonics **8**, 595 (2014).
[6] S. Pirandola, U. L. Andersen, L. Banchi, M. Berta, D. Bunandar, R. Colbeck, D. Englund, T. Gehring, C. Lupo, C. Ottaviani, J. L. Pereira, M. Razavi, J. Shamsul Shaari, M. Tomamichel, V. C. Usenko, G. Vallone, P. Villoresi, and P. Wallden, Adv. Opt. Photonics **12**, 1012 (2020).
[7] F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, Rev. Mod. Phys. **92**, 025002 (2020).
[8] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Phys. Rev. Lett. **23**, 880 (1969).
[9] I. Šupić and J. Bowles, Quantum **4**, 337 (2020).
[10] S. Pironio, A. Acín, S. Massar, A. B. de la Giroday, D. N. Matsukevich, P. Maunz, S. Olmschenk, D. Hayes, L. Luo, T. A. Manning, and C. Monroe, Nature (London) **464**, 1021 (2010).
[11] L. J. Stephenson, D. P. Nadlinger, B. C. Nichol, S. An, P. Drmota, T. G. Ballance, K. Thirumalai, J. F. Goodwin, D. M. Lucas, and C. J. Ballance, Phys. Rev. Lett. **124**, 110501 (2020).
[12] D. P. Nadlinger, P. Drmota, B. C. Nichol, G. Araneda, D. Main, R. Srinivas, D. M. Lucas, C. J. Ballance, K. Ivanov, E. Y.-Z. Tan, P. Sekatski, R. L. Urbanke, R. Renner, N. Sangouard, and J.-D. Bancal, Nature (London) **607**, 682 (2022).
[13] W. Zhang, T. van Leent, K. Redeker, R. Garthoff, R. Schwonnek, F. Fertig, S. Eppelt, W. Rosenfeld, V. Scarani, C. C.-W. Lim, and H. Weinfurter, Nature (London) **607**, 687 (2022).
[14] B. G. Christensen, K. T. McCusker, J. B. Altepeter, B. Calkins, T. Gerrits, A. E. Lita, A. Miller, L. K. Shalm, Y. Zhang, S. W. Nam, N. Brunner, C. C. W. Lim, N. Gisin, and P. G. Kwiat, Phys. Rev. Lett. **111**, 130406 (2013).
[15] L. K. Shalm, E. Meyer-Scott, B. G. Christensen, P. Bierhorst, M. A. Wayne, M. J. Stevens, T. Gerrits, S. Glancy, D. R. Hamel, M. S. Allman, K. J. Coakley, S. D. Dyer, C. Hodge, A. E. Lita, V. B. Verma, C. Lambrocco, E. Tortorici, A. L. Migdall, Y. Zhang, D. R. Kumor *et al.*, Phys. Rev. Lett. **115**, 250402 (2015).
[16] Y. Liu, Q. Zhao, M.-H. Li, J.-Y. Guan, Y. Zhang, B. Bai, W. Zhang, W.-Z. Liu, C. Wu, X. Yuan, H. Li, W. J. Munro, Z. Wang, L. You, J. Zhang, X. Ma, J. Fan, Q. Zhang, and J.-W. Pan, Nature (London) **562**, 548 (2018).
[17] L. Shen, J. Lee, L. P. Thinh, J.-D. Bancal, A. Cerè, A. Lamas-Linares, A. Lita, T. Gerrits, S. W. Nam, V. Scarani, and C. Kurtsiefer, Phys. Rev. Lett. **121**, 150402 (2018).
[18] M. Giustina, M. A. M. Versteegh, S. Wengerowsky, J. Handsteiner, A. Hochrainer, K. Phelan, F. Steinlechner, J. Kofler, J.-A. Larsson, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, J. Beyer, T. Gerrits, A. E. Lita, L. K. Shalm, S. W. Nam, T. Scheidl, R. Ursin *et al.*, Phys. Rev. Lett. **115**, 250401 (2015).
[19] V. Caprara Vivoli, P. Sekatski, J.-D. Bancal, C. C. W. Lim, B. G. Christensen, A. Martin, R. T. Thew, H. Zbinden, N. Gisin, and N. Sangouard, Phys. Rev. A **91**, 012107 (2015).
[20] V. Zapatero, T. van Leent, R. Arnon-Friedman, W.-Z. Liu, Q. Zhang, H. Weinfurter, and M. Curty, npj Quantum Inf. **9**, 10 (2023).
[21] W.-Z. Liu, Y.-Z. Zhang, Y.-Z. Zhen, M.-H. Li, Y. Liu, J. Fan, F. Xu, Q. Zhang, and J.-W. Pan, Phys. Rev. Lett. **129**, 050502 (2022).
[22] M. A. Horne, A. Shimony, and A. Zeilinger, Phys. Rev. Lett. **62**, 2209 (1989).
[23] K. Banaszek and K. Wódkiewicz, Phys. Rev. Lett. **82**, 2009 (1999).
[24] R. García-Patrón, J. Fiurášek, N. J. Cerf, J. Wenger, R. Tualle-Brouri, and P. Grangier, Phys. Rev. Lett. **93**, 130409 (2004).
[25] S. Tanzilli, A. Martin, F. Kaiser, M. D. Micheli, O. Alibart, and D. Ostrowsky, Laser Photonics Rev. **6**, 115 (2012).
[26] E. Pelucchi, G. Fagas, I. Aharonovich, D. Englund, E. Figueroa, Q. Gong, H. Hannes, J. Liu, C.-Y. Lu, N. Matsuda, J.-W. Pan, F. Schreck, F. Sciarrino, C. Silberhorn, J. Wang, and K. D. Jöns, Nat. Rev. Phys. **4**, 194 (2021).
[27] P. Sekatski, J.-D. Bancal, X. Valcarce, E. Y.-Z. Tan, R. Renner, and N. Sangouard, Quantum **5**, 444 (2021).
[28] E. Woodhead, A. Acín, and S. Pironio, Quantum **5**, 443 (2021).
[29] P. Brown, H. Fawzi, and O. Fawzi, arXiv:2106.13692.
[30] E. Y.-Z. Tan, R. Schwonnek, K. T. Goh, I. W. Primaatmaja, and C. C.-W. Lim, npj Quantum Inf. **7**, 158 (2021).
[31] J. Kołodyński, A. Máttar, P. Skrzypczyk, E. Woodhead, D. Cavalcanti, K. Banaszek, and A. Acín, Quantum **4**, 260 (2020).
[32] J.-D. Bancal, L. Sheridan, and V. Scarani, New J. Phys. **16**, 033011 (2014).
[33] O. Nieto-Silleras, S. Pironio, and J. Silman, New J. Phys. **16**, 013035 (2014).

[34] Y. LeCun, Y. Bengio, and G. Hinton, Nature (London) **521**, 436 (2015).
[35] J. Schmidhuber, Neural Networks **61**, 85 (2015).
[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, Nature (London) **518**, 529 (2015).
[37] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, Science **362**, 1140 (2018).
[38] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger, Phys. Rev. Lett. **116**, 090405 (2016).
[39] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature (London) **549**, 195 (2017).
[40] V. Dunjko and H. J. Briegel, Rep. Prog. Phys. **81**, 074001 (2018).
[41] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).
[42] M. Krenn, M. Erhard, and A. Zeilinger, Nat. Rev. Phys. **2**, 649 (2020).
[43] M. Krenn, J. S. Kottmann, N. Tischler, and A. Aspuru-Guzik, Phys. Rev. X **11**, 031044 (2021).
[44] A. A. Melnikov, P. Sekatski, and N. Sangouard, Phys. Rev. Lett. **125**, 160401 (2020).
[45] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, Cambridge, MA, 2018).
[46] J. A. Nelder and R. Mead, Comput. J. **7**, 308 (1965).
[47] I. Devetak and A. Winter, Proc. R. Soc. London A **461**, 207 (2005).
[48] S. Pironio, A. Acín, N. Brunner, N. Gisin, S. Massar, and V. Scarani, New J. Phys. **11**, 045021 (2009).
[49] M. Ho, P. Sekatski, E. Y.-Z. Tan, R. Renner, J.-D. Bancal, and N. Sangouard, Phys. Rev. Lett. **124**, 230502 (2020).
[50] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, Rev. Mod. Phys. **84**, 621 (2012).
[51] G. Adesso, S. Ragy, and A. R. Lee, Open Syst. Inf. Dyn. **21**, 1440001 (2014).
[52] J. B. Brask, arXiv:2102.05748.
[53] X. Valcarce, QuantumOpticalCircuits.jl, https://github.com/xvalcarce/QuantumOpticalCircuits.jl (2021).
[54] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, SIAM Rev. **59**, 65 (2017).
[55] L. Weng, lilianweng.github.io (2018).
[56] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, arXiv:1707.06347.
[57] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, in *Proceedings of the 32nd International Conference on Machine Learning, Lille, France*, edited by F. Bach and D. Blei (PMLR, 2015), Vol. 37.
[58] W. Vogel and D.-G. Welsch, *Quantum Optics* (Wiley, New York, 2006).
[59] J. Tian *et al.*, https://github.com/JuliaReinforcementLearning/ReinforcementLearning.jl (2020).