npj | Quantum Information

www.nature.com/npjqi

ARTICLE OPEN

Check for updates

A neural network oracle for quantum nonlocality problems in networks

Tamás Kriváchy[1]✉, Yu Cai[1], Daniel Cavalcanti[2], Arash Tavakoli[3], Nicolas Gisin[1] and Nicolas Brunner[1]

Characterizing quantum nonlocality in networks is a challenging, but important problem. Using quantum sources one can achieve distributions which are unattainable classically. A key point in investigations is to decide whether an observed probability distribution can be reproduced using only classical resources. This causal inference task is challenging even for simple networks, both analytically and using standard numerical techniques. We propose to use neural networks as numerical tools to overcome these challenges, by learning the classical strategies required to reproduce a distribution. As such, a neural network acts as an oracle for an observed behavior, demonstrating that it is classical if it can be learned. We apply our method to several examples in the triangle configuration. After demonstrating that the method is consistent with previously known results, we give solid evidence that a quantum distribution recently proposed by Gisin is indeed nonlocal as conjectured. Finally we examine the genuinely nonlocal distribution recently presented by Renou et al., and, guided by the findings of the neural network, conjecture nonlocality in a new range of parameters in these distributions. The method allows us to get an estimate on the noise robustness of all examined distributions.

npj Quantum Information (2020)6:70 ; https://doi.org/10.1038/s41534-020-00305-x

INTRODUCTION

The possibility of creating stronger than classical correlations between distant parties has deep implications for both the foundations and applications of quantum theory. These ideas have been initiated by Bell[1], with subsequesnt research leading to the theory of Bell nonlocality[2]. In the Bell scenario multiple parties jointly share a single classical or quantum source, often referred to as local and nonlocal sources, respectively. Recently, interest in more decentralized causal structures, in which several independent sources are shared among the parties over a network, has been on the rise[3–6]. Contrary to the Bell scenario, in even slightly more complex networks the boundary between local and nonlocal correlations becomes nonlinear and the local set non-convex, greatly perplexing rigorous analysis. Though some progress has been made[7–23], we still lack a robust set of tools to investigate generic networks from an analytic and numerical perspective.

Here, we explore the use of machine learning in these problems. In particular we tackle the membership problem for causal structures, i.e., given a network and a distribution over the observed outputs, we must decide whether it could have been produced by using exclusively local resources. We encode the causal structure into a neural network and ask the network to reproduce the target distribution. By doing so, we approximate the question "does a local causal model exist?" with "is a local causal model learnable?". Neural networks have proven to be useful ansätze for generic nonlinear functions in terms of expressivity, ease of learning and robustness, both in- and outside the domain of physical sciences[24–28]. Machine learning has also been used in the study of nonlocality[29,30]. However, while the techniques of ref. [30] can only suggest if a distribution is local or nonlocal, the method employed here is more generative in spirit and provides a certificate that a distribution is local once it is learned, by actually constructing the local model.

In our approach we exploit that both causal structures and feedforward neural networks have their information flow determined by a directed acyclic graph. For any given distribution over observed variables and an ansatz causal structure, we train a neural network which respects that causal structure to reproduce the target distribution, as in Fig. 1. This is equivalent to having a neural network learn the local responses of the parties to their inputs. With this constraint, if the target distribution is inside the local set, then a sufficiently expressive neural network should be able to learn the appropriate response functions and reproduce it. For distributions outside the local set, we should see that the machine can not approximate the given target. This gives us a criterion for deciding whether a target distribution is inside the local set or not. In particular, if a given distribution is truly outside the local set, then by adding noise in a physically relevant way we should see a clear transition in the machine's behavior when entering the set of local correlations.

We explore the strength of this method by examining a notorious causal structure, the so-called "triangle" network, i.e., the causal structure in Fig. 1. The triangle configuration is among the simplest tripartite networks, yet it poses immense challenges theoretically and numerically. We use the triangle with quaternary outcomes as a test-bed for our neural network oracle. After checking for the consistency of our method with known results, we examine the so-called Elegant distribution, proposed in ref. [31], and the distribution proposed by Renou et al. in ref. [20]. Our method gives solid evidence that the Elegant distribution is outside the local set, as originally conjectured. The family of distributions proposed by Renou et al. was shown to be nonlocal in a certain regime of parameters. When examining the full range of parameters we not only recover the nonlocality in the already known regime, but also get a conjecture of nonlocality from the machine in another range of the parameters. Finally, we use our

[1]Department of Applied Physics, University of Geneva, CH-1211 Geneva, Switzerland. [2]ICFO, The Institute of Photonic Sciences, 08860 Castelldefels, Barcelona, Spain. [3]Dyson School of Design Engineering, Imperial College London, London SW7 2AZ, UK. ✉email: tamas.krivachy@gmail.com

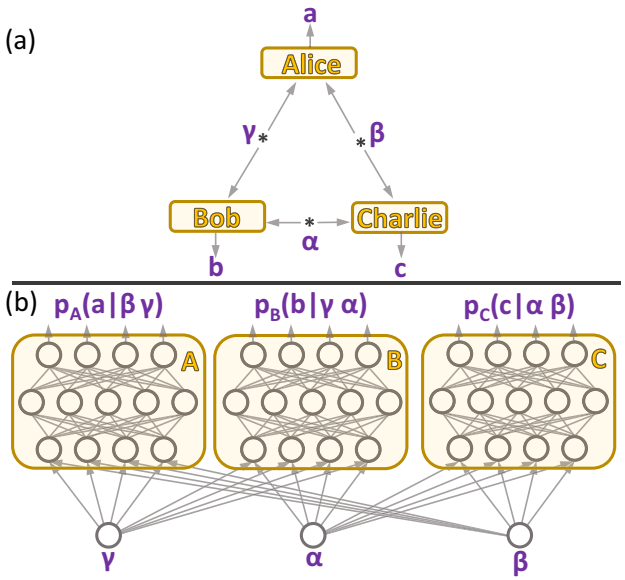

Fig. 1 Triangle network and its neural network encoding. a Triangle network configuration. **b** Neural network which reproduces distributions compatible with the triangle configuration.

method to get estimates of the noise robustness of these nonlocal distributions, and to gain insight into the learned strategies.

RESULTS

Encoding causal structures into neural networks

The methods developed in this work are in principle applicable to any causal structure. Here, we demonstrate how to encode a network nonlocality configuration into a neural network on the highly nontrivial example of the triangle network with quaternary outputs and no inputs. In this scenario three sources, α, β, γ, send information through either a classical or a quantum channel to three parties, Alice, Bob, and Charlie. Flow of information is constrained such that the sources are independent from each other, and each one only sends information to two parties of the three, as depicted in Fig. 1. Alice, Bob, and Charlie process their inputs with arbitrary local response functions, and they each output a number a, b, $c \in \{0, 1, 2, 3\}$, respectively. Under the assumption that each source is independent and identically distributed from round to round, and that the local response functions are fixed (though possibly stochastic), such a scenario is well characterized by the probability distribution $p(abc)$ over the random variables of the outputs.

If quantum channels are permitted from the sources to the parties then the set of distributions is larger than that achievable classically. Due to the nonlocal nature of quantum theory, these correlations are often referred to as nonlocal ones, as opposed to local behaviors arising from only using classical channels. In the classical case, the scenario is equivalent to a causal structure, otherwise known as a Bayesian network[32,33].

For the classical setup we can assume without loss of generality that the sources each send a random variable drawn from a uniform distribution on the continuous interval between 0 and 1 (any other distribution can be reabsorbed by the parties' response functions, e.g., via the inverse transform sampling method). Given the network constraint, the probability distribution over the parties' outputs can be written as

$$p(abc) = \int_0^1 da d\beta d\gamma \; p_A(a|\beta\gamma) p_B(b|\gamma a) p_C(c|a\beta), \tag{1}$$

where the conditional probability $p_X(x|\,\cdot\,,\cdot)$ is the response function of party X.

We now construct a neural network which is able to approximate a distribution of the form (1). We use a feedforward neural network, since it is described by a directed acyclic graph, similarly to a causal structure[32–34]. This allows for a seamless transfer from the causal structure to the neural network model. On a practical level, we represent each party's response function by a fully connected multilayer perceptron, one of the simplest artificial neural network architectures[34]. In our case, the inputs to the three perceptrons are the hidden variables, i.e., uniformly drawn random numbers α, β, γ. So as to respect the communication constraints of the triangle, inputs are routed to the three perceptrons in a restricted manner, as shown in Fig. 1. The outputs are the conditional probabilities conditioned on the respective inputs, $p_A(a|\beta\gamma)$, $p_B(b|\gamma\alpha)$, and $p_C(c|\alpha\beta)$, i.e., three normalized vectors, each of length 4. This restructuring can also be viewed as having one large, not fully connected multilayer perceptron, outputting the three probability vectors $p_A(a|\beta\gamma)$, $p_B(b|\gamma\alpha)$, $p_C(c|\alpha\beta)$ for a given input α, β, γ. Due to the restricted architecture, the output conditional probabilities will obey the causal network constraints, i.e., by construction only local models can be generated by such a neural network.

We evaluate the neural network for N_{batch} values of α, β, γ in order to approximate the joint probability distribution (1) with a Monte Carlo approximation,

$$p_M(abc) = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} p_A(a|\beta_i\gamma_i) p_B(b|\gamma_i a_i) p_C(c|a_i\beta_i). \tag{2}$$

Note that before summing over the batch, we take the Cartesian product of the conditional probability vectors. In our implementation each of these three conditional probability functions is modeled by a multilayer perceptron, with rectified linear or tangent hyperbolic activations, except at the last layer, where we have a softmax layer to impose normalization. Note, however, that any feedforward network can be used to model these conditional probabilities. The loss function can be any differentiable measure of discrepancy between the target distribution p_t and the neural network's output p_M, such as the Kullback–Leibler divergence of one relative to the other, namely

$$L(p_M) = \sum_{abc} p_t(abc) \log\left(\frac{p_t(abc)}{p_M(abc)}\right). \tag{3}$$

In order to train the neural network we synthetically generate uniform random numbers for the hidden variables, the inputs. We then adjust the weights of the network after evaluating the loss function on a minibatch of size N_{batch}, using conventional neural network optimization methods[34]. The minibatch size is chosen arbitrarily and can be increased in order to increase the neural network's precision. For the triangle with quaternary outputs an N_{batch} of several thousands is typically satisfactory.

By encoding the causal structure in a neural network like this, we can train the neural network to try to reproduce a given target distribution. The procedure generalizes in a straight-forward manner to any causal structure, and is thus in principle applicable to any quantum nonlocality network problem. We provide specific code online for the triangle configuration, as well as for the standard Bell scenario, which has inputs as well (see Section "Code availability"). After finishing this work we realized that related ideas have been investigated in causal inference, though in a different context, where network architectures and weights are simultaneously optimized to reproduce a given target distribution over continuous outputs, as opposed to discrete ones examined here[35]. In addition, due to the strict constraint of having a single fixed causal structure we evaluate results differently, by examining transitions in compatibility with the causal structure at hand, as we will soon demonstrate.

Evaluating the output of the neural network

Given a target distribution p_t, the neural network provides an explicit model for a distribution p_M, which is, according to the machine, the closest local distribution to p_t. The distribution p_M is guaranteed to be from the local set by construction. When can we confidently deduce that the target distribution is local (i.e., if we see $p_t \approx p_M$), or nonlocal ($p_t \neq p_M$)? At first sight the question is difficult, since the neural network will almost never exactly reproduce the target distribution since p_M is evaluated by sampling the model a finite number of times, and additionally the learning techniques do not guarantee convergence to the global optimum. A first approach could be to define some confidence level for the similarity between p_M and p_t. This would, however, be somewhat arbitrary, and would give only limited insight into the problem. A central notion in this work is to search for qualitative changes in the machine's behavior when transitioning from the local set to the nonlocal one. We believe this to be much more robust and informative for deciding nonlocality than a confidence level approach.

In order to find such a "phase transition", we typically define a family of target distributions $p_t(v)$ by taking a distribution which is believed to be nonlocal and by adding some noise controlled by the parameter v, with $p_t(v=0)$ being the completely noisy (local) distribution and $p_t(v=1)$ being the noiseless, "most nonlocal" one. By adding noise in a physically meaningful way we guarantee that at some parameter value, v^*, we will enter the local set and stay in it for $v < v^*$. For each noisy target distribution we retrain the neural network and obtain a family of learned distributions $p_M(v)$ (see Fig. 2 for an illustration). Observing a qualitative change in the machine's performance at some point is an indication of traversing the local set's boundary. In this work we extract information from the learned model through

- the distance between the target and the learned distribution,
$$d(p_t, p_M) = \sqrt{\sum_{abc} [p_t(abc) - p_M(abc)]^2},$$
- the learned distributions $p_M(v)$, in particular by examining the local response functions of Alice, Bob, and Charlie.

Observing a clear liftoff of the distance $d_M(v) := d(p_t(v), p_M(v))$ at some point is a signal that we are leaving the local set. Somewhat surprisingly, we can deduce even more from the distance $d_M(v)$. Though the shape of the local set and the threshold value v^* are unknown, in some cases, under mild assumptions, we can extract from $d_M(v)$ not only v^*, but also the angle at which the curve $p_t(v)$ exits the local set, and in addition gain confidence in the proper

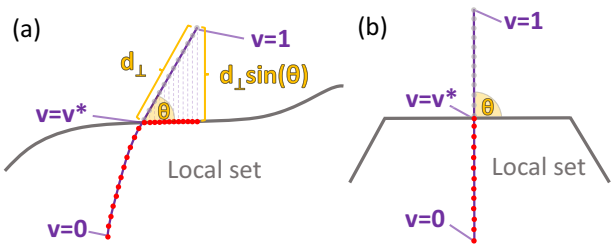

Fig. 2 Geometric considerations when evaluating neural network results. Visualization of target distributions $p_t(v)$ leaving the local set at an angle θ for **a** a generic noisy distribution and for **b** the specific case of the Fritz distribution with a two-qubit Werner state shared between Alice and Bob. The gray dots depict the target distributions, while the red dots depict the distributions which the neural network would find. In the generic case we additionally depict the distance $d_\perp := d(p_t(v), p_t(v^*))$ introduced in Eq. (4), for the special case of $v = 1$, as well as $d_\perp \sin\theta$. Given an estimate for v^*, the distance $d_\perp$ can be evaluated analytically, which (for an appropriate θ) allows us to compare $d_\perp \sin\theta$ with the distance that the machine perceives.

functioning of the algorithm. To do this, let us first assume that the local set is flat near $p_t(v^*)$ and that $p_t(v)$ is a straight curve. Then the true distance from the local set is

$$d(v) = \begin{cases} 0 & \text{if } v \leq v^* \\ d(p_t(v), p_t(v^*)) \sin(\theta) & \text{if } v > v^*, \end{cases} \quad (4)$$

where θ is the angle between the curve $p_t(v)$ and the local set's hyperplane (see Fig. 2 for an illustration). In the more general setting Eq. (4) is still approximately correct even for $v > v^*$, if $p_t(v)$ is almost straight and the local set is almost flat near v^*. We denote this analytic approximation of the true distance form the local set as $\hat{d}(v)$. We use Eq. (4) to calculate it but keep in mind that it is only an approximation. After having trained the machine, we fit $\hat{d}(v)$ to $d_M(v)$ by adjusting v^* and θ. Finding a good fit of the two distance functions gives us strong evidence that indeed the curve $p_t(v)$ exits the local set at $\hat{v}^*$ at an angle $\hat{\theta}$, where the hat is used to signify the obtained estimates. Acquiring such a fit gives us more confidence in the machine since now we do not just observe a qualitative phase transition, but we can also model it quantitatively with just two free parameters, v^* and θ.

In addition, we get information out of the learned model by looking at the local responses of Alice, Bob and Charlie. Recall that the shared random variables, the sources, are uniformly distributed, hence the response functions encode the whole problem. We can visualize, for example, Bob's response function $p_B(b|a, \gamma)$ by sampling several thousand values of $\{a, \gamma\} \in [0, 1]^2$. In order to capture the stochastic nature of the responses, for each pair a, γ we sample from $p_B(b|a, \gamma)$ 30 times and color-code the results $b \in$ {red, blue, green, and yellow}. By scatter plotting these points with a finite opacity we gain an impression of the response function, such as in Fig. 3b.

These figures are already interesting in themselves and can guide us towards analytic guesses of the ideal response functions. However, they can also be used to verify our results in some special cases. For example, if $\theta = 90°$ and the local set is sufficiently flat, then the response functions should be the same for all $v \geq v^*$, as it is in Fig. 3b. On the other hand if $\theta < 90°$ then we are in a scenario similar to that of panel (a) in Fig. 2 and the response functions should differ for different values of v. Finally, note that for any target distribution there is no unique closest local response function, so the visualized response functions could vary greatly. As a result, in order to have visually more similar response functions and to smooth the results, after running the algorithm for the full range of v, for each v we check whether the models at other v' values perform better for $p_t(v)$ (after allowing for small adjustments) and update the model for v accordingly.

Fritz distribution

In order to benchmark the method, we first consider the quantum distribution proposed by Fritz[5], which can be viewed as a Bell scenario wrapped into the triangle topology, and its nonlocality is thus well understood. Alice and Bob share a singlet, i.e., $|\psi\rangle_{AB} = |\psi^-\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)$, while Bob and Charlie share either a maximally entangled or a classically correlated state with Charlie, such as $\rho_{BC} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$ and similarly for ρ_{AC}. Alice measures the shared state with Charlie in the computational basis and, depending on this random bit, she measures either the Pauli X or Z observable. Bob does the same with his shared state with Charlie and measures either $\frac{X+Z}{\sqrt{2}}$ or $\frac{X-Z}{\sqrt{2}}$. They then both output the measurement result and the bit which they used to decide the measurement. Charlie measures both sources in the computational basis and announces the two bits. As a noise model we introduce a finite visibility for the singlet shared by Alice

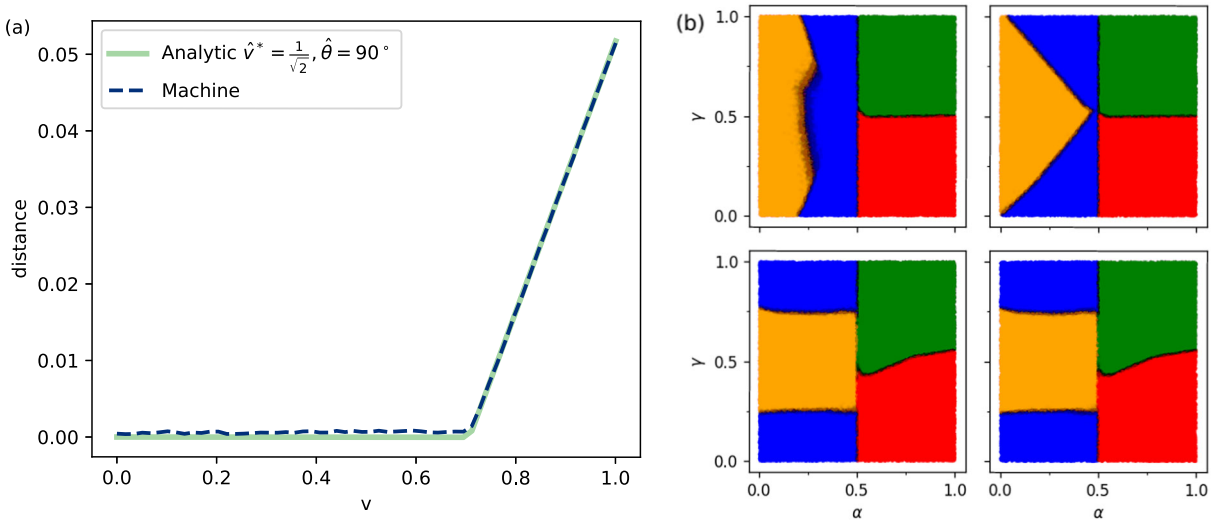

Fig. 3 Fritz distribution results. a Plot of the distance perceived by the machine, $d_M(v)$ and the analytic distance $\hat{d}(v)$ for $\hat{v}^* = 1/\sqrt{2}$ and $\hat{\theta} = 90°$. **b** Visualization of response functions of Bob as a function of α, γ for $v = 0$, 0.44, 0.71, 1, from top left to bottom right, respectively. Note how the responses for $v > \hat{v}^*$ are the same.

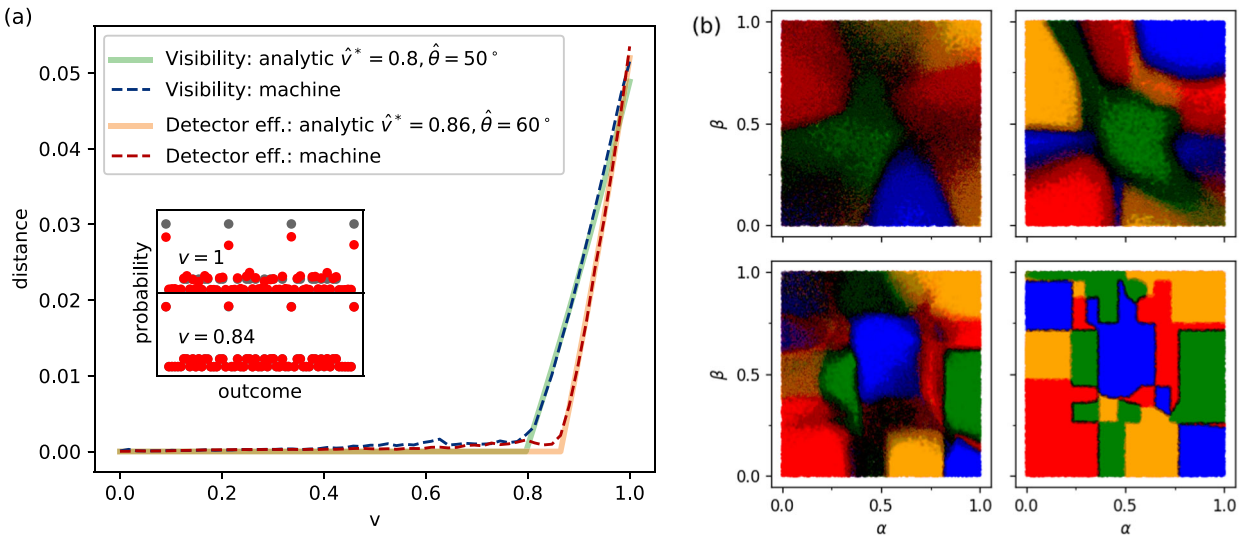

Fig. 4 Elegant distribution results. a Comparison of the distance perceived by the machine, $d_M(v)$ and the analytic distance $\hat{d}(v)$. Both visibility and detector efficiency model results are shown. Inset: The target (gray) and learned (red) distributions visualized by plotting the probability of each of the 64 possible outcomes, for detector efficiency $v = 1$ and $v = 0.84$. Note that for $v = 0.84$ most gray dots are almost fully covered by the corresponding red dots. **b** Responses of Charlie illustrated as a function of α, β. Detector efficiency values (top left to bottom right): $v = 0.5$, 0.72, 0.76, 1.

and Bob, thus we examine a Werner state,

$$\rho(v) = v|\psi^-\rangle\langle\psi^-| + (1 - v)\frac{\mathbb{I}}{4}, \tag{5}$$

where $\mathbb{I}/4$ denotes the maximally mixed state of two qubits. For such a state we expect to find a local model below the threshold of $v^* = \frac{1}{\sqrt{2}}$.

In Fig. 3a we plot the learned $d_M(v)$ and analytic $\hat{d}(v)$ distances discussed previously, for $\hat{\theta} = 90°$ and $\hat{v}^* = \frac{1}{\sqrt{2}}$. The coincidence of the two curves is already good evidence that the machine finds the closest local distributions to the target distributions. Upon examining the response functions of Alice, Bob and Charlie, in Fig. 3b, we see that they do not change above $\hat{v}^*$, which means that the machine finds the same distributions for target distributions outside the local set. This is in line with our expectations. Due to the connection with the standard Bell scenario (where the local set is actually a polytope), we believe the curve $p_t(v)$ exits the local set perpendicularly, as it is depicted on panel (b) in Fig. 2. These results confirm that our algorithm functions well.

Elegant distribution

Next we turn our attention to a more demanding distribution, as neither its locality or nonlocality has been proven to date, and is lacking a proper numerical analysis due to the intractability of conventional optimization over local models (see the "Discussion" section). Compared to the Fritz distribution, it is also more native to the triangle structure, as it combines entangled states and entangled measurements. We examine the Elegant distribution, which is conjectured in ref. [31] to be outside the local set. The three

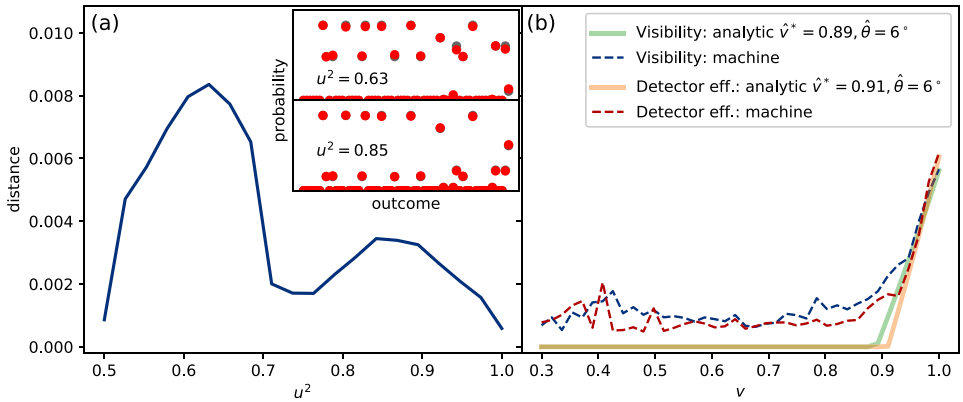

Fig. 5 Renou et al. distribution results. a The distance perceived by the machine, d_M, as a function of u^2, with no added noise. Inset: The target (gray) and learned (red) distributions visualized by plotting the probability of each of the 64 possible outcomes, for $u^2 = 0.63$ and $u^2 = 0.85$. These u^2 values approximately correspond to the two peaks in the scan. Note that most gray dots are almost fully covered by the corresponding red dots. This is an excellent confirmation that searching for transitions is more informative than using predefined thresholds for locality, as the learned and target distribution are nearly indistinguishable for a human eye for $u^2 = 0.85$, even though we know that the target is nonlocal here. **b** Noise scans, i.e., the analytic $\hat{d}(v)$ (see Eq. (4)) and the learned $d_M(v)$, for the target distribution of $u^2 = 0.85$, with v being visibility (green and blue) or detector efficiency (orange and red).

parties share singlets and each perform a measurement on their two qubits, the eigenstates of which are

$$|\Phi_j\rangle = \sqrt{\frac{3}{2}}|m_j, -m_j\rangle + i\frac{\sqrt{3}-1}{2}|\psi^-\rangle, \tag{6}$$

where the $|m_j\rangle$ are the pure qubit states with unit length Bloch vectors pointing at the four vertices of the tetrahedron for $j = 1, 2, 3, 4$, and $|-m_j\rangle$ are the same for the inverted tetrahedron.

We examine two noise models—one at the sources and one at the detectors. First we introduce a visibility to the singlets such that all three shared quantum states have the form (5). Second, we examine detector noise, in which each detector defaults independently with probability $1-v$ and gives a random output as a result. This is equivalent to adding white noise to the quantum measurements performed by the parties, i.e., the positive operator-valued measure elements are $\mathcal{M}_j = v|\Phi_j\rangle\langle\Phi_j| + (1-v)\frac{\mathbb{I}}{4}$.

For both noise models we see a transition in the distance $d_M(v)$, depicted in Fig. 4a, giving us strong evidence that the conjectured distribution is indeed nonlocal. Through this examination we gain insight into the noise robustness of the Elegant distribution as well. It seems that for visibilities above $\hat{v}^* \approx 0.80$, or for detector efficiency above $\hat{v}^* \approx 0.86$, the distribution is still nonlocal. The curves exit the local set at approximately $\hat{\theta} \approx 50°$ and $\hat{\theta} \approx 60°$, respectively. Note that for both distribution families, by looking at the unit tangent vector, one can analytically verify that the curves are almost straight for values of v above the observed threshold. This gives us even more confidence that it is legitimate to use the analytic distance $\hat{d}(v)$ as a reference (see Eq. (4)). In Fig. 4b, we illustrate how the response function of Charlie changes when adding detector noise. It is peculiar how the machine often prefers horizontal and vertical separations of the latent variable space, with very clean, deterministic responses, similarly to how we would do it intuitively, especially for noiseless target distributions.

Renou et al. distribution

The authors of ref. [20] recently introduced the first distribution in the triangle scenario which is not directly inspired by the Bell scenario and is proven to be nonlocal. To generate the distribution take all three shared states to be the entangled states $|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Each party performs the same measurement, characterized by a single parameter $u \in [\frac{1}{\sqrt{2}}, 1]$, with eigenstates $|01\rangle, |10\rangle, u|00\rangle + \sqrt{1-u^2}|11\rangle, \sqrt{1-u^2}|00\rangle - u|11\rangle$. The authors prove that for $u^2_{\max} < u^2 < 1$ this distribution is nonlocal, where $u^2_{\max} \approx 0.785$ and also show that there exist local models for $u^2 \in \{0.5, u^2_{\max}, 1\}$. Though they argue that there must be some noise tolerance of the distribution, they lack a proper estimation of it.

First we examine these distributions as a function of u^2, without any added noise. The results are depicted in Fig. 5a. To start with, note how the distances are numerically much smaller than in the previous examples, i.e., the machine finds distributions which are extremely close to the targets. See the inset in Fig. 5a for examples which exhibit how close the learned distributions are to the targets even for the points which have large distances ($u^2 = 0.63$, 0.85). We observe, consistently with analytic findings, that for $u^2_{\max} < u^2 < 1$, the machine finds a nonzero distance from the local set. It also recovers the local models at $u^2 \in \{0.5, u^2_{\max}, 1\}$, with minor difficulties around $u^2_{\max}$. Astonishingly, the machine finds that for some values of $0.5 < u^2 < u^2_{\max}$, the distance from the local set is even larger than in the provenly nonlocal regime. This is a somewhat surprising finding, as one might naively assume that between 0.5 and $u^2_{\max}$ distributions are local, especially when one looks at the nonlocality proof used in the other regime. However, this is not what the machine finds. Instead it gives us a nontrivial conjecture about nonlocality in a new range of parameters u^2. Though extracting precise boundaries in terms of u^2 for the new nonlocal regime would be difficult from the results in Fig. 5a alone, they strongly suggest that there is some nonlocality in this regime.

Finally, we have a look at the noise robustness of the distribution with $u^2 = 0.85$, which is approximately the most distant distribution from the local set, from within the provenly nonlocal regime. For the detector efficiency and visibility noise models we recover $\hat{v}^* \approx 0.91$, $\hat{v}^* \approx 0.89$ respectively, and $\hat{\theta} \approx 6°$ for both. Note that these estimates are much more crude than those obtained for the Elegant distributions, primarily due to the target distributions being so much closer to the local set and the neural network getting stuck in local optima. This increases the variations in independent runs of the learning algorithm. E.g. in Fig. 5a, at $u^2 = 0.85$ the distance is about 0.0034, whereas in Fig. 5b, in an independent run, the distance for this same point ($v = 1$) is around 0.0055. The absolute difference is small, however the relative changes can have an impact in extracting noise thresholds. Given that the local set is so close to the target distributions (exemplified in the inset in Fig. 5a), it is easily

possible that the noise tolerance is smaller than that obtained here.

DISCUSSION

Let us contrast the presented method to known techniques. Among analytic methods, the technique of inflation, introduced in[19], is known to converge toward a perfect oracle[14]. This method consists of a hierarchy of linear programs, which can be implemented on a computer. However, so far it could not witness nonlocality of the Elegant distribution, and can only witness the nonlocality of the distribution presented by Renou et al. in the provenly nonlocal regime by a small margin, making it difficult to extract noise tolerance results. Another approach that has been taken is to consider the entropy of the distribution, which makes the independence condition a linear one[9,17]. However, such techniques are typically relatively weak at detecting nonlocality and are not particularly useful for examining the distributions studied here.

The analytical difficulties in proving nonlocality and extracting noise robustness motivate us to look at numerical techniques. The standard method for tackling the membership problem in network nonlocality is nonlinear numerical optimization. For a fixed number of possible outputs per party, o, without loss of generality one can take the hidden variables to be discrete with a finite alphabet size, and the response functions to be deterministic. In fact the cardinality of the hidden variables can be upper bounded as a function of o[15]. Specifically for the triangle this upper bound is $o^3 - o$. This results in a straightforward optimization over the probabilities of each hidden variable symbol and the deterministic responses of the observers, giving $3(o^3 - o - 1)$ continuous parameters and a discrete configuration space of size $12(o^3 - o)^2$ to optimize over jointly. Note that this is a non-convex optimization space, making it a terribly difficult task. For binary outputs, i.e., $o = 2$, this means only 15 continuous variables and a discrete configuration space of 432 possibilities, and is feasible. However, already for the case of quaternary outputs, $o = 4$, this optimization is a computational nightmare on standard CPUs with a looming 177 continuous parameters and a discrete configuration space of size 43,200. Even when constraining the response functions to be the same for the three parties, $p_A = p_B = p_C$, and the latent variables to have the same distributions, $p_\alpha = p_\beta = p_\gamma$, the problem becomes intractable around a hidden variable cardinality of 8, which is still much lower than the current upper bound of 60 that needs to be examined. Standard numerical optimization tools quickly become infeasible even for the triangle configuration—not to mention larger networks!

The causal modeling and Bayesian network communities examine scenarios similar to those relevant for quantum information[32,33]. The core of both lines of research are directed acyclic graphs and probability distributions generated by them. In these communities there exist methods for this so-called "structure recovery" or "structure learning" task. However, these methods are either not applicable to our particular scenarios or are also approximate learning methods which make many assumptions on the hidden variables, including that the hidden variables are discrete. Hence, even if these learning methods are quicker than standard optimization for current scenarios of interest, they will run into the scaling problem of the latent variable cardinality.

The method demonstrated in this paper attacks the problem from a different angle. It relaxes both the discrete hidden variable and deterministic response function assumptions which are made by the numerical methods mentioned previously. The complexity of the problem now boils down to the response function of the observers —each of which is represented by a feedforward neural network. Though our method is an approximate one, one can increase its precision by increasing the size of the neural network, the number of samples we sum over (N_{batch}) and the amount of time provided for learning. Due to universal approximation theorems we are guaranteed to be able to represent essentially any function with arbitrary precision[36–38]. For the first two distributions examined here we find that there is no significant change in the learned distributions after increasing the neural network's width and depth above some moderate level, i.e., we have reached a plateau in performance. Regarding the Elegant distribution, for example, we used depth 5 and width 30 per party. However, we did not do a rigorous analysis in the minimum required size, perhaps an even smaller network would have worked. We were satisfied with the current complexity, since getting a local model for a single target distribution takes a few minutes on a standard computer, using a mini-batch size of $N_{\text{batch}} \approx 8000$. For the Renou et al. distribution there is still space for improvement in terms of the neural network architecture and the training procedure. The question of what the minimal required complexity of the response functions for a given target distribution is, is in itself interesting enough for a separate study, and can become a tedious task since the amount of time that the machine needs to learn typically increases with network size.

We have demonstrated how, by adding noise to a distribution and examining a family of distributions with the neural network, we can deduce information about the membership problem. For a single target distribution the machine finds only an upper bound to the distance from the local set. By examining families of target distributions, however, we get a robust signature of nonlocality due to the clear transitions in the distance function, which match very well with the approximately expected distances.

In conclusion, we provide a method for testing whether a distribution is classically reproducible over a directed acyclic graph, relying on a fundamental connection to neural networks. The simple, yet the effective method can be used for arbitrary causal structures, even in cases where current analytic tools are unavailable and numerical methods are futile, allowing quantum information scientist to test their conjectured quantum, or post-quantum, distributions to see whether they are locally reproducible or not, hopefully paving the way to a deeper understanding of quantum nonlocality in networks.

To illustrate the relevance of the method, we have applied it to two open problems, giving firm numerical evidence that the Elegant distribution is nonlocal on the triangle network, and getting estimates for the noise robustness of both the Elegant and the Renou et al. distribution, under physically relevant noise models. Additionally, we conjecture nonlocality in a surprising range of the Renou et al. distribution. Our work motivates finding proofs of the nonlocality for both these distributions.

The obtained results on nonlocality are insightful and convincing, but are nonetheless only numerical evidence. Examining whether a certificate of nonlocality can be obtained from machine learning techniques would be an interesting further research direction. In particular, it would be fascinating if a machine could derive, or at least give a good guess for a (nonlinear) Bell-type inequality which is violated by the Elegant or Renou et al. distribution. In general, seeing what insight can be gained about the boundary of the local set from machine learning would be interesting. Perhaps a step in this direction would be to understand better what the machine learned, for example by somehow extracting an interpretable model from the neural network analytically, instead of by sampling from it. A different direction for further research would be to apply similar ideas to networks with quantum sources, allowing a machine to learn quantum strategies for some target distributions. Moreover, the method introduced here could be straightforwardly applied to other networks, such as the Bell scenario with more inputs, outputs and/or parties, or to bilocality[4].

DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available within the paper.

CODE AVAILABILITY

Our implementation of the method for the triangle network and for the two-party Bell scenario can be found at https://www.github.com/tkrivachy/neural-network-for-nonlocality-in-networks.

Received: 13 March 2020; Accepted: 15 July 2020;
Published online: 21 August 2020

REFERENCES

1. Bell, J. S. On the Einstein Podolsky Rosen paradox. *Phys. Phys. Fiz.* **1**, 195–200 (1964).
2. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell nonlocality. *Rev. Mod. Phys.* **86**, 419–478 (2014).
3. Branciard, C., Gisin, N. & Pironio, S. Characterizing the nonlocal correlations created via entanglement swapping. *Phys. Rev. Lett.* **104**, 170401 (2010).
4. Branciard, C., Rosset, D., Gisin, N. & Pironio, S. Bilocal versus nonbilocal correlations in entanglement-swapping experiments. *Phys. Rev. A* **85**, 032119 (2012).
5. Fritz, T. Beyond Bells theorem: correlation scenarios. *New J. Phys.* **14**, 103001 (2012).
6. Pusey, M. F. Viewpoint: quantum correlations take a new shape. *Physics* **12**, 106 (2019).
7. Henson, J., Lal, R. & Pusey, M. F. Theory-independent limits on correlations from generalized Bayesian networks. *New J. Phys.* **16**, 113043 (2014).
8. Tavakoli, A., Skrzypczyk, P., Cavalcanti, D. & Acín, A. Nonlocal correlations in the star-network configuration. *Phys. Rev. A* **90**, 062109 (2014).
9. Chaves, R., Luft, L. & Gross, D. Causal structures from entropic information: geometry and novel scenarios. *New J. Phys.* **16**, 043001 (2014).
10. Chaves, R. et al. Inferring latent structures via information inequalities. *Proc. 30th Conference on Uncertainty in Artificial Intelligence*, 112–121, Quebec, Canada (2014).
11. Chaves, R., Majenz, C. & Gross, D. Information-theoretic implications of quantum causal structures. *Nat. Commun.* **6**, 5766 (2015).
12. Rosset, D. et al. Nonlinear Bell inequalities tailored for quantum networks. *Phys. Rev. Lett.* **116**, 010403 (2016).
13. Chaves, R. Polynomial Bell inequalities. *Phys. Rev. Lett.* **116**, 010402 (2016).
14. Navascues, M. & Wolfe, E. *The Inflation Technique Completely Solves the Causal Compatibility Problem*. Preprintat http://arxiv.org/abs/1707.06476 (2017).
15. Rosset, D., Gisin, N. & Wolfe, E. Universal bound on the cardinality of local hidden variables in networks. *Quantum Inform. Comp.* **18**, 0910–0926 (2017).
16. Fraser, T. C. & Wolfe, E. Causal compatibility inequalities admitting quantum violations in the triangle structure. *Phys. Rev. A* **98**, 022113 (2018).
17. Weilenmann, M. & Colbeck, R. Non-Shannon inequalities in the entropy vector approach to causal structures. *Quantum* **2**, 57 (2018).
18. Luo, M.-X. Computationally efficient nonlinear Bell inequalities for quantum networks. *Phys. Rev. Lett.* **120**, 140402 (2018).
19. Wolfe, E., Spekkens, R. W. & Fritz, T. The inflation technique for causal inference with latent variables. *J. Causal Inference* **7**, 20170020 (2019).
20. Renou, M.-O. et al. Genuine quantum nonlocality in the triangle network. *Phys. Rev. Lett.* **123**, 140401 (2019).
21. Gisin, N. et al. Constraints on nonlocality in networks from no-signaling and independence. *Nat. Commun.* **11**, 2378 (2020).
22. Renou, M.-O. et al. Limits on correlations in networks for quantum and no-signaling resources. *Phys. Rev. Lett.* **123**, 070403 (2019).
23. Pozas-Kerstjens, A. et al. Bounding the sets of classical and quantum correlations in networks. *Phys. Rev. Lett.* **123**, 140503 (2019).
24. Melko, R. G., Carleo, G., Carrasquilla, J. & Cirac, J. I. Restricted Boltzmann machines in quantum physics. *Nat. Phys.* **15**, 887–892 (2019).
25. Iten, R., Metger, T., Wilming, H., del Rio, L. & Renner, R. Discovering physical concepts with neural networks. *Phys. Rev. Lett.* **124**, 010508 (2020).
26. Melnikov, A. A. et al. Active learning machine learns to create new quantum experiments. *Proc. Nat. Acad. Sci. USA* **115**, 1221–1226 (2018).
27. van Nieuwenburg, E. P. L., Liu, Y.-H. & Huber, S. D. Learning phase transitions by confusion. *Nat. Phys.* **13**, 435–439 (2017).
28. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
29. Deng, D.-L. Machine learning detection of bell nonlocality in quantum many-body systems. *Phys. Rev. Lett.* **120**, 240402 (2018).
30. Canabarro, A., Brito, S. & Chaves, R. Machine learning nonlocal correlations. *Phys. Rev. Lett.* **122**, 200401 (2019).
31. Gisin, N. Entanglement 25 years after quantum teleportation: testing joint measurements in quantum networks. *Entropy* **21**, 325 (2019).
32. Pearl, J. *Causality: Models Reasoning and Inference* (Cambridge University Press, 2000).
33. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
34. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
35. Goudet, O. et al. Learning functional causal models with generative neural networks. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 39–80 (Springer International Publishing, Cham, 2018).
36. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signal.* **2**, 303–314 (1989).
37. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).
38. Lu, Z., Pu, H., Wang, F., Hu, Z. & Wang, L. The expressive power of neural networks: a view from the width. *Adv. Neural Inf. Process. Syst.* **30** 6231–6239 (2017).

ACKNOWLEDGEMENTS

The authors thank Raban Iten, Tony Metger, Elisa Bäumer, Marc-Olivier Renou, Elie Wolfe, and Askery Canabarro for discussions. T.K., Y.C., N.G., and N.B. acknowledge financial support from the Swiss National Science Foundation (Starting grant DIAQ and QSIT), and the European Research Council (ERC MEC). D.C. acknowledges support from the Ramon y Cajal fellowship, Spanish MINECO (QIBEQI, Project No. FIS2016-80773-P, and Severo Ochoa SEV-2015-0522) and Fundació Cellex, Generalitat de Catalunya (SGR875 and CERCA Program). A.T. acknowledges financial support from the UK Engineering and Physical Sciences Research Council (EPSRC DTP).

AUTHOR CONTRIBUTIONS

N.B. and T.K. had the idea to connect the nonlocality with neural networks. T.K. developed the concept in detail, wrote and ran the code. A.T. and T.K. explored other approaches to the problem which turned out to be less efficient and were not included. All authors discussed and analyzed the results extensively and contributed to writing the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to T.K.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.