Quantum Science and Technology

PAPER • **OPEN ACCESS**

Witnessing entanglement in experiments with correlated noise

To cite this article: Bas Dirkse *et al* 2020 *Quantum Sci. Technol.* **5** 035007

Quantum Science and Technology

OPEN ACCESS

RECEIVED
17 March 2020

REVISED
16 April 2020

ACCEPTED FOR PUBLICATION
27 April 2020

PUBLISHED
28 May 2020

PAPER

Witnessing entanglement in experiments with correlated noise

Bas Dirkse[1,2,3,5], Matteo Pompili[1,2], Ronald Hanson[1,2], Michael Walter[3,4] and Stephanie Wehner[1,2]

[1] QuTech, Delft University of Technology, The Netherlands
[2] Kavli Institute of Nanoscience, Delft University of Technology, The Netherlands
[3] QuSoft, University of Amsterdam, The Netherlands
[4] Korteweg-de Vries Institute for Mathematics, Institute for Theoretical Physics, Institute for Logic, Language, and Computation, University of Amsterdam, The Netherlands
[5] Author to whom any correspondence should be addressed.

E-mail: b.dirkse@tudelft.nl

Keywords: entanglement witness, witness experiment, estimation statistics, confidence interval

Abstract

The purpose of an entanglement witness experiment is to certify the creation of an entangled state from a finite number of trials. The statistical confidence of such an experiment is typically expressed as the number of observed standard deviations of witness violations. This method implicitly assumes that the noise is well-behaved so that the central limit theorem applies. In this work, we propose two methods to analyze witness experiments where the states can be subject to arbitrarily correlated noise. Our first method is a *rejection experiment*, in which we certify the creation of entanglement by rejecting the hypothesis that the experiment can only produce separable states. We quantify the statistical confidence by a *p*-value, which can be interpreted as the likelihood that the observed data is consistent with the hypothesis that only separable states can be produced. Hence a small *p*-value implies large confidence in the witnessed entanglement. The method applies to general witness experiments and can also be used to witness genuine multipartite entanglement. Our second method is an *estimation experiment*, in which we estimate and construct confidence intervals for the average witness value. This confidence interval is statistically rigorous in the presence of correlated noise. The method applies to general estimation problems, including fidelity estimation. To account for systematic measurement and random setting generation errors, our model takes into account device imperfections and we show how this affects both methods of statistical analysis. Finally, we illustrate the use of our methods with detailed examples based on a simulation of NV centers.

1. Introduction

Entanglement is a fundamental property of quantum mechanical systems [1] and an important resource for many quantum information processing tasks. In quantum computing, coherently creating entanglement between several qubits is necessary for computational speedups [1–3]. In quantum networks, remote entanglement is an essential resource for quantum cryptography [4–6] and distributed computing applications [7–9]. Entanglement also plays a crucial role in quantum sensing and metrology [10–12], enabling more precise measurement of physical quantities. With the rapid experimental advances in the manipulation and control of quantum systems, much progress had been made toward the generation of entangled states in various physical platforms [13–19]. Yet, the creation of high-quality many-body entanglement is still a significant challenge. It is therefore important to have good tools to certify the creation of entanglement. The main tools used for this purpose are entanglement witnesses.

An *entanglement witness* is an observable W on a quantum system that can certify entanglement of a state ρ^* under investigation [20]. By definition, the witness W satisfies

$$\mathrm{Tr}[\rho W] \geqslant 0 \quad \forall \rho \in \mathcal{S}, \tag{1}$$

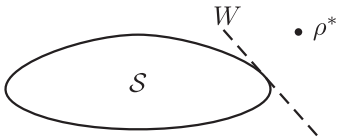


Figure 1. Geometric interpretation of a witness W as a hyperplane that separates the state ρ^* from the convex set $\mathcal{S}$. If $\mathcal{S}$ is the set of separable states, then W certifies that ρ^* is entangled.

for all separable states $\rho \in \mathcal{S}$. As a consequence, it can be used to certify entanglement: If a state ρ^* has negative witness expectation value, $\mathrm{Tr}[W\rho^*] < 0$, then it is necessarily entangled. If the expectation value is non-negative, the test is inconclusive (the state can either be separable or not). The witness method applies more generally than just for separating entangled from separable states. If $\mathcal{S}$ is an arbitrary convex set of states and $\rho^* \notin \mathcal{S}$, then there always exists a witness W such that equation (1) holds, while $\mathrm{Tr}[W\rho^*] < 0$. For example, a witness can be used to certify that states are genuinely multipartite entangled. Geometrically, the witness W can be interpreted as a hyperplane that separates the convex set $\mathcal{S}$ from the state $\rho^* \notin \mathcal{S}$. This is illustrated in figure 1. In general, finding an appropriate witness W for a state ρ^* is a difficult problem that has been studied extensively in literature [21–23]. For the remainder of this article, we will assume that W is chosen and fixed.

Often, a witness W is a non-local observable of the system for which entanglement is to be certified. Such measurements are typically hard to perform, particularly in a network setting. Therefore, in experiments, W is usually decomposed into a sum of locally measurable observables which are then measured individually on the constituent subsystems. The *witness expectation value* $\mathrm{Tr}[W\rho^*]$ is then the sum of the expectation values of the locally measurable observables. Each of these expectation values can then only be estimated to some finite precision, since in any experiment only a finite number n of data points can be collected. As a consequence, the *witness estimate* $\hat{w}_n$ obtained from n measurement outcomes can differ from the true value $\mathrm{Tr}[W\rho^*]$. Therefore, it is an important question how to quantify the confidence in the experimentally determined estimate $\hat{w}_n$.

1.1. Prior work and motivation

In many experiments, the confidence in the estimate $\hat{w}_n$ of the true witness expectation value $\mathrm{Tr}[W\rho^*]$ is expressed by the standard error $\hat{\sigma}$ (the empirical standard deviation) [13–19, 24]. These experiments typically claim the certification of entanglement if the estimate $\hat{w}_n$ is a number of $\hat{\sigma}$'s below zero. This approach is simple and pragmatic, but may suffer from statistical and practical challenges (see [25] for similar objections to using this method for quantifying Bell violations). We give a concrete example in section 4.4 (see figure 6) where this approach could potentially be problematic.

Certification of entanglement by the number of sigma's of witness violation is most easily justified under the assumption that in each round i the state ρ_i is independent and identically distributed (iid assumption). This is equivalent to assuming each round a fixed state ρ^* is produced. Under this assumption, the estimate $\hat{w}_n$ is considered a realization of a Gaussian random variable $\hat{W}_n$ with mean $\mathbb{E}[\hat{W}_n] = \mathrm{Tr}[W\rho^*]$ and standard deviation $\sigma \sim \frac{1}{\sqrt{n}}$ (for sufficiently large n). This is justified by the central limit theorem. The empirically obtained $\hat{w}_n$ and $\hat{\sigma}$ are appropriate estimates of the mean $\mathrm{Tr}[W\rho^*]$ and standard deviation σ. Thus, the reported $\hat{w}_n \pm \hat{\sigma}$ is a complete and accurate characterization of the distribution (and hence leads to meaningful confidence intervals etc.)

However, if the states $\rho_1, \ldots, \rho_n$ produced in an n round experiment are not iid, several difficulties may arise. This starts with what is even to be estimated. Most natural is to estimate the *average witness value*

$$\langle W\rangle_n := \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}[\rho_i W] = \mathrm{Tr}[(\frac{1}{n}\sum_{i=1}^{n}\rho_i)W], \tag{2}$$

which can also be interpreted as the witness expectation value of the average state $\rho^* = \frac{1}{n}\sum_{i=1}^{n}\rho_i$. There are versions of the central limit theorem that relax the iid assumption. They can be used to argue that the estimate $\hat{w}_n$ is still an observation of a Gaussian random variable $\hat{W}_n$ with mean $\mathbb{E}[\hat{W}_n] = \langle W\rangle_n$ for sufficiently large n (Gaussian assumption). But in practical experiments it is not always clear when these theorems can be applied, so that the convergence of $\hat{W}_n$ to a Gaussian with mean $\langle W\rangle_n$ is not guaranteed for any n. Moreover, under non-iid states ρ_i it is unclear whether the observed standard error $\hat{\sigma}$ is an appropriate estimator of the true standard deviation σ. Finally, even if the central limit theorem applies, the convergence of $\hat{W}_n$ to a normal distribution can be extremely slow in n (especially when the witness

violation is large [25]), so that too small n will still cause $\hat{W}_n$ not to be Gaussian. Hence, in practice it can be difficult to justify the Gaussian assumption.

When the iid assumption (or more generally the Gaussian assumption) fails, several different problems can arise. First, it can lead to overestimation of the confidence in the witness violation based on the observed data. This happens when $\Pr[\hat{W}_n < 0]$ is smaller under the true distribution of $\hat{W}_n$ than under the estimated distribution, based on the observed $\hat{w}_n$ and standard error $\hat{\sigma}$ (i.e., based on the Gaussian assumption). Second, the reported numbers $\hat{w}_n \pm \hat{\sigma}$ lack rigorous interpretation. The empirical standard deviation $\hat{\sigma}$ may no longer be an appropriate estimate of the true standard deviation σ. Moreover, the mean and standard deviation do not necessarily describe the distribution of W_n completely (if $\hat{W}_n$ is not Gaussian, the true distribution may depend on more than 2 free parameters). Because of these two effects, the number of $\hat{\sigma}$'s of witness violation in relation to the estimate $\hat{w}_n$ will depend on the way $\hat{W}_n$ fails to be Gaussian or on how $\hat{\sigma}$ fails to estimate σ. This also makes the results between different experiments and physical platforms become incomparable, because the actual distribution of W_n may be influenced by experimental parameters, such as the distribution of ρ_i, measurement settings, hardware imperfections or the choice of witness. Hence the number of $\hat{\sigma}$'s of violation may also be influenced by these experimental details.

Finally, we note that measurement noise (systematic errors) can also lead to overestimation of the confidence in the witness violation. This is because any measurement noise leads to the imperfect implementation $\tilde{W}$ of the witness W. In case that $\langle \tilde{W} \rangle_n < \langle W \rangle_n$, this again leads to an overconfidence in the witness violation. In fact, it can even happen that $\langle \tilde{W} \rangle_n < 0$ while $\langle W \rangle_n \geqslant 0$, leading to falsely concluding entanglement [26]. This overconfidence persists independent of the number of samples n taken, since the error is systematic.

1.2. Our contribution

We propose a new method of carrying out and analyzing witness experiments that addresses all of the aforementioned issues. This method applies without any assumption on the states produced by the experiment (i.e., they may be arbitrarily correlated). Moreover, it has a simple and clear interpretation, and yields figures of merit that are easily comparable between different experiments. Finally, our method takes into account imperfections of the measurement device and random setting generation, avoiding systematic overestimation of confidence.

In our method, we view the source of the states as a black box that produces a quantum state ρ_i on demand. The source can produce multiple states sequentially. All of these states are modeled by random variables that can be arbitrarily distributed and which may depend arbitrarily on the history of the experiment. That is, we allow the source to have memory. We now model the experiment as a sequential process of $i = 1, \ldots, n$ rounds. In each round, a state ρ_i and a random measurement setting (determined by the decomposition of W into locally measurable observables) are requested. The appropriate measurement is performed on the state and the outcomes are recorded for data processing. In this model of the experiment, we additionally allow for arbitrary bounded-strength noise in the measurement devices and random measurement setting generator. That is, we assume that the noise in these devices is smaller than a quantified maximum amount. Witnessing entanglement without any assumptions on the devices, an area known as self-testing [27, 28], is based on Bell-type inequalities, which are typically tighter than witness inequalities (and therefore harder to violate experimentally). Thus, our model is very general and has minimal assumption to be as widely applicable as possible for analyzing witness experiments.

The main contribution of this work is two different types of data analysis for witness experiments. Both methods are valid under these extremely general assumptions (in particular the state assumptions). In the first method, we quantify the confidence that the source has *the capability to produce an entangled state* (i.e., a state outside $\mathcal{S}$). This means that we rigorously determine the confidence that $\mathrm{Tr}[\rho_i W] < 0$ for at least one i. We do this by applying the framework of hypothesis testing, in which a null hypothesis is to be rejected based on the observed evidence in experiment. In witness experiments, the null hypothesis is that the source only produces separable states (i.e. $\forall i : \rho_i \in \mathcal{S}$). To reject this null hypothesis means that at least one entangled state must have been produced by the source (i.e. $\exists i : \rho_i \notin \mathcal{S}$). The figure of merit to quantify the confidence in rejecting the null hypothesis is the *p-value*. Intuitively, the p-value is the probability of obtaining data at least as extreme as the observed data in an experiment *if the experiment were governed by the null hypothesis*, i.e., if the source was only able to produce separable states. A small p-value is then considered strong evidence against the null hypothesis: the observed data is very unlikely to be explained by a model that includes the null hypothesis. If p is smaller than some significance level α, the null hypothesis is rejected and we conclude that entanglement must have been produced at least once with confidence $1 - \alpha$. We shall refer to this entire procedure as a *witness rejection experiment* and the data analysis as the *rejection analysis*. This method is different from the standard methods, in the sense that here we can make a

statement about *at least one* state, whereas typically one makes a statement about the *average produced states* (e.g. when estimating the average witness value $\langle W\rangle_n$ as defined in equation (2)). We emphasize that our method and analysis applies in complete generality to arbitrary witness experiments (with an arbitrary convex set of states $\mathcal{S}$ and witness W). For concreteness we will focus in this work on entanglement witnessing, but see section 5.1 for other examples.

In the second method we aim to estimate the *average witness value* $\langle W\rangle_n$ and we provide a confidence interval around this estimate. The main contribution of our confidence interval method is that it is valid without any assumptions on the produced states and therefore always applies. We will refer to this method as the witness *estimation method* and the data analysis as the *estimation analysis*, since the objective here is to estimate $\langle W\rangle_n$. This method is generally applicable to estimate any Hermitian observable, not just witness operators (i.e., it is not necessary that there is a set $\mathcal{S}$ such that equation (1) holds). Thus our estimation method even applies to fidelity estimation and other estimation experiments.

The contributions of our work are presented in the following way. First, we formulate the round-by-round witness experiment as an entanglement witness game, expand on this description and present a formal model that governs the experiment in section 2. In the model description we incorporate imperfections in the measurement device and random setting generation in a quantitative way. Based on this model, we give a step-by-step description how to set the parameters and carry out the witness experiment in section 3.1. Then, we show how to calculate a witness correction from the quantification of the imperfect experimental devices (section 3.2, theorem 1). It is used in both the rejection and estimation experiments to account for systematic device errors and prevent possible overconfidence in the experiment outcomes (rejection and estimation). Next, we provide the main result to perform the rejection analysis. This is an easy-to-compute bound on the p-value (section 3.3, theorem 2). The bound is simply evaluated from the measurement outcomes. By comparing this bound to a predetermined significance level α, we can determine whether the experiment rejects the null hypothesis with statistical significance. This allows us to rigorously conclude that the source has the capability to produce entangled states with confidence $1-\alpha$. Finally, we provide the main result to perform the estimation analysis. This is a direct method to compute and estimate and confidence interval for the average witness value $\langle W\rangle_n$ (section 3.4, theorem 3). The estimate and this confidence interval are also directly and easily computable from the observed measurement data. We illustrate these contributions with several detailed numerical examples in section 4. Two of our examples are based on the simulation of nitrogen vacancy centers. The focus of these examples is to detect genuine multipartite entanglement between three qubits (i.e., not a convex combination of biseparable states, states separable over some bipartition of the three subsystems). Our third example (figure 6) shows an explicit case where the Gaussian assumption fails to be applicable and where our methods are still applicable.

The technical ingredients of this work are summarized as follows. Both results are obtained by viewing the witness experiment as a game [29], similar to Bell tests being viewed as nonlocal games. This allows us to construct (super)martingale sequences and use a concentration inequality to upper bound the tail probabilities (we use Bentkus' inequality [30, 31], which is slightly tighter than the more commonly used Hoeffding–Azuma inequality [32]). This method is inspired by the analysis of Bell inequalities of reference [32]. By choosing the appropriate (super)martingale sequences, we obtain the p-value bound for the rejection analysis and the confidence interval for the estimation analysis.

1.3. Relation to other work

In this work, we model the measurement noise as general as possible via the POVM formalism and determine a witness correction from this model using analytical methods to guarantee we never overestimate the confidence. Our measurement model can be viewed as a generalization of the model studied in reference [26], where imperfect qubit measurements are modeled by Bloch vectors that are misaligned by at most some fixed angle. In reference [26] a witness correction factor is computed under this more restricted noise model. However, they compute the witness correction via numerical optimization (see section 5.3.2 for why we opt for an analytical bound and how the witness correction factor can alternatively be calculated using numerical optimization for our noise model).

The witness rejection experiment and analysis is new for entanglement witness experiments, but is inspired by the use of this technique for testing local realism through nonlocal games [32]. We emphasize that this rejection method aims to rigorously certify that a machine *has the capability of* producing entanglement. This is different than typical witness experiments in literature where the objective is to estimate the average witness value [13–19, 24]. The estimation method we study here also aims to estimate the average witness value. The main difference is that most works implicitly assume that the states are iid (or at least that the estimator is Gaussian by some type of central limit theorem argument), whereas our

work applies in the most general case with arbitrary noise on the state. This makes our method more generally applicable.

Closely related to the confidence interval we construct here, reference [33] provides a method to construct a Bayesian credible interval for an estimate of the average fidelity of experimentally prepared states to a fixed entangled state (note that is equivalent to a particular choice of witness). The model is similar in the sense that the states can be arbitrarily correlated, but the estimation objective is different: the goal of reference [33] is to estimate the average fidelity of the unmeasured states from the measurement of a subset of all available states. Similarly, reference [34] derives an efficient method to verify the production of graph states by measuring all but one copy of the state. In contrast, we measure all available states and only aim to make a statement about all the created states (after the fact). The work in reference [35] is related to this by giving general lower bounds on the size of a credible regions for quantum parameter estimation.

An alternative method to estimate a property of a quantum system, is by using quantum state tomography to collect measurement data, estimate a figure of merit (fidelity or witness value) and determine a confidence interval [36]. However, this typically requires more measurement data than partial state characterization since the complete state is reconstructed.

Finally, we mention that there is also a way of witnessing entanglement without the need to trust the measurement devices at all (measurement-device-independent entanglement witnessing, MDI-EW) [37, 38]. This, however, requires each party to hold auxiliary local quantum states in each round and perform a joint measurement between the auxiliary quantum state and the quantum state under investigation. This method has been implemented in an experiment under the iid assumption [38].

2. Formulation and model of witness experiments

In this section, we will discuss the formulation and modeling of witness experiments. We will start with a brief review of entanglement witness games as known in the literature in section 2.1. Next, we will generalize the game formulation to handle two additional things: (1) multiple terms in the decomposition of the witness operator may be inferred from a single measurement; and (2) measurements are allowed to be implemented by arbitrary POVMs. We explain how to do this and introduce notation in section 2.2. Finally, in section 2.3 we give a complete description of the experimental model that underpins our experiment. This includes the characterization of noisy measurement and random setting generation devices.

2.1. Entanglement witness games

In this section, we will recap entanglement witness games from the literature. We will start from the assumption that a choice of witness W has been made. The quantum system under investigation is decomposed into m subsystems on which local measurements can be performed (e.g., $m = 2$ for bipartite entanglement witnessing). The witness operator then admits a decomposition into locally measurable observables of the form

$$W = cI + \sum_x w_x M_x^{(1)} \otimes \cdots \otimes M_x^{(m)}, \tag{3}$$

where each $M_x^{(j)}$ is a locally measurable observable on subsystem j and where x runs over the terms in the decomposition. Note that such a decomposition is always possible. A decomposition is minimal if the number of terms over which x runs is minimal. In practice, the locally measurable observables $M_x^{(j)}$ will often be Pauli observables. The decomposition in equation (3) is chosen such that each locally measurable observable can be easily measured in the experiment. Measurement of $M_x^{(j)}$ yields one of the possible outcomes labeled by a_j (in the case of Pauli observables, the outcomes are simply ± 1). We shall denote the vector of all outcomes of the m subsystems as

$$\mathbf{a} = (a_1, \ldots, a_m). \tag{4}$$

With this decomposition, an entanglement witness experiment can be formulated as a game [29]. This is similar to how Bell experiments are often formulated as nonlocal games. See figure 2 for an illustration of an entanglement witness game. The game consists of n rounds. There are m players, one for each subsystem. At the start of each round, each player receives a subsystem of a quantum state ρ_i, as well as a random measurement setting X_i (we will use the conventional notation of writing random variables as capital letters and their realizations as lowercase letters). This random setting X_i dictates which measurements the players should perform on their local subsystems (according to the decomposition equation (10)). Hence, upon

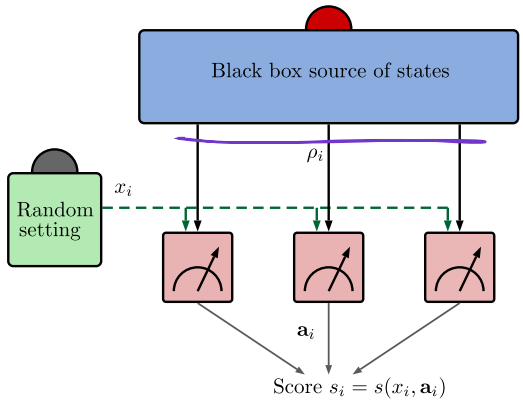


Figure 2. Round i of the entanglement witness game. By pressing the red button, a source produces a quantum state ρ_i and sends its subsystems to the players. We model the source as a black box, meaning it can produce a state by an arbitrary process. The state in round i can arbitrarily depend on everything that happened earlier in the experiment, i.e. the source is allowed to show memory effects. Then, by pressing the gray button, the random setting generator produces a measurement setting x_i (almost independently from ρ_i). The players each perform a measurement according to setting x_i and send their outcomes $\mathbf{a}_i$ to a referee, who computes the score of that round. Afterward, round $i+1$ starts.

receiving $X_i = x$ in round i, each player j perform the local measurement labeled by x and j. They then report their respective outcomes $\mathbf{a}$ to a referee, who assigns a score to the round according to

$$s(x, \mathbf{a}) = -\frac{w_x}{p_x}\prod_{j=1}^{m} a_j, \tag{5}$$

where p_x is the desired probability of realizing measurement setting $X_i = x$. *A priori*, any choice of p_x defines a valid game. However, the choice of p_x has a significant influence on finite statistics in an experiment. We suggest a reasonable choice in equation (17) and discuss the issue further in section 5.2.3. The negative sign in equation (5) is added conform the common convention that games aim to maximize score. The score can be interpreted as the contribution of round i to the witness value. Note that the score itself is a random variable $S_i := s(X_i, \mathbf{A}_i)$, since it is a function of the random measurement setting X_i and the random measurement outcomes $\mathbf{A}_i$. The score is constructed in such a way that the expected value of the score (in the ideal scenario with perfect measurements and randomness) satisfies

$$\mathrm{Tr}[\rho_i W] = c - \mathbb{E}[S_i|\rho_i], \tag{6}$$

for all possible states ρ_i. Thus, the witness expectation value is affinely related to the expected score of each round.

2.2. Generalization of the game formulation

In this section, we will expand on the game formulation as discussed in the previous section. In particular, we will make two generalizations. First, we will explain and introduce notation to easily infer the expectation value of multiple terms in the witness decomposition equation (3) from a single measurement. Doing this typically requires fewer measurements for fixed confidence so it is advantageous to do so when possible. Second, to be as general as possible in our measurement model, we shall allow every local measurement on a individual subsystem to be described by a POVM. Let us make these things more precise.

Sometimes it is not needed to measure all terms in equation (3) separately [13, 22]. For example, with $m = 3$ single-qubit subsystems, Pauli-Z measurements on each subsystem would allow to one infer the expectation values of all operators (omitting the tensor symbol)

$$ZZZ, ZZI, ZIZ, IZZ, ZII, IZI, IIZ. \tag{7}$$

This holds in general. Measurement of m non-identity operators on all of the subsystems, would allow one to infer $2^m - 1$ expectation values. We shall refer to the non-identity operators that are measured (ZZZ in this example) as the *measurement setting* [13, 22] and refer to one or more of the possible $2^m - 1$ operators whose expectation value can be computed (operators from equation (7) in this example) as *observables*. Throughout the rest of this work, we will denote measurement settings as $M_x^{(1)} \otimes \cdots \otimes M_x^{(m)}$, where every

$M_x^{(j)} \neq I$ is not the identity, and index them with a subscript x. We will denote observables as $O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)}$ and index them with the subscript ξ. Note that ξ may run over different (more) terms that x. Using this new notation, the witness decomposition (equation (3)) is thus written as

$$W = cI + \sum_\xi w_\xi O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)}. \tag{8}$$

To keep track of which observables (labeled by ξ) are related to which measurement setting (labeled by x), we define $f(\xi) = x$ if the observable $O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)}$ can be measured by the measurement setting $M_x^{(1)} \otimes \cdots \otimes M_x^{(m)}$. Furthermore, we define $b(\xi) \in \{0,1\}^m$ as the bitstring of length m such that

$$O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)} = (M_{f(\xi)}^{(1)})^{b_1(\xi)} \otimes \cdots \otimes (M_{f(\xi)}^{(m)})^{b_m(\xi)}. \tag{9}$$

In this way, each term in equation (8) is related to a measurement setting from which it can be obtained. For example, the observables IZZ, ZIZ, ZZI can all be measured by the setting ZZZ, and the corresponding bitstrings b are 011, 101, 110, respectively. Note that if all observables require a different setting, then $\xi = f(\xi) = x$, $b(\xi) = 11\cdots1$ and $O_\xi^{(j)} = M_x^{(j)}$, thus reducing to the simple case discussed in section 2.1. Using this notation, we can write equation (8) alternatively as

$$W = cI + \sum_\xi w_\xi \bigotimes_{j=1}^m \left(M_{f(\xi)}^{(j)}\right)^{b_j(\xi)}. \tag{10}$$

To allow for the most general model of measurements, we will allow each $M_x^{(j)}$ in a measurement setting to be measured by a POVM $\{\Pi_a^{(j),x}\}_{a\in\Omega_x^{(j)}}$ with outcomes labeled by a (which take values in the finite set $\Omega_x^{(j)}$). That is, we will write

$$M_x^{(j)} = \sum_{a\in\Omega_x^{(j)}} a\,\Pi_a^{(j),x}. \tag{11}$$

For a standard measurement of the observable $M_x^{(j)}$, this decomposition is simply given by the spectral decomposition, so that the POVM elements are the eigenprojections and the outcomes are simply the eigenvalues of $M_x^{(j)}$. However, this is not the only option: the decomposition is not unique. In particular, the POVM need not be a projective measurement. This allows for the modeling of known non-unitary measurement noise. Suppose for example that we wish to implement the measurement setting $M_x^{(j)} = Z$, the Pauli Z-operator. Its standard implementation would be by a projective measurement in the $|0\rangle$, $|1\rangle$-basis. This corresponds to the decomposition $Z = |0\rangle\langle 0| - |1\rangle\langle 1|$ in equation (11). However, suppose that we cannot perfectly discriminate $|0\rangle$ from $|1\rangle$ and you incorrectly obtain the opposite outcome in 1% of the measurements. Such a situation is modeled, e.g., by the POVM $\{\begin{bmatrix} 0.99 & 0 \\ 0 & 0.01 \end{bmatrix}, \begin{bmatrix} 0.01 & 0 \\ 0 & 0.99 \end{bmatrix}\}$. Nevertheless, this POVM allows us to implement the desired measurement setting if we choose $a = \pm\frac{1}{0.98}$. Indeed,

$$Z = \frac{1}{0.98}\begin{bmatrix} 0.99 & 0 \\ 0 & 0.01 \end{bmatrix} - \frac{1}{0.98}\begin{bmatrix} 0.01 & 0 \\ 0 & 0.99 \end{bmatrix}. \tag{12}$$

Our results take into account the additional statistical uncertainty introduced by non-projective measurements in estimating the expectation value from finite single-shot measurements on the level of single-shot outcomes. Requiring that equation (11) holds, ensures the expectation value of this non-projective measurement equals the expectation value of observable to be implemented.

With the generalizations just discussed, the score function in equation (5) needs to be generalized to

$$s(x,\mathbf{a}) = -\frac{1}{p_x}\sum_{\xi:f(\xi)=x} w_\xi \prod_{j=1}^m (a_j)^{b_j(\xi)}, \tag{13}$$

The score now sums over all observables ξ obtained from the same setting x. The fact that the outcomes are raised to the power $b_j(\xi)$ reflects the fact that $O_\xi^{(j)} = I$ if $b_j(\xi) = 0$ (in which case all outcomes are 1). Note that the weights w_ξ are labeled by ξ as they appear in equation (8), whereas the probabilities p_x are labeled by the measurement setting x. This generalized score function still satisfies equation (6) and is related to the witness decomposition equation (10) via (see appendix A for details)

$$W = cI - \sum_x p_x \sum_{\mathbf{a}} s(x,\mathbf{a})(\Pi_{a_1}^{(1),x} \otimes \cdots \otimes \Pi_{a_m}^{(m),x}). \tag{14}$$

We give an overview of the introduced notation in table 1.

Table 1. Summary of notation to allow multiple observables per measurement setting. The objects in the top half relate only to the choice of witness and its decomposition into observables (labeled by ξ). The objects in the bottom half relate to the implementation of the witness using measurement settings (labeled by x), which are implemented by POVMs.

Object	Symbol(s)	Definition/constraint
Witness	W	$\mathrm{Tr}[\rho W] \geqslant 0, \quad \forall \rho \in \mathcal{S}$
Observable	$O_\xi^{(j)}$	$W = cI + \sum_\xi w_\xi \otimes_{j=1}^m O_\xi^{(j)}$
Weight	w_ξ, c	
Setting	$M_x^{(j)}, f, b$	$O_\xi^{(j)} = (M_{f(\xi)}^{(j)})^{b_j(\xi)}$
POVM	$\{\Pi_a^{(j),x}\}_{a\in\Omega_x^{(j)}}$	$M_x^{(j)} = \sum_{a\in\Omega_x^{(j)}} a\,\Pi_a^{(j),x}$
Distribution	p_x	Equation (17) recommended

Table 2. List of model assumptions on the experimental devices and the nature of the experiment. These assumptions should plausibly hold in the real experiment. The validity of our results depends on these assumption holding. We give a mathematically rigorous definition of the model in appendix B.

Model Assumptions

(I) **Sequentiality.** Rounds of the witness game are played sequentially. At the start of each round i, each player j receives one part of a joint state ρ_i generated by the black box source, as well as a random measurement setting x_i. Each player j performs a POVM measurement that depends on the setting x_i, and reports the outcome a_j to a referee, who computes the score s_i of that round using equation (13). The next round $i+1$ only starts after the referee received all players' measurement outcomes for round i. The experiment is allowed to depend arbitrarily on the past

(II) **Trusted randomness.** The random setting generator produces in each round i a random setting X_i, whose distribution $\tilde{p}_{i,x}$ (conditioned on the history of the experiment and the state produced) is close to the desired distribution p_x:

$$|\tilde{p}_{i,x} - p_x| \leqslant \tau \quad \forall i, x. \tag{15}$$

We assume $\tau < p_x$ for all x, so that each setting has nonzero probability of being realized

(III) **Trusted measurements.** In each round i, each player j performs a noisy POVM measurement $\{\tilde{\Pi}_{i,a}^{(j)}\}_{a\in\Omega_{X_i}^{(j)}}$ that is close to the ideal POVM from equation (11):

$$\|\tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),X_i}\|_\infty \leqslant \delta_j \quad \forall i, x, j, a. \tag{16}$$

The noisy measurements have the same outcomes $a \in \Omega_{X_i}^{(j)}$ as the ideal measurements. Measurement outcomes follow Born's rule

2.3. Model of the experiment

Above we explained that an entanglement witness experiment can naturally be interpreted as a game carried out by m players in n rounds (cf figure 2). We now summarize the key properties of our model in more detail—see table 2. A mathematically rigorous formulation is given in appendix B.

Assumption (I) states that the experiment is performed sequentially. Importantly, we do *not* assume that the states ρ_i are independently and identically distributed (iid). In fact, we allow that the ith round depends arbitrarily on the previous rounds. Thus, the state ρ_i as well as the measurement setting X_i and the noisy POVMs elements of round i can be arbitrarily correlated and depend arbitrarily on the state, settings, POVMs, and outcomes of the previous rounds, as long as assumptions (I) and (III) are satisfied. This takes into account any possible systematic error in the experiment and in particular closes the detection loophole for entanglement witness experiments (the possibility of violating the witness due to classical correlations in POVM measurements) [39].

Assumptions (II) and (III) model the devices used to perform the measurements in the experiment. We assume that the random setting generator is characterized up to some precision τ and that the measurement devices are characterized up to some δ_j, as defined in equations (15) and (16) respectively. In principle, τ and the δ_j may all depend on the round number i, and the δ_j may also depend on the measurement setting x. However, in practice, these dependencies will be small and we may safely take a maximum. Moreover, it would be extremely impractical to characterize the devices for each round specifically. The parameters τ and δ_j are later used to calculate the witness correction. With this we ensure that the confidence in the witness violation is never overestimated, even in the presence of systematic device errors.

Finally, in the rejection experiment, we also need to formalize the null hypothesis which we wish to reject. In the case of entanglement witnessing, the null hypothesis is that all states produced ρ_i are separable in every round i. We formulate this more generally, by letting $\mathcal{S}$ a convex subset of states (e.g. the separable states) such that W is a witness operator for $\mathcal{S}$. This means that $\mathrm{Tr}[W\rho] \geqslant 0$ for all $\rho \in \mathcal{S}$. Then we can finally state the null hypothesis for $\mathcal{S}$ mathematically as the following assumption:

($\mathcal{H}_0$) **Null hypothesis.** In every round i, the source produces a state $\rho_i \in \mathcal{S}$.

This assumption is to be rejected with statistical confidence by the experiment, as we will describe in the next section.

3. Results

In this section we present the main results of our work. In section 3.1 we start by giving a step-by-step description of our method and give an outline on how to apply the rejection analysis and estimation analysis. Then, in section 3.2, we compute the witness correction as a function of the model parameters that quantify the maximum device noise (τ and the δ_j's). Next, in section 3.3 we provide an easy-to-compute upper bound to the p-value, which is used to perform the witness rejection experiment. Finally, in section 3.4 we state how to compute a confidence interval for the average witness value equation (2). These results apply in the presence of arbitrary, possibly correlated noise on the states.

3.1. The design and analysis of an experiment

In this section, we detail all steps necessary to apply our framework—from the design of the experiment to the analysis of the data. In table 3 we summarize our method. We now explain each step in detail.

In this work, we assume that the specific observable W has been chosen. Of course, this choice is part of defining the entire experiment. For the rejection analysis, W should be a witness for some set $\mathcal{S}$ (as defined by property equation (1)). With entanglement witnessing in mind, $\mathcal{S}$ is the convex set of separable states and W is some entanglement witness. See section 5.2.1 for a discussion on how to choose a suitable W for witness experiments.

Step 1. Define the experiment. With a choice of W fixed, we now choose a decomposition of the witness W as in equation (10) (such a decomposition is not unique). A good decomposition minimizes the number of terms, while keeping each term simple to measure.

Then we choose an ideal model for the implemented measurements by describing each measurement as a POVM $\{\Pi_a^{(j),x}\}_{a\in\Omega_x^{(j)}}$ that satisfies equation (11). These POVMs should model the real implementation of the local measurements as accurately as possible (we will quantify the deviation of the real measurement devices in the second step). Note that these POVMs can simply be projective measurements.

Next, we choose the desired probability distribution p_x of the random measurement settings in each round. In principle, this can be chosen arbitrarily and the method will still work, but it has significant influence on the finite statistics of the experiment. We propose to choose p_x as

$$p_x = \frac{\sum_{\xi:f(\xi)=x}|w_\xi|}{\sum_\xi |w_\xi|}. \tag{17}$$

Here w_ξ are the weights appearing in the decomposition equation (10). This equation can be interpreted as choosing p_x proportional to the sum of absolute values of the weights $|w_\xi|$ of all observables ξ that correspond to setting x. Hence, heavy-weight terms are measured more frequently to increase the precision of estimating that term. See section 5.2.3 for a more detailed discussion on choice of p_x. The choices made so far define the score function equation (13) that assigns a score to each round i of the game.

Finally, we fix the number of rounds n to play in the entanglement witness game, as well as a significance level α (typical values are $\alpha = 0.05, 0.01, 0.001$). In the rejection experiment, the significance level determines how small the observed p_{bound} on the p-value must be in order for us to reject hypothesis ($\mathcal{H}_0$). In the estimation experiment, the significance level α determines the confidence of the constructed confidence intervals around the estimate. For the entanglement rejection experiment, it is important that all these parameters, especially α and n are set before the experiment is carried out (see section 5.4).

Step 2. Characterize devices. In this step, we need to characterize the measurement devices that aim to implement the POVM elements of equation (11), and the random setting generator that aims to implement p_x. This characterization is done by determining suitable τ and δ_j's such that equations (15) and (16) hold (ensuring that assumptions (II) and (III) plausibly hold). In practice, this process requires calibration and characterization of the real experimental devices. From the numbers τ and δ_j's obtained in this characterization, we can compute the so-called *witness correction* $\gamma = \gamma_1 + \gamma_2$ using theorem 1,

Table 3. Outline of our method for designing, performing and analyzing witness experiments. We assume that the experiments are guided by the model assumptions of table 2 and that an appropriate operator W is fixed.

Outline of experiment design and analysis
1 *Define the experiment.* Choose (a) (a) Decomposition of W of the form equation (10); (b) Measurement model of the form equation (11); (c) Probability p_x of measurement settings (e.g., using equation (17)); (d) Number of rounds n; (e) Significance level α For a rejection experiment, W should be a witness for some set $\mathcal{S}$. Choose the null hypothesis to be hypothesis ($\mathcal{H}_0$)
2 *Characterize devices* w.r.t. the model assumptions. Determine suitable τ and δ_j (see equations (15) and (16)) for the hardware devices. From this *compute the witness correction* γ using theorem 1, equations (25) and (26)
3 *Carry out the experiment.* In each round i, record the obtained score s_i using equation (13)
4a *Perform the rejection analysis.* From the recorded scores, compute the total normalized score t_n using equation (18). Evaluate the upper bound $p_{\text{bound}}(t_n, n, \gamma)$ using theorem 2, equation (30). If $p_{\text{bound}} \leqslant \alpha$, reject ($\mathcal{H}_0$) and conclude that the source is *capable of producing states* $\rho \notin \mathcal{S}$ with confidence at least $1-\alpha$. Otherwise, the test was inconclusive
4b *Perform the estimation analysis.* From the recorded scores, compute the witness estimate $\hat{w}_n$ using equation (21). From the significance level α, compute the confidence interval radius ε using theorem 3, equation (34). Then, $\mathcal{I} = [\hat{w}_n - \varepsilon, \hat{w}_n + \varepsilon]$ is a $(1-2\alpha)$ two-sided confidence interval and $\mathcal{J} = (-\infty, \hat{w}_n + \varepsilon]$ is a $(1-\alpha)$ one-sided confidence interval for the unknown quantity $\langle W \rangle_n$ as defined in equation (2)

equations (25) and (26), in section 3.2. When appropriate, one can use a first-order approximation for γ_2 given in equation (29). The witness correction γ is defined such that it bounds $\|W - \bar{W}_i\|_\infty$, where $\bar{W}_i$ is the effectively implemented operator in round i and W is the ideal target operator. It is a function of the parameters τ and δ_j, which quantify the device imperfections. The witness correction γ is used to protect against the largest possible systematic error in the experiment under the model assumptions of table 2.

Step 3. Carry out the experiment. Play n rounds of the witness game. Each round i, receive a state ρ_i from the source and measurement setting $X_i = x$ from the random setting generator. Then each subsystem j performs the POVM measurement $\{\tilde{\Pi}^{(j)}_{i,a}\}_{a \in \Omega^{(j)}_x}$ corresponding to setting x and obtains one of the possible outcomes labeled by a_j. Collect all the obtained outcomes $\mathbf{A}_i = \mathbf{a}$, compute and the score $s_i = s(x, \mathbf{a})$ using the score function in equation (13) and record s_i. After the data collection has completed, one can do the analysis. We differentiate between the *rejection analysis* and *estimation analysis*. Both can be done using the same recorded data.

Step 4a. The rejection analysis. After the data collection has completed, we can determine if the experiment successfully rejected the null hypothesis ($\mathcal{H}_0$) with confidence $1-\alpha$. To do so, we compute the *total normalized score* t_n, defined by

$$t_n = \sum_{i=1}^{n} \frac{s_i - s_{\min}}{\Delta s}, \tag{18}$$

where $\Delta s := s_{\max} - s_{\min}$ and

$$s_{\min} := \min_{x,\mathbf{a}} s(x, \mathbf{a}), \quad s_{\max} := \max_{x,\mathbf{a}} s(x, \mathbf{a}), \tag{19}$$

are the algebraic minimum and maximum value of the score function, respectively. Note that $t_n \in [0, n]$. This total normalized score is the our test statistic for the hypothesis test. We can reject the null hypothesis if the p-value is at most α. The p-value is defined as the probability

$$p := \Pr\left[T_n \geqslant t_n \,|\mathcal{H}_0\right] \tag{20}$$

of obtaining a total normalized score T_n under the null hypothesis ($\mathcal{H}_0$) that is at least as large as the observed total normalized score t_n. To determine if $p \leqslant \alpha$, we compute an upper bound $p \leqslant p_{\text{bound}}(t_n, n, \gamma)$ to the p-value in theorem 2, equation (30), and compare p_{bound} to α. If $p_{\text{bound}} \leqslant \alpha$ then we can reject the null hypothesis ($\mathcal{H}_0$) with confidence at least $1-\alpha$. We can therefore conclude that at least one state $\rho \notin \mathcal{S}$ must have been produced and therefore the source has the *capability of producing* such states. In the context where $\mathcal{S}$ is the set of separable states, this is interpreted as concluding that the source is capable of producing entangled states. This logical reasoning is only valid if all the model assumptions (I) to (III) in table 2 hold. If these fail then one may incorrectly reject $\mathcal{H}_0$.

Step 4b. The estimation analysis. From the collected data, we can also estimate the average witness value $\langle W \rangle_n$ as defined in equation (2) and construct a rigorous confidence interval for around the estimate

in the presence of arbitrary noise. The average witness value is estimated by the estimator

$$\hat{w}_n = c - \frac{1}{n}\sum_{i=1}^{n} s_i. \tag{21}$$

This estimator can, in the absence of noise on the measurements and random number generation, be seen as an unbiased estimator of $\langle W\rangle_n$ by equation (6). In the presence of unknown noise, the bias of the estimate can only be bounded by the witness correction γ (see the discussion in section 3.2). Using γ, we can compute the radius ε of the confidence interval using equation (34). By theorem 3, the interval $\mathcal{I}(\hat{w}_n) = [\hat{w}_n - \varepsilon, \hat{w}_n + \varepsilon]$ is a $(1-2\alpha)$ two-sided confidence interval and $\mathcal{J}(\hat{w}_n) = (-\infty, \hat{w}_n + \varepsilon]$ is a $(1-\alpha)$ one-sided confidence interval for $\langle W\rangle_n$. If W is a witness for the set $\mathcal{S}$ in the sense of equation (1) and if $\hat{w}_n + \varepsilon < 0$, then one can conclude that $\langle W\rangle_n < 0$, meaning that *on average* states outside $\mathcal{S}$ must have been produced, i.e.

$$\rho^* = \frac{1}{n}\sum_{i=1}^{n} \rho_i \notin \mathcal{S}, \tag{22}$$

with confidence at least $1-\alpha$. We emphasize that the intervals $\mathcal{I}, \mathcal{J}$ are corrected for systematic (measurement and random setting generation) errors within the model assumptions via the witness correction γ of theorem 1 (since ε depends on γ) and that it is statistically rigorous for arbitrary state noise.

3.2. Computing the witness correction

In this section we present theorem 1 to compute the witness correction γ as a function of the randomness and measurement imperfection parameters τ, δ_j determined in step 2 of table 3. The imperfect implementation of the measurements and random number generator will lead to an effectively implemented operator

$$\bar{W}_i = cI - \sum_x \tilde{p}_{i,x} \sum_{\mathbf{a}} s(x,\mathbf{a})\, \bar{\Pi}^x_{i,\mathbf{a}}. \tag{23}$$

Here $\bar{\Pi}^x_{i,\mathbf{a}}$ is the expected implemented joint POVM in round i, conditioned on the history of the experiment, the state produced, and the event that $X_i = x$ (which happens with probability $\tilde{p}_{i,x}$). See appendix C for a precise definition. Note that this effectively implemented operator $\bar{W}_i$ is closely related by the ideal witness operator W by comparing to equation (14). Indeed, the ideal random setting distribution p_x is replaced with the implemented distribution $\tilde{p}_{i,x}$ (which differ little by assumption (II)) and the ideal POVM elements $\Pi_a^{(1),x} \otimes \cdots \otimes \Pi_a^{(m),x}$ are replaced with the conditional expected implemented joint POVM elements $\bar{\Pi}^x_{i,\mathbf{a}}$ (which differ little by assumption (III)). The witness correction γ we derive in theorem 1 precisely captures how much $\bar{W}_i$ can deviate from W within the model assumptions of table 2.

Theorem 1. *Let W be a Hermitian operator (not necessarily a witness in the sense of equation (1)) with decomposition and ideal implementation given by equations (10) and (11). Suppose the experiment is modeled by the model assumptions in table 2. Define the effectively implemented operator $\bar{W}_i$ by equation (23). Then, in every round i,*

$$\|W - \bar{W}_i\|_\infty \leqslant \gamma, \tag{24}$$

where the witness correction $\gamma := \gamma_1 + \gamma_2$ *is the sum of the* random number generation correction γ_1 *and the* measurement correction γ_2 *defined by*

$$\gamma_1 := \tau \sum_x \max_{\mathbf{a}} |s(x,\mathbf{a})| \tag{25}$$

$$\gamma_2 := \sum_\xi |w_\xi| \sum_{j=1}^{m} \left(\prod_{k=1}^{j-1} (\lambda_\xi^{(k)} + \epsilon_\xi^{(k)}) \right) \epsilon_\xi^{(j)} \prod_{k=j+1}^{m} \lambda_\xi^{(k)}, \tag{26}$$

respectively, in terms of

$$\epsilon_\xi^{(j)} := b_j(\xi)\delta_j \sum_{a \in \Omega^{(j)}_{f(\xi)}} |a| \quad \text{and} \quad \lambda_\xi^{(j)} := \|O_\xi^{(j)}\|_\infty. \tag{27}$$

The proof is given in appendix C. Let us first explain why we call the quantity γ the witness correction. An important consequence of equation (24) is that

$$|\mathrm{Tr}[W\rho_i] - \mathrm{Tr}[\bar{W}_i\rho_i]| \leqslant \gamma \tag{28}$$

for all ρ_i. This means the witness inequality $\mathrm{Tr}[W\rho_i] \geqslant 0$ implies that $\mathrm{Tr}[\bar{W}_i\rho_i] \geqslant -\gamma$. Thus, if W is a witness, then the operator $\bar{W}_i + \gamma I$ is also a witness. Hence, γ is the witness correction in the sense that any effectively implemented witness $\bar{W}_i$ corrected by γ is still a witness. We emphasize that this result does not say anything about the effects of finite statistics, but is solely about the required correction of expectation values due to imperfect devices. That is, the factor γ protects against potential systematic errors in an experiment.

The witness correction γ has two terms, γ_1 and γ_2. The term γ_1 quantifies the correction due to imperfect random number generation. The constant γ_2 quantifies the correction due to measurement errors. Thus, γ can be interpreted as the total correction required if the witness W is implemented with noisy measurements and with an imperfect number generator. Note that the choice of p_x influences the correction γ_1, as the score function equation (13) depends on p_x.

The measurement correction γ_2 has a simple first-order approximation under the assumption that $\lambda_\xi^{(j)} = 1$, making it easier to compute. This assumption means that all measurement operators have eigenvalues in the interval $[-1, 1]$ and is satisfied for example by all Pauli operators. Then a first-order approximation for γ_2 is

$$\gamma_2 = \sum_\xi |w_\xi| \sum_{j=1}^{m} \epsilon_\xi^{(j)} + O(\epsilon^2), \tag{29}$$

where ϵ is a constant such that $\epsilon_\xi^{(j)} \leqslant \epsilon$ for all ξ, j. Hence, this is a good approximation if $\epsilon \ll 1$. This is typically the case when $\delta_j \ll 1$, which means that the measurement devices have been well-characterized. In section 5.3.2, we discuss a possible alternative method for deriving γ.

3.3. Bound on the *p*-value for witness rejection experiments

In this section, we give the main result to perform the rejection analysis in theorem 2. The theorem provides an easy-to-compute upper bound on the p-value under the null hypothesis ($\mathcal{H}_0$). Recall that the p-value is the probability of observing a total normalized score T_n under the Null Hypothesis ($\mathcal{H}_0$) that is at least as large as the observed total normalized score t_n in the experiment, $p = \Pr[T_n \geqslant t_n|\mathcal{H}_0]$. If the p-value is smaller than a previously chosen significance level α, then we may consider the Null Hypothesis ($\mathcal{H}_0$) to be statistically unlikely to explain the observed t_n, and we may reject the model at significance level α. To determine if $p \leqslant \alpha$, we put an upper bound p_{bound} on p in theorem 2, which can be compared to the significance level α.

Theorem 2. *Let W be a witness operator (satisfying equation (1)) for the set $\mathcal{S}$ with decomposition and ideal implementation given by equations (10) and (11). Suppose that the experiment is governed by the model assumptions of table 2 and consider the null hypothesis ($\mathcal{H}_0$) with respect to $\mathcal{S}$. Let t_n denote the observed total normalized score after n rounds in the experiment. Then, the p-value as defined in equation (20) is upper-bounded by*

$$p_{bound} := eF^{\circ}_{n,\beta}(t_n), \tag{30}$$

where

$$F^{\circ}_{n,\beta}(x) := F_{n,\beta}(\lfloor x \rfloor)^{1-(x-\lfloor x \rfloor)} F_{n,\beta}(\lfloor x \rfloor + 1)^{x-\lfloor x \rfloor} \tag{31}$$

is the log-linear interpolation of the survival function of a binomial distribution with parameters n and β,

$$F_{n,\beta}(k) = \sum_{l=k}^{n} \binom{n}{l} \beta^l (1-\beta)^{n-l}, \tag{32}$$

and where

$$\beta = \min\left(1, \frac{c + \gamma - s_{\min}}{\Delta s}\right). \tag{33}$$

Finally, $\lfloor x \rfloor$ is the largest integer less than or equal to x.

We give a detailed theorem 2 in appendix D. We construct a supermartingale sequence from the total normalized scores up to each round i. We then apply Bentkus' inequality [30, 31] (a concentration inequality for bounded difference supermartingale sequences, similar to, but tighter than, the Hoeffding–Azuma inequality) to obtain an upper bound for the p-value. Our proof is inspired by the approach of [32] to certify Bell violations.

3.4. Confidence intervals for average witness estimation experiments

In this section, we give the main result for the witness estimation analysis in theorem 3. The theorem provides confidences interval for the average witness expectation value $\langle W\rangle_n$ as defined in equation (2). The point estimate for $\langle W\rangle_n$ is given in equation (21), and is a function of the scores recorded in the experiment. We construct a $(1-2\alpha)$ two-sided confidence interval $\mathcal{I}$ and a $(1-\alpha)$ one-sided confidence interval $\mathcal{J}$.

Theorem 3. *Let W be a Hermitian operator (not necessarily a witness in the sense of equation (1)) with decomposition and ideal implementation given by equations (10) and (11). Suppose that the experiment is governed by the model assumptions in table 2. Let $\hat{W}_n$ denote the average witness estimate as defined in equation (21). Fix the significance level $\alpha \in [0,1]$. If $\alpha < e(\frac{1}{2})^n$, define $\varepsilon = \Delta s$, otherwise define $\varepsilon \in [\gamma, \gamma+\Delta s]$ implicitly via*

$$\alpha = eF^{\circ}_{n,\frac{1}{2}}\left(\frac{n}{2}(1+\frac{\varepsilon-\gamma}{\Delta s})\right). \tag{34}$$

Here γ is defined in equations (25) and (26), and $F^{\circ}_{n,\beta}$ is defined in equation (31) (with $\beta = \frac{1}{2}$ and $e \approx 2.72$). Then, the intervals

$$\mathcal{I}(\hat{w}_n) := [\hat{w}_n - \varepsilon, \hat{w}_n + \varepsilon] \tag{35}$$

$$\mathcal{J}(\hat{w}_n) := (-\infty, \hat{w}_n + \epsilon] \tag{36}$$

are a $(1-2\alpha)$ two-sided and a $(1-\alpha)$ one-sided confidence interval respectively for the average witness value $\langle W\rangle_n$ as defined in equation (2). That is

$$\Pr[\langle W\rangle_n \in \mathcal{I}(\widehat{W}_n)] \geqslant 1-2\alpha, \tag{37}$$

$$\Pr[\langle W\rangle_n \in \mathcal{J}(\widehat{W}_n)] \geqslant 1-\alpha. \tag{38}$$

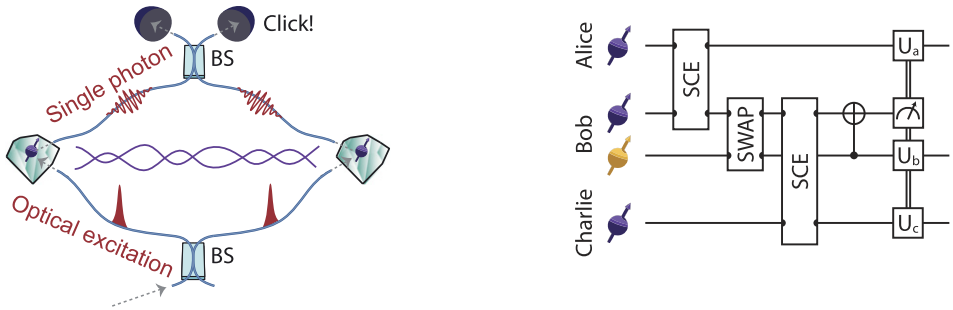
The proof of this theorem is given in appendix E. The confidence interval is also based on the construction of a martingale sequence and the application of Bentkus' inequality. The techniques are very similar to the theorem 2. We chose to use Bentkus' inequality because it is tighter than the more standard Hoeffding–Azuma inequality [32]. The radius of the interval ε is however slightly more difficult because it involves (numerically) solving equation (34). See section 5.3.3 for a brief discussion on this.

4. Examples and illustration

In this section, we will illustrate our results with two examples based on simulations of a proposed entanglement witness experiment in nitrogen vacancy (NV) centers. Moreover, we will give a concrete example in which the iid and Gaussian assumptions fail. Finally, we shall illustrate how the function $eF^{\circ}_{n,\beta}$ of equation (31) scales in its arguments and parameters. This function determines the p-value bound and the confidence interval size in our results. Before we present these examples, we briefly describe the physical system that we aim to simulate in section 4.1. This section serves as a motivation for our simulation, but the examples can also be understood without knowledge of the physical system we simulate. Then we present the two examples. In the first example (section 4.2), we describe how to apply our method in detail, outlining all the steps in section 3.1 in a concrete example. For this example, we simulate a single experiment with identically distributed states ρ. In the second example (section 4.3), we illustrate our method for non-iid states. To do so, we will perform a large Monte Carlo simulation of many independent experiments. In each experiment, we use a sequence of three-qubit states ρ_i that are neither independent nor identically distributed. Then, in section 4.4, we give an artificial example of non-iid states in which the Gaussian assumption fails considerably. This example shows that a Gaussian assumption, on which the central limit theorem relies in prior work, need not always be justified (cf the discussion in section 1.1). Our method applies regardless of the validity of a Gaussian assumption. Finally, in the section 4.5 we illustrate how the function $eF^{\circ}_{n,\beta}(x)$ defined in equation (31) (which directly determines the p-value bound equation (30), and the confidence interval equation (34)) scales with n, β and x. Note that β scales linearly with the witness correction γ that captures device imperfections.

4.1. Simulation details of nitrogen vacancy systems

Both examples in sections 4.2 and 4.3 are based on a scheme for generating tripartite GHZ states in three physically separated nitrogen vacancy (NV) centers in diamond (see reference [40] for a review of this



(a) Singe click entanglement (SCE) between two NV centers.

(b) Schematic of tripartite GHZ generation.

Figure 3. Schematic illustration of tripartite entanglement generation using diamond nitrogen vacancy (NV) center systems. (a) The single click entanglement (SCE) scheme generates a single EPR pair between two NV centers. (b) Two EPR pairs are generated using the SCE scheme and combined into a GHZ state by interfering and measurement. The classically conditioned operations U_a, U_b and U_c are Pauli operations.

system). In these NV centers, the electronic spin associated with the defect can be used as qubit. This qubit is optically accessible and can be entangled with the presence or absence of a photon, which can be used as a flying qubit. Surrounding the NV center there are several carbon-13 atoms (1.1% natural abundance). Their nuclear spins can be used as additional qubits, which can be controlled via the hyperfine interaction between the nuclear and electronic spins.

Two NV centers are entangled in the following way [19]: first, each NV center produces a spin-photon entangled pair, where the qubit state is encoded in the absence/presence of a photon. The joint state of the spin-photon pair is then given by

$$\sqrt{z}|\uparrow\rangle|1\rangle + \sqrt{1-z}|\downarrow\rangle|0\rangle, \tag{39}$$

where z is a tunable parameter. By coupling the photons into single mode fibers and interfering them using a beam splitter, the two electronic spins can become entangled. In essence, this amounts to detecting the presence of a photon but erasing the information about which arm the photon came from. This single click entanglement (SCE) scheme is illustrated in figure 3(a). The joint state between the two electronic spins is now (ideally)

$$|z\uparrow\uparrow\rangle\langle\uparrow\uparrow| + (1-z)|\Psi_\theta^+\rangle\langle\Psi_\theta^+|, \tag{40}$$

where $|\Psi_\theta^+\rangle = |\uparrow\downarrow\rangle + e^{i\theta}|\downarrow\uparrow\rangle$. Here, θ is a relative phase that needs to be characterized and controlled experimentally to create useful entanglement.

To generate a tripartite GHZ state, two EPR pairs are combined into a single GHZ state in the following way: first create one EPR pair between two NV centers using the electronic qubits. Then, one node swaps the state of the electronic spin with a nuclear spin qubit, so that the electronic spin becomes free again for entanglement production. At this point a second EPR pair is produced between the now-free electronic spin of this node and a third node. The GHZ state is then created by coupling the nuclear spin and electronic spin in the middle node and measuring the electronic qubit. This results into a state that is equivalent to a GHZ state under local Pauli operations (the Pauli operations can depend on the observed measurement outcome). This procedure is sketched in figure 3(b).

When simulating this procedure, we account for several noise processes. First, we include noise in the generation of the EPR pairs. Our model for EPR generation follows the noise model for SCE generation developed in [19]. This model incorporates several independent noise parameters: the single photon detection efficiency p_{det} (the probability of detecting a photon in the heralding station, conditioned on it being emitted from the NV center), the distinguishability V of the emitted photons, the double excitation probability $p_{2\text{ph}}$ (when more than one photon is emitted by an NV center in single entanglement attempt), the probability of dark counts p_{dc}, as well as an uncertainty in the relative phase θ that is modeled by applying a Pauli-Z to one of the qubits with probability p_θ. We assume that the detection efficiency is the same for all three setups, and furthermore assume that in each SCE scheme symmetric values for the free parameter z (see equation (39)) are used. However, this parameter may be different for the first and second EPR pair. We refer the reader to reference [19] for full details on the SCE generation model.

On top of the detailed SCE noise, our model assumes dephasing noise on the first EPR pair while it is kept in memory, waiting for the successful generation of the second EPR pair. The off-diagonal terms in the density matrix of the first EPR pair are multiplied with dephasing parameter $q = 1 - \exp(-N_{\text{max}}/\nu)$, where N_{max} is a free parameter determining the maximal number of attempts before we discard everything and

Table 4. Values of the parameters used in the simulated creation of a tripartite GHZ state in NV centers as discussed in section 4.1. The resulting state is described in table 5.

Noise parameter	Value	Free parameter	Value
p_{det}	1.5×10^{-3}	z_1	0.016
V	0.9	z_2	0.080
$p_{2\text{ph}}$	0.02	$N_{\max}$	468
p_{dc}	4.0×10^{-7}		
p_θ	0.030		
ν	1500		

start over (because the first EPR pair has decohered too much) and ν is a parameter quantifying the strength of the dephasing noise. Finally, we assume that all single- and two-qubit gates are performed with unit fidelity. In section 4.2 we instantiate this model with representative numerical values for all model parameters (see table 4) to produce the simulated experimental data.

4.2. Step-by-step application of our method

In our first example, we illustrate our method on data produced by a simulation of iid states on three qubits ($m = 3$). The aim is to witness a genuine tripartite entanglement by producing a GHZ state:

$$|\text{GHZ}\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle). \tag{41}$$

Therefore, we let $\mathcal{S}$ be the set of biseparable states, i.e., the set of all states ρ that are a mixture of separable states on any bipartition of the three subsystems. To witness a state not in $\mathcal{S}$ (and reasonable close to a GHZ state), we use is the projection witness, given by

$$W = \frac{1}{2}I - |\text{GHZ}\rangle\langle\text{GHZ}|. \tag{42}$$

This factor $\frac{1}{2}$ is known to be optimal for the GHZ projection witness [20]. Note that $c \neq \frac{1}{2}$ in equation (8). In fact, it is easily observed that $c = \frac{1}{2} - \frac{1}{d} = \frac{3}{8}$ here (since $d = 2^m = 8$ in this example). We will now describe all steps of section 3.1 to illustrate how to fully define the experiment, obtain the (simulated) data, and calculate the resulting p-value and confidence interval of the experiment.

Step 1a. The first step is choosing a decomposition of our choice of W (equation (42)) of the form equation (10). This witness has a four-setting and five-setting decomposition. We will use the five-setting decomposition into local Pauli observables:

$$W_{\text{GHZ}} = \frac{1}{8}\left(3III - IZZ - ZIZ - ZZI - XXX + XYY + YXY + YYX\right), \tag{43}$$

where $\{I, X, Y, Z\}$ are the Pauli operators (including the identity operator) and the tensor symbol is omitted for clarity. Thus, equation (43) is a decomposition of the form equation (8) with $c = \frac{3}{8}$. We shall label the seven non-identity, traceless observables by $\xi = 1, \ldots, 7$ in the order of their appearance in equation (43).

There are only five measurement settings needed for the decomposition in equation (43). These are $\{ZZZ, XXX, XYY, YXY, YYX\}$, since the first three observables ($\xi = 1, 2, 3$) can all be computed from the first measurement setting ZZZ ($x = 1$), as discussed in section 3.2. Therefore we have $x = 1, \ldots, 5$, indexing the settings. The formal mapping between observables and measurement settings is given by

$$f(\xi) = \begin{cases} 1, & \text{if } \xi = 1, 2, 3, \\ \xi - 2, & \text{if } \xi = 4, 5, 6, 7, \end{cases} \tag{44}$$

and

$$b(\xi) = \begin{cases} 011, & \text{if } \xi = 1, \\ 101, & \text{if } \xi = 2, \\ 110, & \text{if } \xi = 3, \\ 111, & \text{if } \xi = 4, 5, 6, 7. \end{cases} \tag{45}$$

Step 1b. Next, we specify a model for measuring each of the Pauli observables that occur in the chosen measurement settings. In our simulation, we shall model each measurement as systematically imperfect,

meaning that X, Y and Z is not implemented by the usual projective measurements. Instead, we model all Pauli measurement by POVM elements, parameterized by two parameters $u, v \in [0, 1]$, which characterize the efficiency of detecting the $+1$ and -1 eigenstate of the Pauli operator, respectively. In experiments, these numbers are referred to as the readout fidelity [19, 41]. Concretely, for the Pauli-Z measurement on each subsystem, we model (dropping the subsystem index j for notational compactness) the measurement by the POVM elements

$$\Pi^Z_{a^+} = \begin{bmatrix} u & 0 \\ 0 & 1-v \end{bmatrix}, \quad \Pi^Z_{a^-} = \begin{bmatrix} 1-u & 0 \\ 0 & v \end{bmatrix}. \tag{46}$$

The X and Y POVM elements are defined by

$$\Pi^X_{a^\pm} = H\Pi^Z_{a^\pm}H^\dagger, \quad \Pi^X_{a^\pm} = K\Pi^Z_{a^\pm}K^\dagger, \tag{47}$$

where the H and K gate are the gates that rotate the Z to the X and Y basis respectively. The two outcomes of all Pauli measurements are

$$a^\pm = \frac{v - u \pm 1}{u + v - 1}. \tag{48}$$

These values are chosen in such a way that

$$a^+\Pi^P_{a^+} + a^-\Pi^P_{a^-} = P, \tag{49}$$

for all Pauli's $P = X, Y, Z$, so that the measurement operators correspond to the desired observables (according to equation (11)). In our model, we will set $u = 0.95$ and $v = 0.99$. This results in $a^+ \approx 1.1064$ and $a^- \approx -1.0213$.

Step 1c. Next, we choose p_x according to equation (17). That is, we choose

$$p_x = \begin{cases} \frac{3}{7} & \text{if } x = 1, \\ \frac{1}{7} & \text{if } x = 2, 3, 4, 5 \end{cases} \tag{50}$$

This now fully defines the score function (per equation (13)):

$$s(x, \mathbf{a}) = \begin{cases} \frac{7}{24}(a_2a_3 + a_1a_3 + a_1a_2), & \text{if } x = 1, \\ \frac{7}{8}a_1a_2a_3, & \text{if } x = 2, \\ -\frac{7}{8}a_1a_2a_3, & \text{otherwise.} \end{cases} \tag{51}$$

Note how the observables IZZ, ZIZ and ZZI are combined into one measurement setting ZZZ which directly contributes to the score.

Step 1d–e. Finally, we fix the total number of rounds to be $n = 600$ and set $\alpha = 0.05$.

Step 2. We characterize our (simulated) devices to have an RNG bias $\tau = 10^{-6}$ and measurement imperfection as compared to the model described in the previous step of $\delta_j = \delta = 2 \times 10^{-3}$ for all parties $j = 1, 2, 3$. The value of δ is determined from the uncertainty in the measurement characterization of the NV system. The value of τ is chosen sufficiently large for any practical implementation of randomness.

With these values of τ and δ, and the score function equation (51), the witness correction γ can be computed from theorem 1. The random number generation correction γ_1 is computed using equation (25) to be

$$\gamma_1 = \tau \sum_{x=1}^{5} \max_{\mathbf{a}} |s(x, \mathbf{a})| \approx 5.8 \times 10^{-6}. \tag{52}$$

The measurement correction γ_2 is computed using equation (26) and (27). First, we compute $\epsilon^{(j)}_\xi$ from equation (27). We find that

$$\epsilon^{(j)}_\xi = \begin{cases} 0, & \text{if } \xi = j = 1, 2, 3, \\ \delta(|a^+| + |a^-|) \approx 4.26 \times 10^{-3}, & \text{otherwise.} \end{cases} \tag{53}$$

Furthermore, from equation (43), it is clear that $\lambda^{(j)}_\xi = 1$ for all ξ, j, since all local observables $O^{(j)}_\xi$ are any of the four Pauli operators I, X, Y, Z, which have operator norm 1. Since $\lambda^{(j)}_\xi = 1$ and $\epsilon^{(j)}_\xi \ll 1$, we use the

Table 5. Nonzero components of the state ρ used in each round of the simulation of the entanglement witness experiment described in section 4.2. The expected witness value of ρ is $\mathrm{Tr}[W\rho] = -0.172$.

M	$\mathrm{Tr}[M\rho]$	M	$\mathrm{Tr}[M\rho]$	M	$\mathrm{Tr}[M\rho]$
III	1	*XXX*	0.782	*IIX*	−0.077
IZZ	0.787	*XYY*	−0.737	*XXI*	0.072
ZIZ	0.478	*YXY*	−0.478	*YYI*	−0.047
ZZI	0.608	*YYX*	−0.507	*ZZX*	−0.047

approximation equation (29) instead of the full equation (26). We compute that

$$\gamma_2 \approx \sum_{\xi=1}^{7} |w_\xi| \sum_{j=1}^{3} \epsilon_\xi^{(j)} \approx 9.6 \times 10^{-3}. \tag{54}$$

Hence we find that

$$\gamma = \gamma_1 + \gamma_2 \approx 9.6 \times 10^{-3} \leqslant 0.01. \tag{55}$$

Step 3. In this step, the experiment is simulated. We play n rounds of the entanglement witness game. In our simulation, we take the same state $\rho_i = \rho$ in each round, corresponding to an iid situation. This state is computed using the model of tripartite GHZ generation in remote NV centers as discussed in section 4.1. The valued of the parameters we used in the simulation are given in table 4. The resulting state ρ is given in table 5. In addition to the state $\rho_i = \rho$, we also generate a random setting $x_i \in \{1, 2, 3, 4, 5\}$ with probability given by equation (50). The randomness is generated by a standard pseudo-random number generator. For efficiency reasons, we use the ideal POVM elements to simulate the measurement outcomes. We nevertheless use a nonzero value of δ to illustrate the effect of measurement noise on the witness correction. The set of measurement outcomes $\mathbf{a}_i$ is obtained according to Born's rule. From this, we compute and record the score of this round $s_i = s(x_i, \mathbf{a}_i)$ using the score function equation (51).

Step 4a. In this step, we calculate the upper bound p_{bound} to the p-value using theorem 2. This is straightforwardly done using γ, n, c, $s_{\min}$, $s_{\max}$ and the total normalized score t_n by evaluating equations (30)–(33) of theorem 2. The quantities $s_{\min}$ and $s_{\max}$ as defined in equation (19), can be computed from the score function equation (51). In our example we find $s_{\max} = -s_{\min} \approx 1.185$. The total normalized score t_n is computed from the recorded scores s_i for $i = 1, \ldots, n$, according to equation (18). To compute p_{bound} from concrete numbers, suppose the a single simulation of the experiment yields a total normalized score of $t_n = 440.97$. To compute the p-value bound, we first calculate β using equation (33). We find (using $\gamma \leqslant 0.01$) that $\beta \approx 0.662$. Hence, we can evaluate our p-value bound, equations (30) and (31), by

$$p_{\mathrm{bound}} = eF_{n,\beta}(440)^{0.03} F_{n,\beta}(441)^{0.97} \approx 2.1 \times 10^{-4}. \tag{56}$$

Since $p \leqslant p_{\mathrm{bound}} \leqslant \alpha$, we have rejected the null hypothesis ($\mathcal{H}_0$) at significance level $\alpha = 5\%$. As we trust that the model assumptions (I) to (III) hold, we can reject hypothesis ($\mathcal{H}_0$) and conclude that our (simulated) source of quantum states is capable of producing genuinely multipartite entangled states $\rho^* \notin \mathcal{S}$. This is indeed the case, since ρ defined by table 5 is not biseparable (i.e., not in $\mathcal{S}$, since $\mathrm{Tr}[W\rho] < 0$).

To illustrate how the p-value evolves as more rounds are played (more data is taken), we plot in figure 4 the running value of p_{bound} computed with the total normalized score up to round i, as a function of i. Up until round $i \approx 520$, the bound remains constant at its maximal value $e \approx 2.72$, due to the prefactor of e in equation (30). In this regime there is insufficient data to draw any conclusions. Then our p-value bound starts decreasing roughly exponentially in i. The jitter can be explained by the randomness due to measurement settings and outcomes.

Step 4b. Finally, we illustrate how to estimate $\langle W \rangle_n$ as defined in equation (2) and compute the confidence intervals around this estimate using theorem 3. The witness estimate is directly computed from the observed scores and c using equation (21). Suppose that a run of the experiment yielded scores such that $\hat{w}_n = -0.182$ (this is consistent with $t_n = 440.97$ by comparing equations (18) and (21)). Now the radius of the confidence interval can be computed from $\alpha = 0.05$, $n = 600$, $\gamma = 0.01$ and $\Delta s = 2.370$ by solving equation (34) numerically. We find that $\varepsilon \approx 0.216$ in this example. Hence, we find the 90% two-sided confidence interval and the 95% one-sided confidence interval

$$\mathcal{I}_{0.9} = [-0.398, 0.034], \quad \mathcal{J}_{0.95} = (-\infty, 0.034], \tag{57}$$

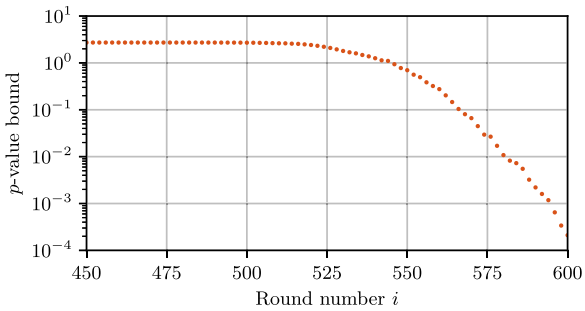


Figure 4. Running p-value bound for our simulated GHZ witness experiment with parameters as defined in the main text. Up until round $i = 520$, the upper bound is the maximal value of $e \approx 2.72$, due to the prefactor e in equation (30). The final p-value bound after $n = 600$ rounds is 2.1×10^{-4}.

respectively. This is just insufficient to conclude that on average ρ was not in $\mathcal{S}$ (i.e., genuinely multipartite entangled). However, we can compare these intervals to the true value $\langle W\rangle_n = -0.172$ (see table 5 and recall that $\rho_i = \rho$ in this example). Clearly $\langle W\rangle_n \in \mathcal{I}_{0.9}$ and $\langle W\rangle_n \in \mathcal{J}_{0.95}$. Moreover, the point estimate $\hat{w}_n$ is not far off the true value.

4.3. Illustration of method for correlated noise

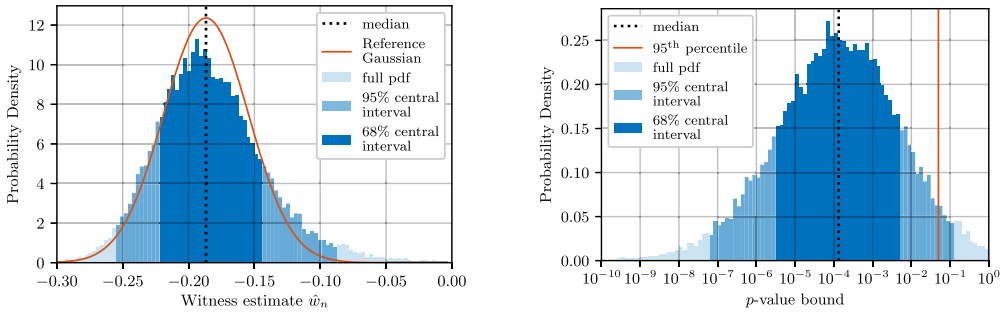
In the previous example we gave a step-by-step illustration of how our method is applied to data gathered in a single (simulated) experiment with identical states each round. Since our method is applicable in general, we now illustrate its use for states ρ_i that are neither independent nor identically distributed. As before, we will consider states ρ_i that are created from two EPR pairs, each of which has a relative phase θ (see equation (40) in section 4.1), but we now assume that the relative phase θ drifts between subsequent rounds following a random walk. That is, the relative phase in round i depends on its value in round $i-1$, which creates correlation between the rounds. This modeled phase drift is motivated by the observation in experiments that the relative phase θ changes over time due to fluctuation in the optical path length [19]. The simulation uses the same parameters as in section 4.2, table 4, except that now $p_\theta = 0$. Instead, in the first round, we model (both EPR pairs) with known initial phase $\theta_0 = 0°$. Then, each round, the phase drifts with step size $\Delta\theta = \pm 0.98°$ (for each EPR pair the drift is an independent random walk). This creates dependence and correlation between the rounds. Furthermore, we use the same witness (equation (42)) and employ the same measurement model with witness correction $\gamma \leqslant 0.01$ (equation (55)) as in section 4.2.

Instead of performing a single simulation with this noise model, we perform a Monte Carlo simulation with $N = 2 \times 10^4$ repetitions of the experiment, each with $n = 600$ rounds. For each repetition, we calculate the witness estimate from the observed score using equation (21). We emphasize that in this simulation, each repetition also a different $\langle W\rangle_n$ is realized, because the collection of $\rho_1, \ldots, \rho_n$ can be different every time. We also compute a bound on the p-value for each repetition of the experiment by using theorem 2. We thus obtain $N = 2 \times 10^4$ witness estimates and bounds on the p-value from our Monte-Carlo simulation.

In figure 5 we plot histograms of the witness estimates $\hat{w}_n$ and p-value bounds p_{bound}, both of which are directly computed from the observed scores $s_1, \ldots, s_n$ (via equation (21) and theorem 2, respectively). We emphasize that all s_i are realizations of a random variable S_i, determined by the random measurement settings X_i, the random measurement outcomes $\mathbf{A}_i$ (following Born's rule), and the correlated random states ρ_i produced by the source. The produced states were not necessarily biseparable (i.e., not necessarily in $\mathcal{S}$).

In figure 5(a) we plot a histogram of the witness estimates $\hat{w}_n$ and compare to a Gaussian reference curve. We clearly see that this noise process can lead to a non-Gaussian witness distribution. The skew in the plot is mainly due to the fact that different true average witness values $\langle W\rangle_n = \frac{1}{n}\mathrm{Tr}[W\rho_i]$ are realized in each run of the experiment. The figure illustrates that under non-iid noise, the standard Gaussian prediction $\hat{w}_n \pm \hat{\sigma}_{\hat{w}_n}$ does not accurately predict future repetitions of the experiment. We emphasize that the plot in figure 5(a) cannot be interpreted as the probability distribution of $\hat{w}_n$ under a particular, fixed sequence of states $\rho_1, \ldots, \rho_n$ (because each Monte Carlo iterate used a different sequence of states), and hence no inference could be made from this plot about the uncertainty of $\hat{w}_n$.

In figure 5(b) we plot a histogram of the p-value bounds. Note that the p-value is plotted logarithmically on the x-axis, making the bins of unequal size. We show the significance level $\alpha = 0.05$ as a red line, which



(a) Histogram of witness estimate

(b) Histogram of p-value bound

Figure 5. Histograms for a Monte-Carlo simulation of $N = 2 \times 10^4$ independent witness experiments (each with $n = 600$ rounds). Here, the noise model is not iid but rather follows a stochastic process, as described in section 4.3. (a) Histogram of witness estimates. The blue histogram shows the distribution of the witness estimate $\hat{w}_n$ (cf equation (21)). The distribution is non-Gaussian, as is obvious from comparison with the red Gaussian reference curve. Thus, the standard Gaussian prediction $\hat{w}_n \pm \hat{\sigma}_{\hat{w}_n}$ does not accurately predict future repetitions of the experiment. The skew in the distribution is primarily due to skew in the true average witness value $\langle W \rangle_n$ across different runs of the non-iid experiment (due to the changing sequence of n states). (b) Histogram of p-value bounds. The blue histogram shows the distribution of p-value upper bounds (cf theorem 2). The red line indicates the significance level $\alpha = 0.05$, the number below which we reject the null hypothesis ($\mathcal{H}_0$).

in this example turns out to be the 95th percentile. Thus, in 95% of the simulated experiments, the null hypothesis ($\mathcal{H}_0$) was rejected with statistical confidence. Note that it is merely a coincidence that $\alpha = 0.05$ marks the 95% percentile of the distribution of p_{bound} over the N simulated experiments. We emphasize that the p-value itself (as defined in equation (20)) cannot be inferred from the simulation results. This is because the p-value is a statement about the distribution of the total normalized score T_n assuming that the null hypothesis ($\mathcal{H}_0$) holds. But here the states produced are not necessarily in $\mathcal{S}$, so hypothesis ($\mathcal{H}_0$) is violated. See section 5.4 for more detailed discussion.

4.4. Example where iid assumption fails

In the previous section we saw an example where the states within each experiment were generated by a non-iid noise process, and we discussed the corresponding distribution of the witness estimate and p-value bound across different runs of the non-iid experiment. While the witness estimate $\hat{w}_n$ was non-Gaussian, this was largely explained by skew in true average witness value $\langle W \rangle_n$. In this section we give an explicit example where the iid assumption (more generally the Gaussian assumption) is inappropriate even when the average witness value $\langle W \rangle_n$ is fixed. This example is not based on simulation of NV centers, but is based on a noise model that is designed to clearly exhibit non-iid behavior.

In this example, we generate the states according to a noise process in which the source of states produces a perfect GHZ state in a fraction F of the n rounds and produces an orthogonal state in the remaining $(1 - F)n$ rounds. Importantly, the fraction of good states F is held constant in this model. This is the only source of correlations between the states in a single run of the experiment. This model is an extreme case of the more realistic scenario in which the source intermittently produces good and bad quality states. These bad quality states can for example be produced when a heralding signal (a signal that indicates entanglement was created) is falsely triggered. Note that the fraction F of good states equals the true average fidelity to the GHZ state,

$$F = \frac{1}{n}\sum_{i=1}^{n} \text{Tr}\left[\rho_i |\text{GHZ}\rangle\langle\text{GHZ}|\right]. \tag{58}$$

and with our choice of witness operator W (according to equation (42)), the true witness value is

$$\langle W \rangle_n = \frac{1}{n}\sum_{i=1}^{n} \text{Tr}[(\frac{1}{2}I - |\text{GHZ}\rangle\langle\text{GHZ}|)\rho_i] = \frac{1}{2} - F. \tag{59}$$

In this example, we fix $F = 0.672$, which yields a true average witness value is $\langle W \rangle_n = -0.172$.

In figure 6 we plot a histogram of the witness estimate $\hat{w}_n$, computed using equation (21), under the described noise model. The histogram represents $N = 2 \times 10^4$ independent simulated experiments (each with $n = 600$ rounds). We also plot the Gaussian distribution that would be obtained under the iid assumption on the median run. For this run, the mean and standard deviation $\hat{w}_n \pm \hat{\sigma}_{\hat{w}_n}$ are computed from the observed scores $s_1, \ldots, s_n$ in that run. In particular, for the red curve we observed

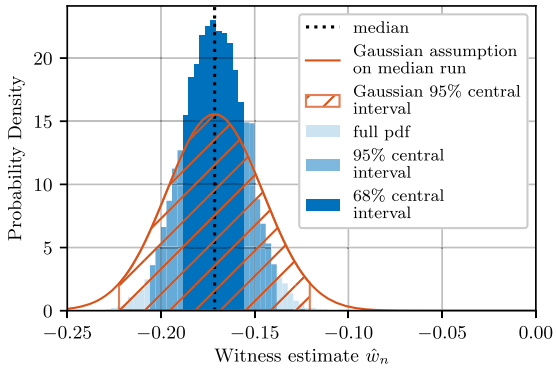


Figure 6. Histogram of a Monte-Carlo simulation of $N = 2 \times 10^4$ witness experiments (each with $n = 600$ rounds). The histogram shows the distribution of the witness estimate $\hat{w}_n$ computed from the observed scores using equation (21) in blue. This approximates the true distribution of $\hat{w}_n$. The 95% interval is shaded in blue. The true average witness value is $\langle W \rangle_n = -0.172$ in each run of the experiment, which coincides with the mean (and median) $\hat{w}_n$. The red curve is shows the inferred distribution from the data (the scores) observed in the median run when the data is assumed to be iid. That is, the red curve is a Gaussian with mean $\hat{w}_n$ and standard deviation $\hat{\sigma}_{\hat{w}_n}$ as computed from the observed scores under the iid assumption in this particular run. The central 95% interval of this Gaussian is shaded with red lines. Clearly, the distribution inferred from the iid assumption does not match the true distribution of $\hat{w}_n$.

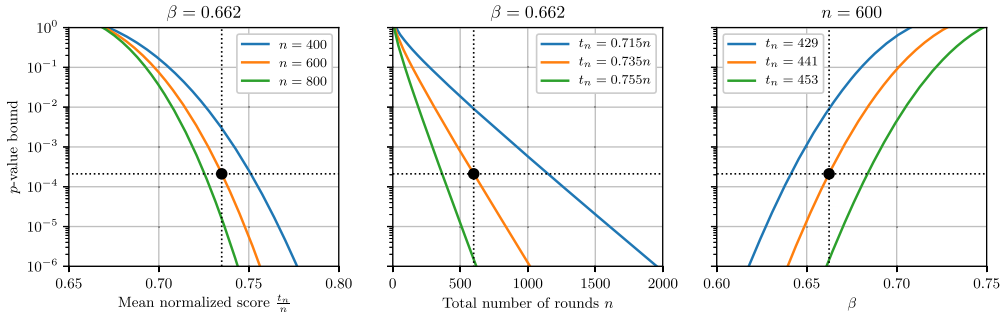


Figure 7. Scaling of the function $p_{\text{bound}} = eF^{\circ}_{n,\beta}(t_n)$, which determines the bound on the p-value theorem 2 and the radius of the confidence interval theorem 3, with the total normalized score t_n (equation (18)), the number of rounds n and β (which depends on γ, see equation (33)). The black dots correspond to the simulation discussed in section 4.2, where $n = 600$, $\beta = 0.662$ (using $\gamma = 0.01$), and $t_n = 440.97$, yielding $p_{\text{bound}} = 2.1 \times 10^{-4}$.

$\hat{w}_n \pm \hat{\sigma}_{\hat{w}_n} = -0.171 \pm 0.026$. The 95% central interval is shown as the shaded region under the red curve. This region is $[\hat{w}_n - 2\hat{\sigma}, \hat{w}_n + 2\hat{\sigma}] = [-0.223, -0.120]$. However, the 95% central interval found from the Monte Carlo simulation is only $[-0.205, -0.138]$. Hence, the 95% central interval estimated under iid assumption (red shaded region) is roughly 1.5 times larger than the 95% central interval found from the Monte Carlo simulation of $\hat{w}_n$ (the medium blue region). This example clearly shows that in this scenario the iid assumption fails.

4.5. Scaling of the p-value bound

Finally, we analyze how our bound $p_{\text{bound}} = eF^{\circ}_{n,\beta}(t_n)$ (see theorem 2, equation (30)) scales with its three free parameters: the total normalized score t_n, the number β (which depends on the witness correction γ, see equation (33)) and the number of rounds n. Note that this also directly illustrates how the size of the confidence interval ϵ scales with the significance level α according to $\alpha = eF^{\circ}_{n,\frac{1}{2}}(\frac{n}{2}(1 + \frac{\varepsilon - \gamma}{\Delta s}))$ (see theorem 3, equation (34)). We discuss the scaling here from the perspective of the rejection analysis. Then it is qualitatively clear that a larger observed score t_n and smaller γ (and hence smaller β) should lead to statistically more significant results (meaning a smaller p_{bound}). Similarly, a larger number of rounds n should lead to a smaller uncertainty and hence make it less likely to observe extreme data under the null hypothesis ($\mathcal{H}_0$). We show that this is indeed the case in figure 7. To compute p_{bound} numerically, we used $c = \frac{3}{8}$ and $s_{\max} = -s_{\min} = 1.185$, just as in the example of section 4.2 as the fixed values. The black dots in the figure correspond to $n = 600$, $\beta = 0.662$ and $t_n = 440.97$ as used for illustration in section 4.2. The corresponding bound is $p_{\text{bound}} = 2.1 \times 10^{-4}$.

In the left plot of figure 7, we see that it appears that $p_{\text{bound}} \propto \mathrm{e}^{-t_n^2}$ for fixed n in the regime $t_n \geqslant n\beta$. Even if this is not exact, p_{bound} at least appears to decrease (super)exponentially in t_n. The p-value bound is

trivial in the regime $t_n \leqslant n\beta$. This is understood to mean that $n\beta$ is a total normalized score that is likely to be achieved under the null hypothesis. In the middle plot, we see that p_{bound} seems to decrease exponentially with n as well and that the asymptotic decay rate depends on the mean normalized score $\frac{t_n}{n}$. This is also expected behavior, which for example also holds in the iid case. In the right plot, we focus on the effects of increasing β. The p-value bound increases (sub)exponentially with β. It also seems that an increment of β (which is due to increased γ from noisy devices) can be directly compensated by observing an increased t_n, since the three curves seem to be identical but displaced in the x-axis.

5. Discussion

5.1. Applicability of our work

The methods we have discussed throughout this work were frequently motivated by and illustrated with entanglement witness experiments. Here we elaborate how generally applicable all of our results are. First, the estimation experiment can in principle be done for any Hermitian observable W. In theorems 1 and 3 we do not require that the operator is a witness; we only need to write W in the form of equations (10) and (11). Note that this is in principle always possible (although for general W, the number of terms labeled by ξ might grow exponentially with the total system dimension). This allows one to perform the experiment and do the estimation analysis of step 4b in table 3. This yields a point estimate and confidence interval for estimating the average value $\langle W\rangle_n$ as defined in equation (2), which is valid without any assumption about the sequence of states $\rho_1, \ldots, \rho_n$ produced sequentially in the experiment.

Second, the rejection analysis of the experiment is valid for any witness W that satisfies equation (1) for some convex subset of states $\mathcal{S}$. The null hypothesis we aim to reject is always the hypothesis that the source can only produce states $\rho \in \mathcal{S}$, conform hypothesis $(\mathcal{H}_0)$. Therefore, the experiment can witness the creation of any state $\rho^* \notin \mathcal{S}$, given suitable choices of W and $\mathcal{S}$. Examples include witnessing different types of entanglement, defined by non-membership of a particular separability class $\mathcal{S}$. A frequently encountered separability class is the set $\mathcal{S}_k$ of k-separable states. These states are a convex combination of pure states that are separable over some k-partition of the subsystems [42]. Our examples in section 4 focused on $k = 2$, the class of biseparable states. In certain applications it might also be important to certify entanglement across a particular partition of subsystems, which leads to a different definition of $\mathcal{S}$. In this case, the set $\mathcal{S}$ describes separability between any specific partitioning of the system. Corresponding witness operators for graph states have been identified in [43]. Finally, $\mathcal{S}$ can also be chosen to include specific classes of entangled states, so that non-membership signals that the entanglement is not of that particular class. As an example of this, $\mathcal{S}$ can be chosen as the convex hull of all biseparable states and W-class states [42]. All of this can be tested with our method by accordingly defining $\mathcal{S}$ in hypothesis $(\mathcal{H}_0)$. Of course, this also requires one to find a suitable witness operator W that separates $\mathcal{S}$ from some particular state ρ^* that one aims to witness (according to equation (1)).

5.2. Choosing the free parameters of the experiment

5.2.1. The witness operator W

The choice of witness operator is another integral part of the design of the experiment. It affects the number of trials n needed in an experiment to attain a certain p-value given some target state ρ^*. It affects the (expected) p-value bound or interval size ε one would observe in an experiment given finite n. There are two different mechanisms behind this. First, the choice of W influences the decomposition equation (8), and in particular the number of terms that appear in this decomposition. A smaller number of terms directly implies more statistically significant results. This is intuitively understood since the budget of n data points can now be used to measure less observables in the decomposition, meaning each can be estimated more accurately. Second, for witness rejection experiments where some (average) ρ^* is assumed to be produced, one would ideally like to minimize the p_{bound} over all witness operators W for fixed ρ^* and n. This task is extremely difficult and impractical. Instead, one often employs the heuristic of minimizing the witness violation. That is, one seeks to find the minimizer W of the following optimization problem

$$\min_{\substack{W:\|W\|_\infty \leqslant 1,\\ \mathrm{Tr}[W\rho] \geqslant 0\ \forall \rho \in \mathcal{S}}} \mathrm{Tr}[W\rho^*]. \tag{60}$$

This witness will then give large contributions to the witness violation, hence making it statistically less likely to observe such a violation with finite number of states from $\mathcal{S}$. The downside of this is that it only optimizes for the magnitude of the violation in case of infinite statistics. In particular, it does not take into account the number of terms in the decomposition.

Many different methods for entanglement witness design have been discussed extensively in literature [20, 22, 23]. Most methods rely on some form of numerical optimization to find a witness with particular attractive properties such as efficient decomposition, maximal violation with a given state or large noise tolerance. For pure states, the most widely used witness operator for genuine multipartite entanglement of states close to $|\psi\rangle$ is the projection witness $W = \lambda^2 I - |\psi\rangle\langle\psi|$, where λ is the maximum of the Schmidt coefficients of $|\psi\rangle$ when all bipartitions are considered [20, 22]. This is also the choice that we made in section 4 for $|\psi\rangle = |\mathrm{GHZ}\rangle$ (see equations (42) and (43)).

5.2.2. The number of rounds n

An important design parameter of a witness experiment is the number n of rounds to be played. This is often delicate, as too small n renders the entire experiment yielding no conclusion (see figure 4) but too large n can be overly costly in experiment resources. To get a good estimate of n that produces small enough p_{bound} without being unnecessarily expensive, the experimenter must have a good understanding of the quantum states being produced and the measurements being performed (e.g., from a theoretical model of the experiment). With this, the experiment can be simulated, so that one can get an estimate of the distribution of the total normalized score T_n. Then pick n such that you are very likely (e.g., in 90% of the experiments) to realize a t_n that evaluates with known or estimated γ to a p_{bound} less than α (for the rejection experiment) or that evaluates given α to a confidence interval that is sufficiently small (e.g. such that the entire interval is smaller than zero).

5.2.3. The target distribution of measurement settings p_x

In principle, the choice of probability distribution over the measurement settings p_x is left free in this work. Each choice leads to a valid witness experiment from which the conclusion can be drawn rigorously. However, bad choices of p_x will lead to suboptimal bounds on the p-value or suboptimal confidence intervals for equal experimental cost. An interesting question is what choice of p_x yields the smallest p-value bound (resp. confidence interval) given fixed n and γ. This question is hard to address, since p_x influences both the distribution of T_n and the value of Δs (which enters in theorem 2 via β and theorem 3 via equation (34)). Based on the underlying concentration inequality (Bentkus' inequality, see lemma 1 in appendix D for details), we conjecture that p_x should optimally be chosen to minimize Δs. This conjecture holds in the simulation of our experiments. Intuitively, this requirement ensures that $S_i - S_{i-1}$ is often large compared to Δs. However, finding a distribution p_x that minimizes Δs is not a simple problem. In small instances, it can be solved by brute force search. If all possible outcome sets are $\Omega_x^{(j)} = \{1, -1\}$, then the p_x that minimizes Δs is given by our recommended choice in equation (17).

5.3. Possible modifications of our method

5.3.1. Model modifications

It is not hard to incorporate different ways of characterization the hardware devices [modifying assumptions (II) and (III)]. For example, the bias in random measurement setting generation can be quantified by different ℓ^q-norms than the ℓ^∞-norm considered here. Similarly, there are many other ways of characterizing noisy measurement devices. For example, one may model and characterize noisy measurements by a small misalignment angle, defined as the angle between the Bloch vectors of the ideal and the noisy measurements [26]. Any modification of the way hardware devices are characterized, requires that theorem 1 is modified accordingly to calculate an appropriate correction γ. We emphasize that this is the only part that needs modification. Theorems 2 and 3 can directly be applied with the new bound γ.

5.3.2. Alternative method for computing γ

In the theorem 1, we applied an analytical method to find the witness correction γ. This bound is not necessarily tight in all cases. If the correction γ is too large for a practical experiment, one can see if tighter bounds can be found numerically. The optimization problem for gamma under and of the model takes the form

$$\gamma = -\min_{\tilde{p}_x} \min_{\bar{W}} \min_{\rho\in\mathcal{S}} \mathrm{Tr}[\bar{W}\rho], \tag{61}$$

where the minimization over $\tilde{p}_x$ is constrained by assumption (II) and the minimization over $\bar{W}$ is constraint by assumption (III). This optimization is not always easy to carry out. For example, optimizing over $\mathcal{S}$ is hard in general (e.g., when $\mathcal{S}$ is the set of separable states), although in low dimensions one can often resort to using PPT relaxations [44, 45]. Furthermore, the set of feasible $\bar{W}$ is not convex, and the objective function $\mathrm{Tr}[\tilde{W}\rho]$ is bilinear in the variables. For non-convex problems, general nonlinear

optimization methods usually only find local optima which may result in witness corrections γ that are smaller than justified.

5.3.3. Using Hoeffding–Azuma instead of Bentkus

In this work, we have chosen to use Bentkus' inequality to bound the relevant tail probabilities in theorems 2 and 3, because this is often slightly tighter than the more common Hoeffding–Azuma inequality. We have stated Bentkus' inequality in lemma 1 in appendix D. Whenever this lemma applies, the Hoeffding–Azuma inequality also applies. So if $R_0, \ldots, R_n$ is a supermartingale with $R_0 = 0$ and $-\beta \leqslant R_i - R_{i-1} \leqslant 1 - \beta$, then for all $x \geqslant 0$ Hoeffding–Azuma states that $\Pr[R_n \geqslant x] \leqslant \mathrm{e}^{-\frac{2x^2}{n}}$. By replacing this in the right places in the proofs of appendices D and E, we can modify the statements of theorems 2 and 3. In particular, equation (30) can be modified to

$$\Pr[T_n \geqslant t_n | \mathcal{H}_0] \leqslant \mathrm{e}^{-2\frac{\max(t_n - n\beta, 0)^2}{n}}, \tag{62}$$

with β still defined as in equations (33) and (34) can be modified to

$$\alpha = \mathrm{e}^{-\frac{n}{2}(\frac{\varepsilon-\gamma}{\Delta s})^2}. \tag{63}$$

Especially this latter modification can be useful, since this makes it easier to compute the confidence interval in theorem 3. This is because the above equation can be explicitly solved for ε to find

$$\varepsilon = \gamma + \Delta s \sqrt{\frac{2}{n} \ln(\alpha^{-1})}. \tag{64}$$

One may choose to use these equation instead of the theorems as stated, but generally this could come at a slight cost in the tightness of the bounds.

5.4. General remarks regarding hypothesis testing and the *p*-value

Our rejection analysis of entanglement witness experiments fits into the general framework for hypothesis testing in statistics. As always, it is important that all relevant parameters are fixed prior to the experiment being carried out. In our case, this concerns the quantities defined in steps 1 and 2 of section 3.1. Most notably, this includes the number of trials n and the significance level α, but also the witness operator, its decomposition, and all other parameters of the model assumptions. This also implies that one may not simply combine data from multiple experiments (e.g., extending experiments by further trials until a desirable outcome is achieved). Instead, p-values of different experiments can be combined using known statistical methods [32, 46], or one can carry out a larger, independent experiment instead.

In the framework of hypothesis testing, the p-value is defined as the probability of observing a test statistic T (in our case, the total normalized score T_n) under the null hypothesis (in our case hypothesis ($\mathcal{H}_0$)) that is at least as large as the observed value t:

$$p = \Pr[T \geqslant t | \mathcal{H}_0]. \tag{65}$$

If the p-value is smaller than a previously chosen significance level α, then the null hypothesis is considered to be statistically unlikely to explain the observed t, and we reject the null hypothesis at significance level α. This means that the p-value is a statement about the observed data in relation to a hypothesis about the distribution of T. In particular, it does not give any information about the distribution of T if the null hypothesis does not hold. Neither should the p-value be interpreted as 'the probability that the null hypothesis was valid' or as 'when the experiment is repeated many times, the null hypothesis is rejected in a $1 - p$ fraction of experiments'. The p-value is only quantifies the likelihood of a particular assumption about the distribution of T, namely that it is governed by the null hypothesis. Therefore, it is this a hypothesis that we may plausibly reject upon observing a small p-value. Finally, we emphasize that the p-value itself can be seen as a random variable, as it is a function of the realization of a randomly observed test statistic t (whether or not governed by the null hypothesis). Each experiment is simply a realization of a particular p-value.

6. Conclusions

In this work, we proposed two new methods to analyze witness experiments. Both methods are applicable to any source of quantum states that produces a sequence of states on demand. These states can be arbitrary distributed and correlated—unlike previous works, which often implicitly assume that the noise

is well-behaved so that the central limit theorem applies. With our rejection analysis, the experimenter can rigorously certify that a device has the capability of producing entangled quantum states. The statistical confidence is quantified by a bound on the p-value, the probability that an experiment yields data as extreme as the observed data under the assumption that the source produces only separable states. Hence a small p-value is directly interpreted as strong evidence for the production of entangled states. The rejection method applies more generally to any witness experiment that is defined by a set of states and a corresponding witness operator. In particular, it can be used to witness genuine multipartite entanglement. With our estimation analysis, the experimenter can estimate the average witness value and construct a statistically rigorous confidence interval around this estimate. This method allows the experimenter to conclude whether entanglement was produced on average. This confidence is valid regardless of the type of noise present. The estimation method applies to any general estimation problem. This means that it is not necessary that W is a witness in the sense of equation (1). Fidelity estimation is therefore also covered by this method.

Both methods we derived are simple to use. We provide a step-by-step recipe to choose all relevant parameters, collect the necessary experimental data and compute both a bound on the p-value and a average witness estimate with confidence intervals from this collected data. Our method requires no modification compared to prior methods of performing witness experiments, except possibly the requirement of measuring different settings in random order instead of a fixed, predetermined order. This requirement is however inevitable if one wishes to deal with arbitrarily correlated noise. Both of our methods yields a figure of merit—a bound on the p-value or a point estimate with a confidence interval—that has a concrete interpretation and is comparable between experiments. We illustrated our methods with simulations of an experiment in nitrogen vacancy centers.

In this work we chose to treat the witness operator as a fixed object that has been chosen with a state to be witnessed in mind. However, other methods exist in which the witness is modified adaptively based on the measurement data collected so far [24]. The objective of adaptive witnessing is to tune the witness operator (perhaps within a parameterized family) to have maximal detection efficiency, based on the partial knowledge of the noise obtained from prior measurement data. In particular, one can attempt to identify and counteract coherent components of the noise. Indeed, coherent and local noise preserve entanglement, and adaptive protocols have been shown to enhance entanglement detection [47]. An interesting open question is whether the statistical methods we develop in this work can be combined with adaptive witness methods. This would open ways to increase the detection efficiency of the witness method, while maintaining the robustness against correlated noise.

Acknowledgments

The authors thank Jérémy Ribeiro for very informative discussions on concentration inequalities and martingale sequences and Liangzhong Ruan for inspiring discussion on estimation problems with non-iid states. The authors would also like to thank Sophie Hermans and Simon Baier for helpful feedback on earlier versions of this manuscript, and Jonas Helsen for proofreading the manuscript. BD, MW, and SW acknowledge the NWA Startimpuls. BD and SW are supported by the Netherlands Organisation for Scientific Research (NWO) through a VIDI grant and by the European Research Council (ERC) trough an ERC Starting Grant QINTERNET. B D, S W and R H are also supported by the NWO Zwaartekracht Grant Quantum Software Consortium. RH furthermore acknowledges financial support from NWO through a VICI grant and from the ERC through an ERC Consolidator Grant. MW is supported by an NWO VENI grant (No. 680-47-459). MP acknowledges the Marie Skłodowska-Curie Action 'Nanoscale solid-state spin systems in emerging quantum technologies' Spin-NANO (grant agreement no. 676108). Finally, B D, M P, R H and S W acknowledge the EU Flagship on Quantum Technologies and the Quantum Internet Alliance (funding from EU Horizon 2020 research and innovation programme, Grant No. 820445).

Appendix A. Game formulation

In this first appendix, we expand on sections 2.1 and 2.2. We will redefine all objects introduced in the main text, while being slightly more formal and precise. We will also show that equation (14) holds, which relates the definition of the score function to the witness decomposition. Let us start of with recapturing the notion of a witness operator (or simply witness). A witness is defined as a Hermitian observable W that, for some convex subset of states $\mathcal{S}$, satisfies

$$\mathrm{Tr}[\rho W] \geqslant 0 \quad \forall \rho \in \mathcal{S}. \tag{A1}$$

As described in the main text (equation (8)), we assume the witness admits a decomposition into the sum of local observables:

$$W = cI + \sum_{\xi} w_\xi O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)}. \tag{A2}$$

Often, it is possible to compute the expectation value of multiple observables $\{O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)}\}_\xi$ from a single measurement setting $M_x^{(1)} \otimes \cdots \otimes M_x^{(m)}$ (we shall label measurement settings by a label x). To keep track of which observables (labeled by ξ) are related to which measurement setting (labeled by x), we define $f(\xi) = x$ if the observable $O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)}$ can be measured by the measurement setting $M_x^{(1)} \otimes \cdots \otimes M_x^{(m)}$. This means that there exists a bitstring $b(\xi) \in \{0,1\}^m$ of length m such that

$$O_\xi^{(1)} \otimes \cdots \otimes O_\xi^{(m)} = (M_{f(\xi)}^{(1)})^{b_1(\xi)} \otimes \cdots \otimes (M_{f(\xi)}^{(m)})^{b_m(\xi)}. \tag{A3}$$

To allow for the most general model of measurements, we allow each $M_x^{(j)}$ in a measurement setting to be measured by a POVM $\{\Pi_a^{(j),x}\}_{a\in\Omega_x^{(j)}}$ with outcomes labeled by elements a in some finite set $\Omega_x^{(j)}$. That is, we write

$$M_x^{(j)} = \sum_{a\in\Omega_x^{(j)}} a\,\Pi_a^{(j),x}. \tag{A4}$$

The tensor product POVM for setting x has outcomes $\mathbf{a} = (a_1, \ldots, a_m)$ in $\Omega_x := \Omega_x^{(1)} \times \cdots \times \Omega_x^{(m)}$. We denote it by

$$\Pi_{\mathbf{a}}^x := \Pi_{a_1}^{(1),x} \otimes \cdots \otimes \Pi_{a_m}^{(m),x}. \tag{A5}$$

Sometimes we also need $\Omega^{(j)} := \bigcup_x \Omega_x^{(j)}$ and $\Omega := \Omega^{(1)} \times \cdots \times \Omega^{(m)}$. We formally define $\Pi_a^{(j),x} := 0$ for $a \in \Omega^{(j)} \backslash \Omega_x^{(j)}$ and $\Pi_{\mathbf{a}}^x := 0$ for $\mathbf{a} \in \Omega \backslash \Omega_x$. This is useful in case we want to sum over $a \in \Omega^{(j)}$ or $\mathbf{a} \in \Omega$ without knowing x.

Using equation (A4), we can write each local observable $O_\xi^{(j)}$ as

$$O_\xi^{(j)} = \left(M_{f(\xi)}^{(j)}\right)^{b_j(\xi)} = \left(\sum_{a\in\Omega^{(j)}} a\,\Pi_a^{(j),f(\xi)}\right)^{b_j(\xi)} = \sum_{a\in\Omega^{(j)}} a^{b_j(\xi)}\,\Pi_a^{(j),f(\xi)}, \tag{A6}$$

where the last equality follows from the facts that $b_j(\xi) \in \{0,1\}$ and POVM elements sum to the identity I, and we use the convention that $a^0 = 1$, $Q^0 = I$ for any number a and operator Q. Hence by plugging equation (A6) into equation (A2), the witness operator can be further decomposed as

$$W = cI + \sum_{\xi} w_\xi \bigotimes_{j=1}^{m} \sum_{a\in\Omega^{(j)}} a^{b_j(\xi)}\,\Pi_a^{(j),f(\xi)} = cI + \sum_{\mathbf{a}\in\Omega}\sum_{x} \left(\sum_{\xi:f(\xi)=x} w_\xi \prod_{j=1}^{m} a_j^{b_j(\xi)}\right) \Pi_{\mathbf{a}}^x. \tag{A7}$$

The last equality follows from rearrangement and equation (A5). Table 1 in the main text summarizes all relevant objects.

As discussed in sections 2.1 and 2.2, the witness experiment now proceeds in each round by selecting a measurement setting x at random, measuring the corresponding POVMs on each subsystem, and, upon obtaining outcomes $\mathbf{a} \in \Omega_x$, assigning a score according to equation (13):

$$s(x,\mathbf{a}) = -\frac{1}{p_x} \sum_{\xi:f(\xi)=x} w_\xi \prod_{j=1}^{m} a_j^{b_j(\xi)}. \tag{A8}$$

It is useful to define $s(x,\mathbf{a}) := 0$ if $\mathbf{a} \notin \Omega_x$. Now it becomes apparent why we had written W in the form equation (A7): it allows us to plug in the definition of the score function equation (A8). As a consequence, we find that

$$W = cI - \sum_{x,\mathbf{a}} p_x\, s(x,\mathbf{a})\, \Pi_{\mathbf{a}}^x, \tag{A9}$$

showing that equation (14) holds. Finally, we define the algebraic minimum and maximum score, and their difference, as

$$s_{\min} := \min_{x,\mathbf{a}} s(x,\mathbf{a}), \quad s_{\max} := \max_{x,\mathbf{a}} s(x,\mathbf{a}), \quad \Delta s := s_{\max} - s_{\min}. \tag{A10}$$

Table B1. Summary of the random variables present in the model and analysis

Object	Random variable	Type	Restrictions
History	H_i	Arbitrary	Assumption (I)
States	ρ_i	Arbitrary	Hypothesis ($\mathcal{H}_0$)
Measurement setting	X_i	Discrete	Assumption (II)
POVM elements	$\tilde{\Pi}_{i,a}^{(j)}$	Arbitrary	Assumption (III)
Measurement outcomes	$\mathbf{A}_i$	Discrete	Born's rule (equation (B7))
Score	S_i	Discrete	Equation (B8)
Total normalized score	T_i	Discrete	Equation (D5)

Appendix B. Formal definition of the model assumptions and null hypothesis

In this appendix, we expand on the model as formulated in section 2.3 and formalize further the model assumptions and null hypothesis that enter our description of the experiment. This allows us to rigorously draw conclusions from the witness experiment. The model assumptions are the collection of the following three assumptions:

(I) **Sequentiality.** The experiment is performed in rounds $i = 1, 2, \ldots, n$. For each round, there are random variables that model the state ρ_i, measurement setting X_i, POVM $\{\tilde{\Pi}_{i,a}^{(j)}\}_a$ and measurement outcome $\mathbf{A}_i$ (defined in the remaining assumptions and summarized in table B1). Apart from the restrictions implied by the remaining assumptions, all these random variables in round i can depend arbitrarily on all previous rounds. We formalize this by random variables H_i that models the history of the experiment up until, but excluding, round i. We assume that H_i contains at least

(a) the history of the preceding round, i.e, the σ-algebra of H_{i-1} is contained in the σ-algebra of H_i:

$$\sigma(H_{i-1}) \subseteq \sigma(H_i) \quad \forall i = 2, \ldots, n. \tag{B1}$$

(b) the state, measurement setting and outcome of round $i-1$, i.e.,

$$\sigma(X_{i-1}, \mathbf{A}_{i-1}, \rho_{i-1}) \subseteq \sigma(H_i) \quad \forall i = 2, \ldots, n, \tag{B2}$$

By Assumption (I)a, this implies that $X_1, \mathbf{A}_1, \rho_1, \ldots, X_{i-1}, \mathbf{A}_{i-1}, \rho_{i-1}$ are all contained in the history H_i.

(II) **Trusted randomness.** The measurement setting for the ith round of the experiment is modeled by a random variable X_i. We assume that the distribution of X_i given the history H_i and the state ρ_i (we denoted this informally as $\tilde{p}_{i,x}$ in the main text) is close to a known distribution p_x, i.e., there is $\tau > 0$ such that

$$|\Pr[X_i = x | H_i, \rho_i] - p_x| \leqslant \tau \quad \forall i, x \tag{B3}$$

holds almost surely. We further assume that

$$\tau < \min_x p_x. \tag{B4}$$

(III) **Trusted measurements.** The measurement in the ith round of the experiment is modeled by random POVMs $\{\tilde{\Pi}_{i,a}^{(j)}\}_a$ that are close to known POVMs $\{\Pi_a^{(j),x}\}_{a \in \Omega_x^{(j)}}$, where we recall that x labels the measurement setting, $j = 1, \ldots, m$ the subsystem, and $\Omega_x^{(j)}$ is the set of possible outcomes (we formally set $\Pi_a^{(j),x} := 0$ for $a \notin \Omega_x^{(j)}$). We model this by assuming that for each of these POVMs $\{\tilde{\Pi}_{i,a}^{(j)}\}_a$ there exists a constant $\delta_j > 0$ such that for each j, almost surely,

$$\left\| \tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),X_i} \right\|_\infty \leqslant \delta_j \quad \forall i, j, a. \tag{B5}$$

We further assume that

$$\tilde{\Pi}_{i,a}^{(j)} = 0 \quad \forall a \notin \Omega_{X_i}^{(j)} \forall i, j, a, \tag{B6}$$

which means the noisy measurements have the same sets of possible outcome as the ideal measurements. Finally, we assume that the measurement outcomes $\mathbf{A}_i$ follow Born's rule, so we demand that

$$\Pr\left[\mathbf{A}_i = \mathbf{a} \,\middle|\, H_i, \rho_i, X_i, \{\tilde{\Pi}_{i,a}^{(j)}\}\right] = \mathrm{Tr}\left[\rho_i \left(\tilde{\Pi}_{i,a_1}^{(1)} \otimes \cdots \otimes \tilde{\Pi}_{i,a_m}^{(m)}\right)\right] \quad \forall i, \mathbf{a}. \tag{B7}$$

In terms of the above data, the score of the ith round is given by the following random variable:

$$S_i := s(X_i, \mathbf{A}_i), \tag{B8}$$

where $s(x, \mathbf{a})$ is the score function as defined in equation (A8).

The model assumptions (I) to (III) are sufficient for analyzing a witness estimation experiment. To analyze a witness rejection experiment, we will also need to define the null hypothesis. We formally define this as follows (an informal definition was already given in section 2.3):

($\mathcal{H}_0$) **Null hypothesis: states in $\mathcal{S}$.** The quantum states ρ_i almost surely take values in the set $\mathcal{S}$.

The model assumptions together with this null hypothesis are sufficient for performing the witness rejection experiment, with the objective of rejecting hypothesis ($\mathcal{H}_0$).

Appendix C. Theorem 1—correcting for imperfect randomness and measurements

In this appendix, we discuss how the witness inequality must be corrected in the case of imperfect measurements (noisy POVM elements) and imperfect randomness. To do so, we first introduce shorthand notation for the noisy POVM elements analogously to equation (A5):

$$\tilde{\Pi}_{i,\mathbf{a}} := \tilde{\Pi}_{i,a_1}^{(1)} \otimes \cdots \otimes \tilde{\Pi}_{i,a_m}^{(m)} \tag{C1}$$

Moreover, we define the effectively implemented operator $\bar{W}_i$ for each round i as (cf equations (23) and (A9))

$$\bar{W}_i = cI - \sum_{x,\mathbf{a}} \Pr[X_i = x | H_i, \rho_i] s(x, \mathbf{a})\, \bar{\Pi}_{i,\mathbf{a}}^x, \tag{C2}$$

in terms of the following random variables

$$\bar{\Pi}_{i,\mathbf{a}}^x := \frac{\mathbb{E}\left[\tilde{\Pi}_{i,\mathbf{a}} \mathbf{1}_{[X_i = x]} | H_i, \rho_i\right]}{\Pr[X_i = x | H_i, \rho_i]}. \tag{C3}$$

We note that the denominator is nonzero almost surely by equation (B4), so that $\{\bar{\Pi}_{i,\mathbf{a}}^x\}$ is almost surely well-defined and a POVM for each i and x. These POVM elements can be interpreted as the expected implemented POVM elements in round i, conditioned on the event that setting $X_i = x$ is realized. We now bound the deviation of the effectively implemented operator $\bar{W}_i$ from the target operator W as a result of imperfect measurements and randomness.

Theorem 1 (restated). *Let W be a Hermitian operator of the form of equation (A7) (not necessarily a witness in the sense of equation (A1)). Suppose the experiment is modeled by the model assumptions (I) to (III) in appendix B. Define the score function as in equation (A8). Denote by $\bar{W}_i$ the effectively implemented operators as in equation (C2). Then, in every round i,*

$$\|\bar{W}_i - W\|_\infty \leqslant \gamma \tag{C4}$$

holds almost surely, where the witness correction

$$\gamma := \gamma_1 + \gamma_2 \tag{C5}$$

is the sum of the random number generation correction *and the* measurement correction *defined by*

$$\gamma_1 := \tau \sum_x \max_{\mathbf{a}} |s(x, \mathbf{a})|, \quad \textit{and} \quad \gamma_2 := \sum_\xi |w_\xi| \gamma_2(\xi), \tag{C6}$$

respectively, in terms of

$$\gamma_2(\xi) := \sum_{j=1}^{m} \left(\prod_{k=1}^{j-1} (\lambda_\xi^{(k)} + \epsilon_\xi^{(k)}) \right) \epsilon_\xi^{(j)} \left(\prod_{k=j+1}^{m} \lambda_\xi^{(k)} \right), \quad \epsilon_\xi^{(j)} := b_j(\xi)\delta_j \sum_{a \in \Omega_{f(\xi)}^{(j)}} |a| \quad \textit{and} \quad \lambda_\xi^{(j)} := \| O_\xi^{(j)} \|_\infty. \tag{C7}$$

□

Proof. Define

$$\check{W}_i := cI - \sum_x p_x \sum_{\mathbf{a}} s(x, \mathbf{a})\, \bar{\Pi}_{i,\mathbf{a}}^x. \tag{C8}$$

We will show that, almost surely, $\| \check{W}_i - \bar{W}_i \|_\infty \leqslant \gamma_1$ and $\| W - \check{W}_i \|_\infty \leqslant \gamma_2$, which together imply the theorem by the triangle inequality. For the former, we find that by comparing the definitions of $\bar{W}_i$ and $\check{W}_i$, equations (C2) and (C8) respectively, that

$$\| \check{W}_i - \bar{W}_i \|_\infty = \| \sum_x \left(\Pr[X_i = x | H_i, \rho_i] - p_x \right) \sum_{\mathbf{a}} s(x, \mathbf{a}) \bar{\Pi}_{i,\mathbf{a}}^x \|_\infty \tag{C9}$$

$$\leqslant \tau \sum_x \| \sum_{\mathbf{a}} s(x, \mathbf{a}) \bar{\Pi}_{i,\mathbf{a}}^x \|_\infty \tag{C10}$$

$$\leqslant \tau \sum_x \max_{\mathbf{a}} |s(x, \mathbf{a})| = \gamma_1, \tag{C11}$$

using assumption (II) in the first inequality and using in the second inequality that the operators $\{\bar{\Pi}_{i,\mathbf{a}}^x\}$ form a POVM for each i and x.

To show that $\| W - \check{W}_i \|_\infty \leqslant \gamma_2$, we start by comparing their respective definitions equations (A9) and (C8), which gives

$$W - \check{W}_i = \sum_x p_x \sum_{\mathbf{a}} s(x, \mathbf{a}) \left(\bar{\Pi}_{i,\mathbf{a}}^x - \Pi_{\mathbf{a}}^x \right) = \sum_\xi w_\xi \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\Pi_{\mathbf{a}}^{f(\xi)} - \bar{\Pi}_{i,\mathbf{a}}^{f(\xi)} \right), \tag{C12}$$

where we inserted the definition of the score function equation (A8) in the second step. Thus,

$$\| W - \check{W}_i \|_\infty \leqslant \sum_\xi |w_\xi| \left\| \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\Pi_{\mathbf{a}}^{f(\xi)} - \bar{\Pi}_{i,\mathbf{a}}^{f(\xi)} \right) \right\|_\infty, \tag{C13}$$

so it remains to prove that

$$\left\| \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\Pi_{\mathbf{a}}^{f(\xi)} - \bar{\Pi}_{i,\mathbf{a}}^{f(\xi)} \right) \right\|_\infty \leqslant \gamma_2(\xi), \tag{C14}$$

for all ξ to conclude the proof. Using the definition of $\bar{\Pi}_{i,\mathbf{a}}^{f(\xi)}$, equation (C3), we find that

$$\sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\Pi_{\mathbf{a}}^{f(\xi)} - \bar{\Pi}_{i,\mathbf{a}}^{f(\xi)} \right) = \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\Pi_{\mathbf{a}}^{f(\xi)} - \frac{\mathbb{E}\left[\tilde{\Pi}_{i,\mathbf{a}} \mathbf{1}_{[X_i = f(\xi)]} | H_i, \rho_i \right]}{\Pr[X_i = f(\xi) | H_i, \rho_i]} \right) \tag{C15}$$

$$= \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \frac{\mathbb{E}\left[(\Pi_{\mathbf{a}}^{f(\xi)} - \tilde{\Pi}_{i,\mathbf{a}}) \mathbf{1}_{[X_i = f(\xi)]} | H_i, \rho_i \right]}{\Pr[X_i = f(\xi) | H_i, \rho_i]} \tag{C16}$$

$$= \frac{\mathbb{E}\left[\sum_{\mathbf{a}} (\prod_{j=1}^{m} a_j^{b_j(\xi)}) (\Pi_{\mathbf{a}}^{f(\xi)} - \tilde{\Pi}_{i,\mathbf{a}}) \mathbf{1}_{[X_i = f(\xi)]} | H_i, \rho_i \right]}{\Pr[X_i = f(\xi) | H_i, \rho_i]}. \tag{C17}$$

Therefore, we find that

$$\left\| \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\Pi_{\mathbf{a}}^{f(\xi)} - \bar{\Pi}_{i,\mathbf{a}}^{f(\xi)} \right) \right\|_\infty \leqslant \frac{\mathbb{E}\left[\| \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) (\Pi_{\mathbf{a}}^{f(\xi)} - \tilde{\Pi}_{i,\mathbf{a}}) \|_\infty \mathbf{1}_{[X_i = f(\xi)]} | H_i, \rho_i \right]}{\Pr[X_i = f(\xi) | H_i, \rho_i]}. \tag{C18}$$

To established the desired equation (C14), it therefore remains to show that, almost surely,

$$\left\| \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) (\Pi_{\mathbf{a}}^{f(\xi)} - \tilde{\Pi}_{i,\mathbf{a}}) \right\|_\infty \mathbf{1}_{[X_i = f(\xi)]} \leqslant \gamma_2(\xi) \mathbf{1}_{[X_i = f(\xi)]}. \tag{C19}$$

We will show equation (C19) in the remainder of the proof. We expand $\tilde{\Pi}_{i,\mathbf{a}} - \Pi_{\mathbf{a}}^x$ using a telescoping sum of the form

$$C_1 C_2 C_3 - B_1 B_2 B_3 = (C_1 - B_1) B_2 B_3 + C_1 (C_2 - B_2) B_3 + C_1 C_2 (C_3 - B_3). \tag{C20}$$

In this case, with the tensor product and m terms, we get

$$\tilde{\Pi}_{i,\mathbf{a}} - \Pi_{\mathbf{a}}^x = \sum_{j=1}^{m} \left(\bigotimes_{k=1}^{j-1} \tilde{\Pi}_{i,a_k}^{(k)} \right) \otimes \left(\tilde{\Pi}_{i,a_j}^{(j)} - \Pi_{a_j}^{(j),x} \right) \otimes \left(\bigotimes_{k=j+1}^{m} \Pi_{a_k}^{(k),x} \right). \tag{C21}$$

Using this, and by distributing the product and sum in the below expression over the tensor factors, we find that

$$\sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) \left(\tilde{\Pi}_{i,\mathbf{a}} - \Pi_{\mathbf{a}}^{f(\xi)} \right) = \sum_{j=1}^{m} \left(\bigotimes_{k=1}^{j-1} \tilde{O}_{i,\xi}^{(k)} \right) \otimes \left(\tilde{O}_{i,\xi}^{(j)} - O_\xi^{(j)} \right) \otimes \left(\bigotimes_{k=j+1}^{m} O_\xi^{(k)} \right), \tag{C22}$$

where

$$\tilde{O}_{i,\xi}^{(j)} := \left(\sum_{a \in \Omega^{(j)}} a \tilde{\Pi}_{i,a}^{(j)} \right)^{b_j(\xi)} = \sum_{a \in \Omega^{(j)}} a^{b_j(\xi)} \tilde{\Pi}_{i,a}^{(j)}, \tag{C23}$$

by the fact that $b_j(\xi) \in \{0, 1\}$ and POVM elements sum to the identity, in analogy to their ideal counterparts $O_\xi^{(j)}$ given in equation (A6). Therefore, using the fact that the operator norm $\| \cdot \|_\infty$ is multiplicative with respect to tensor products (i.e. $\| C \otimes B \|_\infty = \| C \|_\infty \| B \|_\infty$), we find that

$$\left\| \sum_{\mathbf{a}} \left(\prod_{j=1}^{m} a_j^{b_j(\xi)} \right) (\tilde{\Pi}_{i,\mathbf{a}} - \Pi_{\mathbf{a}}^{f(\xi)}) \right\|_\infty \leqslant \sum_{j=1}^{m} \left(\prod_{k=1}^{j-1} \left\| \tilde{O}_{i,\xi}^{(k)} \right\|_\infty \right) \left\| \tilde{O}_{i,\xi}^{(j)} - O_\xi^{(j)} \right\|_\infty \left(\prod_{k=j+1}^{m} \left\| O_\xi^{(k)} \right\|_\infty \right). \tag{C24}$$

We continue by bounding $\| \tilde{O}_{i,\xi}^{(j)} - O_\xi^{(j)} \|_\infty$ and $\| \tilde{O}_{i,\xi}^{(j)} \|_\infty$. We start with bounding the first, by considering the two cases $b_j(\xi) = 0$ and $b_j(\xi) = 1$ separately. If $b_j(\xi) = 0$, then

$$\tilde{O}_{i,\xi}^{(j)} - O_\xi^{(j)} = \sum_{a \in \Omega^{(j)}} a^{b_j(\xi)} \left(\tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),f(\xi)} \right) = \sum_{a \in \Omega^{(j)}} \left(\tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),f(\xi)} \right) = I - I = 0, \tag{C25}$$

so then $\| \tilde{O}_{i,\xi}^{(j)} - O_\xi^{(j)} \|_\infty = 0$. On the other hand, if $b_j(\xi) = 1$, then, almost surely,

$$\left\| \tilde{O}_{i,\xi}^{(j)} - O_\xi^{(j)} \right\|_\infty \mathbf{1}_{[X_i = f(\xi)]} = \left\| \sum_{a \in \Omega^{(j)}} a (\tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),f(\xi)}) \right\|_\infty \mathbf{1}_{[X_i = f(\xi)]} \tag{C26}$$

$$\leqslant \sum_{a \in \Omega^{(j)}} |a| \left\| \tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),f(\xi)} \right\|_\infty \mathbf{1}_{[X_i = f(\xi)]} \tag{C27}$$

$$= \sum_{a \in \Omega_{f(\xi)}^{(j)}} |a| \left\| \tilde{\Pi}_{i,a}^{(j)} - \Pi_a^{(j),X_i} \right\|_\infty \mathbf{1}_{[X_i = f(\xi)]} \tag{C28}$$

$$\leqslant \delta_j \sum_{a \in \Omega_{f(\xi)}^{(j)}} |a| \mathbf{1}_{[X_i = f(\xi)]}. \tag{C29}$$

For the equality, note that we can assume that $X_i = f(\xi)$ by the indicator function; since then both POVMs have outcomes in $\Omega^{(j)}_{f(\xi)}$ we can also restrict the summation (cf equation (B6)). The last step holds by equation (B5) in assumption (III). Both cases are neatly summarized by

$$\left\|\tilde{O}^{(j)}_{i,\xi} - O^{(j)}_{\xi}\right\|_\infty \mathbf{1}_{[X_i=f(\xi)]} \leqslant \epsilon^{(j)}_{\xi} \mathbf{1}_{[X_i=f(\xi)]}, \tag{C30}$$

with $\epsilon^{(j)}_{\xi}$ defined as in equation (C7). By the triangle inequality, this in turn implies that

$$\left\|\tilde{O}^{(j)}_{i,\xi}\right\|_\infty \mathbf{1}_{[X_i=f(\xi)]} \leqslant \left\|O^{(j)}_{\xi}\right\|_\infty \mathbf{1}_{[X_i=f(\xi)]} + \left\|\tilde{O}^{(j)}_{i,\xi} - O^{(j)}_{\xi}\right\|_\infty \mathbf{1}_{[X_i=f(\xi)]} \leqslant \left(\lambda^{(j)}_{\xi} + \epsilon^{(j)}_{\xi}\right) \mathbf{1}_{[X_i=f(\xi)]} \tag{C31}$$

with $\lambda^{(j)}_{\xi} = \| O^{(j)}_{\xi} \|_\infty$ defined as in equation (C7). Thus, plugging equations (C30) and (C31) into (C24), we obtain the desired bound equation (C19), completing the proof. □

As a consequence of this theorem, we obtain two corollaries that related the expected score $\mathbb{E}[S_i|H_i, \rho_i]$ to the witness value $\mathrm{Tr}[W\rho_i]$. These corollaries are used in theorems 2 and 3. We state and proof both below.

Corollary 1.1. *Let W be a Hermitian operator of the form of equation (A7) (not necessarily a witness in the sense of equation (A1)). Suppose that the experiment is modeled by the model assumptions (I) to (III) in appendix B. Let the score function be defined as in equation (A8). Then, in every round i,*

$$|\mathrm{Tr}[W\rho_i] - (c - \mathbb{E}[S_i|H_i, \rho_i])| \leqslant \gamma \tag{C32}$$

holds almost surely, where c and γ are defined in equations (A7) and (C5).

Proof. We first compute

$$\begin{aligned}
\mathbb{E}\left[S_i|H_i, \rho_i, X_i, \{\tilde{\Pi}^{(j)}_{i,a}\}\right] &= \mathbb{E}\left[s(X_i, \mathbf{A}_i)|H_i, \rho_i, X_i, \{\tilde{\Pi}^{(j)}_{i,a}\}\right] && \text{(C33)}\\
&= \sum_{\mathbf{a}} \mathbb{E}\left[s(X_i, \mathbf{a})\mathbf{1}_{[\mathbf{A}_i=\mathbf{a}]}|H_i, \rho_i, X_i, \{\tilde{\Pi}^{(j)}_{i,a}\}\right] && \text{(C34)}\\
&= \sum_{\mathbf{a}} s(X_i, \mathbf{a}) \mathbb{E}\left[\mathbf{1}_{[\mathbf{A}_i=\mathbf{a}]}|H_i, \rho_i, X_i, \{\tilde{\Pi}^{(j)}_{i,a}\}\right] && \text{(C35)}\\
&= \sum_{\mathbf{a}} s(X_i, \mathbf{a}) \Pr\left[\mathbf{A}_i = \mathbf{a} \,\middle|\, H_i, \rho_i, X_i, \{\tilde{\Pi}^{(j)}_{i,a}\}\right] && \text{(C36)}\\
&= \sum_{\mathbf{a}} s(X_i, \mathbf{a}) \mathrm{Tr}[\rho_i \tilde{\Pi}_{i,\mathbf{a}}], && \text{(C37)}
\end{aligned}$$

using Born's rule (B7) in the last equation. This implies that

$$\begin{aligned}
\mathbb{E}\left[S_i|H_i, \rho_i\right] &= \sum_{\mathbf{a}} \mathbb{E}\left[s(X_i, \mathbf{a})\mathrm{Tr}[\rho_i \tilde{\Pi}_{i,\mathbf{a}}]\,|H_i, \rho_i\right] && \text{(C38)}\\
&= \sum_{x,\mathbf{a}} s(x, \mathbf{a}) \mathbb{E}\left[\mathbf{1}_{[\mathbf{X}_i=x]}\mathrm{Tr}[\rho_i \tilde{\Pi}_{i,\mathbf{a}}]\,|H_i, \rho_i\right] && \text{(C39)}\\
&= \mathrm{Tr}\left[\rho_i \sum_{x,\mathbf{a}} s(x, \mathbf{a}) \mathbb{E}\left[\mathbf{1}_{[\mathbf{X}_i=x]}\tilde{\Pi}_{i,\mathbf{a}}\,|H_i, \rho_i\right]\right] && \text{(C40)}\\
&= c - \mathrm{Tr}[\bar{W}_i \rho_i], && \text{(C41)}
\end{aligned}$$

where in equation (C40) we pull can pull ρ_i out of the conditional expectation value since it is conditioned on ρ_i and in equation (C41) we use the definition of $\bar{W}_i$ (given in equations (C2) and (C3)). From this, it follows that

$$|\mathrm{Tr}[W\rho_i] - (c - \mathbb{E}[S_i|H_i, \rho_i])| = |\mathrm{Tr}[(W - \bar{W}_i)\rho_i]| \leqslant \| W - \bar{W}_i \|_\infty \leqslant \gamma, \tag{C42}$$

where the last inequality follows from the theorem, equation (C4). □

Corollary 1.2. *Let W be an operator of the form of equation (A7) that is a witness for the set $\mathcal{S}$ (i.e., W satisfies equation (A1)). Suppose that the experiment is modeled by the model assumptions (I) to (III) in appendix B. Furthermore, assume that the Hypothesis ($\mathcal{H}_0$) holds with respect to this $\mathcal{S}$. Then, almost surely, $\mathbb{E}[S_i|H_i] \leqslant c + \gamma$ for all rounds $i = 1, \ldots, n$, where c and γ are defined in equations (A7) and (C5).*

Proof. From corollary 1.1, we directly obtain the one-sided inequality $\mathbb{E}[S_i|H_i, \rho_i] \leqslant \gamma + c - \mathrm{Tr}[\rho_i W]$. Now, hypothesis ($\mathcal{H}_0$) and equation (A1) imply that $\mathrm{Tr}[\rho_i W] \geqslant 0$ holds almost surely, so that $\mathbb{E}[S_i|H_i, \rho_i] \leqslant \gamma + c - \mathrm{Tr}[\rho_i W] \leqslant c + \gamma$. Thus, $\mathbb{E}[S_i|H_i] = \mathbb{E}[\mathbb{E}[S_i|H_i, \rho_i]|H_i] \leqslant \mathbb{E}[c + \gamma|H_i] = c + \gamma$. □

Appendix D. Theorem 2—bounding the *p*-value

The goal of this section is to prove theorem 2. In this theorem, we derive a bound on the p-value. The main ingredient for the proof is a concentration inequality that bounds the tail probabilities of (super)martingales. Here, we use a concentration inequality due to Bentkus [30, 31], which was used in reference [32] for the analysis of Bell violation experiments. We state Bentkus' inequality in the form of theorem 1 in reference [32].

Lemma 1 (Bentkus' inequality [30, 31]). *Let $R_0, R_1, \ldots, R_n$ be a supermartingale (with respect to an arbitrary filtration) with $R_0 = 0$ and differences bounded as $-\beta_i \leqslant R_i - R_{i-1} \leqslant 1 - \beta_i$ almost surely for constants $\beta_i \in [0, 1]$. Let $\beta := (\beta_1 + \cdots + \beta_n)/n$. Then, for all $x \in \mathbb{R}$,*

$$\Pr[R_n \geqslant x - n\beta] \leqslant eF^{\circ}_{n,\beta}(x), \tag{D1}$$

where

$$F^{\circ}_{n,\beta}(x) = F_{n,\beta}(\lfloor x \rfloor)^{1-(x-\lfloor x \rfloor)} F_{n,\beta}(\lfloor x \rfloor + 1)^{x-\lfloor x \rfloor} \tag{D2}$$

is the log-linear interpolation of the survival function of a binomial distribution with parameters n and β,

$$F_{n,\beta}(k) = \Pr[X \geqslant k | X \sim \mathrm{Binom}(n, \beta)] \tag{D3}$$

$$= \sum_{l=k}^{n} \binom{n}{l} \beta^l (1-\beta)^{n-l}. \tag{D4}$$

In equation (D2), we define $0^0 := 1$.

With this main ingredient in hand, we now state and prove a bound on the p-value $p := \Pr[T_n \geqslant t_n|\mathcal{H}_0]$, where

$$T_n := \sum_{i=1}^{n} \frac{S_i - s_{\min}}{\Delta s} \tag{D5}$$

is the total normalized score after n rounds (see equations (A8) and (B8) for the definition of the score). We recall that $s_{\min}$ and $s_{\max}$ are defined in equation (A10) as the minimum and maximum value, respectively, that can be attained by the score function.

Theorem 2 (restated). *Let W be an operator of the form of equation (A7) that is a witness for the set $\mathcal{S}$ (i.e., which satisfies equation (A1)). Assume that the experiment is modeled by the model assumptions (I) to (III) in appendix B and consider the null hypothesis ($\mathcal{H}_0$) with respect to $\mathcal{S}$. Then, for all $t_n \in \mathbb{R}$,*

$$\Pr[T_n \geqslant t_n|\mathcal{H}_0] \leqslant eF^{\circ}_{n,\beta}(t_n), \quad \text{where } \beta := \min\left(1, \frac{c + \gamma - s_{\min}}{\Delta s}\right), \tag{D6}$$

and where c, γ, and $F^{\circ}_{n,\beta}$ are defined in equations (A7), (C5), and (D2), respectively. □

Proof. We first verify that $\beta \geqslant 0$, so that $F^{\circ}_{n,\beta}$ is well-defined and we can later apply lemma 1. Now, hypothesis ($\mathcal{H}_0$) and equation (A1) imply that $\mathrm{Tr}[\rho_i W] \geqslant 0$. Therefore by the decomposition of W from equation (A9), we find that

$$0 \leqslant \mathrm{Tr}[W\rho_i] = c - \sum_{x,\mathbf{a}} p_x \mathrm{Tr}\left[\rho_i \Pi^x_{\mathbf{a}}\right] s(x, \mathbf{a}) \leqslant c - s_{\min}. \tag{D7}$$

Since $\gamma \geqslant 0$, it follows that indeed $\beta \geqslant 0$.

Next, we verify that the bound in equation (D6) holds for $\beta = 1$. Indeed, $\Pr[T_n \geqslant t_n] > 0$ only if $t_n \leqslant n$, since $T_n \in [0, n]$. But in this case, $F^{\circ}_{n,1}(t_n) = 1$, so the bound reads $\Pr[T_n \geqslant t] \leqslant e$, which holds trivially.

Now assume that $\beta \in [0, 1)$, so that

$$\beta = \frac{c + \gamma - s_{\min}}{\Delta s}. \tag{D8}$$

The proof comes down to constructing a suitable supermartingale and applying lemma 1. For $i = 0, 1, \ldots, n$, define

$$R_i := \sum_{l=1}^{i} \frac{S_l - c - \gamma}{\Delta s}. \tag{D9}$$

By convention of the empty sum, this means $R_0 = 0$. Note that R_i is affinely related to T_i, the total normalized score after round i.

We first show that $R_0, R_1, \ldots, R_n$ is a supermartingale with respect to $H_1, \ldots, H_n$, i.e.,

$$\mathbb{E}[R_i | H_1, \ldots, H_i] \leqslant R_{i-1} \quad \forall i = 1, \ldots, n. \tag{D10}$$

Note that, for any $i = 1, \ldots, n$, we have the recursion formula

$$R_i = R_{i-1} + \frac{S_i - c - \gamma}{\Delta s}. \tag{D11}$$

In view of equations (B8) and (D9), R_{i-1} depends only on the measurement settings and outcomes of the first $i - 1$ rounds. Thus, $\mathbb{E}[R_{i-1} | H_1, \ldots, H_i] = R_{i-1}$ by assumption (I)b. Moreover, $\mathbb{E}[S_i | H_1, \ldots, H_i] = \mathbb{E}[S_i | H_i]$ by assumption (I)a. Therefore,

$$\mathbb{E}[R_i | H_1, \ldots, H_i] = R_{i-1} + \frac{\mathbb{E}[S_i | H_i] - c - \gamma}{\Delta s} \leqslant R_{i-1}. \tag{D12}$$

where the last inequality holds by corollary 1.2, which asserts that $\mathbb{E}[S_i | H_i] \leqslant c + \gamma$ almost surely. Thus, $R_0, R_1, \ldots, R_n$ is a supermarginale as claimed.

Next we bound the differences. Since $s_{\min} \leqslant S_i \leqslant s_{\max}$, we find using equation (D11) that

$$\frac{s_{\min} - c - \gamma}{\Delta s} \leqslant R_i - R_{i-1} = \frac{S_i - c - \gamma}{\Delta s} \leqslant \frac{s_{\max} - c - \gamma}{\Delta s}. \tag{D13}$$

Thus, it holds that $-\beta \leqslant R_i - R_{i-1} \leqslant 1 - \beta$ since β is given by equation (D8). Indeed,

$$-\beta = \frac{s_{\min} - c - \gamma}{\Delta s}, \quad 1 - \beta = \frac{s_{\max} - c - \gamma}{\Delta s}. \tag{D14}$$

Thus we can apply lemma 1, which gives

$$\Pr[R_n \geqslant x - n\beta] \leqslant e F^{\circ}_{n,\beta}(x) \quad \forall x \in \mathbb{R}. \tag{D15}$$

To complete the proof, we need to relate T_n to R_n and evaluate at the appropriate value of x. By comparing equations (D5) and (D9), we find that

$$R_n = \sum_{l=1}^{n} \frac{S_l - s_{\min} - (c + \gamma - s_{\min})}{\Delta s} = T_n - n\beta. \tag{D16}$$

If we combine this with equation (D15), we arrive at

$$\Pr[T_n \geqslant t_n | \mathcal{H}_0] = \Pr[R_n + n\beta \geqslant t_n | \mathcal{H}_0] = \Pr[R_n \geqslant t_n - n\beta | \mathcal{H}_0] \leqslant e F^{\circ}_{n,\beta}(t_n). \tag{D17}$$

This completes the proof. □

Appendix E. Theorem 3—confidence intervals for the witness estimate

In this appendix we will prove theorem 3 in the main text. This theorem establishes a $(1-2\alpha)$ two-sided and a $(1-\alpha)$ one-sided confidence interval around an estimate $\hat{w}_n$ of the average witness value $\langle W\rangle_n$. Recall (from equation (21)) that the witness estimate is computed from the score as

$$\widehat{W}_n := c - \frac{1}{n}\sum_{i=1}^{n} S_i. \tag{E1}$$

It is a point estimate of the average realized witness value $\langle W\rangle_n$, which is defined as (recall equation (2))

$$\langle W\rangle_n := \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}[\rho_i W]. \tag{E2}$$

We start with a lemma that establishes the some properties of the function $F^{\circ}_{n,\frac{1}{2}}$ from Bentkus' inequality, so that it has a well-defined inverse.

Lemma 2. *Let $F^{\circ}_{n,\frac{1}{2}}(x)$ be defined as in equation (D2) for any $n \in \mathbb{N}$ (with $\beta = \frac{1}{2}$). Then for any $n \in \mathbb{N}$ the function $x \mapsto F^{\circ}_{n,\frac{1}{2}}(x)$ has the following properties:*

(a) $F^{\circ}_{n,\frac{1}{2}}(x)$ is a strictly decreasing and continuous in x on the interval $[\frac{n}{2}, n]$; and
(b) $F^{\circ}_{n,\frac{1}{2}}([\frac{n}{2}, n]) \supseteq [\frac{1}{2^n}, \frac{1}{2}]$.

Hence, for any $y \in [\frac{1}{2^n}, \frac{1}{2}]$, there is a unique $x \in [\frac{n}{2}, n]$ such that $F^{\circ}_{n,\frac{1}{2}}(x) = y$.

Proof. Recall from equation (D2) that $F^{\circ}_{n,\frac{1}{2}}$ interpolates the survival function $F_{n,\frac{1}{2}}(k) = \Pr[X \geqslant k]$ of a binomial random variables $X \sim \mathrm{Binom}(n, \frac{1}{2})$ at non-integer points by the log-linear function

$$F^{\circ}_{n,\frac{1}{2}}(x) = F_{n,\frac{1}{2}}(\lfloor x\rfloor)^{1-(x-\lfloor x\rfloor)} F_{n,\frac{1}{2}}(\lfloor x\rfloor + 1)^{x-\lfloor x\rfloor}. \tag{E3}$$

(*a*) Since $F_{n,\frac{1}{2}}(k)$ is strictly decreasing in k and strictly positive for all $k = 0, \ldots, n$, and since the logarithm is strictly monotonic and continuous, it is clear that $F^{\circ}_{n,\frac{1}{2}}$ is strictly decreasing and continuous in x for all $x \in [\frac{n}{2}, n]$.

(*b*) By the convention that $0^0 = 1$ in the log-linear interpolation, it follows that $F^{\circ}_{n,\frac{1}{2}}(n) = F_{n,\frac{1}{2}}(n) = \frac{1}{2^n}$. Now we show that $F^{\circ}_{n,\frac{1}{2}}(\frac{n}{2}) \geqslant \frac{1}{2}$. To do so, we use the symmetry of the binomial distribution with parameter half. We observe that

$$F_{n,\frac{1}{2}}(k) = \sum_{i=k}^{n} \binom{n}{i} \frac{1}{2^n} = \sum_{i=k}^{n} \binom{n}{n-i} \frac{1}{2^n} = \sum_{i=0}^{n-k} \binom{n}{i} \frac{1}{2^n} = 1 - F_{n,\frac{1}{2}}(n-k+1). \tag{E4}$$

So, for even n, we find that

$$2F^{\circ}_{n,\frac{1}{2}}(\frac{n}{2}) = 2F_{n,\frac{1}{2}}(\frac{n}{2}) = F_{n,\frac{1}{2}}(\frac{n}{2}) + 1 - F_{n,\frac{1}{2}}(\frac{n}{2}+1) \geqslant 1 \tag{E5}$$

since $F_{n,\frac{1}{2}}(\frac{n}{2}+1) > F_{n,\frac{1}{2}}(\frac{n}{2})$. And for odd n is odd, then we find [by property (a)]

$$2F^{\circ}_{n,\frac{1}{2}}(\frac{n}{2}) \geqslant 2F^{\circ}_{n,\frac{1}{2}}(\frac{n+1}{2}) = F_{n,\frac{1}{2}}(\frac{n+1}{2}) + 1 - F_{n,\frac{1}{2}}(n - \frac{n+1}{2} + 1) = 1. \tag{E6}$$

Hence in either case $F^{\circ}_{n,\frac{1}{2}}(\frac{n}{2}) \geqslant \frac{1}{2}$. By property (a) and $F^{\circ}_{n,\frac{1}{2}}(n) = \frac{1}{2^n}$ statement (b) follows. □

With this lemma in hand, we can state and proof the theorem. The main ingredient in the proof is again lemma 1 applied to a suitably chosen martingale.

Theorem 3 (restated). *Let W be a Hermitian operator of the form of equation (A7) (not necessarily a witness in the sense of equation (A1)). Suppose that the experiment is modeled by the model assumptions (I) to (III) in appendix B. Let $\hat{W}_n$ denote the average witness estimate as defined in equation (E1). Fix the significance level $\alpha \in [0,1]$. If $\alpha < \frac{e}{2^n}$, define $\varepsilon = \Delta s$, otherwise define $\varepsilon \in [\gamma, \gamma + \Delta s]$ as the unique solution to*

$$\alpha = eF^{\circ}_{n,\frac{1}{2}}\left(\frac{n}{2}(1+\frac{\varepsilon-\gamma}{\Delta s})\right). \tag{E7}$$

Here γ and $F^{\circ}_{n,\beta}$ are defined in equations (C5) and (D2), respectively (with $\beta = \frac{1}{2}$ here and $e \approx 2.72$). Then,

(a) $\Pr[|\langle W\rangle_n - \hat{W}_n| \leqslant \Delta s] = 1$;
(b) $\Pr[\langle W\rangle_n - \hat{W}_n \leqslant \varepsilon] \geqslant 1-\alpha$ *and* $\Pr[\langle W\rangle_n - \hat{W}_n \geqslant -\varepsilon] \geqslant 1-\alpha$;
(c) $\Pr[|\langle W\rangle_n - \hat{W}_n| \leqslant \varepsilon] \geqslant 1-2\alpha$.

That is, if $\hat{w}_n$ is the average witness estimate after n rounds, then $\mathcal{I}(\hat{w}_n) := [\hat{w}_n - \varepsilon, \hat{w}_n + \varepsilon]$ is a $(1-2\alpha)$ two-sided confidence interval and $\mathcal{J}(\hat{w}_n) := [\hat{w}_n - \Delta s, \hat{w}_n + \epsilon]$ is a $(1-\alpha)$ one-sided confidence interval for the average witness value $\langle W\rangle_n$ as defined in equation (E2). □

Proof.

(*a*) We show that $-\Delta s \leqslant \langle W\rangle_n - \hat{W}_n \leqslant \Delta s$ holds almost surely by

$$\langle W\rangle_n - \widehat{W}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{Tr}[\rho_i W] - c + S_i) \tag{E8}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(S_i - \sum_x\sum_{\mathbf{a}} p_x \mathrm{Tr}\left[\bigotimes_{j=1}^{m}\Pi_{a_j}^{(j),x}\rho_i\right] s(x,\mathbf{a})\right) \tag{E9}$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}\left(s_{\max} - \sum_x\sum_{\mathbf{a}} p_x \mathrm{Tr}\left[\bigotimes_{j=1}^{m}\Pi_{a_j}^{(j),x}\rho_i\right] s_{\min}\right) = \Delta s, \tag{E10}$$

since p_x, $\mathrm{Tr}\left[\otimes_{j=1}^{m}\Pi_{a_j}^{(j),x}\rho_i\right] \geqslant 0$ and sum to one. Similarly $\langle W\rangle_n - \hat{W}_n \geqslant -\Delta s$, so that $\Pr[|\langle W\rangle_n - \hat{W}_n| \leqslant \Delta s] = 1$.

(*b*) For $\alpha < \frac{e}{2^n}$, both statements in (b) follow immediately from (a), since then $\varepsilon = \Delta s$. So from now on, assume that $\alpha \in [\frac{e}{2^n}, 1]$. First we show that ε is well-defined. This follows from lemma 2, which states that for all $y = \frac{\alpha}{e} \in [\frac{1}{2^n}, \frac{1}{e}] \subset [\frac{1}{2^n}, \frac{1}{2}]$, there exists a unique $x = \frac{n}{2}(1+\frac{\varepsilon-\gamma}{\Delta s}) \in [\frac{n}{2}, n]$ such that $F^{\circ}_{n,\frac{1}{2}}(\frac{n}{2}(1+\frac{\varepsilon-\gamma}{\Delta s})) = \frac{\alpha}{e}$. Hence for all $\alpha \in [\frac{e}{2^n}, 1]$ there is a unique $\varepsilon \in [\gamma, \gamma+\Delta s]$ such that equation (E7) holds.

Now, we construct a suitable martingale sequence Z_i so that we can apply lemma 1 to Z_i and $-Z_i$ (to get both statements). For $i = 0, \ldots, n$ we define

$$Z_i := \frac{1}{2}\sum_{l=1}^{i}\frac{S_l - \mathbb{E}[S_l|H_l,\rho_l]}{\Delta s} \tag{E11}$$

This is a martingale sequence with respect to the sequence $(H_1,\rho_1),\ldots,(H_n,\rho_n)$, since

$$\mathbb{E}[Z_i|H_1,\rho_1,\ldots,H_i,\rho_i] = \frac{1}{2\Delta s}\sum_{l=1}^{i}\mathbb{E}\left[S_l - \mathbb{E}[S_l|H_l,\rho_l]\,|H_1,\rho_1,\ldots,H_i,\rho_i\right] \tag{E12}$$

$$= \frac{1}{2\Delta s}\sum_{l=1}^{i}\left(\mathbb{E}\left[S_l\,|H_1,\rho_1,\ldots,H_i,\rho_i\right] - \mathbb{E}[S_l|H_l,\rho_l]\right) \tag{E13}$$

$$= \frac{1}{2\Delta s}\left(\mathbb{E}\left[S_i\,|H_1,\rho_1,\ldots,H_i,\rho_i\right] - \mathbb{E}[S_i|H_i,\rho_i]\right) \tag{E14}$$

$$+ \frac{1}{2\Delta s}\sum_{l=1}^{i-1}\left(\mathbb{E}\left[S_l\,|H_1,\rho_1,\ldots,H_i,\rho_i\right] - \mathbb{E}[S_l|H_l,\rho_l]\right) \tag{E15}$$

$$= \frac{1}{2\Delta s}\left(\mathbb{E}[S_i|H_i,\rho_i] - \mathbb{E}[S_i|H_i,\rho_i]\right) + \frac{1}{2\Delta s}\sum_{l=1}^{i-1}\left(S_l - \mathbb{E}[S_l|H_l,\rho_l]\right) \tag{E16}$$

$$= 0 + Z_{i-1}. \tag{E17}$$

In equation (E13) we used that $\mathbb{E}[\mathbb{E}[S_l|H_l,\rho_l]|H_1,\rho_1,\ldots,H_i,\rho_i] = \mathbb{E}[S_l|H_l,\rho_l]$ for all $l = 1,\ldots,i$. Equation (E16) holds since $\mathbb{E}[S_i|H_1,\rho_1,\ldots,H_i,\rho_i] = \mathbb{E}[S_i|H_i,\rho_i]$ and $\mathbb{E}[S_l|H_1,\rho_1,\ldots,H_i,\rho_i] = S_l$ for all $l = 1,\ldots,i-1$ by assumptions (I)a and (I)b. Moreover, Z_i is bounded difference, where the difference is bounded by

$$|Z_i - Z_{i-1}| = \frac{1}{2}\frac{|S_i - \mathbb{E}[S_i|H_i,\rho_i]|}{\Delta s} \leqslant \frac{1}{2}\frac{\Delta s}{\Delta s} = \frac{1}{2}, \tag{E18}$$

since $s_{\min} \leqslant S_i \leqslant s_{\max}$ for all $i = 1,\ldots,n$. Hence lemma 1 applies with $\beta = \frac{1}{2}$ (to both Z_i and $-Z_i$), which yields

$$\Pr[Z_n \geqslant x - \frac{n}{2}] \leqslant eF^{\circ}_{n,\frac{1}{2}}(x) \quad \text{and} \quad \Pr[-Z_n \geqslant x - \frac{n}{2}] \leqslant eF^{\circ}_{n,\frac{1}{2}}(x). \tag{E19}$$

Next, we invoke corollary 1.1 to relate $\langle W\rangle_n - \hat{W}_n$ to Z_n. We find that

$$\left|\langle W\rangle_n - \widehat{W}_n - \frac{2\Delta s}{n}Z_n\right| = \left|\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}[W\rho_i]\right) - \left(c - \frac{1}{n}\sum_{i=1}^{n}S_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n}S_i - \mathbb{E}[S_i|H_i,\rho_i]\right)\right| \tag{E20}$$

$$= \left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}[W\rho_i] - (c - \mathbb{E}[S_i|H_i,\rho_i])\right| \tag{E21}$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}|\mathrm{Tr}[W\rho_i] - (c - \mathbb{E}[S_i|H_i,\rho_i])| \leqslant \gamma \tag{E22}$$

almost surely. In equation (E22), we applied corollary 1.1 for each $i = 1,\ldots,n$. Hence by using equations (E19) and (E22), we find that

$$\Pr[\langle W\rangle_n - \widehat{W}_n \geqslant \varepsilon] \leqslant \Pr[\gamma + \frac{2\Delta s}{n}Z_n \geqslant \varepsilon] \tag{E23}$$

$$= \Pr[Z_n \geqslant \frac{n(\varepsilon-\gamma)}{2\Delta s}] \tag{E24}$$

$$\leqslant eF^{\circ}_{n,\frac{1}{2}}\left(\frac{n}{2}(1 + \frac{\varepsilon-\gamma}{\Delta s})\right) = \alpha, \tag{E25}$$

by evaluating equation (E19) at the appropriate value of x and using the implicit definition of ε in equation (E7). Similarly,

$$\Pr[\widehat{W}_n - \langle W\rangle_n \geqslant \varepsilon] \leqslant \Pr[\gamma - \frac{2\Delta s}{n}Z_n \geqslant \varepsilon] \tag{E26}$$

$$= \Pr[-Z_n \geqslant \frac{n(\varepsilon-\gamma)}{2\Delta s}] \tag{E27}$$

$$\leqslant eF^{\circ}_{n,\frac{1}{2}}\left(\frac{n}{2}(1 + \frac{\varepsilon-\gamma}{\Delta s})\right) = \alpha. \tag{E28}$$

Equations (E25) and (E28) imply the one-sided intervals

$$\Pr[\langle W\rangle_n - \widehat{W}_n \leqslant \varepsilon] \geqslant 1-\alpha, \quad \text{and} \quad \Pr[\langle W\rangle_n - \widehat{W}_n \geqslant -\varepsilon] \geqslant 1-\alpha, \tag{E29}$$

respectively, as claimed.

(*c*) Combining equations (E25) and (E28) with the union bound yields

$$\Pr[|\langle W\rangle_n - \widehat{W}_n| \geqslant \varepsilon] \leqslant \Pr[\langle W\rangle_n - \widehat{W}_n \geqslant \varepsilon] + \Pr[\widehat{W}_n - \langle W\rangle_n \geqslant \varepsilon] \leqslant 2\alpha. \tag{E30}$$

This implies that

$$\Pr[|\langle W\rangle_n - \widehat{W}_n| \leqslant \varepsilon] \geqslant 1 - 2\alpha, \tag{E31}$$

as claimed. □

ORCID iDs

Bas Dirkse https://orcid.org/0000-0002-4391-6962

References

[1] Horodecki R, Horodecki P, Horodecki M and Horodecki K 2009 *Rev. Mod. Phys.* **81** 865
[2] Vidal G 2003 *Phys. Rev. Lett.* **91** 147902
[3] Nielsen M A and Chuang I L 2010 *Quantum Computation and Quantum Information* 10th edn (Cambridge: Cambridge University Press) https://doi.org/10.1017/CBO9780511976667
[4] Shenoy-Hejamadi A, Pathak A and Radhakrishna S 2017 *Quanta* **6** 1
[5] Broadbent A and Schaffner C 2016 *Des. Codes, Cryptogr.* **78** 351
[6] Pirandola S *et al* 2019 arXiv:1906.01645
[7] Ben-Or M and Hassidim A 2005 *Proc. ACM STOC* (2005) pp 481–5
[8] Tani S, Kobayashi H and Matsumoto K 2012 *ACM Trans. Comput. Theory* **4** 1
[9] Jozsa R, Abrams D S, Dowling J P and Williams C P 2000 *Phys. Rev. Lett.* **85** 2010
[10] Bollinger J J, Itano W M, Wineland D J and Heinzen D J 1996 *Phys. Rev.* A **54** R4649
[11] Gottesman D, Jennewein T and Croke S 2012 *Phys. Rev. Lett.* **109** 070503
[12] Kómár P, Kessler E M, Bishof M, Jiang L, Sørensen A S, Ye J and Lukin M D 2014 *Nat. Phys.* **10** 582
[13] Bourennane M, Eibl M, Kurtsiefer C, Gaertner S, Weinfurter H, Gühne O, Hyllus P, Bruß D, Lewenstein M and Sanpera A 2004 *Phys. Rev. Lett.* **92** 087902
[14] Filgueiras J G, Maciel T O, Auccaise R E, Vianna R O, Sarthour R S and Oliveira I S 2012 *Quantum Inf. Process.* **11** 1883
[15] Song C *et al* 2017 *Phys. Rev. Lett.* **119** 180511
[16] Friis N *et al* 2018 *Phys. Rev.* X **8** 021012
[17] Zhong H S *et al* 2018 *Phys. Rev. Lett.* **121** 250505
[18] Wang X L *et al* 2018 *Phys. Rev. Lett.* **120** 260502
[19] Humphreys P C, Kalb N, Morits J P, Schouten R N, Vermeulen R F, Twitchen D J, Markham M and Hanson R 2018 *Nature* **558** 268
[20] Gühne O and Tóth G 2009 *Phys. Rep.* **474** 1
[21] Jungnitsch B, Moroder T and Gühne O 2011 *Phys. Rev.* A **84** 032310
[22] Tóth G, Wieczorek W, Krischek R, Kiesel N, Michelberger P and Weinfurter H 2009 *New J. Phys.* **11** 083002
[23] Ryu S and Lee S-s B 2012 *Phys. Rev.* A **86** 042324
[24] Dai J, Len Y L, Teo Y S, Englert B G and Krivitsky L A 2014 *Phys. Rev. Lett.* **113** 170402
[25] Zhang Y, Glancy S and Knill E 2011 *Phys. Rev.* A **84** 062118
[26] Rosset D, Ferretti-Schöbitz R, Bancal J D, Gisin N and Liang Y C 2012 *Phys. Rev.* A **86** 062325
[27] Šupić I, Coladangelo A, Augusiak R and Acin A 2018 *New J. Phys.* **20** 083041
[28] Šupić I and Bowles J 2019 arXiv:1904.10042
[29] Chen X, Hu X and Zhou D L 2017 *Phys. Rev.* A **95** 052326
[30] Bentkus V 2004 *Ann. Probab.* **32** 1650
[31] Bentkus V, Kalosha N and van Zuijlen M 2006 *Lith. Math. J.* **46** 1
[32] Elkouss D and Wehner S 2016 *npj Quantum Inf* **2** 16026
[33] Ruan L, Dirkse B and Wehner S 2020 In preparation
[34] Takeuchi Y, Mantri A, Morimae T, Mizutani A and Fitzsimons J F 2019 *npj Quantum Inf* **5** 1
[35] Walter M and Renes J M 2014 *IEEE Trans. Inf. Theory* **60** 8007
[36] Faist P and Renner R 2016 *Phys. Rev. Lett.* **117** 010404
[37] Branciard C, Rosset D, Liang Y-C and Gisin N 2013 *Phys. Rev. Lett.* **110** 060405
[38] Nawareg M, Muhammad S, Amselem E and Bourennane M 2015 *Sci. Rep.* **5** 1
[39] Skwara P, Kampermann H, Kleinmann M and Bruß D 2007 *Phys. Rev.* A **76** 012312
[40] Awschalom D D, Hanson R, Wrachtrup J and Zhou B B 2018 *Nat. Photonics* **12** 516
[41] Bultink C C *et al* 2016 *Phys. Rev. Appl.* **6** 034008
[42] Walter M, Gross D and Eisert J 2016 arXiv:1612.02437
[43] Zhou Y, Zhao Q, Yuan X and Ma X 2019 *npj Quantum Inf.* **5** 83
[44] Peres A 1996 *Phys. Rev. Lett.* **77** 1413
[45] Horodecki M, Horodecki P and Horodecki R 1996 *Phys. Lett.* A **223** 1
[46] Loughin T M 2004 *Comput. Stat. Data Anal.* **47** 467
[47] Zhou Y 2020 *Phys. Rev.* A **101** 012301