



PAPER • OPEN ACCESS

## Memory cost of temporal correlations

To cite this article: Costantino Budroni *et al* 2019 *New J. Phys.* **21** 093018

View the [article online](#) for updates and enhancements.

### You may also like

- [Optimizing the optical and electrical properties of graphene ink thin films by laser-annealing](#)  
Sepideh Khandan Del, Rainer Bornemann, Andreas Bablich et al.
- [Load identification approach based on basis pursuit denoising algorithm](#)  
D Ginsberg, M Ruby and C P Fritzen
- [Electro-magnetoencephalography for a spherical multiple-shell model: novel integral operators with singular-value decompositions](#)  
S Leweke, V Michel and A S Fokas



## PAPER

# Memory cost of temporal correlations

## OPEN ACCESS

RECEIVED  
24 June 2019ACCEPTED FOR PUBLICATION  
19 August 2019PUBLISHED  
10 September 2019

Original content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Costantino Budroni<sup>1,2,5</sup>, Gabriel Fagundes<sup>3</sup> and Matthias Kleinmann<sup>4</sup><sup>1</sup> Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria<sup>2</sup> Faculty of Physics, University of Vienna, Boltzmanngasse 5, A-1090 Vienna, Austria<sup>3</sup> Departamento de Física, Universidade Federal de Minas Gerais UFMG, PO Box 702, 30123-970, Belo Horizonte, MG, Brazil<sup>4</sup> Naturwissenschaftlich-Technische Fakultät, Universität Siegen, Walter-Flex-Straße 3, D-57068 Siegen, Germany<sup>5</sup> Author to whom any correspondence should be addressed.E-mail: [costantino.budroni@oeaw.ac.at](mailto:costantino.budroni@oeaw.ac.at), [gabrielf@fisica.ufmg.br](mailto:gabrielf@fisica.ufmg.br) and [matthias.kleinmann@uni-siegen.de](mailto:matthias.kleinmann@uni-siegen.de)**Keywords:** temporal correlations, Leggett–Garg inequalities, finite-state machines, quantum contextuality

## Abstract

A possible notion of nonclassicality for single systems can be defined on the basis of the notion of memory cost of classically simulating probabilities observed in a temporal sequence of measurements. We further explore this idea in a theory-independent framework, namely, from the perspective of general probability theories (GPTs), which includes classical and quantum theory as special examples. Under the assumption that each system has a finite memory capacity, identified with the maximal number of states perfectly distinguishable with a single measurement, we investigate what are the temporal correlations achievable with different theories, namely, classical, quantum, and GPTs beyond quantum mechanics. Already for the simplest nontrivial scenario, we derive inequalities able to distinguish temporal correlations where the underlying system is classical, quantum, or more general.

## 1. Introduction

Given a single quantum system, in what sense can we say that it has some nonclassical properties? The most celebrated phenomena where quantum systems depart from their classical counterpart involve notions such as entanglement [1, 2] and nonlocality [3, 4], which can be defined only in terms of multipartite systems. What if we are able to perform experiments only on a single, indivisible, system? Can we still say that the observed statistics has some ‘nonclassical properties’? Some notion of nonclassicality have been proposed for single systems, such as contextuality [5] and nonmacrorealism [6, 7]. One may argue that such notions are limited to specific measurement procedures and hence are not fully satisfactory. Contextuality restricts the set of possible operations to compatible measurements, which in many cases need to be (approximately) projective or at least satisfy some analogous notion of repeatability and nondisturbance [8, 9], in order to avoid the so-called ‘compatibility loophole’ [8] or other similar classical explanations. Macrorealism has similar strong restrictions on the set of allowed measurements, namely, they must be noninvasive to avoid the clumsiness loophole or other forms of classical interpretation of the results [10].

A strong motivation for developing such a notion of nonclassicality for single systems also arises from quantum information theory. Notions such as entanglement and nonlocality have been proved to play a role in quantum information tasks related to communication, such as, e.g. device-independent quantum key distribution [11]. That such notions should play a role also for tasks involving only single systems, such as, e.g. quantum computation, is less evident. Several recent results connected quantum contextuality with models of quantum computation such as, e.g. quantum computation via magic state injection or measurement based quantum computation [12–18]. However, a natural question arises of whether this connection is fundamental or just related to the particular model used for quantum computation [19]. If one moves from compatible projective measurements to general instruments, it is no longer clear whether the notion of quantum contextuality make sense at all, due to the compatibility loophole mentioned above [8].

In this paper, we go beyond such notions and introduce a notion of nonclassicality for the measurement statistics of a single system which is not restricted to specific measurement operations. The main tool of this investigation is the notion of memory cost of simulating temporal correlations. By temporal correlations we mean the observed statistics arising from sequences of measurements on a single system and memory roughly refers the amount of classical information that can be stored in the physical system.

The notion of memory cost has been explored in connection with classical simulations of quantum contextuality [20, 21], quantum simulation of classical stochastic processes [22] memory asymmetry between prediction and retrodiction [23], and in relation with the accuracy of classical and quantum clocks [63]. A related notion, i.e. that of communication cost, has been explored in relation to both Bell nonlocality [24, 25] and temporal correlations [26, 27]. Similar notions have been explored also in the prepare-and-measure scenario [28–32] and in connection with quantum information tasks such as random access codes [33–35].

In our approach, we go beyond the prepare-and-measure scenario by exploring arbitrary long sequences of measurements and we remove any restriction on the type of measurement by considering arbitrary quantum instruments. Our analysis is not only restricted to the differences between classical and quantum theory, but is extended to general probabilistic theories (GPTs) [36–39], which embrace also the former theories. In particular, we derive inequalities on the observed probabilities that are able to discriminate between classical, quantum, and genuine GPT correlations. Moreover, as a further development of the ideas presented in [20, 21], we show that in the framework of finite-state machines it is impossible to simulate contextual correlations on a qubit system, for a fixed initial state and arbitrary instruments.

The paper is organized as follows. In section 2, we will introduce the basic notions and tools necessary for our analysis, namely, temporal correlations and the arrow of time polytope. In section 3, we will introduce finite-state machines in GPTs, in particular, also in classical and quantum theory. In section 4, we will discuss the existence of nontrivial temporal bounds for such theories and the impossibility of simulating contextual correlations on a qubit. Finally, we present the conclusions and an outlook of the paper.

## 2. Temporal correlations

We consider a box that accepts certain inputs from an input alphabet  $\mathcal{X}$  and produces outputs from an output alphabet  $\mathcal{A}$ . The box is operated in a sequential fashion, see figure 1(a), such that, for instance, it first receives an input labeled by  $x \in \mathcal{X}$  yielding an output labeled by  $a \in \mathcal{A}$ , subsequently it receives  $y$  yielding  $b$ , and finally it receives  $z$  yielding  $c$ . Prior to this sequence the box is initialized, such that its behavior is independent of anything except the input sequence  $xyz$ . Consequently, for a fixed input sequence  $xyz \in \mathcal{X}^3$ , the admissible output sequences  $abc \in \mathcal{A}^3$  are governed by a probability distribution. If we now consider all possible inputs, we obtain the correlations  $p(abc|xyz)$ . Due to the time ordering of the inputs and outputs, these correlations must satisfy the arrow of time constraints [40]

$$\sum_c p(abc|xyz) = \sum_c p(abc|xyz'), \text{ for all } a, b \in \mathcal{A} \text{ and all } x, y, z, z' \in \mathcal{X}, \quad (1)$$

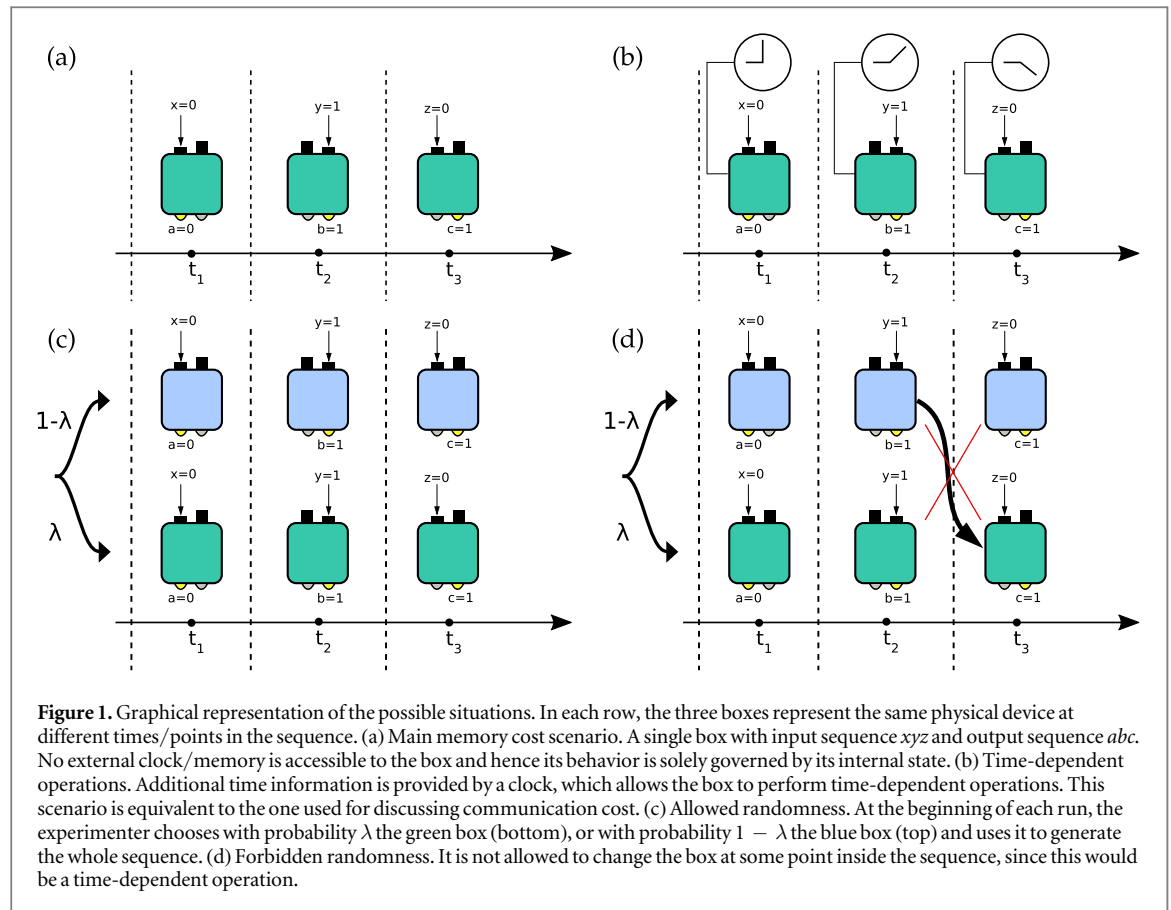
$$\sum_{bc} p(abc|xyz) = \sum_{bc} p(abc|xy'z'), \text{ for all } a \in \mathcal{A} \text{ and all } x, y, y', z, z' \in \mathcal{X}. \quad (2)$$

These constraints encode the fact that a future choice of an input, e.g.  $z$  or  $z'$  in equation (1), must not influence previous outputs of the box, e.g.  $a$  or  $b$ . This is in analogy to the nonsignaling conditions in the usual Bell scenario [41]. The arrow of time constraints come solely from causality and hence, they must be satisfied not only in classical and quantum theory, but in any GPT.

We can represent the correlations  $p(abc|xyz)$  as a vector with coordinates labeled by the possible sequences  $abc$  and  $xyz$ . Due to the linearity of the arrow of time constraints, the set of correlations satisfying those forms a polytope. Its extremal points have been recently characterized [42–44]. It is instructive to briefly sketch the central steps for the simple case of sequences of length three. All correlations in the corresponding polytope can be decomposed as

$$p(abc|xyz) = p(a|x)p(b|a; xy)p(c|ab; xyz), \quad (3)$$

since the marginals on the right-hand side are well defined (for the pathological cases where  $p(ab|xy) = 0$  we define the right-hand side to be zero). Vice versa, taking valid probability distributions  $p(a|x)$ ,  $p(b|a; xy)$ ,  $p(c|ab; xyz)$  over  $a, b, c$ , respectively, one always obtains an element of the polytope. Its extremal points are obtained by deterministic strategies, i.e. where each of the probability distributions on the right-hand side of equation (3) consists only of probabilities 0 or 1. It easily follows that classical and quantum models can reach extremal points if enough memory is available. In more precise terms, each deterministic strategy can be reached if the box internally keeps a record of all previous inputs and outputs. Storing this record then requires the box to have memory. Of course, the notion of memory needs clarification, in particular if the box is described using



quantum theory or a GPT, for details see section 3. Clearly, storing the full record of previous inputs and outputs is not necessarily memory optimal and gives rise to the question: what is the minimal number of states necessary to obtain certain correlations? How does such a number depend on the specific theory we use to describe the internals of the box?

An important element, in order to be able to speak about the memory cost of temporal correlations, is the requirement that all time-dependent information used to produce the outputs must be stored within the physical system used to implement the box. This implies that the physical operations performed to produce an output must be time-independent, e.g. the experimenter is not allowed to look at the wall clock and decide to implement in a different way the operation associated with a certain input  $x$ , as this will result in an additional source of memory, i.e. the clock keeping track of time. It is interesting to notice that the case where such time-dependent operations are admissible is equivalent to the case of quantum communication scenarios such as quantum random access codes or the scenario described by Brierley *et al* [26]. In fact, the latter scenario can be modeled as a network with ordered nodes, where a single physical system is transmitted through the nodes, and at each time step one of the nodes receives the system, performs a local operation, and transmit the system to the subsequent node. Since for each node it is known in advance in which part of the sequence it is situated, its local operations can be adapted to maximize a certain figure of merit defined in terms of probabilities of outcomes. This scenario covers the notion of ‘communication cost’ and it must be distinguished from the notion of ‘memory cost’ that is considered here. Moreover, even though in the memory cost scenario we are not allowed to change the operations throughout the sequence, it still makes sense to use classical randomness at the beginning of a sequence: at each experimental run, the experimenter can flip a coin and decide to perform the whole sequence of with one box or another. The resulting correlations will be a convex combination of the correlations obtained from either box. A graphical representation of the above ideas is presented in figure 1. These intuitive notions are made more rigorous in the next section.

### 3. Finite-state machines

In this section, we formally define the classical, quantum, and GPT models for the box used in the previous section. In this model we assume that the box is implemented as a machine which acts on an internal state. Upon receiving an input  $x$ , the box operates on the internal state and produces the output  $a$ . The internal state is the

specific model of the memory from the previous section. More precisely, we use the finite number of perfectly distinguishable states as a measure for the memory and for this reason we call this model a finite-state machine.

In a first step we need to describe the internal state  $\omega$  and the operations  $\mathcal{I}_{a|x}$  of the machine. We choose ordered vector spaces to describe the machine, which is an appropriate framework for a wide range of GPTs. In appendix A we give a brief summary of this mathematical formalism. In brief, a GPT is then described by a real vector space  $V$  with partial order ' $\leq$ ' and an order unit  $e \in V$ . In quantum theory  $V$  would be the set of Hermitian operators,  $A \leq B$  would correspond to  $B - A$  being positive semidefinite, and  $e$  to the identity operator. Measurement outcomes are represented by effects  $f \in V$  with  $0 \leq f \leq e$  and a measurement  $M_x$  is represented by a collection of effects  $M_x = (f_{a|x})_a$  with  $\sum_a f_{a|x} = e$ . The set of states  $\mathcal{S}$  is a subset of the dual space of  $V$  such that the probability of outcome  $a$  in the measurement  $M_x$  is given by  $p(a|x) = \omega f_{a|x}$ . Therefore  $\omega e = 1$  and  $\omega f \geq 0$  for all  $f \in V$  with  $f \geq 0$ . The operations  $\mathcal{I}_{a|x}$  represent a specific way to implement a measurement, taking into account the change of the internal state  $\omega$ . More precisely, the linear map  $\mathcal{I}_{a|x}: V \rightarrow V$  is such that  $f_{a|x} = \mathcal{I}_{a|x}e$  is the effect describing the output  $a$ . In addition the positivity condition  $\mathcal{I}_{a|x}f \geq 0$  for any  $f \in V$  with  $f \geq 0$  needs to be satisfied and further restrictions to  $\mathcal{I}_{a|x}$  may apply depending on the specific GPT. If we group together the transformations  $\mathcal{I}_x = (\mathcal{I}_{a|x})_{a \in \mathcal{A}}$  for a fixed input  $x$ , then  $\mathcal{I}_x$  is called an instrument. If we ignore the outcome  $a$ , then the instrument maps states to states, in the sense that  $\omega \sum_a \mathcal{I}_{a|x} \in \mathcal{S}$  for any state  $\omega$ .

Given the initial internal state  $\omega$  of the finite-state machine and the instrument  $\mathcal{I}_x = (\mathcal{I}_{a|x})_a$ , the probabilities associated with a sequence of measurement are given by

$$p(a|x) = \omega \mathcal{I}_{a|x}e, \quad p(ab|xy) = \omega \mathcal{I}_{a|x} \mathcal{I}_{b|y}e, \quad \text{etc.} \quad (4)$$

Note, that we write the transformations in the Heisenberg picture, so that the time ordering proceeds from the left to the right. For a general sequence of inputs  $x_1 x_2 \dots x_n = \vec{x}$  and outputs  $a_1 a_2 \dots a_n = \vec{a}$  we write

$$p(\vec{a}|\vec{x}) \equiv p(a_1 \dots a_n | x_1 \dots x_n) = \omega \mathcal{I}_{a_1|x_1} \dots \mathcal{I}_{a_n|x_n}e \equiv \omega \mathcal{I}_{\vec{a}|\vec{x}}e. \quad (5)$$

We exemplify in the next sections how this expression is specialized to the classical and quantum case.

As we discussed previously, we exclude any external source of memory, such as a clock keeping track of time. This is formalized by the fact that all instruments solely depend on the input and in particular by the fact that all transformations are time-independent. In general, for a fixed GPT this requirement makes the set of achievable correlations nonconvex. Nevertheless, we can recover convexity by allowing the use of convex mixtures as follows. Before starting the experiment we use a random variable  $\lambda$ , distributed according to some probability distribution  $q(\lambda)$ , to decide which finite-state machine to use subsequently. Since the machine is characterized by the initial state  $\omega_\lambda$  and the instruments  $\mathcal{I}_x^\lambda$ , this yields the correlations

$$p(\vec{a}|\vec{x}) = \sum_\lambda q(\lambda) \omega_\lambda \mathcal{I}_{\vec{a}|\vec{x}}^\lambda e. \quad (6)$$

The above procedure allows us to generate all correlations from the convex hull of correlations obtainable from a family of finite-state machines parametrized by  $\lambda$ .

Finally, we define the memory of the system using the GPT notion of capacity (see [45]), i.e. the size of the maximal set of perfectly distinguishable states. More precisely, we say that a GPT defines a  $d$ -state machine if  $d$  is the maximal integer such that there exists a collection of  $d$  states  $(\omega_k)_k$  and  $d$  effects  $(f_k)_k$  such that

$$\sum_k f_k \leq e \text{ and } \omega_i f_j = \delta_{ij} \text{ for all } i, j. \quad (7)$$

Namely, all effects are part of the same measurement, which is able to perfectly (i.e. probability one) discriminate among the states. This notion of capacity corresponds to the dimension of the Hilbert space in quantum mechanics and with the number of extremal points of the state simplex in classical probability theory (see, e.g. [45]).

It is instructive to discuss in more detail the classical and quantum case, which may be more familiar to the reader. We subsequently introduce a particular class of capacity-2 GPTs, the dichotomic norm cones [46].

### 3.1. Classical finite-state machines

A classical finite-state machine [47] is described by its internal rules for state transitions and output probabilities. Given the classical state  $\mathcal{C} = \{1, 2, \dots, d\}$ , the observed probability distribution  $p(\vec{a}|\vec{x})$  for an input sequence  $\vec{x}$  of length  $n$  can be written as

$$p(\vec{a}|\vec{x}) = \sum_{s_0, \dots, s_n \in \mathcal{C}} r(s_0) q(a_1, s_1|s_0, x_1) \cdots q(a_n, s_n|s_{n-1}, x_n). \quad (8)$$

Here,  $r(s_0)$  describes the probability of preparing the initial state  $s_0$  of the machine<sup>6</sup> and  $q(a, s'|s, x)$  describes the probability that the machine yields the output  $a$  and transition to the state  $s'$ , given that the internal state is  $s$  and the input is  $x$ . As in equation (6), those machines can depend on a random variable  $\lambda$  generated at the beginning of each sequence, i.e.

$$p(\vec{a}|\vec{x}) = \sum_{s_0, \dots, s_n \in \mathcal{C}, \lambda} p(\lambda) r^\lambda(s_0) q^\lambda(a_1, s_1|s_0, x_1) \cdots q^\lambda(a_n, s_n|s_{n-1}, x_n). \quad (9)$$

For clarity reasons, we use only equation (8) in the following. The correlations  $p(\vec{a}|\vec{x})$  can be rewritten as

$$p(\vec{a}|\vec{x}) = \pi^\dagger T(a_1|x_1) \cdots T(a_n|x_n) \eta \equiv \pi^\dagger T(\vec{a}|\vec{x}) \eta, \quad (10)$$

where  $\eta = (1, 1, \dots, 1)^\dagger$  is the  $d$ -dimensional vector of ones,  $\pi$  is the vector representing the initial state, and  $T(a|x)$  is the  $d \times d$  transition matrix. Hence,  $\pi_s = r(s)$  and  $[T(a|x)]_{s,s'} = q(a, s'|s, x)$ . The rules for probabilities that constrain  $q(a, s'|s, x)$  translate to  $[T(a|x)]_{s,s'} \geq 0$  for all  $s, s', a, x$ , and  $\sum_a [T(a|x)]_s = 1$  for all  $s, x$ .

Translating the above in the languages of GPTs, we let  $V = \mathbb{R}^d$  and set the order unit  $e$  to  $\eta$ . The partial order is such that  $\nu \leq w$  if  $\nu_s \leq w_s$  for all  $s$ . Then the set of states is given by the canonical  $(d - 1)$ -dimensional simplex

$$\mathcal{S} = \{\varpi \in \mathbb{R}^d | \varpi \geq 0 \text{ and } \varpi^\dagger \eta = 1\}. \quad (11)$$

In particular  $\pi$  is a state. Analogously, the transition matrix  $T(a|x)$  corresponds to the instruments  $\mathcal{I}_{a|x}$ , whereas the effects can be obtained as  $f_{a|x} := T(a|x)\eta$ . It can be easily seen that  $d$  correspond exactly to the capacity defined according to equation (7).

### 3.1.1. Classical finite-state machines and Leggett–Garg’s macrorealist models

It is interesting at this point to briefly compare the model in equation (9) with the macrorealist model of Leggett and Garg [6]. A macrorealist model can be simply obtained by reducing the set of possible internal states to a single one, i.e.  $d = 1$ , and re-introducing the time-dependence of operations

$$p(\vec{a}|\vec{x}) = \sum_{\lambda} p(\lambda) q_{t_1}^\lambda(a_1|x_1) \cdots q_{t_n}^\lambda(a_n|x_n), \quad (12)$$

where the dependency on  $s_0, \dots, s_n$  becomes trivial and is then removed. We recall that macrorealist models are based on two assumptions: *macrorealism per se*, i.e. the existence of a classical probability, and *noninvasive measurability*, i.e. the assumption that the measurement has no effect on the subsequent evolution of the system. The finite-state machine model can be seen as arising from the macrorealist model via a relaxation of the assumption of a noninvasive measurement: the measurement can be invasive up to a certain amount quantified by the internal memory of the system, e.g. for a two state-machine the measurement can encode at most one bit of information in the system. Notice that, however, usually Leggett–Garg assumptions allow the operations to be time-dependent.

It is interesting to remark that similar ideas have been already employed in Leggett–Garg tests to tighten the clumsiness loophole. Under the assumption of a classical model with two internal states, Knee *et al* [?] were able to quantify the measurement invasivity via a control experiment, and consequently modify the classical bound for the Leggett–Garg inequality. In agreement with our argument above, the work of Knee *et al* shows how the notion of finite memory can be used as a relaxation of the assumption of a noninvasive measurement.

### 3.2. Quantum finite-state machines

The quantum case is perhaps the most familiar to readers from quantum information. The probability distribution is obtained by sequences of generalized measurements  $M_x = (E_{a|x})_a$  on a single system described by a Hilbert space of fixed dimension  $d$ . The outcomes of the measurement are described by positive semidefinite operators  $E_{a|x} \geq 0$  with  $\sum_a E_{a|x} = \mathbb{1}$ .

In order to discuss sequential measurements, however, we need to know the post-measurement state, or, better, the transformation induced by the measurements. This information is provided by a quantum instrument  $\mathcal{I}_x$ , defined as a collection of completely positive maps  $\mathcal{I}_x = (\mathcal{I}_{a|x})_a$ , from the space of linear operators into itself, that sum up to a unital map, i.e.  $\sum_a \mathcal{I}_{a|x}(\mathbb{1}) = \mathbb{1}$ , corresponding to the rule of preservation of probability in the Heisenberg picture, see, e.g. [48]. Each instrument defines a generalized measurement through the formula  $E_{a|x} = \mathcal{I}_{a|x}(\mathbb{1})$ . Similarly to the previous cases, we can shorten the notation by defining

<sup>6</sup> Without loss of generality, we could assume a fixed pure initial state  $s_0$ , since we allow for convex mixtures of different machines. Nevertheless, we keep the notation with an initial distribution  $r(s_0)$  over all pure states  $\mathcal{C}$ , i.e. a mixed state, to keep the analogy with the standard notation for GPT states ( $\omega$ ) and quantum states ( $\rho$ ).



$\mathcal{I}_{\vec{a}|\vec{x}} := \mathcal{I}_{a_1|x_1} \circ \dots \circ \mathcal{I}_{a_n|x_n}$ , where  $\circ$  denotes the composition of maps and write

$$p(\vec{a}|\vec{x}) = \text{tr}[\rho \mathcal{I}_{\vec{a}|\vec{x}}(\mathbb{1})]. \quad (13)$$

As mentioned before, quantum theory is a particular case of a GPT, where the vectors space  $V$  is the set of Hermitian operators, the partial order is defined through positive semidefiniteness and the order unit  $e$  is given by  $\mathbb{1}$ . The set of states is given by the density operators, identified by the Hilbert–Schmidt inner product with the elements of the dual space of  $V$

$$\mathcal{S} = \{X \mapsto \text{tr}(\rho X) | \rho \geq 0 \text{ and } \text{tr}(\rho) = 1\}. \quad (14)$$

Hence equations (13) and (5) are equivalent. It is then clear that the capacity of the system, defined as the number of perfectly distinguishable state [44, 49] precisely corresponds to the dimension of the Hilbert space. It is important to remark that we need to consider the general formalism of quantum instruments, since if the measurement devices would merely act projectively, there would be nontrivial limitations on the achievable correlations that are valid for arbitrary dimensions [50, 51].

### 3.3. GPT two-state machines

We already provided a definition of GPT finite-state machines at the beginning of section 3. In this section, we specialize this definition by considering a class GPTs where the effects belong to a dichotomic norm cone. These theories are a generalization of the classical bit (cbit) and quantum bit (qubit), in the sense that they have capacity two, i.e. they allow for a set of perfectly distinguishable states, in the sense of equation (7), of at most size two. We then specialize our discussion to the case of hyperbits (hbits) [52] and generalized bits (gbits) [53]. The former are a generalization of the Bloch sphere to dimension higher than three, whereas the latter are the local part of a Popescu–Rohrlich box [41]. We also provide a more detailed discussion of GPTs in appendix A.

Consider the vector space  $V := \mathbb{R} \times \mathbb{R}^n$ , and the partial order where  $(t, \mathbf{x}) \geq 0$  if  $t \geq |\mathbf{x}|$ . Here,  $|\mathbf{x}|$  is any norm in  $\mathbb{R}^n$ . We define the order unit  $e := (1, 0)$ . This implies that effects are vectors  $f = (t, \mathbf{x})$  such that  $|\mathbf{x}| \leq \min\{t, 1 - t\}$ . The states for a dichotomic norm cone are the maps  $\omega: (t, \mathbf{x}) \mapsto t + \mathbf{w}^T \mathbf{x}$  with the condition  $|\mathbf{w}|_* \leq 1$ , where  $|\mathbf{w}|_* := \sup\{\mathbf{w}^T \mathbf{y} | |\mathbf{y}| \leq 1\}$  is the dual norm of  $|\cdot|$ . A peculiarity of this GPT is that it has exactly capacity two, independent of  $n$  or the choice of the norm  $|\cdot|$ . We provide a proof of this fact in appendix C.

Depending on the norm chosen and on  $n$  we have different GPTs. If we take  $|\mathbf{x}|$  to be the Euclidean (or  $\ell_2$ ) norm, i.e.  $|\mathbf{x}|^2 = \sum_i x_i^2$ , we obtain hbits, and specifically cbits for  $n = 1$ , qubits for  $n = 3$  and more general hbits for  $n > 3$ . If we take  $n = 2$  and the Manhattan (or  $\ell_1$ ) norm, i.e.  $|\mathbf{x}| = \sum_i |x_i|$ , we obtain a gbit. For the case of the Euclidean norm, the dual norm is also the Euclidean norm itself, whereas the dual of the Manhattan norm is the supremum (or  $\ell_\infty$ ) norm, i.e.  $|\mathbf{w}|_* = \max_i |w_i|$ .

## 4. Bounds on temporal correlations

In this section, we consider the simplest nontrivial scenario, a sequence of two measurements, with inputs  $x, y$  and outputs  $a, b$ , with  $a, b, x, y = 0, 1$ . We are interested in bounds on the sum of correlations

$$S = p(01|00) + p(10|10) + p(10|11). \quad (15)$$

Similar expressions have been considered in [43, 44, 54]. Clearly, the trivial bound  $S \leq 3$  holds. For hbits the value  $S = 3$  cannot be reached and therefore there must exist a nontrivial bound  $S \leq \Omega_{\text{hbit},n}$  for any dimension  $n$  of the hbit, in particular for the cbit ( $n = 1$ ) and the qubit ( $n = 3$ ). A simple analytical proof of  $\Omega_{\text{hbit},n} < 3$  is presented in appendix B.

### 4.1. Measure-and-prepare strategies

The analysis of the case of sequences of length two can be greatly simplified using measure-and-prepare instruments. These are instruments of the form  $\mathcal{T}_x = (f_{a|x} \sigma_{a|x})_a$ , where  $\mathbf{M}_x = (f_{a|x})_a$  is a measurement and  $(\sigma_{a|x})_a$  is a collection of states. Hence  $\mathcal{T}_x$  can be implemented by first measuring  $\mathbf{M}_x$  and then, depending on the outcome  $a$ , preparing the state  $\sigma_{a|x}$ .

Now, for a sequence of length two, the correlations are given by

$$p(ab|xy) = \sum_{\lambda} p(\lambda) \omega^{\lambda} \mathcal{I}_{a|x}^{\lambda} f_{b|y}^{\lambda}, \quad (16)$$

where  $\omega^{\lambda}$  is given by the initialization procedure of the individual finite-state machines participating in the mixture of machines. Clearly, the extremal values  $S$  can be achieved by a single finite-state machine and hence in the following we will omit the index  $\lambda$  and the summation of  $\lambda$ .

The instruments  $\mathcal{T}_x$  can be replaced by measure-and-prepare instruments, by letting  $f_{a|x} = \mathcal{I}_{a|x} e$  and  $\sigma_{a|x} = \omega \mathcal{I}_{a|x} / \omega(f_{a|x})$  if the denominator is nonzero, or  $\sigma_{a|x} = \omega$ . Then  $p(ab|xy) = \omega f_{a|x} \sigma_{a|x} f_{b|y}$ . Hence we can

equivalently replace  $\mathcal{I}_{a|x}$  by the prepare-and-measure strategy  $\mathcal{T}_{a|x} = f_{a|x} \sigma_{a|x}$ . Using this simplification, we obtain

$$\begin{aligned} S &= p(0|0) + p(1|1) + p(10|10) - p(00|00) - p(11|11) \\ &= p(0|0)[1 - p(0|0; 00)] + p(1|1)[1 + p(0|1; 10) - p(1|1; 11)] \\ &= \omega(f_{0|0})[1 - \sigma_{0|0}(f_{0|0})] + \omega(f_{1|1})[1 + \sigma_{1|1}(f_{0|0} - f_{1|1})], \end{aligned} \quad (17)$$

where we used the notation  $p(b|a; xy)$  for the probabilities conditioned on previous outputs.

#### 4.2. Analytical and numerical bounds

Since  $S = 3$  cannot be reached with hbits, there must be a finite gap between the actual bound for cbits, qubits, and hbits with a Bloch sphere of fixed dimension. In fact, the sets of states and effects are compact, and the expression  $S$  can be written as a continuous function from the set of states and effects into the interval  $[0, 3]$ , so its image must be compact. In this section, we explore in more detail the bounds for cbits, qubits, and hbits via numerical methods.

##### 4.2.1. Classical bit

For the cbit case, we use the representation from section 3.1, specifically,  $\omega$  is represented by  $(1, 0)$ ,  $\sigma_{i|i}$  by  $(s_i, 1 - s_i)$ , and  $f_{i|i}$  by  $(a_i, b_i)^\dagger$ , where  $s_i, a_i, b_i \in [0, 1]$ . Then equation (17) reads

$$S = a_0[1 - s_0 a_0 - (1 - s_0) b_0] + a_1[1 + s_1(a_0 - a_1) + (1 - s_1)(b_0 - b_1)]. \quad (18)$$

Only  $a_0$  and  $a_1$  appear nonlinearly in this expression. Therefore, the maximum of  $S$  is attained when all remaining parameters are either 0 or 1. This leaves us with a two-dimensional, at most quadratic optimization, which can be performed at once. For the maximal value  $\Omega_{\text{cbit}}$  of  $S$  using classical bits we then obtain

$$\Omega_{\text{cbit}} = \frac{9}{4}. \quad (19)$$

This maximum occurs at a unique point, where  $s_1 = b_1 = 0$ ,  $b_0 = s_0 = a_1 = 1$ , and  $a_0 = \frac{1}{2}$ . Hence, an optimal machine is given by the initial state  $\pi^\dagger = (1, 0)$  and the transition matrices

$$T(0|0) = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix}, \quad T(1|0) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}, \quad T(1|1) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = T(0|1)^\dagger. \quad (20)$$

Note, that while the solution for the chosen parametrization is unique, the transition matrices are not unique.

##### 4.2.2. Quantum bit

For the qubit case, we can proceed similarly to [44]. First we note that in equation (17), the initial state  $\omega$  can be replaced by a pure state, so that  $\omega: X \mapsto \langle 0|X|0\rangle$ . The expression  $S$  can then be written as

$$S = \langle 0|E_{0|0}|0\rangle[1 - \text{tr}[\sigma_0 E_{0|0}]] + \langle 0|E_{1|1}|0\rangle[1 + \text{tr}[\sigma_1(E_{0|0} - E_{1|1})]], \quad (21)$$

where  $0 \leq E_{i|i} \leq \mathbb{1}$  are effects and  $\sigma_0$  and  $\sigma_1$  are density operators. Since the latter occur only linearly in  $S$ , we can substitute them with pure states  $|\psi_0\rangle$  and  $|\psi_1\rangle$ , respectively. The maximum of  $S$  for qubits is hence given by

$$\Omega_{\text{qubit}} = \max_{\substack{|\psi_0\rangle, |\psi_1\rangle \\ E_{0|0}, E_{1|1}}} [\langle 0|E_{0|0}|0\rangle(1 - \langle \psi_0|E_{0|0}|\psi_0\rangle) + \langle 0|E_{1|1}|0\rangle(1 + \langle \psi_1|E_{0|0} - E_{1|1}|\psi_1\rangle)]. \quad (22)$$

By parametrizing  $E_{0|0}$ ,  $E_{1|1}$ ,  $|\psi_0\rangle$ ,  $|\psi_1\rangle$  with real parameters, one can write the expression in equation (22) as fourth degree polynomial. This can be further simplified, by taking  $E_{0|0}$ ,  $E_{1|1}$ ,  $|\psi_0\rangle$ ,  $|\psi_1\rangle$  as real expression, which lowers the number of parameters to ten<sup>7</sup>. The reduction to the real part of a qubit does not affect the optimality as we show in the next section, see equation (28).

It is always possible to obtain a lower bound  $\Omega_{\text{qubit}}^{\text{feas}}$  on  $\Omega_{\text{qubit}}$  by guessing appropriate values for the free parameters. An upper bound,  $\Omega_{\text{qubit}}^{\text{Lass}}$ , can be obtained via Lasserre's method [55] of polynomial optimization based on moment matrices and semidefinite programming [56], which provides analytical upper bounds up to the numerical precision. That is

$$\Omega_{\text{qubit}}^{\text{feas}} \leq \Omega_{\text{qubit}} \leq \Omega_{\text{qubit}}^{\text{Lass}}. \quad (23)$$

With the simplifications used above, the upper and lower bounds coincide up to the numerical precision of  $10^{-5}$ . We have

<sup>7</sup> Since the upper bound is calculated by polynomial optimization methods, it is more convenient to keep the expression and constraints in polynomial form, rather than minimizing the number of variables. For example, a parametrization of a pure state as  $\cos \theta|0\rangle + \sin \theta|1\rangle$  removes one variable and one constraint, but it is no long a polynomial in the parameters.



$$\Omega_{\text{qubit}}^{\text{feas}} \approx \Omega_{\text{qubit}}^{\text{Lass}} \approx 2.355\,70, \quad (24)$$

showing a gap between the cbit and qubit case. A feasible solution is given by the post-measurement states and effects

$$|\psi_0\rangle \approx 0.408|0\rangle - 0.913|1\rangle, \quad |\psi_1\rangle \approx 0.640|0\rangle + 0.768|1\rangle, \quad (25)$$

and the effects

$$E_{0|0} = \mathbb{1} - |\psi_0\rangle\langle\psi_0|, \text{ and } E_{1|1} = |\phi\rangle\langle\phi|, \text{ where } |\phi\rangle \approx 0.971|0\rangle - 0.238|1\rangle. \quad (26)$$

#### 4.2.3. Hyperbit

For the case of hbits, and also the more general dichotomic norm cones, we use the parametrization  $\omega: (t, \mathbf{x}) \mapsto t + \mathbf{w}^\dagger \mathbf{x}$  and  $\sigma_{i|i}: (t, \mathbf{x}) \mapsto t + \mathbf{w}_i^\dagger \mathbf{x}$  for the states and  $f_{i|i} = (t_i, \mathbf{f}_i)$  for the effects. Then equation (17) reads

$$S = (t_0 + \mathbf{w}^\dagger \mathbf{f}_0)[1 - t_0 - \mathbf{w}_0^\dagger \mathbf{f}_0] + (t_1 + \mathbf{w}^\dagger \mathbf{f}_1)[1 + t_0 - t_1 + \mathbf{w}_1^\dagger (\mathbf{f}_0 - \mathbf{f}_1)]. \quad (27)$$

When maximizing  $S$ , we can eliminate the maximization over  $\mathbf{w}_0$  and  $\mathbf{w}_1$ , by choosing appropriate vectors with  $|\mathbf{w}_i|_* = 1$  such that  $\mathbf{w}_0^\dagger \mathbf{f}_0 = |\mathbf{f}_0|$  and  $\mathbf{w}_1^\dagger (\mathbf{f}_0 - \mathbf{f}_1) = |\mathbf{f}_0 - \mathbf{f}_1|$ . The maximal value of  $S$  for a given dichotomic norm cone is hence

$$\Omega_{\text{dnc}} = \max_{\substack{\mathbf{w}, t_0, \mathbf{f}_0 \\ t_1, \mathbf{f}_1}} \{ (t_0 + \mathbf{w}^\dagger \mathbf{f}_0)[1 - t_0 + |\mathbf{f}_0|] + (t_1 + \mathbf{w}^\dagger \mathbf{f}_1)[1 + t_0 - t_1 + |\mathbf{f}_0 - \mathbf{f}_1|] \}, \quad (28)$$

where the constraints of the optimization are  $|\mathbf{w}|_* \leq 1$  and  $|\mathbf{f}_i| \leq \min\{t_i, 1 - t_i\}$ . For the case of hbits, both  $|\cdot|_*$  and  $|\cdot|$  correspond to the  $\ell_2$  norm, hence the conditions are invariant under orthogonal transformations as it is the case for the function to be maximized, which depends only on the norm of  $\mathbf{f}_i$  and the scalar products between  $\mathbf{w}$  and  $\mathbf{f}_i$ . Since the only contribution for  $\mathbf{w}$  comes from the component in the span of  $\mathbf{f}_0, \mathbf{f}_1$ , the problem reduces to a two-dimensional one. This is equivalent to the qubit case with the Bloch ball restricted to the  $xz$ -plane, both for states and effects. This implies that the bound for hbits coincide with the bound for qubits. We thus have

$$\Omega_{\text{hbit}} \approx 2.355\,70, \quad (29)$$

as in equation (24).

#### 4.2.4. Generalized bit

The case of gbits differs from the previous one because we can actually reach  $S = 3$  already for a two-state machine, namely the dichotomic norm cone with  $n = 2$  and the  $\ell_1$  norm. This model corresponds to the local part of a Popescu–Rohrlich box [41, 53]. The space of effects is a polytope with extremal effects given by the extremal point of the two-dimensional  $\ell_1$  norm, i.e.  $a_{\pm i} = \frac{1}{2}(1, \pm \mathbf{e}_i)$ , with  $\mathbf{e}_i$  the canonical vectors in  $\mathbb{R}^2$ . Then, the states are the  $\omega = (1, \mathbf{w})$  with  $\mathbf{w}$  in the square  $[-1, 1] \times [-1, 1]$ , i.e. the unit ball with respect to the  $\ell_\infty$  norm. The choices

$$\mathbf{w}^\dagger = (1, -1), \quad \mathbf{w}_0^\dagger = (-1, 1), \quad \mathbf{w}_1^\dagger = (1, 1) \quad (30)$$

and

$$\mathbf{f}_0 = \mathbf{e}_1, \quad \mathbf{f}_1 = -\mathbf{e}_2 \quad (31)$$

yield, according to equation (28), the algebraic maximum for  $S$ , i.e.  $S = 3$ . We thus have

$$\Omega_{\text{gbit}} = 3 \quad (32)$$

for gbits and hence also for the set of all dichotomic norm cones with the same norm and arbitrary  $n$ .

### 4.3. Impossibility of simulating contextual correlations with general instruments on a qubit

In this section, we investigate whether qubit machines are able to simulate some contextual correlations that arise in higher dimensional quantum systems. In [20] it was proved that in order to simulate all deterministic predictions associated with the observables of the Peres–Mermin square [57, 58], a classical machine with at least 4 states is necessary. This result was obtained in the framework of tests of contextuality involving sequential measurements [8], in which the relevant compatibility notion is given by the nondisturbance among compatible measurements and repeatability of outcomes, e.g. if  $M_x$  and  $M_y$  are compatible measurements in the measurement sequence  $M_x M_y M_x$ , the outcome for the first measurement of  $M_x$  will be repeated in the second measurement of  $M_x$ .

We derive here a related result by showing that even a qubit is not sufficient to exhibit contextual correlations. For this we use a rather broad notion of contextuality. Consider a box with inputs from an alphabet

$\mathcal{X}$  and outputs from an alphabet  $\mathcal{A}$  as before. The input sequences are restricted such that a sequence  $\vec{x}$  is admissible if and only if all inputs are from the same context  $\mathcal{C} \subset \mathcal{X}$ , i.e.  $\{x_i|i\} \subset \mathcal{C}$ . A context  $\mathcal{C}$  is a set of inputs, such that  $p(\vec{a}|\vec{x}) = p[\pi(\vec{a})|\pi(\vec{x})]$  for any inputs sequence  $\vec{x}$  from  $\mathcal{C}$ , any output sequence  $\vec{a}$ , and any permutation  $\pi$ . In addition we assume that any input is repeatable, i.e.  $p(\vec{a}b|\vec{x}x_i) = p(\vec{a}|\vec{x})\delta_{b,a_i}$  for any position  $i$  in any admissible sequence.

Such a box is noncontextual, if all correlations of the box (using only admissible input sequences) can be reproduced by a box without memory, i.e. by a noncontextual model. We claim that any such box implemented on a qubit is noncontextual.

We start the proof of this statement by determining those inputs, which cannot require the use of memory. First, if an input  $z$  ever produces only the output  $c$ , within all admissible input sequences, then we can eliminate this input from our considerations. This is the case, because in any sequence we can permute  $z$  to the end of the sequence. Then

$$p(\vec{a}c|\vec{x}z) = p(\vec{a}|\vec{x})p(c|\vec{a}; \vec{x}z) = p(\vec{a}|\vec{x}), \quad (33)$$

where the first equality is due to equation (3) and the second due to the assumption that only the output  $c$  ever occurs. Second, assume that for a certain input  $z$ , whenever it occurs in an admissible sequence, the internal state of the machine before the input  $z$  is only ever the state  $\rho$ . Again we can eliminate this input from our considerations, because the output for  $z$  and the state after the output can be determined without considering the state. Third, we can ignore the pathological cases of inputs, which are not member of any context. In the following we assume without loss of generality, that the box does not have any input falling under the those three cases just discussed.

Next, we show that for any input  $z$  the instrument  $(\mathcal{I}_{c|z})_c$  must be a measure-and-prepare instrument of the form

$$\mathcal{I}_{c|z}: X \mapsto |\psi_{c,z}\rangle\langle\psi_{c,z}|X|\psi_{c,z}\rangle\langle\psi_{c,z}| \text{ with } \langle\psi_{c,z}|\psi_{c,z}\rangle \in \{0, 1\}. \quad (34)$$

This can be seen as follows. According to the assumptions, there are two input sequences  $\vec{x}z$  and  $\vec{y}z$  and corresponding output sequences  $\vec{a}c$  and  $\vec{b}c$ , so that the state before the input  $z$  is  $\rho$  and  $\rho'$ , respectively, with  $\rho \neq \rho'$ . Using equations (3) and (13) we have

$$p(\vec{a}c|\vec{x}z)\delta_{c,c'} = p(\vec{a}cc'|\vec{x}zz) = p(\vec{a}|\vec{x})p(cc'|\vec{a}; \vec{x}zz) = p(\vec{a}|\vec{x})\text{tr}[\rho \mathcal{I}_{c|z}\mathcal{I}_{c'|z}\mathbb{1}] \text{ and} \quad (35)$$

$$p(\vec{b}c|\vec{y}z)\delta_{c,c'} = p(\vec{b}cc'|\vec{y}zz) = p(\vec{b}|\vec{y})p(cc'|\vec{b}; \vec{y}zz) = p(\vec{b}|\vec{y})\text{tr}[\rho' \mathcal{I}_{c|z}\mathcal{I}_{c'|z}\mathbb{1}], \quad (36)$$

where  $p(\vec{a}|\vec{x}) > 0$  and  $p(\vec{b}|\vec{y}) > 0$ . Therefore for  $c \neq c'$

$$\text{tr}[\bar{\rho} \mathcal{I}_{c|z}\mathcal{I}_{c'|z}\mathbb{1}] = 0 \quad (37)$$

with  $\bar{\rho} = (\rho + \rho')/2$ . Since  $\rho \neq \rho'$  and we assume a qubit system, the mixture  $\bar{\rho}$  has necessarily rank two, i.e.  $\bar{\rho} \geq \epsilon \mathbb{1}$  for some  $\epsilon > 0$ . We arrive at the condition

$$\sum_{i,j} \text{tr}[K_i Q_j Q_j^\dagger K_i^\dagger] = 0, \quad (38)$$

where  $K_i$  and  $Q_j$  are the Kraus operators associated, respectively, with the instruments  $\mathcal{I}_{c'|z}$  and  $\mathcal{I}_{c|z}$ , e.g.  $\mathcal{I}_{c'|z}X = \sum_j K_j^\dagger X K_j$ . Then  $K_i Q_j = 0$  for all  $i, j$ . Similarly, exchanging  $c$  with  $c'$ , we obtain  $Q_j K_i = 0$  for all  $i, j$ . This implies that  $K_i$  and  $Q_j$  are of rank one and that  $K_i$  is proportional to  $K_{i'}$  as well as  $Q_j$  being proportional to  $Q_{j'}$ , for all  $i, i'$  and  $j, j'$ . Hence we can omit the indices  $i, j$  and consider simply  $K$  and  $Q$ . Note that from  $\sum_c \mathcal{I}_{c|z}\mathbb{1} = \mathbb{1}$ , the condition  $Q^\dagger Q \leq \mathbb{1}$  follows which allows us to write  $Q = |\alpha\rangle\langle\beta|$  with  $\langle\alpha|\alpha\rangle = 1$  and  $\langle\beta|\beta\rangle \leq 1$ . Now, for  $c = c'$  we obtain

$$\text{tr}(\bar{\rho} \mathcal{I}_{c|z}\mathcal{I}_{c|z}\mathbb{1}) = \text{tr}(\bar{\rho} \mathcal{I}_{c|z}\mathbb{1}), \quad (39)$$

which implies  $(Q^\dagger)^2 Q^2 = Q^\dagger Q$ . It follows that either  $|\beta\rangle = 0$  or  $|\alpha\rangle$  and  $|\beta\rangle$  are equal up to a phase and hence  $\mathcal{I}_{c|z}$  is as stated in equation (34).

As final step we need to show that there is no contextuality for projective qubit instruments. Given an admissible input sequence  $\vec{x}yz$ , and an output sequence  $\vec{a}bc$  such that  $p(\vec{a}b|\vec{x}y) > 0$ , we have

$$p(\vec{a}bc|\vec{x}yz) = p(\vec{a}b|\vec{x}y)|\langle\psi_{b,y}|\psi_{c,z}\rangle|^2 \text{ and } p(\vec{a}bcb|\vec{x}yzy) = p(\vec{a}b|\vec{x}y)|\langle\psi_{b,y}|\psi_{c,z}\rangle|^4. \quad (40)$$

The left-hand side of both expressions has to be equal, yielding  $|\langle\psi_{b,y}|\psi_{c,z}\rangle| \in \{0, 1\}$ .

Consequently, any two inputs within a context are realized by the same projective instrument, except for some relabeling of the outcomes. We choose a specific measurement within one context, say  $y$ , so that  $\mathcal{I}_{a|x} = \sum_b \mathcal{I}_{b|y} f^b(a|x)$  with some coefficients  $f^b(a|x) \in \{0, 1\}$ . This way we can write for any correlations of this context

$$p(\vec{a}|\vec{x}) = \sum_b p(b|y) \prod_i f^b(a_i|x_i), \quad (41)$$

which is exactly the formula for a one-state machine, i.e. a noncontextual model.

This concludes the proof of our statement, due to the following observation. If two contexts share an observable, then our argument already applies and the union of both contexts must admit a noncontextual model and hence the union of both contexts is again a context. Eventually, we can join contexts until all contexts are mutually disjoint. For each disjoint set we can construct a noncontextual model, and since there are no admissible sequence involving two different contexts, we have constructed a noncontextual model for all admissible input sequences.

## 5. Conclusions and outlook

We introduced the notion memory cost of simulating temporal correlations based on the notion of finite-state machine, i.e. a physical system accepting an input at each time instant and generating an outcome and an internal state transition according to probabilistic rules. We investigated the correlations obtainable via such finite-state machines operating according to different probability theories, i.e. classical, quantum, or GPT. Our framework allow us to derive inequalities able to discriminate among different theories for the simplest nontrivial case, i.e. two-state machines, two inputs, two outputs, and sequences of length two. Moreover, we investigated, from the perspective of quantum finite-state machines, the possibility of simulating contextual correlations with a qubit and answered this question in the negative.

Our framework provides a notion of nonclassicality for single systems, which is based solely on observed correlations and does not make any assumption of the type of measurements involved, e.g. compatibility or noninvasiveness. We believe that several problems in quantum foundations and quantum information could be studied in this framework. For instance, a notion of nonclassicality for single systems, i.e. quantum contextuality, has recently been suggested as a resource for quantum computation. On the other hand, memory has been identified as a resource needed to simulate contextual correlations classically [20, 21]. In addition, a different notion of contextuality for sequential operations has been defined and connected to speed-up in quantum computation [59]. Our work could provide a general framework to discuss such different results and understand better the connection between memory cost of (classical) simulations, contextual correlations, and advantages in computation. Moreover, the idea of computation in GPTs, such as Spekkens' toy model [60], that are intermediate between classical and quantum probability has been recently investigated [61, 62]. In particular, this GPT can be exactly simulated with two classical bits.

## Acknowledgments

The authors would like to thank Rafael Chaves, Andrew Garner, Otfried Gühne, Jannik Hoffmann, Niklas Johansson, Jan-Åke Larsson, Nikolai Miklin, Miguel Navascués, Cornelia Spee, Giuseppe Vitagliano, and Mischa Woods for discussions. This work has been supported by the Austrian Science Fund (FWF): M 2107 (Meitner-Programm) and ZK 3 (Zukunftskolleg), FQXi Large Grant 'The Observer Observed: A Bayesian Route to the Reconstruction of Quantum Theory', FIS2015-67161-P (MINECO/FEDER), Basque Government (project IT986-16), and ERC (Consolidator Grant 683107/TempoQ).

## Appendix A. Brief introduction to GPTs

In quantum theory the set of effects is represented by Hermitian operators  $F$  with  $0 \leq F \leq \mathbb{1}$ . This convex set has three characteristic properties. (i) It is a subset of the real vector space of Hermitian operators. (ii) There exists the special operator  $\mathbb{1}$  representing the all-embracing effect. (iii) Its shape is given by the partial order  $A \leq B$  which is defined by the condition that  $B - A$  is positive semidefinite.

In a GPT, the notion of an effect is generalized by considering a straightforward generalization of those properties. We start with an arbitrary real vector space  $V$  with a partial order  $a \leq b$ . This partial order has to be linear in the sense that  $a \leq b$  implies  $\lambda a \leq \lambda b$  for any  $\lambda \in \mathbb{R}^+$  and  $a \leq b$  implies  $a + c \leq b + d$  if also  $c \leq d$ . This turns  $(V, \leq)$  into an ordered vector space.

The all-embracing effect is a distinct element  $e \in V$ . It is required to dominate all of  $V$ , i.e. for any  $x \in V$  there is a positive number  $\lambda$  such that  $x \leq \lambda e$ . This property makes  $e$  an order unit and  $(V, \leq, e)$  an order unit vector space. In addition, it is convenient to assume that the order unit is Archimedean, i.e. if  $x \leq \lambda e$  holds for all  $\lambda > 0$ , then already  $x \leq 0$ . In our paper we implicitly assume that any order unit is Archimedean.

It is sometimes convenient to let  $V^+ = \{x \in V | 0 \leq x\}$ . Since  $a \leq b$  is equivalent to  $b - a \in V^+$ , we then equivalently describe an AOU space by the tuple  $(V, V^+, e)$ . The effects in a GPT are now given by the set  $V_e^+ = V^+ \cap (e - V^+)$ . A measurement  $M$  in a GPT is represented by a collection of elements  $M = (f_k)_k \subset V_e^+$  with  $\sum f_k = e$ , where  $f_k$  represent the outcomes of the measurement.

For the set of states, we note that in quantum theory one can represent a state  $\rho$  equivalently by the linear map  $\omega: X \mapsto \text{tr}(\rho X)$ . Then the normalization of  $\rho$  becomes  $\omega(\mathbb{1}) = 1$  and the condition  $\rho \geq 0$  reads  $\omega(X) \geq 0$  for all  $X \geq 0$ . By analogy, the set of states in a GPT is given by

$$\mathcal{S} = \{\omega \in V^* | \omega(e) = 1 \text{ and } \omega(f) \geq 0 \text{ for all } f \geq 0\}, \quad (\text{A1})$$

where  $V^* = \{\varphi: V \rightarrow \mathbb{R} | \varphi \text{ is linear}\}$  is the dual space of  $V$ . With this definition, the probability for outcome  $k$  of a measurement  $M = (f_k)_k$  is given by  $p_k = \omega(f_k)$ .

## Appendix B. Bound on $S$ for hbits

The proof is by contradiction. Let us assume  $\Omega_{\text{hbit}} = 3$ , we then have  $p(01|00) = p(10|10) = p(10|11) = 1$ , and  $p(0|0) = p(1|1) = 1$ . From  $p(0|0) = 1$ , we have  $\omega(f_{0|0}) = t_{0|0} + \mathbf{w}^\dagger f_{0|0} = 1$ , where  $f_{0|0} = (t_{0|0}, f_{0|0})$ . On the other hand, by the definition of effects and state, we have  $|f_{0|0}| \leq \min(t_{0|0}, 1 - t_{0|0})$  and  $|\mathbf{w}|_* \leq 1$ . We then have

$$t_{0|0} \geq \frac{1}{2} \text{ and } \mathbf{w}^\dagger f_{0|0} = 1 - t_{0|0}. \quad (\text{B1})$$

From  $f_{0|0} \neq e$  (because  $p(00|00) = 0$ ), we have  $f_{0|0} \neq 0$  and hence  $t_{0|0} < 1$ . Then, using again  $|\mathbf{w}|_* \leq 1$  and  $|\cdot| = |\cdot|_*$ , together with Cauchy–Schwarz inequality, we have

$$\mathbf{w} = \frac{f_{0|0}}{|f_{0|0}|}, \text{ with } |f_{0|0}| = 1 - t_{0|0}. \quad (\text{B2})$$

Similarly, we obtain

$$\mathbf{w} = \frac{f_{1|1}}{|f_{1|1}|}, \text{ with } |f_{1|1}| = 1 - t_{1|1}, \quad (\text{B3})$$

and, again,  $t_{1|1} < 1$ .

We need now to characterize the terms of the form  $\omega(\mathcal{I}_{a|x} f_{b|y})$ , corresponding to sequences of length two. We use the constraints that arise from the condition that the transformation must map effects to effects. Then, we use that  $\mathcal{I}_{a|x}$  is a linear transformation that maps the identity element to  $f_{a|x}$ , i.e.  $\mathcal{I}_{a|x} e = f_{a|x}$ . We, thus, have

$$\mathcal{I}_{a|x} = \begin{pmatrix} t_{a|x} & \alpha_{a|x}^\dagger \\ f_{a|x} & B_{a|x} \end{pmatrix}, \quad (\text{B4})$$

where  $\alpha_{a|x}$  is a  $n$ -dimensional vector and  $B_{a|x}$  a  $n \times n$  matrix. The expectation value can then be written as

$$\omega(\mathcal{I}_{a|x} f_{b|y}) = (1, \mathbf{w}^\dagger) \begin{pmatrix} t_{a|x} & \alpha_{a|x}^\dagger \\ f_{a|x} & B_{a|x} \end{pmatrix} \begin{pmatrix} t_{b|y} \\ f_{b|y} \end{pmatrix} \quad (\text{B5})$$

$$= (t_{a|x} + \mathbf{w} \cdot f_{a|x}, \mathbf{w}^\dagger B_{a|x} + \alpha_{a|x}^\dagger) \begin{pmatrix} t_{b|y} \\ f_{b|y} \end{pmatrix} \quad (\text{B6})$$

$$= t_{b|y}(t_{a|x} + \mathbf{w} \cdot f_{a|x}) + f_{b|y} \cdot (B_{a|x}^\dagger \mathbf{w} + \alpha_{a|x}). \quad (\text{B7})$$

We can see the transformation  $\mathcal{I}_{a|x}$ , applied to the left, as a state transformation, i.e. Schrödinger picture and with normalization corresponding to outcome probability. Then, we have that  $|B_{a|x}^\dagger \mathbf{w} + \alpha_{a|x}|_* \leq t_{a|x} + \mathbf{w} \cdot f_{a|x}$ . Notice that such a condition also guarantees that  $p(ab|xy) \geq 0$  and  $p(ab|xy) \leq p(a|x)$ .

This translates to  $|B_{a|x}^\dagger \mathbf{w} + \alpha_{a|x}|_* \leq 1$  for the case  $(a, x) = (0, 0)$  or  $(1, 1)$ . In fact, in those cases we have  $t_{a|x} + \mathbf{w} \cdot f_{a|x} = 1$ , so the dual norm condition guarantee that  $\omega(\mathcal{I}_{a|x} f_{b|y}) \leq 1$  for all  $f_{b|y}$ .

From the conditions  $p(11|11) = 0$ , we obtain

$$t_{1|1}(t_{1|1} + \mathbf{w} \cdot f_{1|1}) + f_{1|1} \cdot (B_{1|1}^\dagger \mathbf{w} + \alpha_{1|1}) \quad (\text{B8})$$

$$= t_{1|1} + f_{1|1} \cdot (B_{1|1}^\dagger \mathbf{w} + \alpha_{1|1}) = 0, \quad (\text{B9})$$

which implies together with equation (B3),  $|B_{1|1}^\dagger \mathbf{w} + \alpha_{1|1}|_* \leq 1$ , and Cauchy–Schwarz inequality, that

$$(B_{1|1}^\dagger \mathbf{w} + \alpha_{1|1}) = -\mathbf{w} \text{ and } t_{1|1} = \frac{1}{2}. \quad (\text{B10})$$

On the other hand, we have  $p(10|10) = 1$  that, by equations (B2) and (B10), implies

$$t_{0|0}(t_{1|1} + \mathbf{w} \cdot \mathbf{f}_{1|1}) + f_{0|0} \cdot (B_{1|1}^\dagger \mathbf{w} + \alpha_{1|1}) = t_{0|0} - f_{0|0} \cdot \mathbf{w} = 2t_{0|0} - 1 = 1, \quad (\text{B11})$$

which implies  $t_{0|0} = 1$ , i.e. a contradiction with  $t_{0|0} < 1$ , which concludes the proof.

## Appendix C. Capacity of dichotomic norm cones

In the following we prove that dichotomic norm cones describe systems of capacity two. For convenience, we repeat equation (7) from the main text.

$$\sum_k f_k \leq e \text{ and } \omega_i f_j = \delta_{ij} \text{ for all } i, j.$$

We first show that a capacity of two is an upper bound.

**Lemma 1.** *In a dichotomic norm cone, let  $(\omega_k)_k$  be a collection of  $d$  states and  $(f_k)_k$  a collection of  $d$  effects, such that equation (7) is satisfied. Then  $d \leq 2$ .*

**Proof.** Any effect  $f = (t, \mathbf{x})$  must satisfy  $0 \leq f \leq e$ , i.e.  $t \geq |\mathbf{x}|$  and  $1 - t \geq |\mathbf{x}|$ . Furthermore, a state  $\omega: (s, \mathbf{y}) \mapsto s + \mathbf{w}^\dagger \mathbf{y}$  must obey  $|\mathbf{w}|_* \leq 1$ . It follows that  $\mathbf{w}^\dagger \mathbf{x} \leq |\mathbf{x}|$  and hence  $\omega f = 1$  requires  $t \geq \frac{1}{2}$ . Thus  $\sum_k f_k \leq e$  implies for  $f_k = (t_k, \mathbf{x}_k)$  the inequalities

$$0 \leq |\sum_k \mathbf{x}_k| \leq 1 - \sum_k t_k \leq 1 - \frac{d}{2}. \quad (\text{C1})$$

Which yields at once the assertion.  $\square$

In addition, if the dimension of the underlying vector space is finite, we can always find vectors  $\mathbf{x}$  and  $\mathbf{w}$ , such that  $|\mathbf{x}| = 1$ ,  $|\mathbf{w}|_* = 1$ , and  $\mathbf{w}^\dagger \mathbf{x} = 1$ . Hence, the states  $\omega_{1,2} = (1, \pm \mathbf{w})$  and effects  $f_{1,2} = (1, \pm \mathbf{x})/2$  obey equation (7). It follows that the capacity of a dichotomic norm cone is always exactly two.

## References

- [1] Horodecki R, Horodecki P, Horodecki M and Horodecki K 2009 Quantum entanglement *Rev. Mod. Phys.* **81** 865–942
- [2] Gühne O and Tóth G 2009 Entanglement detection *Phys. Rep.* **474** 1–75
- [3] Bell J S 1964 On the Einstein–Podolsky–Rosen paradox *Physics* **1** 195–200
- [4] Brunner N, Cavalcanti D, Pironio S, Scarani V and Wehner S 2014 Bell nonlocality *Rev. Mod. Phys.* **86** 419–78
- [5] Kochen S and Specker E P 1967 The problem of hidden variables in quantum mechanics *J. Math. Mech.* **17** 59
- [6] Leggett A J and Garg A 1985 Quantum mechanics versus macroscopic realism: is the flux there when nobody looks? *Phys. Rev. Lett.* **54** 857–60
- [7] Emary C, Lambert N and Nori F 2014 Leggett–Garg inequalities *Rep. Prog. Phys.* **77** 016001
- [8] Gühne O, Kleinmann M, Cabello A, Larsson J-Å, Kirchmair G, Zähringer F, Gerritsma R and Roos C F 2010 Compatibility and noncontextuality for sequential measurements *Phys. Rev. A* **81** 022121
- [9] Kujala J V, Dzharafarov E N and Larsson J-Å 2015 Necessary and sufficient conditions for an extended noncontextuality in a broad class of quantum mechanical systems *Phys. Rev. Lett.* **115** 150401
- [10] Wilde M M and Mizel A 2012 Addressing the clumsiness loophole in a Leggett–Garg test of macrorealism *Found. Phys.* **42** 256–65
- [11] Acín A, Brunner N, Gisin N, Massar S, Pironio S and Scarani V 2007 Device-independent security of quantum cryptography against collective attacks *Phys. Rev. Lett.* **98** 230501
- [12] Howard M, Wallman J, Veitch V and Emerson J 2014 Contextuality supplies the ‘magic’ for quantum computation *Nature* **510** 351
- [13] Delfosse N, Allard Guerin P, Bian J and Raussendorf R 2015 Wigner function negativity and contextuality in quantum computation on rebits *Phys. Rev. X* **5** 021003
- [14] Raussendorf R 2013 Contextuality in measurement-based quantum computation *Phys. Rev. A* **88** 022322
- [15] Bermejo-Vega J, Delfosse N, Browne D E, Okay C and Raussendorf R 2017 Contextuality as a resource for models of quantum computation with qubits *Phys. Rev. Lett.* **119** 120505
- [16] Raussendorf R, Browne D E, Delfosse N, Okay C and Bermejo-Vega J 2017 Contextuality and Wigner-function negativity in qubit quantum computation *Phys. Rev. A* **95** 052334
- [17] Abramsky S, Barbosa R S and Mansfield S 2017 Contextual fraction as a measure of contextuality *Phys. Rev. Lett.* **119** 050504
- [18] Oestereich A L and Galvão E F 2017 Reliable computation from contextual correlations *Phys. Rev. A* **96** 062305
- [19] Markiewicz M, Przysiężna A, Brierley S and Paterek T 2014 Genuinely multipoint temporal quantum correlations and universal measurement-based quantum computing *Phys. Rev. A* **89** 062319
- [20] Kleinmann M, Gühne O, Portillo J R, Larsson J-Å and Cabello A 2011 Memory cost of quantum contextuality *New J. Phys.* **13** 113011
- [21] Fagundes G and Kleinmann M 2017 Memory cost for simulating all quantum correlations from the Peres–Mermin scenario *J. Phys. A: Math. Theor.* **50** 325302

- [22] Garner A J P, Liu Q, Thompson J, Vedral V and Gu M 2017 Provably unbounded memory advantage in stochastic simulation using quantum mechanics *New J. Phys.* **19** 103009
- [23] Thompson J, Garner A J P, Mahoney J R, Crutchfield J P, Vedral V and Gu M 2018 Causal asymmetry in a quantum world *Phys. Rev. X* **8** 031013
- [24] Pironio S 2003 Violations of bell inequalities as lower bounds on the communication cost of nonlocal correlations *Phys. Rev. A* **68** 062102
- [25] Montina A and Wolf S 2016 Information-based measure of nonlocality *New J. Phys.* **18** 013035
- [26] Briertley S, Kosowski A, Markiewicz M, Paterek T and Przysiężna A 2015 Nonclassicality of temporal correlations *Phys. Rev. Lett.* **115** 120404
- [27] Żukowski M 2014 Temporal inequalities for sequential multi-time actions in quantum information processing *Frontiers Phys.* **9** 629–33
- [28] Gallego R, Brunner N, Hadley C and Acín A 2010 Device-independent tests of classical and quantum dimensions *Phys. Rev. Lett.* **105** 230501
- [29] Brunner N, Navascués M and Vértesi T 2013 Dimension witnesses and quantum state discrimination *Phys. Rev. Lett.* **110** 150501
- [30] Dall’Arno M, Brandsen S and Buscemi F 2017 Device-independent tests of quantum channels *Proc. R. Soc. A* **473** 20160721
- [31] Dall’Arno M, Brandsen S, Buscemi F and Vedral V 2017 Device-independent tests of quantum measurements *Phys. Rev. Lett.* **118** 250501
- [32] Rosset D, Buscemi F and Liang Y-C 2018 Resource theory of quantum memories and their faithful verification with minimal assumptions *Phys. Rev. X* **8** 021033
- [33] Bowles J, Brunner N and Pawłowski M 2015 Testing dimension and nonclassicality in communication networks *Phys. Rev. A* **92** 022351
- [34] Tavakoli A, Marques B, Pawłowski M and Bourennane M 2016 Spatial versus sequential correlations for random access coding *Phys. Rev. A* **93** 032336
- [35] Aguilar E A, Farkas M, Martínez D, Alvarado M, Cariñe J, Xavier G B, Barra J F, Cañas G, Pawłowski M and Lima G 2018 Certifying an irreducible 1024-dimensional photonic state using refined dimension witnesses *Phys. Rev. Lett.* **120** 230503
- [36] Ludwig G 1985 *An Axiomatic Basis for Quantum Mechanics* vol 1 (Berlin: Springer)
- [37] Mittelstaedt P 1998 *The Interpretation of Quantum Mechanics and the Measurement Process* (Cambridge: Cambridge University Press)
- [38] Chiribella G, D’Ariano G M and Perinotti P 2010 Probabilistic theories with purification *Phys. Rev. A* **81** 062348
- [39] Acín A, Augusiak R, Cavalcanti D, Hadley C, Korbicz J K, Lewenstein M, Masanes L and Piani M 2010 Unified framework for correlations in terms of local quantum observables *Phys. Rev. Lett.* **104** 140404
- [40] Clemente L and Kofler J 2016 No fine theorem for macrorealism: limitations of the Leggett–Garg inequality *Phys. Rev. Lett.* **116** 150401
- [41] Popescu S and Rohrlich D 1994 Quantum nonlocality as an axiom *Found. Phys.* **24** 379–85
- [42] Abbott A A, Giarmatzi C, Costa F and Branciard C 2016 Multipartite causal correlations: polytopes and inequalities *Phys. Rev. A* **94** 032131
- [43] Hoffmann J 2016 Temporal correlations in quantum theory *MSc Thesis* University of Siegen
- [44] Hoffmann J, Spee C, Gühne O and Budroni C 2018 Structure of temporal correlations of a qubit *New J. Phys.* **20** 102001
- [45] Masanes L and Müller M P 2011 A derivation of quantum theory from physical requirements *New J. Phys.* **13** 063001
- [46] Kleinmann M 2014 Sequences of projective measurements in generalized probabilistic models *J. Phys. A: Math. Theor.* **47** 455304
- [47] Paz A 2003 *Probabilistic Automata* (New York: Wiley)
- [48] Heinosaari T and Ziman M 2011 *The Mathematical Language of Quantum Theory: From Uncertainty to Entanglement* (Cambridge: Cambridge University Press)
- [49] Fritz T 2010 Quantum correlations in the temporal Clauser–Horne–Shimony–Holt (CHSH) scenario *New J. Phys.* **12** 083055
- [50] Budroni C, Moroder T, Kleinmann M and Gühne O 2013 Bounding temporal quantum correlations *Phys. Rev. Lett.* **111** 020403
- [51] Budroni C and Emary C 2014 Temporal quantum correlations and Leggett–Garg inequalities in multilevel systems *Phys. Rev. Lett.* **113** 050401
- [52] Pawłowski M and Winter A 2012 ‘Hyperbits’: the information quasiparticles *Phys. Rev. A* **85** 022331
- [53] Barrett J 2007 Information processing in generalized probabilistic theories *Phys. Rev. A* **75** 032304
- [54] Spee C, Siebeneich H, Gloger T F, Kaufmann P, Johanning M, Wunderlich C, Kleinmann M and Gühne O 2018 *Genuine Temporal Correlations can Certify the Quantum Dimension* arXiv:1811.12259
- [55] Lasserre J B 2001 Global optimization with polynomials and the problem of moments *SIAM J. Optim.* **11** 796–817
- [56] Vandenberghe L and Boyd S 1996 Semidefinite programming *SIAM Rev.* **38** 49–95
- [57] Peres A 1990 Incompatible results of quantum measurements *Phys. Lett. A* **151** 107–8
- [58] Mermin N D 1990 Simple unified form for the major no-hidden-variables theorems *Phys. Rev. Lett.* **65** 3373–6
- [59] Mansfield S and Kashefi E 2018 Quantum advantage from sequential transformation contextuality *Phys. Rev. Lett.* **121** 230401
- [60] Spekkens R W 2007 Evidence for the epistemic view of quantum states: a toy theory *Phys. Rev. A* **75** 032110
- [61] Johansson N and Larsson J-Å 2017 Efficient classical simulation of the Deutsch–Jozsa and Simon’s algorithms *Quantum Inf. Process.* **16** 233
- [62] Johansson N and Larsson J-Å 2017 *Realization of Shor’s Algorithm at Room Temperature* arXiv:1706.03215
- [63] Woods M P, Silva R, Pütz G, Stupar S and Renner R 2018 Quantum clocks are more accurate than classical ones arXiv:1806.00491 [quant-ph]