

Complexity of contextuality

Theodoros Yianni and Farid Shahandeh *Department of Computer Science, Royal Holloway, University of London, United Kingdom*

(Received 4 November 2025; accepted 23 February 2026; published 19 March 2026)

Generalized contextuality is a hallmark of nonclassical theories like quantum mechanics. Yet, three fundamental computational problems concerning its decidability and complexity remain open: first, determining the complexity of deciding if a theory admits a noncontextual ontological model; second, determining the complexity of deciding if such a model is possible for a specific dimension k ; and, third, efficiently computing the smallest such model when it exists, given that finding the smallest ontological model is NP-hard. We address the second problem by presenting an algorithm derived from a geometric formulation and its reduction to the intermediate simplex problem in computational geometry. We find that the complexity of deciding the existence of a noncontextual ontological model of dimension k is at least exponential in the dimension of the theory and at most exponential in k . This, in turn, implies that computing the smallest noncontextual ontological model is inefficient in general. Finally, we demonstrate the fundamental difference between finding the smallest noncontextual ontological model and the smallest ontological model using an explicit example wherein the respective minimum ontic sizes are five and four.

DOI: [10.1103/k9ls-xyps](https://doi.org/10.1103/k9ls-xyps)

I. INTRODUCTION

Quantum contextuality is the property that a quantum experiment cannot be modelled by assigning fixed hidden variable representations to its states and outcome effects [1–3]. This peculiar property is considered to be an important resource for quantum advantages in many tasks such as quantum computation [4–6], communication [7], optimal state discrimination [8,9], and quantum cloning [9,10].

Determining the size of the ontological (hidden-variable) models for quantum theory, and, more broadly, for any generalized probabilistic theory (GPT), is an active field of research [11–14]. Indeed, the principle of parsimony suggests that when searching for an ontological explanation of the data, the simplest model with the fewest ontic states is preferred. In particular, a large number of ontic states necessary for a model compared to the quantum dimensionality indicates quantum advantage, an example of which can be identified in the context of classical versus quantum communication [15]. Notably, when a noncontextual model exists, an upper bound on its size is the square of the GPT dimension [16]. The value of such bounds is that they can be used to explain nonclassical advantages of protocols under mild assumptions [9,13,17]. They also enable proofs of contextuality in cases where the minimum required ontic size exceeds the minimum size of noncontextual models. While it is possible to find the smallest noncontextual model, the currently available method to achieve this is to optimize over all solutions of a linear

program [18]. However, this is challenging because there may be an infinite number of such solutions, rendering the optimization highly inefficient.

With the introduction of generalized contextuality [3], there has been significant progress in theoretical and experimental methods to detect it [16,18,19]. These methods generally rely on linear programs, such as the one presented in Ref. [18], which detects contextuality in arbitrary GPTs. Alternatively, by considering non-negative matrix factorizations (NMFs) of the matrix of conditional outcome probability evaluations (COPE), one arrives at a criterion for noncontextuality, as introduced in Refs. [9,20]. More specifically, noncontextuality implies that the ranks of the COPE decomposition factors must be equal, the violation of which is called the *rank separation*. This allows us to detect contextuality by studying bounds on the ranks of the NMF factors. Here the COPE matrix is assumed to capture the complete probabilistic structure of the operational theory. Consequently, all operational equivalences are implicit in the COPE matrix [20]. Unlike other methods, the rank separation approach does not require explicit analysis of operational equivalences and operational identities.

In this work, we present a method for finding the smallest noncontextual model of the prepare-measure experiments, along with its time complexity. We achieve this by leveraging the rank separation criterion and providing a new geometric interpretation of contextuality. This geometric perspective is used to reduce the problem of finding the smallest noncontextual model to an NMF problem. The latter has a known complexity and can be solved with existing algorithms, such as Moitra’s almost optimal algorithm [21]. In particular, we show that for COPE matrices of fixed rank, the existence of a noncontextual model can be decided in polynomial time. However, for $m \times n$ matrices of arbitrary rank r , the complexity

is at least $(nm)^{O(r)}$ and at most $\text{poly}(b, m, n)^{O(k^2)}$ wherein b is the bit length of the matrix entries and k is the size of the ontological model.

It is already known that the smallest (not necessarily non-contextual) ontological model of a prepare-measure scenario is given by the NMF of its COPE matrix [9,20], the inner dimension k of which determines the cardinality of the ontic space. It is also known that performing an NMF of the smallest inner dimension is NP-hard [22], i.e., computing the smallest ontological model for a prepare-measure scenario is NP-hard in the worst case. An interesting question is thus whether the smallest ontological model and the smallest noncontextual ontological model coincide or a size gap between the two is possible. We present an explicit example of a COPE matrix where such a gap exists. This means that, in general, the two classes of ontological models are different in size, which opens up the possibility that different ontic sizes may be responsible for different types of quantum advantages.

This paper is organized as follows. In Sec. II we give an overview of the preliminary concepts and definitions, including COPE matrix formalism and its underlying models. In Sec. III, we reformulate the problem of finding the smallest noncontextual model as an NMF problem and give an upper bound for its complexity. In Sec. IV, we clarify the relationship between our approach and the existing embeddability results. Finally, in Sec. V we demonstrate an example where there is a size gap between the smallest ontological and non-contextual ontological models. Discussions and conclusions are presented in Sec. VI.

In what follows, the following notation is adopted. We use sans-serif capital Latin letters (**A**) to denote experimental procedures and outcomes. The generic real matrices are denoted by capital Latin letters (A), while bold capital Latin letters (**A**) are used to denote non-negative matrices. The symbol $\mathbb{1}$ represents the identity matrix. Vectors are denoted by bold italic lowercase Latin letters (**a**); vector spaces are denoted by Fraktur capital Latin letters ($\mathfrak{A}$); and $\mathbf{1}$ is used for the vector of ones whose length is determined by the context if not explicitly mentioned. In working with geometrical objects, calligraphic capital Latin letters ($\mathcal{A}$) are used to represent polytopes.

II. PRELIMINARIES

A. Conditional outcome probability evaluations matrix

The primitive elements of any physical experiment are the propositions describing each operational procedure. Those associated with preparing the system are called *preparations* and denoted by $\mathbf{P}$. For example, “turn the laser on AND adjust its intensity as such AND align its beam as such” is a preparation of a photonic system. Similarly, propositions such as “turn the detector on AND align it as such AND open its aperture” associated with measurements of the system are called *measurements*, and outcomes of measurements are *events*. We denote these by $\mathbf{M}$ and $\mathbf{E}$, respectively. For all practical purposes, we can assume that the sets of preparations $\{\mathbf{P}_i\}$, measurements $\{\mathbf{M}_j\}$, and events $\{\mathbf{E}_k\}$ are finite. In the most general scenario, we can associate conditional probabilities to events as $p(\mathbf{E}_k|\mathbf{P}_i, \mathbf{M}_j)$ indicating the probability of event $\mathbf{E}_k$

in the measurement $\mathbf{M}_j$ following the preparation $\mathbf{P}_i$. These data can be put into a matrix where each row represents a particular event, and each column represents a preparation,

$$\mathbf{C}^a := \begin{pmatrix} p(\mathbf{E}_1|\mathbf{P}_1, \mathbf{M}_1) & \cdots & p(\mathbf{E}_1|\mathbf{P}_n, \mathbf{M}_1) \\ \vdots & & \vdots \\ p(\mathbf{E}_{m_1}|\mathbf{P}_1, \mathbf{M}_1) & \cdots & p(\mathbf{E}_{m_1}|\mathbf{P}_n, \mathbf{M}_1) \\ \vdots & & \vdots \\ p(\mathbf{E}_1|\mathbf{P}_1, \mathbf{M}_l) & \cdots & p(\mathbf{E}_1|\mathbf{P}_n, \mathbf{M}_l) \\ \vdots & & \vdots \\ p(\mathbf{E}_{m_l}|\mathbf{P}_1, \mathbf{M}_l) & \cdots & p(\mathbf{E}_{m_l}|\mathbf{P}_n, \mathbf{M}_l) \end{pmatrix} = \begin{pmatrix} \mathbf{C}^{a1} \\ \vdots \\ \mathbf{C}^{a,l} \end{pmatrix}, \quad (1)$$

where we have assumed l measurements with m_j outcomes each and n preparations. Each submatrix $\mathbf{C}^{ai}$ corresponds to the measurement $\mathbf{M}_i$. We call the matrix of COPE [9,20]. We can also represent the same information about the system with another COPE as follows. Construct the measurement procedure $\mathbf{M}$ by choosing a measurement from $\{\mathbf{M}_i\}$ uniformly at random. The set of events for $\mathbf{M}$ is the union of the sets of events of the previous measurements. Thus $\mathbf{M}$ has $m = \sum_{j=1}^l m_j$ possible events, each with probability $p(\mathbf{E}_k|\mathbf{P}_i)$, where we have dropped the dependence on the measurement for brevity. The COPE matrix $\mathbf{C}^b$ obtained from using only the measurement $\mathbf{M}$ and the same set of preparations is given by

$$\mathbf{C}^b := \begin{pmatrix} p(\mathbf{E}_1|\mathbf{P}_1) & \cdots & p(\mathbf{E}_1|\mathbf{P}_n) \\ \vdots & & \vdots \\ p(\mathbf{E}_m|\mathbf{P}_1) & \cdots & p(\mathbf{E}_m|\mathbf{P}_n) \end{pmatrix}. \quad (2)$$

This matrix is a rescaling of $\mathbf{C}^a$ making it column stochastic as shown in Refs. [9,20].

B. Ontological model

One may want to construct an ontological model that explains the statistics of an experiment on a system. Such a model consists of a set of ontic states indexed by λ that represent the different mutually exclusive definite states the system can be in. The set of all ontic states is called the ontic space, which is conventionally denoted by Λ and its cardinality is $|\Lambda|$.

In general, a preparation $\mathbf{P}$ is modelled in an ontological model as some epistemic state $\boldsymbol{\mu}^{(\mathbf{P})}$. An epistemic state is a stochastic vector $(\mu_\lambda)_\lambda$, where μ_λ is the probability of the system being in the ontic state λ . A measurement, on the other hand, is given by a collection of response functions $\{\xi^i\}$ where each ξ^i corresponds to a different event. The response functions for a measurement are vectors with entries in $[0,1]$. The response functions sum to the unit response function $\mathbf{1}$.

The probability of the event $\mathbf{E}_k$ corresponding to the response function ξ^k of $\mathbf{M}$ given that the preparation $\mathbf{P}$ gives λ as a complete ontic description of the state of the system is

$$p(\mathbf{E}_k|\mathbf{P}) = \langle \xi^k, \delta^\lambda \rangle. \quad (3)$$

Here $\delta^\lambda = (\delta^\lambda_\lambda)_\lambda$ with δ^λ_j being the Kronecker delta and $\langle \cdot, \cdot \rangle$ is the inner product over $\mathbb{R}^{|\Lambda|}$. If, instead, our knowledge of

the true state of the system is limited such that we can only describe it as an epistemic state $\mu^{(P)}$, then the probability of the same event E_k conditioned on this preparation of the system is given by

$$p(E_k|P) = \langle \xi^k, \mu^{(P)} \rangle. \quad (4)$$

The stochastic property of the epistemic state and the unit response function guarantee that $\langle \mu^{(P)}, \mathbf{1} \rangle = 1$.

An ontological model of a COPE $\mathbf{C}^a$ composed of l column-stochastic submatrices as in Eq. (1) is given by an NMF of the form $\mathbf{C}^a = \mathbf{R}\mathbf{E}$, where the factors $\mathbf{R}$ and $\mathbf{E}$ are entrywise non-negative. If $\mathbf{R}$ and $\mathbf{E}$ are column sum-1 and column stochastic, respectively, then the columns of $\mathbf{E}$ would have the form of epistemic states, and the rows of $\mathbf{R}$ sum to $\mathbf{1}$. It can always be assumed that these conditions are met because, if not, they can be achieved using a diagonal scaling as

$$\mathbf{C}^a = \mathbf{R}'\mathbf{E}' = (\mathbf{R}'\mathbf{D}^{-1})(\mathbf{D}\mathbf{E}') = \mathbf{R}\mathbf{E}, \quad (5)$$

where $\mathbf{D}^{-1}$ is a diagonal matrix that rescales the columns of $\mathbf{R}'$, and we assume $\mathbf{C}^b$ has no zero rows or columns. Since $\mathbf{C}^a$ and $\mathbf{R}$ are column sum-1, $\mathbf{E}$ is column stochastic. The non-negative rank (NNR) of $\mathbf{C}^a$, i.e., the smallest inner dimension k at which an NMF is possible, determines the size of the smallest ontological model of $\mathbf{C}^a$. Notice that by simply rescaling $\mathbf{R}$, we also get an ontological model of $\mathbf{C}^a$ with the same ontic size, which is why both forms of the COPE matrix can be used to describe the same system.

C. Noncontextual ontological model

The statistics of any physical system can be reproduced by an ontological model. However, a natural criterion for a system to be considered *classically* explainable is that such a model is *noncontextual*. In this work, we adopt the concept of generalized *noncontextuality*, as introduced in Ref. [3]. This framework relies on the operational equivalence of experimental procedures to define contextuality.

The operational equivalence ($\simeq$) of preparations (measurement events) means that there are no measurement events (preparations) in the operational theory that can distinguish them [3], i.e.,

$$\begin{aligned} P_i \simeq P_j & \text{ iff } p(E_k|P_i, M) = p(E_k|P_j, M) \forall E_k, M, \\ E_k \simeq E_l & \text{ iff } p(E_k|P, M) = p(E_l|P, M) \forall P. \end{aligned} \quad (6)$$

Motivated by Leibniz's principle of the (ontological) identity of empirical indiscernibles [23], we may require any model of the data to satisfy the *broad noncontextuality hypothesis* [24]. The latter suggests that our models of physical phenomena should rely on equivalence classes instead of individual contexts. Then, we can define the preparation (measurement) contextuality of an ontological model as the property that there are at least two different preparations (measurement events) that are operationally equivalent but have different ontic representations. Conversely, in a noncontextual model, operationally equivalent procedures are represented by the

same function over the ontic space, i.e.,

$$\begin{aligned} P_1 \simeq P_2 & \Leftrightarrow \mu^{(P_1)} = \mu^{(P_2)}, \\ E_1 \simeq E_2 & \Leftrightarrow \xi^{(E_1)} = \xi^{(E_2)}, \end{aligned} \quad (7)$$

where $\mu^{(P_i)}$ is the epistemic state of the system prepared by P_i and $\xi^{(E_i)}$ is the response function corresponding to the event E_i ($i = 1, 2$).

Let us show how these definitions translate to the COPE picture via an example. Consider the COPE matrix of a GPT known as the *box world*,

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}. \quad (8)$$

It can easily be checked that an NMF of this matrix with inner dimension three does not exist. Hence, it does not admit an ontological model of dimension three or less. The smallest ontological model possible is thus given as

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \mathbf{R}\mathbf{E}, \quad (9)$$

which shows a trivial NMF as $\mathbf{E} = \mathbf{1}$. However, this model is contextual because there are two operationally indistinguishable preparations, which are represented with two different vectors in the ontological model. Specifically, the average of the first and last column of the COPE matrix and the average of its second and third column are equal. However, the average of the first and last column of the epistemic state matrix $\mathbf{E}$ is not equal to the average of its second and third column. Therefore, at least two epistemic states exist in the model that the response functions cannot distinguish. The latter is formally reflected in the rank separation of the matrices, that is, $\text{rank } \mathbf{R} \neq \text{rank } \mathbf{E}$.

As shown in the example above, the representations of the response functions (epistemic states) in a model might not span the full space of ontic states. Hence, in a contextual model, there can be two epistemic states (response functions) whose difference is orthogonal to the span of the response functions (epistemic states), so that they cannot be distinguished. This leads us to the following lemma and corollary from Refs. [9,20].

Lemma 1. An ontological model is preparation (measurement) noncontextual if and only if its set of response functions (epistemic states) separates the points in the linear span of its epistemic states (response functions).

Corollary 1. A noncontextual operational theory with a COPE $\mathbf{C}$ admits an NMF $\mathbf{C} = \mathbf{R}\mathbf{E}$ such that

$$\text{rank } \mathbf{C} = \text{rank } \mathbf{R} = \text{rank } \mathbf{E}. \quad (10)$$

We call such a factorization a *equirank non-negative matrix factorization* (ENMF). Furthermore, we define the *equirank non-negative rank* (ENNR) of $\mathbf{C}$ as the smallest inner dimension where an ENMF is possible, which is equal to the smallest possible ontic size of a noncontextual model for $\mathbf{C}$.

The advantage of recasting noncontextuality in terms of matrix factorizations is that it does not require operational

equivalences and operational identities to be explicitly stated. It also allows us to marry computational problems concerning contextuality with problems and results from linear algebra and computational complexity. For example, we obtain our first complexity result as follows.

Corollary 2. Assuming the exponential time hypothesis, the complexity of computing the size of the smallest noncontextual model of an $n \times m$ COPE of rank r is at least $(nm)^{O(r)}$.

Proof. By computing the size of the smallest noncontextual model, by extension, it is decided whether an ENMF of inner dimension r is possible. This is equivalent to deciding if an NMF of inner dimension r is possible, because any NMF of inner dimension r is also an ENMF. The complexity of the NMF decision is known to be at least $(nm)^{O(r)}$ [25], assuming the exponential time hypothesis [26]. ■

In this paper, we are concerned with a geometric interpretation of ENMFs and the complexity of determining the ENNR of operational theories.

D. Generalized probabilistic theories

A GPT is an alternative model that can be used to explain the observed outcome statistics of experiments. It consists of a set of states $\mathcal{S} := \{s^i\}_i$ which lie in a real vector space $\mathfrak{V}$ and a set of effects $\mathcal{E} := \{e^i, \mathbf{u}\}_i$, with $\mathbf{u}$ being the unit effect, lying in the dual vector space $\mathfrak{V}^*$. Here the dual of a polytope is defined in the following way. Given a polytope $\mathcal{K}$ in a real inner product space $(\mathfrak{V}, \langle \cdot, \cdot \rangle_{\mathfrak{V}})$, its dual $\mathcal{K}^*$ is given by

$$\mathcal{K}^* := \{x \in \mathfrak{V} | \langle x, s \rangle \in [0, 1] \forall s \in \mathcal{K}\}. \quad (11)$$

In finite dimensions, thanks to the Riesz theorem, there is an isomorphism between $\mathfrak{V}^*$ and $\mathfrak{V}$, so we can assume that the set of effects is also in $\mathfrak{V}$.

Within the GPT framework, a state is a representation of a class of operationally equivalent preparation procedures, and an effect is a representation of a class of operationally equivalent events. The effects in a measurement sum to the unit effect $\mathbf{u}$.

Let e^j be the effect corresponding to the class $[E_j]$ of events operationally equivalent to the j th outcome of a particular measurement M . Then the probability of observing this event given that the system was prepared using $[P_i]$, i.e., any preparation procedure operationally equivalent to P_i , is given by

$$p([E_j] | [P_i], [M]) = \langle e^j, s^i \rangle. \quad (12)$$

Here $\langle \cdot, \cdot \rangle$ refers to the inner product over $\mathfrak{V}$. Since the event probabilities of a particular measurement must sum to 1, the states should be normalized as

$$\langle s^i, \mathbf{u} \rangle = 1. \quad (13)$$

A GPT model of a COPE $\mathbf{C}^a$ can be found by simply finding a real factorization of the form,

$$\mathbf{C}^a = AB, \quad (14)$$

where $A \in \mathbb{R}^{m \times k}$, $B \in \mathbb{R}^{k \times n}$ is column stochastic, and they are such that $\text{rank } \mathbf{C}^a = \text{rank } A = \text{rank } B$. Note that, because $\mathbf{C}^a$ has no zero columns, we can rescale any decomposition using a diagonal matrix D as $\mathbf{C}^a = A'B' = (A'D^{-1})(DB') = AB$ so that B is column stochastic. Also, note that since $\mathbf{C}^a$ represents

an experiment with l measurements, it is divided into submatrices C^{ai} , each corresponding to a particular measurement, as illustrated in Eq. (1). These submatrices are each column stochastic, and, therefore, the sum of rows of the corresponding submatrices of A, A^i , are equal.

Now we have a GPT for the COPE where each row of A is an effect vector and the rows within a subset of rows corresponding to a particular measurement sum to the unit effect $\mathbf{u}$, and each column of B is a normalized state vector. Such a factorization of $\mathbf{C}$ gives the simplest (lowest-dimensional) GPT that can reproduce it whenever the inner dimension k of the decomposition equals the rank of $\mathbf{C}$. The equality between ranks guarantees that every equivalence class is represented by a unique vector in the GPT as required.

III. GEOMETRIC INTERPRETATION OF ENMF

It is easy to visualize any non-negative $m \times n$ matrix geometrically: Simply interpret the columns as n vectors in the positive orthant of an m -dimensional Euclidean space, corresponding to the vertices of some geometrical body. In particular, if the columns are convexly independent, then the vertices will form the extreme points of a bounded convex polytope. Building on this simple intuition, one can also devise different geometrical interpretations of the decompositions of such matrices, including NMF and ENMF. In this section, we provide two such interpretations for ENMF. This leads to an algorithm for deciding whether a COPE admits a noncontextual model, as well as an algorithm for finding its smallest noncontextual model.

To do this, we first consider how the matrix factors of the COPE can be represented as polytopes.

Theorem 1. Let $\mathbf{C}$ be an $m \times n$ COPE matrix of rank r , which has l measurements. Let $\mathbf{C} = AB$ be a real matrix factorization of $\mathbf{C}$ with inner dimension r . Let $\mathcal{A}$ be the polyhedron $\mathcal{A} := \{x \in \mathbb{R}^r | Ax \geq 0, \mathbf{1}^\top Ax = l\}$, $\mathcal{B}$ the convex hull of the columns of B , and $\mathcal{B}^* := \{x \in \mathbb{R}^r | x^\top B \leq 1\}$ the dual polytope of $\mathcal{B}$. Then $\mathcal{B} \subseteq \mathcal{A} \subseteq \mathcal{B}^*$. Furthermore, $\mathcal{A}$ is bounded.

Proof. Since $AB = \mathbf{C} \geq 0$ and $\mathbf{C}$ is divided into l column-stochastic subsets of rows, each column of B satisfies $AB_{\cdot i} \geq 0$ and $\mathbf{1}^\top AB_{\cdot i} = l$, and therefore it is contained in $\mathcal{A}$. Furthermore, for any such subset $\mathbf{C}^i$ of rows of $\mathbf{C}$, the rows within corresponding subset of rows A^i of A sum to the unit effect $\mathbf{u}$. Now, we note that for any $x \in \mathcal{A}$, $\mathbf{1}^\top Ax = l$, and $\mathbf{1}^\top A = l\mathbf{u}^\top$, since there are l measurements and the sum of effects within each measurement is $\mathbf{u}$. It follows that $\mathbf{u}^\top x = 1$. Therefore, given that $Ax \geq 0$, we have $0 \leq Ax \leq 1$. This implies that $\mathcal{A}$ is contained in $\mathcal{B}^*$. ■

We observe that $\mathbf{C}_{ij}$ is the slack of vertices of $\mathcal{B}$ with respect to the hyperplanes of $\mathcal{A}$. Also, $\mathcal{A}$ is the set of all logically possible normalized states given the effects.

Given the non-negative matrix $\mathbf{C}$, the following two problems are also shown to be equivalent [22]:

(1) Decide if the non-negative rank and the rank of $\mathbf{C}$ are equal.

(2) Decide if a nested simplex $\mathcal{G}$ exists such that $\mathcal{B} \subseteq \mathcal{G} \subseteq \mathcal{A}$. Here the polytopes $\mathcal{A}$ and $\mathcal{B}$ are defined via a decomposition of the form given in Theorem 1.

We generalize this result to map the problem of deciding whether an ENMF of inner dimension k of $\mathbf{C}$ exists to a nested simplex problem.

Lemma 2. Let $\mathbf{C}$ be a non-negative $m \times n$ matrix of rank r whose columns each sum to l , and suppose that $\mathbf{C} = \mathbf{A}\mathbf{B}$ is a real matrix factorization of $\mathbf{C}$ with inner dimension r . Define the polytopes $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^r$ from $\mathbf{B}$ and $\mathbf{A}$ as in Theorem 1 so that $\mathcal{B} \subseteq \mathcal{A}$. There exists an ENMF of $\mathbf{C}$ of inner dimension k if and only if there exists a polytope $\mathcal{B} \subseteq \mathcal{G} \subseteq \mathcal{A}$ with k vertices, and a rank- r non-negative matrix $\mathbf{E}$ of convex coefficients that maps the matrix of vertices of $\mathcal{G}$ to the matrix of vertices of $\mathcal{B}$.

Proof. We first show the only if direction. When an ENMF $\mathbf{C} = \mathbf{R}\mathbf{E}$ of inner dimension k exists, there exist rank- r matrices $\mathbf{G} \in \mathbb{R}^{r \times k}$ and $\mathbf{F} \in \mathbb{R}^{k \times r}$ such that $\mathbf{R} = \mathbf{A}\mathbf{G}$ and $\mathbf{E} = \mathbf{F}\mathbf{B}$. Without loss of generality, we assume that the columns of $\mathbf{R}$ each sum to l and that $\mathbf{E}$ is column stochastic. We can assume that $\mathbf{F}$ and $\mathbf{G}$ have no zero rows and columns, respectively. This is because if the j th row of $\mathbf{F}$ is zero, then the j th row of $\mathbf{E} = \mathbf{F}\mathbf{B}$ will be zero. The j th column of $\mathbf{G}$ can also be set to zero, making the j th column of $\mathbf{R} = \mathbf{A}\mathbf{G}$ zero without changing the product $\mathbf{A}\mathbf{G}\mathbf{F}\mathbf{B}$. This means, in turn, that we can remove the j th row of $\mathbf{E}$ and the j th column of $\mathbf{R}$ to obtain an ENMF in $k - 1$ dimensions.

Since $\mathbf{R} = \mathbf{A}\mathbf{G}$ is non-negative and has a column sum of l , the columns of $\mathbf{G}$ are in $\mathcal{A}$, and therefore we construct the polytope $\mathcal{G}$ as the convex hull of the columns of $\mathbf{G}$. Given that $\mathbf{A}$ and $\mathbf{B}$ are full rank, $\mathbf{A}\mathbf{G}\mathbf{F}\mathbf{B} = \mathbf{A}\mathbf{B}$ implies that $\mathbf{G}\mathbf{F}\mathbf{B} = \mathbf{B}$. Since $\mathbf{E}$ is column stochastic, the columns of $\mathbf{B}$ are given by convex combinations of the columns of $\mathbf{G}$. Equivalently, $\mathcal{B} \subseteq \mathcal{G}$.

Furthermore, since $\mathbf{R}$ and $\mathbf{E}$ have factors of $\mathbf{A}$ and $\mathbf{B}$, respectively, $\text{rank}(\mathbf{R}), \text{rank}(\mathbf{E}) \leq r$. Then $\mathbf{R}\mathbf{E} = \mathbf{C}$ implies that $\text{rank}(\mathbf{R})\text{rank}(\mathbf{E}) \geq \text{rank}(\mathbf{C}) = r$, and therefore $\text{rank}(\mathbf{R}) = \text{rank}(\mathbf{E}) = r$.

For the reverse direction, suppose there exists a nested polytope $\mathcal{G}$ satisfying $\mathcal{B} \subseteq \mathcal{G} \subseteq \mathcal{A}$, of k vertices, and a rank- r column-stochastic matrix $\mathbf{E}$ that maps the vertices of $\mathcal{G}$ to the vertices of $\mathcal{B}$. Given $\mathcal{G}$ and $\mathcal{B}$ have associated matrices $\mathbf{G}$ and $\mathbf{B}$, respectively, we have $\mathbf{G}\mathbf{E} = \mathbf{B}$ (up to a permutation of the columns of $\mathbf{B}$, which can be arbitrarily changed by permuting the rows of $\mathbf{E}$). Defining $\mathbf{B}^\dagger$ as the pseudoinverse of $\mathbf{B}$, we can define $\mathbf{F} := \mathbf{E}\mathbf{B}^\dagger$. ■

Lemma 2 boils the complexity of ENMF problem down to two parts. First, the complexity of finding the polytope $\mathcal{G}$. Second, the complexity of deciding whether a convex coefficient matrix $\mathbf{R}$ of rank r exists and its computation. Note that, while a non-negative transformation from $\mathbf{G}$ to $\mathbf{A}$ can be found using methods such as non-negative least squares, such generic approaches do not provide rank guarantees. We thus begin by showing that our desired transformation can be decided and found in polynomial time.

Lemma 3. Let $\mathbf{G} \in \mathbb{R}^{r \times k}$ and $\mathbf{B} \in \mathbb{R}^{r \times m}$ be rank- r matrices corresponding to polytopes $\mathcal{G}$ and $\mathcal{B}$, respectively, such that $\mathcal{B} \subseteq \mathcal{G}$. For a bit length b for specifying the entries of $\mathbf{G}$ and $\mathbf{B}$, it can be decided in $\text{poly}(r, k, m, b)$ time whether there exists a rank- r non-negative matrix $\mathbf{E}$ such that $\mathbf{G}\mathbf{E} = \mathbf{B}$. When yes, $\mathbf{E}$ can be found within an element-wise δ error in $\text{poly}(r, k, m, b, \log(\delta^{-1}))$ time.

Proof. Let $\mathbf{G}^\dagger$ be the pseudoinverse of $\mathbf{G}$. Let $\bar{\mathbf{E}} := \mathbf{G}^\dagger \mathbf{B}$. If $\bar{\mathbf{E}}$ is non-negative, then $\mathbf{E} = \bar{\mathbf{E}}$. If not, then we search for

a shear transformation L of $\bar{\mathbf{E}}$ such that $L\bar{\mathbf{E}} \geq 0$ and $\mathbf{G}L\bar{\mathbf{E}} = \mathbf{B}$. The shear matrix L must be of the form,

$$L = \mathbf{1} + \sum_{i=1}^r \sum_{j=1}^{k-r} D_{ij} \mathbf{b}^j \mathbf{a}^{i\top}, \quad (15)$$

wherein $\{\mathbf{a}^{1\top}, \dots, \mathbf{a}^{r\top}\}$ is an orthonormal basis for the rows of $\mathbf{G}$, $\{\mathbf{b}^1, \dots, \mathbf{b}^{(k-r)}\}$ is an orthonormal basis for the right kernel of $\mathbf{G}$, and D is a real matrix to be found via optimization.

Define the objective function $f(D)$ as the modulus of the sum of negative entries in $L\bar{\mathbf{E}}$. Minimizing $f(D)$ leads to a solution when $f(D) = 0$. In Appendix A, we give the details of a linear program for optimizing D . We show that the number of constraints and search space size are polynomial in m, r , and k . The respective complexities of deciding if a solution to a linear program exists, and computing it when it does, are well known to be $\text{poly}(r, k, m, b)$ and $\text{poly}(r, k, m, b, \log(\delta^{-1}))$, respectively [27,28]. ■

Corollary 3. Given $m \times n$ non-negative matrices $\mathbf{C}$ of a fixed rank r , it can be decided in $\text{poly}(m, n, b)$ time whether an ENMF of unspecified inner dimension exists. Here b is the bit length for specifying the entries of $\mathbf{C}$.

Proof. First, find a rank factorization of $\mathbf{C}$ and the corresponding inner and outer polytopes $\mathcal{B}$ and $\mathcal{A}$ as in Lemma 2. The number of vertices of $\mathcal{A}$, which has m facets, as well as the complexity of finding these vertices from the facets, is polynomial in m . This polynomial is of degree $r - 1$ as shown in Sec. 5 of Ref. [29]. Note that $\mathcal{B} \subseteq \mathcal{G} \subseteq \mathcal{A}$. Hence, $\mathcal{G} = \mathcal{A}$ is a possible nested polytope. Then, using Lemma 3, where we choose k to be the number of vertices of $\mathcal{A}$, which is polynomial in n for fixed r , a rank- r non-negative matrix $\mathbf{E}$ that maps the vertices of $\mathcal{A}$ to the vertices of $\mathcal{B}$ can be decided and computed in polynomial time. Since $\mathcal{A}$ is the largest possible nested polytope, such a matrix exists if and only if an ENMF exists. ■

Corollary 3 might tempt us to think that deciding whether an ENMF, and thus a noncontextual ontological model, of a given COPE matrix (without specifying the inner dimension k) exists is efficient. For one thing, according to the proof, we can choose the trivial nested polytope $\mathcal{G} = \mathcal{A}$ and check the existence of $\mathbf{E}$ in polynomial time. However, we emphasize that this observation is only valid if the rank of the input COPE matrices is constant. We believe that the complexity of this simplest check grows exponentially with the rank. The reason is that checking the existence of $\mathbf{E}$ requires specifying *vertices* of $\mathcal{A}$ (known as the *V-representation*), while the description of $\mathcal{A}$ is in terms of its hyperplanes (known as the *H-representation*). The problem of finding the V-representation of a polytope from its H-representation is called the *vertex enumeration problem*. While this is known to be polynomial in special cases, e.g., for fixed dimension or if the polytope is simple or simplicial, the problem is believed to be exponential on general instances because the number of vertices grows exponentially with the number of defining hyperplanes. In particular, the problem is known to be NP-hard for unbounded polyhedra [30].

The problem of finding a polytope $\mathcal{G}$ with k vertices nested between two polytopes $\mathcal{B}$ and $\mathcal{A}$ is well known in computational geometry as the *nested polytope problem* (NPP). A natural approach thus would be to use algorithms solving the

general NPP to find $\mathcal{G}$. Unfortunately, this is not possible for two reasons. First, it is known that NPP is NP-hard, and in fact, $\exists\mathbb{R}$ -complete, in dimensions greater than two [31,32]. Second, when a solution for $\mathcal{G}$ with a specified k exists, it is not guaranteed that the solution's vertices can be mapped to $\mathcal{B}$ by a rank- r non-negative matrix $\mathbf{E}$. This extra restriction calls for devising a task-specific algorithm. In what follows, we will provide such an algorithm through a geometric interpretation.

Lemma 4. Let $\mathbf{C}$ be a non-negative $m \times n$ matrix of rank r whose columns each sum to l and $\mathbf{C} = \mathbf{A}\mathbf{B}$ a real matrix factorization of $\mathbf{C}$ with inner dimension r . For a given integer k , construct $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times k}$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{k \times n}$ by appending the $m \times (k-r)$ and $(k-r) \times n$ zero matrices to the left of $\mathbf{A}$ and top of $\mathbf{B}$, respectively. Define the unbounded polyhedral cone $\tilde{\mathcal{A}}$ as the set $\{\mathbf{x} \in \mathbb{R}^k | \mathbf{A}\mathbf{x} \geq 0, \mathbf{1}^\top \mathbf{A}\mathbf{x} = l\}$. $\tilde{\mathcal{B}}$ is the corresponding polytope of $\tilde{\mathbf{B}}$. There is a solution to ENMF of $\mathbf{C}$ of inner dimension k if and only if there is a $(k-1)$ -simplex $\tilde{\mathcal{G}}$ such that $\tilde{\mathcal{B}} \subseteq \tilde{\mathcal{G}} \subseteq \tilde{\mathcal{A}}$.

Proof. Let $\mathbf{C} = (\mathbf{A}\mathbf{G})(\mathbf{F}\mathbf{B})$ be a solution to the ENMF of $\mathbf{C}$ of inner dimension k as in Lemma 2, where $\mathbf{R} := \mathbf{A}\mathbf{G}$ and $\mathbf{E} = \mathbf{F}\mathbf{B}$ are column sum- l and column stochastic, respectively. Note that G is not square in general; hence, it does not represent a simplex, which is a full-dimensional bounded convex body. Thus, we aim to show that there exists a full-rank square matrix $\tilde{G}'$ such that $\mathbf{C} = \tilde{\mathbf{A}}\tilde{G}'\tilde{B}$. Then, since $\tilde{G}'^{-1}$ is full-rank (k) and $\tilde{\mathbf{B}}$ is rank- r , we have $\mathbf{E} = \tilde{G}'^{-1}\tilde{\mathbf{B}}$ which is rank- r and non-negative.

We begin by observing that due to the zeros in $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$, we can append two $(k-r) \times k$ and $k \times (k-r)$ random real matrices to the top of G and left of F , respectively, so that $\tilde{\mathbf{A}}\tilde{G}\tilde{\mathbf{F}}\tilde{\mathbf{B}} = \mathbf{C}$. Clearly, we also have $\tilde{\mathbf{A}}\tilde{G} \geq 0$ and $\tilde{\mathbf{F}}\tilde{\mathbf{B}} \geq 0$.

Furthermore, $\tilde{G}\tilde{\mathbf{F}}\tilde{\mathbf{B}}$ can only differ from $\tilde{\mathbf{B}}$ in its first $(k-r)$ rows because $\tilde{\mathbf{A}}\tilde{G}\tilde{\mathbf{F}}\tilde{\mathbf{B}} = \mathbf{C}$, the first $(k-r)$ rows of $\tilde{\mathbf{B}}$ are all zeros, and the last r columns and rows of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$, respectively, give a rank factorization of $\mathbf{C}$. Therefore, there exists a full-rank shear matrix L of the form,

$$L = \mathbb{1} + \sum_{i=1}^{k-r} \sum_{j=k-r+1}^k (D_{ij} \mathbf{a}^i (\mathbf{a}^j)^\top), \quad (16)$$

where $\mathbf{a}^i$'s are the columns of $\mathbb{1}_{k \times k}$ and D is a real matrix, such that $\tilde{G}\tilde{\mathbf{F}}\tilde{\mathbf{B}} = L\tilde{\mathbf{B}}$. The inverse of L is given by

$$L^{-1} = \mathbb{1} - \sum_{i=1}^{k-r} \sum_{j=k-r+1}^k (D_{ij} \mathbf{a}^i (\mathbf{a}^j)^\top). \quad (17)$$

We have

$$\begin{aligned} \mathbf{C} &= \tilde{\mathbf{A}}\tilde{G}\tilde{G}^{-1}\tilde{G}\tilde{\mathbf{F}}\tilde{\mathbf{B}}, \\ &= \tilde{\mathbf{A}}\tilde{G}\tilde{G}^{-1}L\tilde{\mathbf{B}}. \end{aligned} \quad (18)$$

Using Eq. (17), $\tilde{\mathbf{A}}\tilde{G} = \tilde{\mathbf{A}}L^{-1}\tilde{G}$ because $L^{-1}\tilde{G}$ only differs from $\tilde{G}$ in the first $(k-r)$ rows which leads to

$$\begin{aligned} \mathbf{C} &= \tilde{\mathbf{A}}L^{-1}\tilde{G}\tilde{G}^{-1}L\tilde{\mathbf{B}} \\ &= (\tilde{\mathbf{A}}\tilde{G}')(\tilde{G}'^{-1}\tilde{\mathbf{B}}), \\ \mathbf{R} &= \tilde{\mathbf{A}}\tilde{G} = \tilde{\mathbf{A}}\tilde{G}', \\ \mathbf{E} &= \tilde{\mathbf{F}}\tilde{\mathbf{B}} = \tilde{G}'^{-1}\tilde{\mathbf{B}}. \end{aligned} \quad (19)$$

Now $\tilde{G}'$ is square and full rank. Hence, its corresponding polytope $\tilde{\mathcal{G}}'$ is a $(k-1)$ -simplex. Furthermore, $\mathbf{R} = \tilde{\mathbf{A}}\tilde{G}'$ and $\mathbf{E} = \tilde{G}'^{-1}\tilde{\mathbf{B}}$ being non-negative and $\mathbf{E}$ being also column stochastic imply that $\tilde{\mathcal{G}}' \subseteq \tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}} \subseteq \tilde{\mathcal{G}}'$, respectively. Hence the claim.

For the reverse direction, assume there exists a full-dimensional nested $(k-1)$ -simplex $\mathcal{G}$ with a corresponding full-rank matrix G . Then $\mathcal{G} \subseteq \tilde{\mathcal{A}}$ implies that $\mathbf{R} := \tilde{\mathbf{A}}G$ is non-negative and has a column sum of l . $\tilde{\mathcal{B}} \subseteq \mathcal{G}$ implies that $\mathbf{E} := \tilde{\mathbf{A}}G^{-1}$ is a matrix of convex coefficients mapping $\mathcal{G}$ to $\tilde{\mathcal{A}}$, so it is non-negative and column stochastic. Finally, $\mathbf{E}$ is unique and rank $\mathbf{E} = \text{rank } \tilde{\mathbf{B}} = r$ because G and G^{-1} are full rank. ■

The advantage of the intermediate simplex approach of Lemma 4 is that, once found, it is guaranteed that the rank of $\mathbf{E}$ is r . As a result, following finding the intermediate simplex, one can use other efficient algorithms such as non-negative least squares to compute $\mathbf{E}$. The main question is thus, what is the complexity of the *intermediate simplex problem*? This was shown by Vavasis [22] to be NP-hard, assuming that the inner polytope is full dimensional and the outer polytope is bounded. In our construction in Lemma 4, however, neither of these conditions are met. Our next task is thus to transform this construction into one wherein $\tilde{\mathcal{B}}$ is full dimensional and $\tilde{\mathcal{A}}$ is bounded. However, we first show that the intermediate simplex $\mathcal{G}$ of Lemma 4 can always be assumed to have a specific orientation in the subspace orthogonal to the affine span of the inner polytope $\tilde{\mathcal{B}}$.

Lemma 5. For a rank- r non-negative matrix $\mathbf{C}$, let $\tilde{\mathcal{B}} \subseteq \tilde{\mathcal{G}} \subseteq \tilde{\mathcal{A}}$ be nested polytopes as in Lemma 4. Then transforming $\tilde{\mathcal{G}}$ by any orthogonal transformation acting on the subspace orthogonal to $\text{span}(\tilde{\mathcal{B}})$ also results in a nested simplex.

Proof. Such a transformation corresponds to multiplying the matrix G on the right by a matrix of the form,

$$O = \begin{pmatrix} O_{(k-r) \times (k-r)} & \mathbf{0} \\ \mathbf{0} & \mathbb{1} \end{pmatrix}, \quad (20)$$

where $O_{(k-r) \times (k-r)}$ is an orthogonal matrix. Note that, for any matrix $\mathbf{Q}$ such that the top $k-r$ rows of $G\mathbf{Q}$ are zeros, the top $k-r$ rows of $GO\mathbf{Q}$ will also be zeros. Then, since O acts as the identity on the subspace $\text{span}(\tilde{\mathcal{B}})$, $\mathbf{Q}G = \tilde{\mathcal{B}}$ would imply $\mathbf{Q}OG = \tilde{\mathcal{B}}$. ■

Theorem 2. For a rank- r non-negative matrix $\mathbf{C}$ with column sum l , let $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ be matrices as in Lemma 4. There exists an $(m+2k-2r) \times (n+k-r)$ rank- k non-negative matrix $\tilde{\mathbf{C}}$ for $k > r$ such that a ENMF of $\mathbf{C}$ of inner dimension k exists if and only if the NNR of $\tilde{\mathbf{C}}$ equals k .

Proof. Recall that $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ are the trivial embeddings of the $(r-1)$ -dimensional polytopes $\mathcal{A}$ and $\mathcal{B}$ in $k-1$ dimensions. Our first goal is to bound $\tilde{\mathcal{A}}$ by adding extra hyperplanes. This can be done as follows: For each extra dimension added to $\mathcal{A}$, extending it to an unbounded prism, we need two hyperplanes to bound the convex body in that direction. To do this, define $\mathbf{h}^i$ as $\mathbf{h}_j^i = \delta_{ij}$ for $1 \leq j \leq k$. Then, add $2(k-r)$ rows of length k to the bottom of $\tilde{\mathbf{A}}$, of the form $-\mathbf{h}^i/2 + \mathbf{u}/2$ for $1 \leq i \leq k-r$ and $\mathbf{h}^i/2 + \mathbf{u}/2$ for $k-r+1 \leq i \leq 2(k-r)$. This corresponds to adding $k-r$ extra dichotomic measurements, meaning $l \rightarrow l+k-r$, but $\mathbf{u}$ does not change. Note that each $\mathbf{h}^i$ is orthogonal to the span of $\tilde{\mathcal{A}}$. Let us denote by $\tilde{\mathcal{A}}_b$ and $\tilde{\mathcal{A}}_b$ the resulting $k \times (n+$

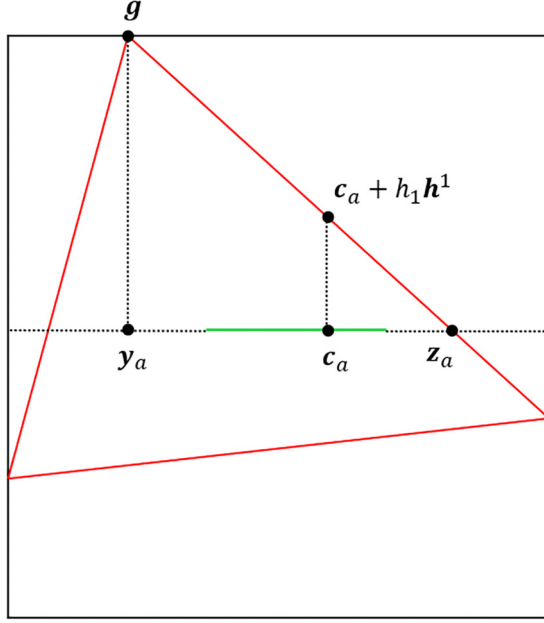


FIG. 1. Two-dimensional cross section of the nested polytopes $\bar{\mathcal{A}}_b$ (solid black), $\bar{\mathcal{A}}$ (solid green), and $\bar{\mathcal{G}}$ (solid red).

$2k - 2r$) matrix and its corresponding $(k - 1)$ -dimensional bounded polytope, respectively.

Assume that there exists a $(k - 1)$ -simplex $\mathcal{G}$ nested between $\bar{\mathcal{B}}$ and $\bar{\mathcal{A}}$. We want the description of $\mathcal{G}$ to be as simple as possible. Therefore, using Lemma 5, we assume that the vertex of $\mathcal{G}$ furthest from the span of $\bar{\mathcal{B}}$ has no $\mathbf{h}^j$ component for $j > 1$. Further, we assume that the vertex of $\mathcal{G}$ furthest from $\text{span}(\bar{\mathcal{B}} \cup_{j=1}^{i-1} \{\mathbf{h}^j\})$ has no $\mathbf{h}^p$ component for $p > i + 1$ and $i = 1, \dots, k - r$.

To obtain a nested simplex between $\bar{\mathcal{B}}$ and $\bar{\mathcal{A}}_b$, the vertices of $\mathcal{G}$ can be rescaled along the $\mathbf{h}^i$ directions to obtain the $(k - 1)$ -simplex $\bar{\mathcal{G}}$ such that $k - r$ of its vertices lie on the hyperplanes defined by $\mathbf{h}^i$ s and $\bar{\mathcal{B}} \subseteq \bar{\mathcal{G}} \subseteq \bar{\mathcal{A}}_b$.

We now aim to extend $\bar{\mathcal{B}}$ to a full-dimensional polytope by adding $k - r$ vertices $\{\mathbf{v}^i \in \bar{\mathcal{G}}\}_{i=1}^{k-r}$ to it. The first vertex can be constructed by adding a $\mathbf{h}^1$ component to a point in $\bar{\mathcal{B}}$. Let this point be $\mathbf{c}^1$, the barycenter of $\bar{\mathcal{B}}$ computed from its vertices, so that $\mathbf{v}^1 = \mathbf{c}^1 + \bar{h}_1 \mathbf{h}^1$. We thus need to estimate the maximum $\bar{h}_1$.

Suppose $\bar{\mathcal{B}}_{\bar{\mathcal{A}}} := \text{span}(\bar{\mathcal{B}}) \cap \bar{\mathcal{A}}$ is the polytope formed by the intersection of the span of $\bar{\mathcal{B}}$ and $\bar{\mathcal{A}}$. Similarly, define $\bar{\mathcal{B}}_{\bar{\mathcal{G}}} := \text{span}(\bar{\mathcal{B}}) \cap \bar{\mathcal{G}}$ as the polytope formed by the intersection of the span of $\bar{\mathcal{B}}$ and $\bar{\mathcal{G}}$. Note that $\bar{\mathcal{B}} \subseteq \bar{\mathcal{B}}_{\bar{\mathcal{G}}} \subseteq \bar{\mathcal{B}}_{\bar{\mathcal{A}}}$.

Consider the vertex $\mathbf{g}$ of $\bar{\mathcal{G}}$ on the hyperplane $\mathbf{x}^\top \mathbf{h}^1 = 1$ and denote its projection along $\mathbf{h}^1$ onto $\bar{\mathcal{B}}_{\bar{\mathcal{A}}}$ as $\mathbf{y}^1$. The line segment ℓ connecting $\mathbf{y}^1$ to $\mathbf{c}^1$ intersects with $\bar{\mathcal{B}}_{\bar{\mathcal{G}}}$ at a point $\mathbf{z}^1$ opposite to $\mathbf{y}^1$. The three points $\mathbf{g}$, $\mathbf{y}^1$, and $\mathbf{z}^1$ form a right triangle with its base $\mathbf{g}\mathbf{y}^1$ having unit length. Using the basic proportionality theorem of triangles, we have

$$h_1 = \frac{h_1}{\mathbf{g}\mathbf{y}^1} = \frac{\bar{z}^1 \mathbf{c}^1}{\bar{z}^1 \mathbf{y}^1}. \quad (21)$$

This is illustrated in Fig. 1. Let $\mathcal{S}_{11}$ be a simplex whose vertices are in $\bar{\mathcal{B}}$ and contains $\mathbf{c}^1$. Since $\mathbf{c}^1$ is the barycenter of $\bar{\mathcal{B}}$, it

must be the case that $\bar{z}^1 \mathbf{c}^1$ is greater than or equal to the closest point of $\partial \mathcal{S}_{11}$ to $\mathbf{c}^1$. Let us call this distance d_1 . Furthermore, note that without loss of generality, we can assume $\bar{\mathcal{A}}$ contains the origin. Let $\mathcal{S}_{21}$ be a simplex that contains $\bar{\mathcal{B}}_{\bar{\mathcal{A}}}$. This can be given by the polar dual of a simplex that is contained in the polar dual of $\bar{\mathcal{A}}$ and contains the origin. Since $\mathbf{y}^1, \mathbf{z}^1 \in \bar{\mathcal{B}}_{\bar{\mathcal{A}}}$, $\bar{z}^1 \mathbf{y}^1$ is upper bounded by the maximum distance between any two vertices of $\mathcal{S}_{21}$. We call this distance d_2 . It follows that

$$h_1 = \frac{\bar{z}^1 \mathbf{c}^1}{\bar{z}^1 \mathbf{y}^1} \geq \frac{d_1}{d_2} := \bar{h}_1. \quad (22)$$

We iterate the above for $i = 1, \dots, k - r$ using the following procedure. After the $(i - 1)$ th iteration, we so far have values for $\bar{h}_1, \dots, \bar{h}_{i-1}$. For the i th iteration, define $\bar{\mathcal{B}}_i = \text{conv}(\bar{\mathcal{B}} \cup_{j=1}^{i-1} \{\mathbf{v}^j\})$ where $\mathbf{v}^j = \mathbf{c}^j + \bar{h}_j \mathbf{h}^j$, $\bar{\mathcal{B}}_{\bar{\mathcal{A}}_i} = \text{span}(\bar{\mathcal{B}}_i) \cap \bar{\mathcal{A}}$, and $\bar{\mathcal{B}}_{\bar{\mathcal{G}}_i} = \text{span}(\bar{\mathcal{B}}_i) \cap \bar{\mathcal{G}}$. Similarly to before, let $\mathbf{c}^i$ be the barycenter of $\bar{\mathcal{B}}_i$, let $\mathcal{S}_{1i}$ be a simplex contained in $\bar{\mathcal{B}}_i$ containing $\mathbf{c}^i$, and let $\mathcal{S}_{2i}$ be a simplex containing $\bar{\mathcal{B}}_{\bar{\mathcal{A}}_i}$. Then, let d_{1i} be the minimum distance from $\mathbf{c}^i$ to $\partial \mathcal{S}_{1i}$ and d_{2i} be the maximum distance between two vertices of $\mathcal{S}_{2i}$. Then, using a similar line of reasoning as above, h_i can be bounded as

$$h_i \geq \bar{h}_i = \frac{d_{1i}}{d_{2i}}. \quad (23)$$

The iteration procedure above ensures that the distance between the boundary of $\bar{\mathcal{G}}$ and $\mathbf{c}_i$ along the $\mathbf{h}^i$ direction is not infinitesimally small.

Since $\mathbf{h}^i$ is chosen orthogonal to $\text{span}(\bar{\mathcal{B}}_i)$, we have $\mathbf{v}^i \notin \text{span}(\bar{\mathcal{B}}_i)$, and therefore $\dim(\text{span}(\bar{\mathcal{B}}_{i+1})) = \dim(\text{span}(\bar{\mathcal{B}}_i)) + 1$. Starting from $\dim(\text{span}(\bar{\mathcal{B}})) = r$, after $k - r$ steps we obtain $\dim(\text{span}(\bar{\mathcal{B}}_b)) = k$. It follows that $\bar{\mathcal{B}}_b = \text{conv}(\bar{\mathcal{B}} \cup_{i=1}^{k-r} \{\mathbf{v}^i\})$ is a full-dimensional bounded polytope in $k - 1$ affine dimensions such that $\bar{\mathcal{B}}_b \subseteq \bar{\mathcal{G}} \subseteq \bar{\mathcal{A}}_b$. This polytope corresponds to an $(m + k - r) \times k$ rank- k matrix $\bar{\mathcal{B}}_b$, which has $k - r$ extra columns added to the right of those of $\bar{\mathcal{A}}$.

Now $\bar{\mathcal{B}}_b$ and $\bar{\mathcal{A}}_b$ are both $(k - 1)$ -dimensional polytopes and we define $\bar{\mathcal{C}} = \bar{\mathcal{A}}_b \bar{\mathcal{B}}_b$ as the $(m + 2k - 2r) \times (n + k - r)$ slack of vertices of $\bar{\mathcal{B}}_b$ with respect to the facets of $\bar{\mathcal{A}}_b$. This construction from $\bar{\mathcal{B}} \subseteq \bar{\mathcal{G}} \subseteq \bar{\mathcal{A}}$ to $\bar{\mathcal{B}}_b \subseteq \bar{\mathcal{G}} \subseteq \bar{\mathcal{A}}_b$ is shown in Fig. 2. We note that $\bar{\mathcal{C}}$ contains $\mathcal{C}$ as a submatrix in its top-left corner.

Using Ref. [22], the existence of the intermediate simplex $\bar{\mathcal{G}}$ is equivalent to the existence of an NMF of inner dimension k of $\bar{\mathcal{C}}$. Since $\text{rank } \bar{\mathcal{C}} = k$, it follows also that the NNR of $\bar{\mathcal{C}}$ is k . ■

The proof of existence of the matrix $\bar{\mathcal{C}}$ in Theorem 2 could have been done using the topological properties of $\mathbb{R}^k$. More precisely, we could extend $\bar{\mathcal{B}}$ by simply choosing $k - r$ random vertices from a sufficiently small ball centered at its barycenter. However, this could have implied unbounded bit lengths for describing the added vertices. Our selection of bounding hyperplanes for $\bar{\mathcal{A}}_b$ and the specific construction of $\bar{\mathcal{B}}_b$ ensure that the extra vertices can be specified using a polynomial number of bits, as shown in the following lemma.

Lemma 6. Let $\mathcal{C}$ be a rank- r $m \times n$ non-negative matrix. The bit length of entries in $\bar{\mathcal{C}}$ as defined in Theorem 2 is $\text{poly}(b, m, n)$, where b is the bit length of entries in $\mathcal{C}$.

Proof. Since for each i $\bar{h}_i$ only has to be a positive lower bound of h_i , it can be rounded down to a positive number when

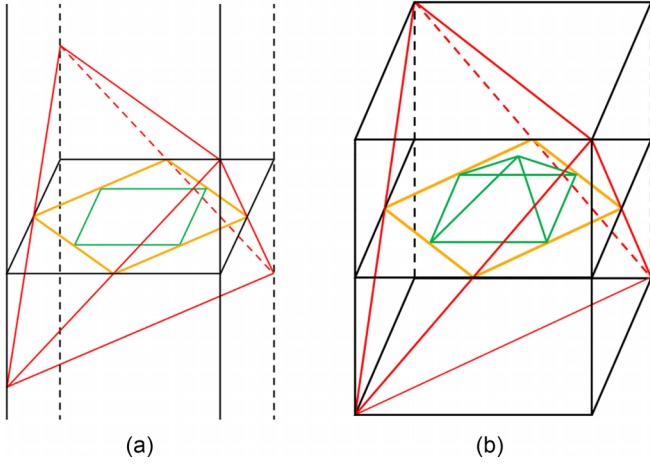


FIG. 2. Illustration of the intermediate simplex problem in Theorem 2. (a) The inner polytope (green) is not full dimensional, and the outer polytope (black) is unbounded. (b) The inner polytope (green) is full dimensional, and the outer polytope (black) is bounded. The intermediate simplex and its intersection with the affine span of the inner polytope are represented in red and orange, respectively.

constructing $\tilde{\mathcal{B}}_b$. Therefore, only the orders of magnitude of $\tilde{h}_i$ s determine the number of extra bits needed for defining $\tilde{\mathcal{C}}$.

Assuming that the entries of $\mathbf{C}$ are rational and defined with b bits, the entries of the rank factors A and B are, without loss of generality, rational and defined with $\text{poly}(b, m, n)$ bits. Therefore, the vertices of $\tilde{\mathcal{B}}$, the point $\mathbf{c}^1$ and the facets of $\tilde{\mathcal{A}}$, as in Theorem 2, are defined with $\text{poly}(b, m, n)$ bits. Furthermore, it follows that $\mathcal{S}_{11}$ and $\mathcal{S}_{21}$, defined in the proof of Theorem 2, can be efficiently chosen such that the same bound on bit length applies to the vertices of the former and facets of the latter. For example, $\mathcal{S}_{11}$ can be found by choosing r approximately evenly spaced unit vectors $\{\mathbf{f}^i\}$ that have $\text{poly}(b, m, n)$ bit representations. Then, for each $\mathbf{f}^i$, a vertex of $\mathcal{S}_{11}$ is given by $\mathbf{c}^1 + f_i \mathbf{f}^i$, where f_i is chosen such that $\mathbf{c}^1 + f_i \mathbf{f}^i \in \tilde{\mathcal{B}}$. The inclusion of $\mathbf{c}^1 + f_i \mathbf{f}^i$ in $\tilde{\mathcal{B}}$ can be decided by solving a linear program which has a number of constraints and parameters linear in the number of vertices of $\tilde{\mathcal{B}}$. Therefore, since the input data to this linear program has $\text{poly}(b, m, n)$ bit length, a solution for f_i with $\text{poly}(b, m, n)$ bit length exists and can be found using a binary search.

We will show that each stage of the computation of $\tilde{h}_i$ s requires only $\text{poly}(b, m, n)$ bits. The distance d_1 from $\mathbf{c}^1$ to $\partial\mathcal{S}_{11}$ is given by the distance to the closest facet of $\partial\mathcal{S}_{11}$. Given that $\mathcal{S}_{11}$ is $(r-1)$ dimensional, only $r-1$ vertices define each of its bounding hyperplanes. Let $\{\mathbf{w}^1, \dots, \mathbf{w}^{r-1}\}$ be the subset of vertices of $\mathcal{S}_{11}$ in one of the bounding hyperplanes. Then, to find the normal vector, define the matrix W as $W_{i,:} = \mathbf{w}^1 - \mathbf{w}^{i+1}$ for $i = 1, \dots, r-2$. Gaussian elimination can be used to find an unnormalized vector $\mathbf{n}$ in the subspace of $\text{span}(\tilde{\mathcal{B}})$. This yields a rational vector and only requires $\text{poly}(b, m, n)$ bits.

The distance from $\mathbf{c}^1$ to the hyperplane is given by $(\mathbf{c}^1 - \mathbf{w}^1)^\top \mathbf{n} / |\mathbf{n}|$. The inner product of two rational vectors each defined with $\text{poly}(b, m, n)$ bits will have a bit length of $\text{poly}(b, m, n)$. Therefore, the minimum and maximum pos-

sible orders of magnitude of such an inner product are $2^{-\text{poly}(b, m, n)}$ and $2^{\text{poly}(b, m, n)}$, respectively.

Now we will consider the maximum distance d_2 between two vertices of $\mathcal{S}_{21}$. The intersection of only $r-1$ hyperplanes of $\tilde{\mathcal{B}}_{\tilde{\mathcal{A}}}$ define a vertex. This point can be found by solving the system of linear equations $H\mathbf{x} = \mathbf{o}$, where H is a rational matrix of bit length $\text{poly}(b, m, n)$ and $\mathbf{o}$ is a rational vector whose entries are the offsets of the hyperplanes, also of bit length $\text{poly}(b, m, n)$. As shown in Ref. [33], inverting a rational matrix increases the bit length only polynomially. It follows that the solution $|\mathbf{x}^*|$ has a maximum order of magnitude of $2^{\text{poly}(b, m, n)}$. Furthermore, we can apply the same reasoning to the polytopes $\mathcal{S}_{2i}$ as defined in the proof of Theorem 2.

Putting this together and using Eq. (22), the minimum order of magnitude of $\tilde{h}_1$ is $2^{-\text{poly}(b, n, m)}$. However, we need to also take into consideration the iterations in the proof of Theorem 2. Starting from $\tilde{\mathcal{B}}$, each iteration $\tilde{\mathcal{B}}_i$ is constructed by adding a vertex of the form $\mathbf{c}_i + \tilde{h}_{i-1} \mathbf{h}^{i-1}$ to $\tilde{\mathcal{B}}_{i-1}$ and taking the convex hull.

Consider the second iteration, in which we find the distance of the barycenter $\mathbf{c}^2$ of $\tilde{\mathcal{B}}_2$ to $\partial\mathcal{S}_{1,2}$, where $\tilde{\mathcal{B}}_2 = \text{conv}(\tilde{\mathcal{B}} \cup \mathbf{c}^1 + \tilde{h}_1 \mathbf{h}^1)$, and $\mathcal{S}_{1,2}$ is a simplex contained in $\tilde{\mathcal{B}}_2$ containing $\mathbf{c}^2$. This can simply be chosen as $\mathcal{S}_{1,2} = \text{conv}(\mathcal{S}_{11} \cup \mathbf{c}^1 + \tilde{h}_1 \mathbf{h}^1)$. Given that $\tilde{\mathcal{B}}$ has m vertices, and assuming $\tilde{h}_1$ is sufficiently small, this distance is given by $\tilde{h}_1/m$, so its order of magnitude is at least $2^{-\text{poly}(b, n, m)}$. Then, by the same reasoning, the minimum distance given in each iteration is also of this form, where division by $m+i$ decreases the exponent by $O(\log(m+i))$. i has a maximum value of $k-r$, so the minimum order of magnitude of $\tilde{h}_i$ s is $2^{-\text{poly}(k, l, n, m)}$.

It follows that $\tilde{h}_i$ s can be rounded down such that their bit lengths are at most $\text{poly}(b, k, n, m)$. Therefore, the bit lengths of the entries of $\tilde{\mathcal{C}}$ are also $\text{poly}(b, k, n, m)$. Since $k \leq r^2$, which is shown in Refs. [16, 18, 34], and $r \leq m, n$, the bit lengths are of the form $\text{poly}(b, n, m)$. ■

Corollary 4. The reduction in Theorem 2 can be computed in $\text{poly}(b, n, m)$ time.

Proof. First, the simplexes $\mathcal{S}_{1i}$ and $\mathcal{S}_{2i}$ for each iteration can be constructed in polynomial time. Then we note that the H-representation of a simplex can be computed in polynomial time from its V-representation and vice versa. Therefore, the distances d_{1i} and d_{2i} can be computed in runtimes that are polynomial in the bit length required for the desired precision of these distances.

Since this bit length is bounded by $\text{poly}(b, n, m)$, the complexity of the whole procedure given in the proof of Theorem 2 has a similar bound. ■

In his seminal work, Moitra [21] showed through an explicit algorithm that the complexity of deciding whether the NNR of an $m \times n$ non-negative matrix is at most k is $\text{poly}(b, m, n)^{O(k^2)}$. Here, as before, b determines the bit length of the entries of the input. Therefore, using Corollary 4, and given that $r \leq m, n$ and $k \leq r^2$, we have the following result.

Corollary 5. The complexity of deciding if a rank- r COPE $\mathbf{C}$ admits a noncontextual ontological model of inner dimension k is at most $\text{poly}(b, m, n)^{O(k^2)}$, where the entries of $\mathbf{C}$ are given with a bit length b .

Corollary 5 provides an upper bound on the complexity of deciding if the ENNR of a matrix is at most k , and Corollary 2

ALGORITHM 1. Finding the ENNR of a COPE.

Input: A rank- r COPE matrix $\mathbf{C}$.**Output:** The ENNR of $\mathbf{C}$.

-
- 1: Compute a real rank factorization $\mathbf{C} = \mathbf{A}\mathbf{B}$.
 - 2: Set $k = r$ and $\text{ENNR} = 0$.
 - 3: Construct the matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$, where the rows of $\bar{\mathbf{A}}$ and the columns of $\bar{\mathbf{B}}$ are the trivial embeddings of the rows of $\mathbf{A}$ and columns of $\mathbf{B}$, respectively, into $\mathbb{R}^k$.
 - 4: Add $2(k - r)$ rows of length k to the bottom of $\bar{\mathbf{A}}$, of the form $-\mathbf{h}^i/2 + \mathbf{u}/2$ for $1 \leq i \leq k - r$ and $\mathbf{h}^i/2 + \mathbf{u}/2$ for $k - r + 1 \leq i \leq 2(k - r)$ to construct $\bar{\mathbf{A}}_b$, where $\mathbf{h}_j^i = \delta_{ij}$ for $1 \leq j \leq k$.
 - 5: Set $i = 1$ and let $\bar{\mathbf{B}}_1 = \bar{\mathbf{B}}$.
 - 6: Compute $\mathbf{c}^i$, the mean of the vertices of $\bar{\mathbf{B}}_i$, and find a simplex $\mathcal{S}_{1i}$ that contains $\mathbf{c}^i$ and is contained in $\bar{\mathbf{B}}_i$. To find such a simplex efficiently, see Lemma 6.
 - 7: Find a simplex $\mathcal{S}_{2i}$ that contains $\bar{\mathbf{B}}_{\mathcal{A}i} := \text{span}(\bar{\mathbf{B}}_i) \cap \bar{\mathcal{A}}$. This can be done by finding a simplex which is contained in the polar dual of $\bar{\mathbf{B}}_{\mathcal{A}i} := \text{span}(\bar{\mathbf{B}}_i) \cap \bar{\mathcal{A}}$ and contains the origin, then taking the polar dual of that simplex.
 - 8: Compute d_{1i} as the distance from $\mathbf{c}^i$ to $\partial\mathcal{S}_{1i}$, and d_{2i} as the diameter of $\mathcal{S}_{2i}$, both to only the precision of the leading respective bit. Round down d_{1i} to the previous power of 2, and round up d_{2i} to the next power of 2. Then, compute $\bar{h}_i = d_{1i}/d_{2i}$.
 - 9: Construct $\bar{\mathbf{B}}_{i+1} = \text{conv}(\bar{\mathbf{B}}_i \cup \mathbf{v}^i)$, where $\mathbf{v}^i = \mathbf{c}^i + \bar{h}_i \mathbf{h}^i$ and move onto the next iteration.
 - 10: If $i < k - r$, set $i = i + 1$ and go to step 6.
 - 11: Set $\bar{\mathbf{B}}_b = \bar{\mathbf{B}}_{k-r+1}$ and compute $\bar{\mathbf{C}} := \bar{\mathbf{A}}_b \bar{\mathbf{B}}_b$.
 - 12: Decide if the NNR of $\bar{\mathbf{C}}$ equals k , which can be done using Moitra's algorithm. If yes, then the ENNR equals k , set $\text{ENNR} = k$, and abort. Else if $k \leq r^2$, then set $k = k + 1$ and go to step 3.
-

provides a lower bound on the complexity of deciding if the ENNR is at least r . However, a lower bound in k remains elusive.

The constructive proofs used in Theorem 2 and Corollary 4 lead to an algorithm for finding the smallest noncontextual model as shown in Algorithm 1.

Indeed, using the algorithm above, if the final value of ENNR is zero, then it means that an ENNR is not found, hence an ENMF does not exist. Our algorithm uses a subroutine that computes NNR. Moitra's algorithm is so far the best for this purpose, and an expert reader may notice that its first step is to guess the ranks of the non-negative matrix factors. Therefore, one may ask if these ranks could be chosen to equal the rank of the input matrix to attempt to find an ENNR of a specified inner dimension k . The limitation of Moitra's algorithm in this case is that it searches for NMFs with an additional restriction, as outlined in Appendix B. This restriction does not rule out the simplest NMF, but, as shown in Appendix B, could rule out the simplest ENMF.

IV. COMPARISON TO SIMPLEX EMBEDDABILITY

As we mentioned in Sec. II B, every NMF $\mathbf{C} = \mathbf{R}\mathbf{E}$ of a COPE matrix, after an appropriate rescaling of its factors, corresponds to an ontological model. We have also seen that any ENMF of $\mathbf{C}$, which is simply an NMF in which all factors

have equal ranks, can be written as $\mathbf{C} = \bar{\mathbf{A}}\bar{\mathbf{G}}\bar{\mathbf{G}}^{-1}\bar{\mathbf{B}}$ so that the factor matrices correspond to the nested polytope $\bar{\mathbf{A}} \mapsto \bar{\mathcal{A}}$, $\bar{\mathbf{G}} \mapsto \bar{\mathcal{G}}$, and $\bar{\mathbf{B}} \mapsto \bar{\mathcal{B}}$ such that $\bar{\mathcal{G}}$ is a $(k - 1)$ -simplex and $\bar{\mathcal{B}} \subseteq \bar{\mathcal{G}} \subseteq \bar{\mathcal{A}}$, as shown in Lemma 4.

Alternatively, it was shown in Refs. [24,34] that noncontextuality has a geometrical interpretation known as *simplex embeddability*. Namely, a GPT admits a noncontextual ontological model if and only if its state and effect spaces can be embedded into a simplex and a hypercube, respectively. As demonstrated in Sec. II, a GPT model can be easily found from the COPE of a prepare-measure experiment. In this section, we characterize the relationship between our geometric interpretation of noncontextuality and simplex embeddability.

We denote the polytopes of states and effects by $\mathcal{Q}$ and $\mathcal{E}$, respectively, which live in an inner product space $(\mathfrak{V}, \langle \cdot, \cdot \rangle_{\mathfrak{V}})$. Then, simplex embeddability is defined as follows.

Definition 1 (simplex embeddability [34]). A GPT with state and effect spaces $\mathcal{S}$ and $\mathcal{E}$ over a vector space $\mathfrak{V}$ with inner product $\langle \cdot, \cdot \rangle_{\mathfrak{V}}$ is simplex-embeddable if and only if there exists a $(k - 1)$ -simplex $\mathcal{W}_k$ (whose affine span does not contain the origin) in a k -dimensional inner product space $\mathfrak{W}_k$ and its dual hypercube $\mathcal{W}_k^*$ (including the zero and unit vectors), and a pair of linear maps $\iota, \kappa : \mathfrak{V} \rightarrow \mathfrak{W}_k$ such that $\iota(\mathcal{S}) \subseteq \mathcal{W}_k$, $\kappa(\mathcal{E}) \subseteq \mathcal{W}_k^*$, and $\langle \kappa(e), \iota(s) \rangle_{\mathfrak{W}_k} = \langle e, s \rangle_{\mathfrak{V}}$ for all GPT states and effects.

This formulation, like ours, requires finding a single polytope, specifically a simplex. Furthermore, Ref. [34] shows that after the simplex embedding is found, there exists a linear transformation T such that its adjoint, $T^\dagger$ maps the vertices $\mathcal{W}_k$ to the positive unit vectors in $\mathbb{R}^k$, and $\langle T(e), T^\dagger(s) \rangle = \langle e, s \rangle$ for all pairs of states and effects. We can identify our matrix $\bar{\mathbf{G}}$ with the transformation $T \circ \kappa$, and $\bar{\mathbf{G}}^{-1}$ with $T^\dagger \circ \iota$. To see the connection between $\bar{\mathcal{G}}$ and $\mathcal{W}_k$, let ι^1 be the trivial embedding into $\mathbb{R}^k$, and define κ^2 and ι^2 such that $\kappa = \kappa^2 \circ \iota^1$ and $\iota = \iota^2 \circ \iota^1$. Then $\iota^2(\bar{\mathcal{G}}) = \mathcal{W}_k$.

Reference [18] suggests using a linear program to decide if the simplex embeddability conditions can be satisfied. Here, each solution to the linear program is a matrix that represents the product of a pair of linear maps ι, κ . Then, the dimensionality of the codomains of the associated maps ι, κ to a solution instance is given by its non-negative rank. The smallest noncontextual model is found by searching over the set of solutions to find the minimum non-negative rank in that set. Since the non-negative rank is a nonsmooth function of matrices, the number of times a non-negative rank of a matrix needs to be computed could be very large, perhaps even exponential in the number of input states and effects. The key utility of our construction is that it allows us to compute the ontic size d by only computing the non-negative rank of $d + 1$ matrices. Using our construction, the size d of the smallest noncontextual model is found by testing increasing ontic sizes starting at the minimum GPT dimension r . For each ontic size, the non-negative rank of only one matrix needs to be computed, so the total number of non-negative rank computations needed is $d - r$.

V. ONTIC SIZE SEPARATION

We have seen that the problem of finding the smallest noncontextual ontological model for a given COPE matrix

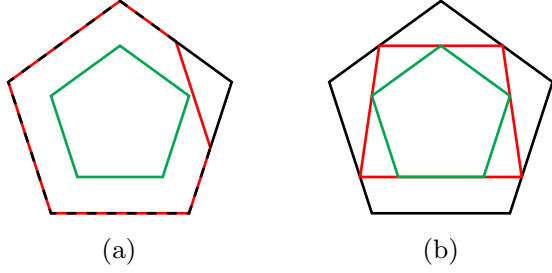


FIG. 3. Inner (green) and outer (black) pentagons with nested polytopes (red) (a) $\mathcal{G}^2$ and (b) $\mathcal{G}^3$.

is equivalent to the problem of finding the smallest (not necessarily noncontextual) ontological model of a different non-negative matrix. We may therefore ask whether a non-contextual ontological model, when exists, is the smallest ontological model. In other words, is there a gap between the ENNR and the NNR whenever an ENMF exists?

In this section, we present a COPE matrix that shows such a gap, indicating that the ENNR cannot be trivially obtained from the NNR and vice versa.

We show this for a 5×5 rank-3 COPE matrix whose first column is given by

$$\mathbf{C}_{:,1}^1 = \frac{1}{10} \begin{pmatrix} 2(\sqrt{5}-1) \\ 8-2\sqrt{5} \\ 2(\sqrt{5}-1) \\ 3-\sqrt{5} \\ 3-\sqrt{5} \end{pmatrix}, \quad (24)$$

and each subsequent column is given by permuting the entries the previous column up by one. This COPE also admits the following real factorization,

$$\mathbf{C}^1 := \mathbf{A}_{\mathcal{U}} \mathbf{B}_{\mathcal{I}}, \quad (25)$$

which has associated inner and outer polygons (pentagons), shown in Fig. 3(a), $\mathcal{B}_{\mathcal{I}}$ and $\mathcal{A}_{\mathcal{U}}$, respectively, as in Lemma 2. These are defined by constructing a pentagon $\mathcal{I}$ and the smallest regular 4-simplex $\mathcal{U}$ that contains it.

We begin with the bounding inequalities of a regular pentagon, given by $K\mathbf{x} + \mathbf{p} \geq 0$ wherein

$$K = \begin{pmatrix} \sin(\frac{2\pi}{5}) & \cos(\frac{2\pi}{5}) \\ \sin(\frac{4\pi}{5}) & \cos(\frac{4\pi}{5}) \\ \sin(\frac{6\pi}{5}) & \cos(\frac{6\pi}{5}) \\ \sin(\frac{8\pi}{5}) & \cos(\frac{8\pi}{5}) \\ \sin(\frac{10\pi}{5}) & \cos(\frac{10\pi}{5}) \end{pmatrix}, \quad (26)$$

and

$$\mathbf{p} = \frac{\sqrt{5}-1}{10} (1 \ 1 \ 1 \ 1 \ 1)^\top. \quad (27)$$

The vertices of this are given by

$$\mathbf{x}^i = \frac{\sqrt{5}-1}{10} \begin{pmatrix} \cos(\frac{2\pi i}{5}) \tan(\frac{2\pi}{10}) + \sin(\frac{2\pi i}{5}) \\ \sin(\frac{2\pi i}{5}) \tan(\frac{2\pi}{10}) - \cos(\frac{2\pi i}{5}) \end{pmatrix}, \quad (28)$$

for $i \in \{1, 2, \dots, 5\}$.

Now, we can use this to construct $\mathcal{I}$ and $\mathcal{U}$. The points $\{\mathbf{y}^i\}$ given by $\mathbf{y}^i = (2/5)(K\mathbf{x}^i + \mathbf{p})$ form a regular pentagon embedded in $\mathbb{R}_{\geq 0}^5$. Let $\mathcal{I}$ and $\mathcal{U}$ be the convex hull of $\{\mathbf{y}^i\}$ and the regular 4-simplex whose vertices are given by the columns of $([\sqrt{5}-1]/5)\mathbb{1}_{5 \times 5}$, respectively. Note that $\mathcal{U}$ can also be defined as

$$\mathcal{U} := \left\{ \mathbf{u} \in \mathbb{R}^5 \mid \mathbf{u} \geq 0, \mathbf{u}^\top \left(\frac{10}{\sqrt{5}-1} \mathbf{p} \right) = \frac{\sqrt{5}-1}{5} \right\}. \quad (29)$$

Then, for any real vector $(a \ b)^\top \in \mathbb{R}^2$,

$$\begin{aligned} (a \ b) K^\top \left(\frac{10}{\sqrt{5}-1} \mathbf{p} \right) \\ = \sum_{i=1}^5 \left(a \cos\left(\frac{2\pi i}{5}\right) + b \sin\left(\frac{2\pi i}{5}\right) \right) = 0. \end{aligned} \quad (30)$$

Therefore, given

$$\frac{2}{5} \mathbf{p}^\top \left(\frac{10}{\sqrt{5}-1} \mathbf{p} \right) = \frac{\sqrt{5}-1}{5}, \quad (31)$$

it is clear that $\mathcal{I} \subset \mathcal{U}$.

Let $\mathcal{P}_{\mathcal{U}}$ be the projection of $\mathcal{U}$ onto the affine span of $\mathcal{I}$, so $\mathcal{P}_{\mathcal{U}}$ is another regular pentagon.

It can be shown that $\mathcal{P}_{\mathcal{U}}$ is a $\times(\sqrt{5}-1)/2$ scaling of $\mathcal{I}$ about its center, and this scaling factor $(\sqrt{5}-1)/2 \approx 0.618$ is derived in Appendix C. Then we choose $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{B}_{\mathcal{I}}$ to be nested pentagons where $\mathcal{A}_{\mathcal{U}}$ is a scaling of $\mathcal{B}_{\mathcal{I}}$ by a scale factor $(\sqrt{5}-1)/2$ about its center. Then, the corresponding matrices $\mathbf{A}_{\mathcal{U}}$ and $\mathbf{B}_{\mathcal{I}}$ are given by

$$(\mathbf{A}_{\mathcal{U}})_{i,:} = \left(\frac{2}{5-\sqrt{5}} \ K_{i,:} \right), \quad (32)$$

$$(\mathbf{B}_{\mathcal{I}})_{:,i} = \left(\frac{5-\sqrt{5}}{10} \ \mathbf{x}^{i\top} \right)^\top. \quad (33)$$

We have $(\mathbf{A}_{\mathcal{U}} \mathbf{B}_{\mathcal{I}})_{:,i} = K\mathbf{x}^i + 2/(\sqrt{5}-1)\mathbf{p}$, which has the form of the slack of a vertex of a regular pentagon with respect to the facets of a factor of $(\sqrt{5}-1)/2$ scaling of that pentagon.

Next, to see the existence of a noncontextual model, we construct the infinite cone $\bar{\mathcal{A}}_{\mathcal{U}} \subset \mathbb{R}^5$ from $\mathcal{A}_{\mathcal{U}}$ and trivially embed the polytope $\bar{\mathcal{B}}_{\mathcal{I}}$ into $\mathbb{R}^5$ as in Lemma 4. It follows from the fact that $\mathcal{I} \subset \mathcal{U}$ that there exists a nested simplex $\mathcal{G}$ between $\bar{\mathcal{B}}_{\mathcal{I}}$ and $\bar{\mathcal{A}}_{\mathcal{U}}$. Therefore, Lemma 4 can be used to show that there is a noncontextual model of ontic size five for $\mathbf{C}^1$, given by $\mathbf{C}^1 = \mathbf{A}^1 \mathbf{B}^1$, where

$$\mathbf{A}^1 = \begin{pmatrix} \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} & 0 & 0 \\ 0 & \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} & 0 \\ 0 & 0 & \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} \\ \frac{5-\sqrt{5}}{10} & 0 & 0 & \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} & 0 & 0 & \frac{5-\sqrt{5}}{10} \end{pmatrix},$$

$$\mathbf{B}^1 = \begin{pmatrix} 0 & \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} & 0 \\ \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} & 0 & 0 \\ \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} & 0 & 0 & \frac{5-\sqrt{5}}{10} \\ \frac{5-\sqrt{5}}{10} & 0 & 0 & \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} \\ 0 & 0 & \frac{5-\sqrt{5}}{10} & \frac{\sqrt{5}}{5} & \frac{5-\sqrt{5}}{10} \end{pmatrix}. \quad (34)$$

Before proving the ontic size separation, we present a geometric lemma, which we will use for the main proof.

Lemma 7. Let $\{\mathbf{v}^i\}_{i=0}^4$ be the set of vertices of a convex pentagon. For each $\mathbf{v}^i$, let the triangle $\mathcal{T}^i$ be $\text{conv}\{\frac{\mathbf{v}^{i-1}+\mathbf{v}^i}{2}, \mathbf{v}^i, \frac{\mathbf{v}^{i+1}+\mathbf{v}^i}{2}\}$, where superscript additions are mod 5. Then it is not possible to inscribe a quadrilateral in the pentagon such that it intersects the interiors of each $\mathcal{T}^i$.

Proof. Since all vertices of the pentagon are convexly independent, the interior of each triangle is convexly independent from the other triangles. Therefore, a line can only intersect the interiors of at most two triangles. Assume there is an inscribed quadrilateral with vertices $\{\mathbf{u}^i\}_{i=0}^3$ that intersects the interiors of all the triangles. Then, at least one vertex of the quadrilateral, $\mathbf{u}^0$ must be in a subset of one of the triangles, say, $\mathcal{T}^1$, given by $\mathcal{T}^1 \setminus \text{conv}\{(\mathbf{v}^0 + \mathbf{v}^1)/2, (\mathbf{v}^1 + \mathbf{v}^2)/2\}$. If the quadrilateral intersects the interior of $\mathcal{T}^2$, then since $\text{conv}\{\mathbf{u}^0, \mathbf{u}^1\}$ can only intersect up to two triangles, $\mathbf{u}^1$ must be in $\mathcal{T}^2 \setminus \text{conv}\{(\mathbf{v}^2 + \mathbf{v}^1)/2, (\mathbf{v}^3 + \mathbf{v}^2)/2\}$. Similarly, the quadrilateral would have to have another vertex, $\mathbf{u}^2$, in $\mathcal{T}^3 \setminus \text{conv}\{(\mathbf{v}^2 + \mathbf{v}^3)/2, (\mathbf{v}^3 + \mathbf{v}^4)/2\}$ to intersect the interior of $\mathcal{T}^3$, and so on, i.e., one quadrilateral vertex for each $\mathcal{T}^i$, which is not possible. ■

We are now ready to prove our second claim.

Proposition 1. There is no ENMF of $\mathbf{C}^1$ of inner dimension four.

Proof. Assume that there is an ENMF of $\mathbf{C}^1$ of inner dimension four. Then there is a nested quadrilateral $\mathcal{G}^1$ between $\mathcal{B}_{\mathcal{I}}$ and $\mathcal{A}_{\mathcal{U}}$ such that there is a rank-3 non-negative matrix $\mathbf{E}^1$ that maps the vertices of $\mathcal{G}^1$ to the vertices of $\mathcal{B}_{\mathcal{I}}$. Note that for any larger polytope $\mathcal{G}^2$ that contains $\mathcal{G}^1$, there is a non-negative rank-3 matrix $\mathbf{E}^2 = \mathbf{W}\mathbf{E}^1$ that maps the vertices of $\mathcal{G}^2$ to $\mathcal{B}_{\mathcal{I}}$, where $\mathbf{W}$ is a non-negative matrix that maps the vertices of $\mathcal{G}^2$ to $\mathcal{G}^1$.

Labelling the vertices of $\mathcal{A}_{\mathcal{U}}$ as $\mathbf{v}^0, \dots, \mathbf{v}^4$ as in Lemma 7, let $\mathcal{G}^2$ be $\text{conv}\{\{\mathbf{v}^1, \dots, \mathbf{v}^4\} \cup \{(\mathbf{v}^0 + \mathbf{v}^1)/2, (\mathbf{v}^4 + \mathbf{v}^0)/2\}\}$, as shown in Fig. 3(a). Note that this polytope is also given by removing one of the triangles $\mathcal{T}^i$ described in Lemma 7. Then, by Lemma 7 and the rotational symmetry of the pentagon, any nested quadrilateral $\mathcal{G}^1$ is contained in $\mathcal{G}^2$, up to a $2n\pi/5$ rotation, where n is an integer.

Now, we use the algorithm in the proof of Lemma 3 to find such a matrix $\mathbf{E}^2$ by solving $\mathcal{G}^2\mathbf{E}^2 = \mathcal{A}_{\mathcal{I}}$. Using the linear program described in the proof of Lemma 3, we find that the minimum value of $f(D)$, as defined in Lemma 3, is about $0.146 > 0$, indicating that there is no solution for $\mathbf{Q}^2$. We also find that the maximum of the dual linear program is the same. Since any value of the dual linear program lower bounds the minimum of the primal linear program, this proves that the minimum of the primal problem is positive. Thus, any matrix $\mathbf{E}$ that satisfies $\mathcal{G}^2\mathbf{E} = \mathcal{A}_{\mathcal{I}}$ must have negative

elements. We outline the calculation of this objective value in Appendix D. ■

The final step is now to show that the COPE $\mathbf{C}^1$ admits an ontological model of dimension four. This model is necessarily contextual.

Lemma 8. There exists an ontological model of $\mathbf{C}^1$ of ontic size four.

Proof. The existence of a nested quadrilateral $\mathcal{G}^3$ as shown in Fig. 3(b) implies an NMF of $\mathbf{C}^1$ of inner dimension four. Constructing $\mathcal{G}^3$ from $\mathcal{G}^2$, $\mathcal{G}^3 \subset \mathcal{A}_{\mathcal{U}}$ implies $\mathcal{A}_{\mathcal{U}}\mathcal{G}^3$ is non-negative. $\mathcal{B}_{\mathcal{I}} \subset \mathcal{G}^3$ implies there is a non-negative matrix that maps the columns of $\mathcal{G}^3$ to the columns of $\mathcal{B}_{\mathcal{I}}$. Such an ontological model $\mathbf{C}^1 = \mathbf{A}^2\mathbf{B}^2$ is given explicitly as

$$\mathbf{A}^2 = \begin{pmatrix} 0 & 1 - \frac{1}{\sqrt{5}} & 0 & \frac{(5-\sqrt{5})}{10} \\ \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 & 0 \\ 1 - \frac{1}{\sqrt{5}} & 0 & \frac{(5-\sqrt{5})}{10} & 0 \\ 0 & 0 & 1 - \frac{1}{\sqrt{5}} & \frac{(3\sqrt{5}-5)}{10} \\ 0 & 0 & \frac{(3\sqrt{5}-5)}{10} & 1 - \frac{1}{\sqrt{5}} \end{pmatrix},$$

$$\mathbf{B}^2 = \begin{pmatrix} \frac{2}{\sqrt{5}} - \frac{1}{2} & \frac{5-\sqrt{5}}{20} & 0 & \frac{3\sqrt{5}-5}{10} & \frac{3(5-\sqrt{5})}{20} \\ \frac{2}{\sqrt{5}} - \frac{1}{2} & \frac{3(5-\sqrt{5})}{20} & \frac{3\sqrt{5}-5}{10} & 0 & \frac{5-\sqrt{5}}{20} \\ 1 - \frac{2}{\sqrt{5}} & 0 & \frac{5-\sqrt{5}}{10} & 1 - \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ 1 - \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 1 - \frac{1}{\sqrt{5}} & \frac{5-\sqrt{5}}{10} & 0 \end{pmatrix}. \quad (35)$$

We have now seen a COPE whose minimum ontic size is four and minimum noncontextual ontic size is five. This proves that the number of ontic states of the smallest ontological model and the smallest contextual model are different in general. This implies that, when studying how the latter depends on properties of the COPE, the former cannot be directly used.

VI. DISCUSSION AND CONCLUSIONS

We introduced the ENMF problem, which is to find a NMF wherein all components have equal ranks. We demonstrated that this is a linear-algebraic formulation of the problem of finding a noncontextual ontological model for explaining a set of observed statistics as characterized by the rank-separation criterion. We proved that the complexity of deciding whether COPE matrices of fixed rank admit a noncontextual model of any size is polynomial in the size of the matrices. We achieved this through a novel geometric reformulation of the problem: We showed that deciding and computing an ENMF is equivalent to deciding and finding an intermediate simplex sandwiched between the preparations and outcome events polytope and dual polytope, respectively.

Our construction also resulted in a reduction from ENMF to the NNR problem, wherein one asks whether the NNR of a non-negative matrix equals its rank. This led to an upper bound on the complexity of deciding and computing noncontextual ontological models in given dimensions. More

specifically, we showed that the latter is at most exponential in the required noncontextual ontic dimension.

From a practical point of view, we would like to know whether a given experiment can be modelled classically, i.e., by a noncontextual ontological model. Furthermore, we would like to determine the minimum number of classical states required for such a model, whenever it exists, in order to estimate the classical resources necessary for emulating the experiment classically. The geometric interpretations, as well as the constructive nature of our proofs, put forward a method for answering these questions. For example, obtaining tight upper bounds and lower bounds on ENNR and NNR, respectively, provides an effective method for detecting contextuality in the data. Specifically, if the lower bound on the NNR of a COPE exceeds the upper bound on its ENNR, then it implies the presence of contextuality. However, our results also provide evidence that addressing these problems is computationally demanding. We conjecture that, contrary to common belief, detecting contextuality is computationally hard. This is further supported by the fact that all currently known algorithms for this task have runtimes that scale exponentially with the rank of the COPE matrix, or its equivalent, the GPT dimension. However, a formal proof of this conjecture remains elusive.

Given that finding ontological models and noncontextual ontological models have geometrical interpretations which reduce to the intermediate simplex problem, it is natural to ask whether they are genuinely different problems. We answered this question in the affirmative by constructing an explicit example of a COPE matrix which exhibits a size gap between its smallest ontological model and smallest noncontextual model. The roles of these minimum model sizes in explaining the quantum advantage of information processing protocols remain unclear. Of course, the latter is only relevant if a noncontextual model exists, which seems to greatly restrict the degree of quantum advantage possible. This can be seen by considering the maximum possible gap between the smallest GPT dimensionality and the smallest noncontextual ontic dimensionality needed to explain a COPE. Since the former is given by the rank r and the latter is at most $O(r^2)$, the advantage of noncontextually explainable protocols would be polynomially bounded.

A better understanding of the features of the COPE matrix that dictate the behavior of NNR and ENNR could make it easier to identify tasks and protocols that exhibit a quantum advantage. For instance, the role of NNR in classical and quantum communication complexity has been studied in Refs. [35,36]. However, to our knowledge, no prior work has explored its connection to time or space complexity of quantum circuits. Furthermore, analyzing communication and computation complexities through the lens of ENNR, hence contextuality, is an interesting direction for future research.

ACKNOWLEDGMENTS

T.Y. acknowledges the support through the Quantum Computing Studentship funded by Royal Holloway, University of London. F.S. gratefully acknowledge the financial support from the Engineering and Physical Sciences Research Council (EPSRC) through the Hub in Quantum Computing and Simulation grant (Grant No. EP/T001062/1).

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: EXPLANATION OF LINEAR PROGRAMS

Our task is to find a matrix D that minimizes the sum of the moduli of negative elements of $L\bar{E}$, where L is defined in Eq. (15). We take $\bar{E}$ to be a $k \times n$ matrix, and D to be a $(k - r) \times r$ matrix. This can be formulated as a linear program, which in general, along with its dual linear program, have the following forms:

Standard primal problem:

$$\begin{aligned} \text{Minimize} \quad & \tilde{\mathbf{c}}^\top \tilde{\mathbf{x}} \\ \text{Subject to} \quad & \tilde{\mathbf{A}}\tilde{\mathbf{x}} \geq \tilde{\mathbf{b}} \\ & \tilde{\mathbf{x}} \geq \mathbf{0} \end{aligned}$$

Standard dual problem:

$$\begin{aligned} \text{Maximize} \quad & \tilde{\mathbf{b}}^\top \tilde{\mathbf{y}} \\ \text{Subject to} \quad & \tilde{\mathbf{A}}^\top \tilde{\mathbf{y}} \leq \tilde{\mathbf{c}} \\ & \tilde{\mathbf{y}} \geq \mathbf{0} \end{aligned}$$

Now we will formulate our minimization task as a linear program.

Primal problem:

We introduce auxiliary variables $\mathbf{S}_{ps} \geq \max[0, (-L\bar{E})_{ps}]$ to represent the magnitude of negative deviations. Since D can have negative entries, we introduce non-negative variables $\mathbf{D}_{ij}^+$ and $\mathbf{D}_{ij}^-$ such that $D_{ij} = \mathbf{D}_{ij}^+ - \mathbf{D}_{ij}^-$. The primal problem is as follows:

$$\begin{aligned} \text{Minimize} \quad & \sum_{p=1}^k \sum_{s=1}^n \mathbf{S}_{ps} \\ \text{Subject to} \quad & \mathbf{S}_{ps} + \sum_{i=1}^{k-r} \sum_{j=1}^r (a^{j\top} \bar{E}_{:,s} b_p^j \mathbf{D}_{ij}^- - a^{j\top} \bar{E}_{:,s} b_p^j \mathbf{D}_{ij}^+) \\ & \geq -\bar{E}_{ps} \quad \forall p = 1, \dots, k; s = 1, \dots, n \\ & \mathbf{S}_{ps} \geq 0 \quad \forall p = 1, \dots, k; s = 1, \dots, n \\ & \mathbf{D}_{ij}^+ \geq 0 \quad \forall i = 1, \dots, k-r; j = 1, \dots, r \\ & \mathbf{D}_{ij}^- \geq 0 \quad \forall i = 1, \dots, k-r; j = 1, \dots, r. \end{aligned}$$

The first constraint represents $\mathbf{S}_{ps} \geq (-L\bar{E})_{ps}$, using the definition of L in Eq. (15).

In this case, the variable $\tilde{\mathbf{x}}$ is given by $(\mathbf{S}_{1,:} \dots \mathbf{S}_{k,:} \mathbf{D}_{1,:}^+ \dots \mathbf{D}_{k-r,:}^+ \mathbf{D}_{1,:}^- \dots \mathbf{D}_{k-r,:}^-)^\top$, which is a concatenation of three flattened matrices. Alternatively, we

can write $\tilde{\mathbf{x}} = ((\tilde{\mathbf{x}}^1)^\top \quad (\tilde{\mathbf{x}}^2)^\top \quad (\tilde{\mathbf{x}}^3)^\top)^\top$, where

$$\begin{aligned}\tilde{\mathbf{x}}_q^1 &= \mathbf{S}_{ps}, \quad p = (q-1 \bmod k) + 1, \quad s = \left\lfloor \frac{q-1}{k} \right\rfloor + 1, \\ q &\in \{1, \dots, kn\}, \\ \tilde{\mathbf{x}}_q^2 &= \mathbf{D}_{ij}^+, \quad i = \left\lfloor \frac{q-1}{r} \right\rfloor + 1, \quad j = (q-1 \bmod r) + 1, \\ q &\in \{1, \dots, (k-r)r\}, \\ \tilde{\mathbf{x}}_q^3 &= \mathbf{D}_{ij}^-, \quad i = \left\lfloor \frac{q-1}{r} \right\rfloor + 1, \quad j = (q-1 \bmod r) + 1, \\ q &\in \{1, \dots, (k-r)r\}.\end{aligned}$$

$\tilde{\mathbf{c}}^\top$ is given by kn ones followed by $2(k-r)r$ zeros. $\tilde{\mathbf{A}}$ can be written in block matrix form as follows:

$$\begin{aligned}\tilde{\mathbf{A}} &= (\mathbb{1}_{kn \times kn} \quad -\mathbf{R} \quad \mathbf{R}), \\ R_{tu} &= a^{j^\top} \tilde{E}_{:s} b_p^i, \quad t \in \{1, \dots, kn\}, \quad u \in \{1, \dots, (k-r)r\}, \\ p &= (t-1 \bmod k) + 1, \quad s = \left\lfloor \frac{t-1}{k} \right\rfloor + 1, \\ i &= \left\lfloor \frac{u-1}{r} \right\rfloor + 1, \quad j = (u-1 \bmod r) + 1.\end{aligned}\tag{A1}$$

The vector $\tilde{\mathbf{b}}$ has length kn , which is given by

$$\tilde{\mathbf{b}} = -(\tilde{E}_{:1}^\top \quad \tilde{E}_{:2}^\top \quad \dots \quad \tilde{E}_{:n}^\top)^\top.$$

Dual problem:

$$\text{Maximize} \quad - \sum_{p=1}^k \sum_{s=1}^n \tilde{E}_{ps} \mathbf{Y}_{ps}$$

Subject to

$$\begin{aligned}\sum_{p=1}^k \sum_{s=1}^n (-a^{j^\top} \tilde{E}_{:s} b_p^i) \mathbf{Y}_{ps} &\leq 0 \\ \forall i &= 1, \dots, k-r; \quad j = 1, \dots, r \\ \sum_{p=1}^k \sum_{s=1}^n (a^{j^\top} \tilde{E}_{:s} b_p^i) \mathbf{Y}_{ps} &\leq 0 \\ \forall i &= 1, \dots, k-r; \quad j = 1, \dots, r \\ \mathbf{Y}_{ps} &\leq 1 \quad \forall p = 1, \dots, k; \quad s = 1, \dots, n \\ \mathbf{Y}_{ps} &\geq 0 \quad \forall p, s,\end{aligned}$$

which we derive using the standard relationship between primal and dual linear programs, and $\tilde{\mathbf{y}} = (\mathbf{Y}_{:1}^\top \quad \mathbf{Y}_{:2}^\top \quad \dots \quad \mathbf{Y}_{:n}^\top)^\top$. Defining $H^{(ij)}$ as $H_{ps}^{(ij)} := a^{j^\top} \tilde{E}_{:s} b_p^i$, the first two constraints of the dual are equivalent to $\sum_{p,s} H_{ps}^{(ij)} \mathbf{Y}_{ps} = 0$.

APPENDIX B: LIMITATIONS OF MOITRA'S ALGORITHM

Moitra's algorithm can only find NMFs of a particular form, referred to as *stable*. Any NMF can be transformed to a stable NMF without changing the inner dimension,

as shown in Lemma 2.6 in Ref. [21], but there is no guarantee that the ranks of the factors will not change. The following terminology is needed to define stable NMFs.

Definition A1 (admissible set). An admissible set of a vector $\mathbf{v}$ is a set of vectors whose conic hull contains $\mathbf{v}$.

Definition A2 (lexicographic ordering). Given two sets $\mathbb{S} = \{s_1, s_2, \dots, s_k\}$ and $\mathbb{T} = \{t_1, t_2, \dots, t_k\}$ of the same size such that $s_1 \leq s_2 \leq \dots \leq s_k$ and $t_1 \leq t_2 \leq \dots \leq t_k$, we say that $\mathbb{S}$ is lexicographically before $\mathbb{T}$ if there is an $i \in [k]$ such that $s_j \leq t_j$ for all $j = 1, 2, \dots, i-1$ and $s_i < t_i$. Moreover, if $\mathbb{S}$ and $\mathbb{T}$ have different sizes, and $|\mathbb{S}| < |\mathbb{T}|$, then we define $\mathbb{S}$ to be lexicographically before $\mathbb{T}$.

The important property of lexicographic ordering that we will use is that a smaller set of always lexicographically before a larger set. Now, stability can be defined.

Definition A3. A non-negative matrix factorization $\mathbf{C} = \mathbf{R}\mathbf{E}$ is called *stable* if

- (1) For each column $\mathbf{C}_{:i}$, the support of $\mathbf{E}_{:i}$ is the lexicographically first admissible subset (of columns of $\mathbf{R}$) for $\mathbf{C}_{:i}$.
- (2) For each row $\mathbf{C}_{j,:}$, the support of $\mathbf{R}_{j,:}$ is the lexicographically first admissible subset (of rows of $\mathbf{E}$) for $\mathbf{C}_{j,:}$.

Now, we will prove that restricting search of factorizations to stable NMFs prevents the algorithm from finding the ENNR in general.

Lemma A1. Not every non-negative matrix that admits an ENMF admits a stable ENMF.

Proof. Consider the COPE $\mathbf{C}^1$ as defined in Eq. (25), where the factors $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{B}_{\mathcal{I}}$ correspond to two-dimensional polytopes $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{B}_{\mathcal{I}}$, respectively, as defined in Sec. V. In this case, we know that a nested polytope $\mathcal{G}$ of four vertices is not possible. Construct a new COPE $\mathbf{C}^2$ by keeping the inner polytope $\mathcal{B}_{\mathcal{I}}$ the same and constructing a new outer polytope $\mathcal{A}'_{\mathcal{U}}$. $\mathcal{A}'_{\mathcal{U}}$ is given by adding an extra vertex to $\mathcal{A}_{\mathcal{U}}$ a distance ϵ from one of the existing vertices, such that this new vertex is convexly independent of the others. Then, $\mathbf{C}^2$ is the slack of the vertices of the inner polytope $\mathcal{B}_{\mathcal{I}}$ with respect to the facets of the outer polytope $\mathcal{A}'_{\mathcal{U}}$. We can see that if ϵ is sufficiently small, then an ENMF of inner dimension 4 is still not possible. This is because the boundary displacement of the nested polytope $\mathcal{G}$ is $O(\epsilon)$. Therefore, for any such $\mathcal{G}$, the objective value of the linear program used in Lemma 3 would change by $O(\epsilon)$. Now let ϵ be sufficiently small to keep the ENNR at 5.

Using this construction, $\mathbf{C}^2$ has six rows and $\mathcal{A}'_{\mathcal{U}}$ has six facets. Then, any vertex of a nested polytope $\mathcal{G}$ of the form in Lemma 2 is a nonzero distance from at least 4 of the facets of $\mathcal{A}'_{\mathcal{U}}$. Now, rescale $\mathbf{C}^2$ to obtain a row-stochastic non-negative matrix $\mathbf{C}'^2$. For an ENMF $\mathbf{C}'^2 = \mathbf{R}'\mathbf{E}'$ of inner dimension $k \geq 5$, where the factors are, without loss of generality, row stochastic. In total, $\mathbf{R}'$ has $6k$ entries and at least $4k$ positive entries, since each vertex of $\mathcal{G}$ is a positive distance from at least four edges of $\mathcal{A}'_{\mathcal{U}}$. Assuming each row of $\mathbf{R}$ has at most three positive entries, $\mathbf{R}$ would have 18 positive entries in total, but this is not possible for $k \geq 5$. Therefore, at least two columns of $\mathbf{E}$ have four supports. Since the rows of $\mathbf{C}'^2$ can be written as convex combinations of the rows of $\mathbf{E}'$, the rows of $\mathbf{E}'$ form a two-dimensional polytope that contains the columns of $\mathbf{C}'^2$. Therefore, any row of $\mathbf{C}'^2$ can be written as the convex combination of only three rows of $\mathbf{E}'$. This means

that if the ENMF is stable, then each row of $\mathbf{R}'$ should have only three supports, which is not possible.

APPENDIX C: SIZE COMPARISON OF INNER AND OUTER PENTAGON

The polytope $\mathcal{I}$ has an affine dimension of 2 and its center is the point $\mathbf{p}$. We will show that when the simplex $\mathcal{U}$ is projected onto the affine span of $\mathcal{I}$, it forms a similar pentagon with the same center and orientation, and only differs in size.

Let M^1 be a matrix whose columns are the displacements of the vertices $\{y^i\}$ of $\mathcal{I}$ from the center $\mathbf{p}$ in a cyclic order of the vertices. Now the affine span of the columns of M^1 equals their span. Using the same order of vertices, let M^2 be the matrix whose i th column is the difference of the two adjacent vertices of y^i , e.g., $M_{:,2}^2 = M_{:,3}^1 - M_{:,1}^1$. Also, the matrix $\frac{\sqrt{5}-1}{2} \mathbb{1}_{5 \times 5}$ has columns equal to the vertices of the simplex $\mathcal{U}$. Define the matrix M^3 as $M_{:,i}^3 = (\mathbb{1}_{5 \times 5})_{:,i} - \mathbf{p}$. The first columns of these matrices are shown below using a particular choice of the order of vertices:

$$\begin{aligned}
 M_{:,1}^1 &= \frac{\sqrt{5}-1}{10} \begin{pmatrix} \sin\left(\frac{4\pi}{5}\right) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{4\pi}{5}\right) \\ \sin\left(\frac{6\pi}{5}\right) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{6\pi}{5}\right) \\ \sin\left(\frac{8\pi}{5}\right) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{8\pi}{5}\right) \\ -1 \\ \sin\left(\frac{2\pi}{5}\right) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{2\pi}{5}\right) \end{pmatrix} \\
 M_{:,1}^2 &= \frac{\sqrt{5}-1}{10} \begin{pmatrix} (\sin\left(\frac{6\pi}{5}\right) - \sin\left(\frac{2\pi}{5}\right)) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{6\pi}{5}\right) + \cos\left(\frac{2\pi}{5}\right) \\ (\sin\left(\frac{8\pi}{5}\right) - \sin\left(\frac{4\pi}{5}\right)) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{8\pi}{5}\right) + \cos\left(\frac{4\pi}{5}\right) \\ -1 - \sin\left(\frac{6\pi}{5}\right) \tan\left(\frac{\pi}{5}\right) + \cos\left(\frac{6\pi}{5}\right) \\ (\sin\left(\frac{2\pi}{5}\right) - \sin\left(\frac{8\pi}{5}\right)) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{2\pi}{5}\right) + \cos\left(\frac{8\pi}{5}\right) \\ \sin\left(\frac{4\pi}{5}\right) \tan\left(\frac{\pi}{5}\right) - \cos\left(\frac{4\pi}{5}\right) + 1 \end{pmatrix} \\
 M_{:,1}^3 &= \frac{\sqrt{5}-1}{10} \begin{pmatrix} 4 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}.
 \end{aligned} \tag{C1}$$

Due to rotational symmetry, each column of M^1 is a cyclic permutation of the first column, specifically, each entry is shifted up by one position to get the next column. Due to the definition of M^2 , it has the same permutation symmetry and it is clear that M^3 also does.

Given that $(M_{:,1}^2)^\top M_{:,1}^3 = 0$, then by the permutation symmetry, $(M_{:,i}^2)^\top M_{:,i}^3 = 0$ for all i . Given that the columns of M^1 are the vertices of a regular pentagon, this implies that when the columns of M^3 are projected onto the span of the columns of M^1 , $M_{:,i}^3$ has no component orthogonal to $M_{:,i}^1$. Also, we find that $(M_{:,i}^1)^\top M_{:,i}^3$ is positive, so $M_{:,i}^1$ is the closest column of M^1 to $M_{:,i}^3$. Therefore, we can conclude that $\mathcal{B}_{\mathcal{I}}$ and $\mathcal{P}_{\mathcal{U}}$ are similar pentagons with the same orientation. The projection of $M_{:,1}^3$ onto the span of the columns of M^1 has no component in the direction orthogonal to $M_{:,1}^1$, so the scale factor can be calculated as $\frac{(M_{:,1}^1)^\top M_{:,1}^3}{(M_{:,1}^1)^\top M_{:,1}^1}$, which simplifies to $\frac{\sqrt{5}-1}{2} \approx 0.618$.

APPENDIX D: EXPLICIT CALCULATION

The matrix G^2 is given by

$$\begin{pmatrix} \sqrt{5-2\sqrt{5}} & \frac{1}{2}\sqrt{\frac{1}{2}(\sqrt{5}+5)} & \frac{1}{4}\sqrt{10-2\sqrt{5}} & 0 & -\sqrt{\frac{1}{2}(5-\sqrt{5})} & -\sqrt{5-2\sqrt{5}} \\ -1 & \frac{1}{4}(1-\sqrt{5}) & \frac{1}{4}(\sqrt{5}+1) & \sqrt{5}-1 & \frac{1}{2}(3-\sqrt{5}) & -1 \\ \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} \end{pmatrix}, \tag{D1}$$

and the matrix $B_{\mathcal{I}}$ is given by

$$\begin{pmatrix} \sqrt{\frac{1}{2}(25-11\sqrt{5})} & \sqrt{5-2\sqrt{5}} & 0 & -\sqrt{5-2\sqrt{5}} & -\sqrt{\frac{1}{2}(25-11\sqrt{5})} \\ \frac{1}{2}(1-\sqrt{5}) & \sqrt{5}-2 & 3-\sqrt{5} & \sqrt{5}-2 & \frac{1}{2}(1-\sqrt{5}) \\ \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} & \frac{5-\sqrt{5}}{10} \end{pmatrix}. \quad (\text{D2})$$

The matrix $\bar{E}^2$ is given by

$$\begin{pmatrix} \frac{3(632\sqrt{5}-463)}{7676} & \frac{11425-7\sqrt{5}}{38380} & \frac{11805-3427\sqrt{5}}{38380} & \frac{528\sqrt{5}-1237}{7676} & \frac{5400-2443\sqrt{5}}{38380} & \frac{4576-\frac{6243}{\sqrt{5}}}{7676} \\ \frac{1}{76}(42-11\sqrt{5}) & \frac{3}{380}(11\sqrt{5}+15) & \frac{3}{380}(11\sqrt{5}+15) & \frac{1}{76}(42-11\sqrt{5}) & \frac{1}{380}(22\sqrt{5}-65) & \frac{1}{380}(22\sqrt{5}-65) \\ \frac{528\sqrt{5}-1237}{7676} & \frac{11805-3427\sqrt{5}}{38380} & \frac{11425-7\sqrt{5}}{38380} & \frac{3(632\sqrt{5}-463)}{7676} & \frac{4576-\frac{6243}{\sqrt{5}}}{7676} & \frac{5400-2443\sqrt{5}}{38380} \\ \frac{-24\sqrt{5}-31}{7676} & \frac{9(27\sqrt{5}-205)}{38380} & \frac{6895-1657\sqrt{5}}{38380} & \frac{7(495-112\sqrt{5})}{7676} & \frac{13(619\sqrt{5}-80)}{38380} & \frac{3440-\frac{2593}{\sqrt{5}}}{7676} \\ \frac{7(495-112\sqrt{5})}{7676} & \frac{6895-1657\sqrt{5}}{38380} & \frac{9(27\sqrt{5}-205)}{38380} & \frac{-24\sqrt{5}-31}{7676} & \frac{3440-\frac{2593}{\sqrt{5}}}{7676} & \frac{13(619\sqrt{5}-80)}{38380} \end{pmatrix}^T. \quad (\text{D3})$$

An orthonormal basis $\{a^1, a^2, a^3\}$ for the rows of G^2 is given by the columns of the following matrix:

$$\begin{pmatrix} 2\sqrt{\frac{1}{281}(19-4\sqrt{5})} & \frac{4(2\sqrt{5}-7)}{\sqrt{6727-2784\sqrt{5}}} & \frac{5}{2}\sqrt{\frac{15\sqrt{5}+74}{4351}} \\ \sqrt{\frac{1}{562}(29\sqrt{5}+73)} & -3\sqrt{\frac{5(5329-693\sqrt{5})}{12998498}} & \sqrt{\frac{1473-113\sqrt{5}}{8702}} \\ \sqrt{\frac{2}{37-7\sqrt{5}}} & \sqrt{\frac{477653\sqrt{5}}{12998498} + \frac{1072449}{12998498}} & \sqrt{\frac{1473-113\sqrt{5}}{8702}} \\ 0 & 2\sqrt{\frac{2623-276\sqrt{5}}{23129}} & \frac{5}{2}\sqrt{\frac{15\sqrt{5}+74}{4351}} \\ -\sqrt{\frac{2}{281}(7\sqrt{5}+37)} & 6\sqrt{\frac{5(4863-1852\sqrt{5})}{6499249}} & \frac{1}{2}\sqrt{\frac{3906-149\sqrt{5}}{4351}} \\ -2\sqrt{\frac{1}{281}(19-4\sqrt{5})} & -\sqrt{\frac{360510\sqrt{5}}{6499249} + \frac{829082}{6499249}} & \frac{1}{2}\sqrt{\frac{3906-149\sqrt{5}}{4351}} \end{pmatrix}, \quad (\text{D4})$$

and an orthonormal basis $\{b^1, b^2, b^3\}$ for the left kernel of G^2 is given by the columns of

$$\begin{pmatrix} \frac{1}{2}\sqrt{\frac{3(67\sqrt{5}+1126)}{1919}} & 0 & 0 \\ -\sqrt{\frac{435559\sqrt{5}}{7472586} + \frac{1261193}{7472586}} & 2\sqrt{\frac{250-46\sqrt{5}}{1947}} & 0 \\ -\sqrt{\frac{21055\sqrt{5}}{7472586} + \frac{48385}{7472586}} & -\sqrt{\frac{685}{3894} + \frac{1187}{3894\sqrt{5}}} & \sqrt{\frac{1}{10}(5-\sqrt{5})} \\ \frac{1}{2}\sqrt{\frac{2723614-1051741\sqrt{5}}{3736293}} & \frac{1}{2}\sqrt{\frac{305-134\sqrt{5}}{9735}} & -\frac{1}{2}\sqrt{1+\frac{2}{\sqrt{5}}} \\ \frac{1}{2}\sqrt{\frac{729766-141769\sqrt{5}}{3736293}} & \frac{1}{2}\sqrt{\frac{15865-427\sqrt{5}}{19470}} & \frac{1}{2}\sqrt{\frac{1}{10}(\sqrt{5}+5)} \\ -\frac{1}{2}\sqrt{\frac{5(459134-22213\sqrt{5})}{3736293}} & -\sqrt{\frac{1541}{15576} + \frac{3307}{15576\sqrt{5}}} & -\frac{1}{2}\sqrt{\frac{1}{10}(5-\sqrt{5})} \end{pmatrix}. \quad (\text{D5})$$

We have found an instance of $\mathbf{Y}$ given by

$$\mathbf{Y}^* = \begin{pmatrix} 0 & 0 & 0 & 5-2\sqrt{5} & \frac{1}{2}(\sqrt{5}-1) & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 5-2\sqrt{5} & 0 & 0 & 0 & 0 & \frac{1}{2}(\sqrt{5}-1) \\ 3-\sqrt{5} & 1 & \frac{1}{4}(3-\sqrt{5}) & 0 & 0 & 0 \\ 0 & \frac{1}{4}(3-\sqrt{5}) & 1 & 3-\sqrt{5} & 0 & 0 \end{pmatrix}^T. \quad (\text{D6})$$

By plugging in $\{a^1, a^2, a^3\}$, $\{b^1, b^2, b^3\}$, $\bar{E}^2$, and $\mathbf{Y}^*$ into the constraints of the dual program, it can be shown that $\mathbf{Y}^*$ is in the feasible region. The dual objective value of $\mathbf{Y}^*$ can be calculated as $\frac{1}{2}(7-3\sqrt{5}) \approx 0.145898$.

APPENDIX E: EXPLICIT MODELS

A noncontextual model $\mathbf{C}^1 = \mathbf{A}^1 \mathbf{B}^1$ of ontic dimension 5 is given by

$$\mathbf{C}^1 = \frac{1}{10} \begin{pmatrix} 2(\sqrt{5}-1) & 8-2\sqrt{5} & 2(\sqrt{5}-1) & 3-\sqrt{5} & 3-\sqrt{5} \\ 8-2\sqrt{5} & 2(\sqrt{5}-1) & 3-\sqrt{5} & 3-\sqrt{5} & 2(\sqrt{5}-1) \\ 2(\sqrt{5}-1) & 3-\sqrt{5} & 3-\sqrt{5} & 2(\sqrt{5}-1) & 8-2\sqrt{5} \\ 3-\sqrt{5} & 3-\sqrt{5} & 2(\sqrt{5}-1) & 8-2\sqrt{5} & 2(\sqrt{5}-1) \\ 3-\sqrt{5} & 2(\sqrt{5}-1) & 8-2\sqrt{5} & 2(\sqrt{5}-1) & 3-\sqrt{5} \end{pmatrix},$$

$$\mathbf{A}^1 = \frac{1}{10} \begin{pmatrix} 5-\sqrt{5} & 2\sqrt{5} & 5-\sqrt{5} & 0 & 0 \\ 0 & 5-\sqrt{5} & 2\sqrt{5} & 5-\sqrt{5} & 0 \\ 0 & 0 & 5-\sqrt{5} & 2\sqrt{5} & 5-\sqrt{5} \\ 5-\sqrt{5} & 0 & 0 & 5-\sqrt{5} & 2\sqrt{5} \\ 2\sqrt{5} & 5-\sqrt{5} & 0 & 0 & 5-\sqrt{5} \end{pmatrix}, \quad (\text{E1})$$

$$\mathbf{B}^1 = \frac{1}{10} \begin{pmatrix} 0 & 5-\sqrt{5} & 2\sqrt{5} & 5-\sqrt{5} & 0 \\ 5-\sqrt{5} & 2\sqrt{5} & 5-\sqrt{5} & 0 & 0 \\ 2\sqrt{5} & 5-\sqrt{5} & 0 & 0 & 5-\sqrt{5} \\ 5-\sqrt{5} & 0 & 0 & 5-\sqrt{5} & 2\sqrt{5} \\ 0 & 0 & 5-\sqrt{5} & 2\sqrt{5} & 5-\sqrt{5} \end{pmatrix}.$$

A quantum model $\mathbf{C}^1 = \mathbf{A}^1 \mathbf{B}^1$ of dimension 3 (subset of rebit theory) is given by

$$\mathbf{A}^1 = \begin{pmatrix} \frac{1}{5}\sqrt{5-2\sqrt{5}} & \frac{1}{\sqrt{5}} - \frac{2}{5} & \frac{1}{5} \\ 0 & \frac{3}{5} - \frac{1}{\sqrt{5}} & \frac{1}{5} \\ -\frac{1}{5}\sqrt{5-2\sqrt{5}} & \frac{1}{\sqrt{5}} - \frac{2}{5} & \frac{1}{5} \\ -\frac{1}{10}\sqrt{50-22\sqrt{5}} & \frac{1}{10}(1-\sqrt{5}) & \frac{1}{5} \\ \frac{1}{10}\sqrt{50-22\sqrt{5}} & \frac{1}{10}(1-\sqrt{5}) & \frac{1}{5} \end{pmatrix}, \quad (\text{E2})$$

$$\mathbf{B}^1 = \begin{pmatrix} 0 & \sqrt{\frac{\sqrt{5}}{8} + \frac{5}{8}} & \sqrt{\frac{5}{8} - \frac{\sqrt{5}}{8}} & -\sqrt{\frac{5}{8} - \frac{\sqrt{5}}{8}} & -\sqrt{\frac{\sqrt{5}}{8} + \frac{5}{8}} \\ 1 & \frac{1}{4}(\sqrt{5}-1) & \frac{1}{4}(-\sqrt{5}-1) & \frac{1}{4}(-\sqrt{5}-1) & \frac{1}{4}(\sqrt{5}-1) \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

This represents using a single measurement given by the POVM whose operators are of the form $A_{i1}^1 \mu_x + A_{i2}^1 \mu_z + A_{i3}^1 \mathbb{1}_{2 \times 2}$ for $i = 1, \dots, 5$, and states given by $\frac{1}{2}(B_{1j}^1 \mu_x + B_{2j}^1 \mu_z + B_{3j}^1 \mathbb{1}_{2 \times 2})$ for $j = 1, \dots, 5$.

An ontological model $\mathbf{C}^1 = \mathbf{A}^2 \mathbf{B}^2$ of ontic dimension 4 is given by

$$\mathbf{A}^2 = \begin{pmatrix} 0 & 1 - \frac{1}{\sqrt{5}} & 0 & \frac{1}{10}(5 - \sqrt{5}) \\ \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 & 0 \\ 1 - \frac{1}{\sqrt{5}} & 0 & \frac{1}{10}(5 - \sqrt{5}) & 0 \\ 0 & 0 & 1 - \frac{1}{\sqrt{5}} & \frac{1}{10}(3\sqrt{5} - 5) \\ 0 & 0 & \frac{1}{10}(3\sqrt{5} - 5) & 1 - \frac{1}{\sqrt{5}} \end{pmatrix},$$

$$\mathbf{B}^2 = \frac{1}{20} \begin{pmatrix} 20\left(\frac{2}{\sqrt{5}} - \frac{1}{2}\right) & 5 - \sqrt{5} & 0 & 2(3\sqrt{5} - 5) & 3(5 - \sqrt{5}) \\ 20\left(\frac{2}{\sqrt{5}} - \frac{1}{2}\right) & 3(5 - \sqrt{5}) & 2(3\sqrt{5} - 5) & 0 & 5 - \sqrt{5} \\ 20\left(1 - \frac{2}{\sqrt{5}}\right) & 0 & 2(5 - \sqrt{5}) & 20\left(1 - \frac{1}{\sqrt{5}}\right) & 4\sqrt{5} \\ 20\left(1 - \frac{2}{\sqrt{5}}\right) & 4\sqrt{5} & 20\left(1 - \frac{1}{\sqrt{5}}\right) & 2(5 - \sqrt{5}) & 0 \end{pmatrix}. \quad (\text{E3})$$

- [1] S. Kochen and E. Specker, The problem of hidden variables in quantum mechanics, *Indiana Univ. Math. J.* **17**, 59 (1967).
- [2] N. D. Mermin, Hidden variables and the two theorems of John Bell, *Rev. Mod. Phys.* **65**, 803 (1993).
- [3] R. W. Spekkens, Contextuality for preparations, transformations, and unsharp measurements, *Phys. Rev. A* **71**, 052108 (2005).
- [4] J. Bermejo-Vega, N. Delfosse, D. E. Browne, C. Okay, and R. Raussendorf, Contextuality as a resource for models of quantum computation with qubits, *Phys. Rev. Lett.* **119**, 120505 (2017).
- [5] F. Shahandeh, Quantum computational advantage implies contextuality, [arXiv:2112.00024](#).
- [6] D. Schmid, H. Du, J. H. Selby, and M. F. Pusey, Uniqueness of noncontextual models for stabilizer subtheories, *Phys. Rev. Lett.* **129**, 120403 (2022).
- [7] S. Gupta, D. Saha, Z.-P. Xu, A. Cabello, and A. S. Majumdar, Quantum contextuality provides communication complexity advantage, *Phys. Rev. Lett.* **130**, 080802 (2023).
- [8] D. Schmid and R. W. Spekkens, Contextual advantage for state discrimination, *Phys. Rev. X* **8**, 011015 (2018).
- [9] F. Shahandeh, T. Yianni, and M. Doosti, Characterizing contextuality via rank separation with applications to cloning, [arXiv:2406.19382](#).
- [10] M. Lostaglio and G. Senno, Contextual advantage for state-dependent cloning, *Quantum* **4**, 258 (2020).
- [11] K. Korzekwa and M. Lostaglio, Quantum advantage in simulating stochastic processes, *Phys. Rev. X* **11**, 021019 (2021).
- [12] V. Havlíček and J. Barrett, Simple communication complexity separation from quantum state antidistinguishability, *Phys. Rev. Res.* **2**, 013326 (2020).
- [13] A. Montina, Exponential complexity and ontological theories of quantum mechanics, *Phys. Rev. A* **77**, 022104 (2008).
- [14] L. Hardy, Quantum ontological excess baggage, *Stud. Hist. Philos. Sci. Part B: Stud. Hist. Philos. Mod. Phys.* **35**, 267 (2004).
- [15] Y. Shitov, Euclidean distance matrices and separations in communication complexity theory, *Discr. Comput. Geom.* **61**, 653 (2019).
- [16] V. Gitten and M. P. Woods, Solvable criterion for the contextuality of any prepare-and-measure scenario, *Quantum* **6**, 732 (2022).
- [17] E. F. Galvão and L. Hardy, Substituting a qubit for an arbitrarily large number of classical bits, *Phys. Rev. Lett.* **90**, 087902 (2003).
- [18] J. H. Selby, E. Wolfe, D. Schmid, A. B. Sainz, and V. P. Rossi, Linear program for testing nonclassicality and an open-source implementation, *Phys. Rev. Lett.* **132**, 050202 (2024).
- [19] D. Schmid, R. D. Baldijão, J. H. Selby, A. B. Sainz, and R. W. Spekkens, Noncontextuality inequalities for prepare-transform-measure scenarios, [arXiv:2407.09624](#).
- [20] F. Shahandeh, T. Yianni, and M. Doosti, A unified linear algebraic framework for physical models and generalized contextuality, [arXiv:2512.10000](#).
- [21] A. Moitra, An almost optimal algorithm for computing nonnegative rank, *SIAM J. Comput.* **45**, 156 (2016).
- [22] S. A. Vavasis, On the complexity of nonnegative matrix factorization, *SIAM J. Opt.* **20**, 1364 (2010).
- [23] R. W. Spekkens, The ontological identity of empirical indiscernibles: Leibniz's methodological principle and its significance in the work of Einstein, [arXiv:1909.04628](#).
- [24] F. Shahandeh, Contextuality of general probabilistic theories, *PRX Quantum* **2**, 010330 (2021).
- [25] S. Arora, R. Ge, R. Kannan, and A. Moitra, Computing a nonnegative matrix factorization—Provably, in *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '12 (Association for Computing Machinery, New York, 2012), pp. 145–162.
- [26] R. Impagliazzo and R. Paturi, On the complexity of k-SAT, *J. Comput. Syst. Sci.* **62**, 367 (2001).
- [27] B. Aspvall and R. E. Stone, Khachiyan's linear programming algorithm, *J. Algorithms* **1**, 1 (1980).
- [28] J. Renegar, A polynomial-time algorithm, based on Newton's method, for linear programming, *Math. Program.* **40**, 59 (1988).
- [29] M. E. Dyer, The complexity of vertex enumeration methods, *Math. Oper. Res.* **8**, 381 (1983).
- [30] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich, Generating all vertices of a polyhedron is hard, *Discr. Comput. Geom.* **39**, 174 (2008).
- [31] G. Das and D. Joseph, The complexity of minimum convex nested polyhedra, in *The 2nd Canadian Conference on Computational Geometry* (CCCG, Ottawa, 1990), pp. 296–301.
- [32] M. G. Dobbins, A. Holmsen, and T. Miltzow, A universality theorem for nested polytopes, [arXiv:1908.02213](#).
- [33] A. Storjohann, On the complexity of inverting integer and polynomial matrices, *Comput. Complex.* **24**, 777 (2015).
- [34] D. Schmid, J. H. Selby, E. Wolfe, R. Kunjwal, and R. W. Spekkens, Characterization of noncontextuality in the framework of generalized probabilistic theories, *PRX Quantum* **2**, 010331 (2021).
- [35] O. Lalonde, N. S. Mande, and R. de Wolf, Tight bounds for the randomized and quantum communication complexities of equality with small error, in *43rd IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, edited by P. Bouyer and S. Srinivasan (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2023), Vol. 284, pp. 32:1–32:18.
- [36] T. Heinosaari, O. Kerppo, L. Leppäjärvi, and M. Plávala, Simple information processing tasks with unbounded quantum advantage, *Phys. Rev. A* **109**, 032627 (2024).