Bounding Classical and Quantum Correlations in Bayesian Networks with Quasiprobabilities

Paul Becsi and Matty J. Hoban
Quantum Group, Dept. of Computer Science
University of Oxford

1 Introduction

Bell's theorem reveals that quantum theory is in tension with classical causal reasoning [1, 2]. Local causality can be defined with respect to a Bayesian network [3], where the nodes of the directed acyclic graph (DAG) represent either the experimental (observed) variables or the hidden (latent) variable, as depicted in Fig. 1. Given this DAG one can derive inequality constraints: the Bell inequalities. What is particularly appealing about this presentation is that it permits a general framework for the study of quantum correlations. That is, we can ask about quantum and classical correlations in different DAGs with observed and latent nodes. This has resulted in a rich literature at the intersection of quantum theory and classical causal inference [4, 5, 6]. One aspect of this is to derive (inequality) constraints for classical models, and another to find quantum violations of these constraints. There are now good techniques for both the former such as the inflation technique [7] and the quantum inflation technique for the latter [8].

It is well established that quantum theory is just one example of a class of generalised probabilistic theories (GPTs) [9]. These GPTs permit violations of Bell inequalities that exceed those permitted by quantum theory; the Popescu-Rohrlich (PR) box is a notable example [10]. Given this observation, Henson, Lal and Pusey in [11] considered a theory-independent approach to studying correlations in Bayesian networks. In this work, they presented a generalisation of Bayesian networks (with latents) that allow one to incorporate elements from any (causal) GPT. That is, latent variables could be replaced by preparations of systems in some (possibly post-quantum) theory. Not only is this of foundational interest, it also led to new methods for finding new scenarios where one could see quantum systems producing non-classical correlations; these are called *interesting* DAGs. Within [11] a condition was proposed for identifying such DAGs and there has been recent progress on determining if it captures all interesting DAGs [12].

Within the field of classical causal inference itself, the study of marginal probability distributions for DAGs with latent variables has seen much progress [6, 13, 14]. Motivated by the *d-separation criterion*, one can describe a couple of outer approximations to the set of classical correlations. These are the *nested Markov* and *ordinary Markov* models, with the former contained in the latter. In [14], it was shown that the set of classical correlations is "algebraically equivalent" to the nested Markov model. Furthermore, the set of distributions in the nested Markov model contains all distributions that can be produced by GPT systems. For the Bell scenario, the nested Markov and ordinary Markov sets coincide with the set of all non-signalling correlations.

Within the study of Bell's theorem and GPTs, there has been some study of *quasiprobabilities*. That is, if one replaces the distribution over hidden variables (and the conditional probabilities for outcomes) with signed real numbers, one can again get correlations beyond quantum. That is, the probabilities over latent variables can become "negative probabilities". It has been established many times independently that such generalisations of local hidden variable models can maximally violate any Bell inequality and even simulate any non-signalling probability distribution [15, 16, 17, 18, 19]. This naturally begs the question of whether this is true for other DAGs. This work addresses the role of quasiprobabilities for Bayesian networks, and what is their relationship to the aforementioned sets of correlations.

For some brief notation, given a DAG $\mathcal{G}$ let $C(\mathcal{G})$, $Q(\mathcal{G})$, $G(\mathcal{G})$ and $W(\mathcal{G})$ be the sets of classical, quantum, GPT[1] and quasiprobabilistic correlations. It should be clear that $C(\mathcal{G}) \subseteq Q(\mathcal{G}) \subseteq G(\mathcal{G})$ Furthermore, define $N(\mathcal{G})$ to be the nested Markov set for $\mathcal{G}$. Henson, Lal and Pusey showed (implicitly) that $G(\mathcal{G})$ is *strictly* contained in $N(\mathcal{G})$ for certain DAGs G. We can ask the following questions:

- Is $G(\mathcal{G}) \subseteq W(\mathcal{G})$?
- Is $W(\mathcal{G}) \subseteq N(\mathcal{G})$?
- Is $N(\mathcal{G}) \subseteq W(\mathcal{G})$?

[1] The union of set of correlations over all GPTs.

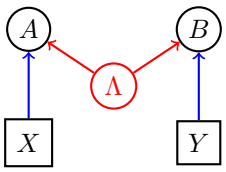


Figure 1: The bipartite Bell Scenario.

Figure 2: The 4-on-line Scenario.

It is relatively straightforward to prove that the answer to the first question is yes following work such as [9]. We can also prove that the second question has a positive response. Finally, we conjecture that the answer to the third question is also yes. Our work makes partial progress on resolving this conjecture for a broad family of DAGs.

2 Contributions

In our work we give a definition of the set of correlations from quasiprobabilistic models $W(\mathcal{G})$, we also call this set the *quasi set*. In full generality, one could all "non-observed" probability distributions to become quasiprobabilistic. That is, the distribution over latent variables can become quasiprobabilistic, as well as the conditional probabilities of the form $p(A|B,\lambda)$, where A and B are observed variables and λ is latent.

In addition to proving that the set $W(\mathcal{G})$ is contained in the Nested Markov model, we prove the following:

- In the context of any DAG, it doesn't matter if one allows negative probabilities for the latent variables, or negative probabilities for the response functions (or both), the resulting set of correlations is the same, i.e. still $W(\mathcal{G})$;
- In the context of correlation scenarios [4] such that the latent projection has no cycles when viewed as an undirected graph[2], the quasi set coincides with the nested Markov set. We call such graphs *tree-structured correlation scenarios*.

While we do not completely resolve our conjecture, the set of tree-structured correlation scenarios is non-trivial. In [4], the 4-on-line scenario (Fig. 2) also exhibits the Clauser-Horne-Shimony-Holt (CHSH) inequalities, thus the setting has non-trivial inequalities that can be violated. Furthermore, there are scenarios related to well studied networks, such as the *star network* [20], for instance.

In the proof of the second result, we make use of an interesting link between tensor network decompositions [21] and the factorisation criterion used in graphical models. Tensor network decompositions are generally used in the context of efficiently storing tensors, such as in machine learning or classical simulation of quantum circuits, but also multi-body quantum physics, and have even been related to undirected graphical models [22]. In our proof, we show they are linked to directed graphical models too, at least in the context of tree-structured correlation scenarios. For example, in the 4-on-line scenario, a distribution in the nested Markov set has a quasiprobabilistic latent variable construction of given size if and only if the same distribution viewed as a tensor has a network decomposition with respect to the graph in Fig. 3 of the same size. This connection between tensor network decompositions and causal networks could be of independent interest.

3 Summary and Future Work

In conclusion, we have studied the behaviour of quasiprobabilities in Bayesian networks. We have introduced the quasi set and showed that it coincides with the nested Markov set for all tree-structured correlation scenarios. As mentioned, we conjecture that for general DAGs, the quasi set coincides with the nested Markov model. Clearly, this general case remains open. We can show that, by an explicit construction, that for the *triangle scenario* [4], again the quasi set and the nested Markov model coincide. This triangle scenario is not tree-structured; the triangle gives a cycle. The triangle scenario is interesting because, for this DAG, $G(\mathcal{G})$ is strictly contained in the nested Markov set, thus we have a separation between $G(\mathcal{G})$ and $W(\mathcal{G})$ here as well.

[2]Note that this condition implies that any latent variable has at most two children. Otherwise, latent projection introduces (bidirected) edges between each pair of children, and hence cycles will exist.

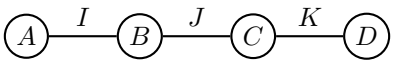


Figure 3: The tensor network $\mathcal{G}^t$ corresponding to the 4-on-line Scenario.

An obvious direction for future work would be to prove the conjecture result for all DAGs. A first step would be to consider scenarios that reduce to a tree-structured correlation scenario, similarly to the 4-on-line scenario. The *unpacking* technique [23] can also relate general networks to correlation scenarios. Along these lines, it would be interesting to see how tools such as the inflation technique [7] can be amended to accommodate quasiprobabilities and whether this gives something non-trivial.

There is also an interesting connection between quasiprobabilities and GPTs, first discussed in [24], where the notion of a *non-free* GPT was introduced: a GPT that is, in some sense, not completely closed under composition. It was shown that there exists a non-free theory that can simulate quantum computation and solve certain problems thought to be hard for quantum computers, but does not have ridiculous computational power. The construction of the non-free theory came from constructing a quasiprobabilistic Turing Machine: a Turing Machine where transitions occur with a possibly negative probability. The separation between $G(\mathcal{G})$ and $W(\mathcal{G})$ could be interpreted as the quasiprobabilistic models being related to non-free GPTs, and as soon as we restrict to *free* GPTs, we restrict the set of possible correlations.

References

1. Bell, J.S.: On the Einstein Podolsky Rosen paradox. Physics Physique Fizika **1**, 195–200 (1964). https://doi.org/10.1103/PhysicsPhysiqueFizika.1.195. https://link.aps.org/doi/10.1103/PhysicsPhysiqueFizika.1.195
2. Wood, C.J., Spekkens, R.W.: The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning. New Journal of Physics **17**(3) (2015). https://doi.org/10.1088/1367-2630/17/3/033002
3. Pearl, J.: Causality. Cambridge University Press (2009)
4. Fritz, T.: Beyond Bell's Theorem: Correlation Scenarios. New Journal of Physics **14** (2012). https://doi.org/10.1088/1367-2630/14/10/103001
5. Fritz, T.: Beyond Bell's Theorem II: Scenarios with Arbitrary Causal Structure. Communications in Mathematical Physics **341**(2), 391–434 (2015). https://doi.org/10.1007/s00220-015-2495-5
6. Evans, R.J.: Graphs for Margins of Bayesian Networks. Scandinavian Journal of Statistics **43**(3), 625–648 (2015). https://doi.org/10.1111/sjos.12194
7. Wolfe, E., Spekkens, R.W., Fritz, T.: The Inflation Technique for Causal Inference with Latent Variables. Journal of Causal Inference **7**(2) (2019). https://doi.org/10.1515/jci-2017-0020
8. Wolfe, E., Pozas-Kerstjens, A., Grinberg, M., Rosset, D., Acín, A., Navascués, M.: Quantum Inflation: A General Approach to Quantum Causal Compatibility. Physical Review X **11**. https://doi.org/10.1103/PhysRevX.11.021043
9. Barrett, J.: Information processing in generalized probabilistic theories, (2006). arXiv: quant-ph/0508211 [quant-ph]. https://arxiv.org/abs/quant-ph/0508211.
10. Popescu, S., Rohrlich, D.: Quantum Nonlocality as an Axiom. Foundations of Physics **24**(3) (1994). https://doi.org/10.1007/BF02058098
11. Henson, J., Lal, R., Pusey, M.F.: Theory-independent limits on correlations from generalized Bayesian networks. New Journal of Physics **16**(11) (2014). https://doi.org/10.1088/1367-2630/16/11/113043
12. Khanna, S., Ansanelli, M.M., Pusey, M.F., Wolfe, E.: Classifying causal structures: Ascertaining when classical correlations are constrained by inequalities. Phys. Rev. Res. **6**, 023038 (2024). https://doi.org/10.1103/PhysRevResearch.6.023038. https://link.aps.org/doi/10.1103/PhysRevResearch.6.023038
13. Shpitser, I., Evans, R.J., Richardson, T.S., Robins, J.M.: Introduction to nested Markov models. Behaviormetrika **41**, 3–39 (2014)
14. Evans, R.J.: Margins of discrete Bayesian networks. The Annals of Statistics **46**(6A) (2018). https://doi.org/10.1214/17-aos1631
15. Fine, A.: Hidden Variables, Joint Probability, and the Bell Inequalities. Phys. Rev. Lett. **48**, 291–295 (1982). https://doi.org/10.1103/PhysRevLett.48.291. https://link.aps.org/doi/10.1103/PhysRevLett.48.291
16. Al-Safi, S.W., Short, A.J.: Simulating all Nonsignaling Correlations via Classical or Quantum Theory with Negative Probabilities. Physical Review Letters **111**(17) (2013). https://doi.org/10.1103/physrevlett.111.170403
17. Degorre, J., Kaplan, M., Laplante, S., Roland, J.: "The Communication Complexity of Non-signaling Distributions". In: Mathematical Foundations of Computer Science 2009. Springer Berlin Heidelberg, 2009, pp. 270–281. ISBN: 9783642038167. https://doi.org/10.1007/978-3-642-03816-7_24.
18. Abramsky, S., Brandenburger, A.: The sheaf-theoretic structure of non-locality and contextuality. New Journal of Physics **13**(11), 113036 (2011). https://doi.org/10.1088/1367-2630/13/11/113036
19. Morris, B., Fiderer, L.J., Lang, B., Goldwater, D.: Witnessing Bell violations through probabilistic negativity. Phys. Rev. A **105**, 032202 (2022). https://doi.org/10.1103/PhysRevA.105.032202. https://link.aps.org/doi/10.1103/PhysRevA.105.032202

20. Tavakoli, A., Skrzypczyk, P., Cavalcanti, D., Acín, A.: Nonlocal correlations in the star-network configuration. Phys. Rev. A **90**, 062109 (2014). https://doi.org/10.1103/PhysRevA.90.062109. https://link.aps.org/doi/10.1103/PhysRevA.90.062109

21. Ye, K., Lim, L.-H.: Tensor network ranks, (2019). arXiv: 1801.02662 [math.NA]. https://arxiv.org/abs/1801.02662.

22. Robeva, E., Seigal, A.: Duality of graphical models and tensor networks. Information and Inference: A Journal of the IMA **8**(2), 273–288 (2018). https://doi.org/10.1093/imaiai/iay009. eprint: https://academic.oup.com/imaiai/article-pdf/8/2/273/28864933/iay009.pdf

23. Navascués, M., Wolfe, E.: The Inflation Technique Completely Solves the Causal Compatibility Problem. Journal of Causal Inference **8**(1), 70–91 (2020). https://doi.org/10.1515/jci-2018-0008

24. Barrett, J., de Beaudrap, N., Hoban, M.J., Lee, C.M.: The computational landscape of general physical theories. npj Quantum Information **5**(1) (2019). https://doi.org/10.1038/s41534-019-0156-9

Bounding Classical and Quantum Correlations in Bayesian Networks with Quasiprobabilities

Paul Becsi and Matty J. Hoban

University of Oxford

It is known that negative probability distributions can generate any correlation in the non-signaling set of the Bell Scenario. Motivated by this, we formalise a graphical model framework in which probabilities are allowed to live in the entirety of $\mathbb{R}$. We conjecture that a generalisation of the aforementioned result holds for arbitrary directed acyclic graphs, and present some initial work towards a proof using tensor network decompositions.

1 Introduction

Directed graphical models are a tool used frequently in all fields related to statistical and causal modeling, such as medicine [1], psychology [2], economics [3], social studies, and machine learning [4]. Under a causal interpretation, they provide an attractive, intuitive formalism for cause-effect experiments. Going back as far as the 1920s with some work carried out by Wright (see for example [5]), they are now a central concept in the field of Causality (see [6], [7]), and extensive research has been carried out to understand their behaviour.

A Directed Acyclic Graph (DAG) consists of a set of vertices V and a set of directed edges $E \subset V \times V$. In DAG modeling, each vertex $v \in V$ is associated with a random variable X_v from a sample space $\mathcal{X}_v$. We call a *Causal Structure* the tuple consisting of the graph $\mathcal{G}$ together with the sample spaces of the vertices. An edge between two vertices can be interpreted as a direct causal influence between the two corresponding variables, following the direction of the edge. Intuitively, this might imply a temporal ordering on the two variables, hence the acyclicity condition: it is less intuitive what it means for a variable to be an (indirect) cause of itself (although such scenarios are the object of current research, see [8], [9] for example). Alternatively, one can interpret the lack of an edge between two variables as the absence of a channel between them. The questions of interest are "Given a distribution, what are the causal structures it is consistent with?" and "Given a causal structure, what is the set of distributions consistent with it?". Of course, one needs to define mathematically what "consistency" means in this context.

In the simplest setting where all variables are observed, consistency is given by the three equivalent Markov properties: the factorisation criterion, the local Markov property, and the global Markov property. If latent variables are present, classically one is interested in the set of observed distributions that can be extended to a full distribution Markov with respect to the graph. For a given graph $\mathcal{G}$, characterising this *marginal model*, which we will call *the classical set* and denote $\mathcal{C}(\mathcal{G})$, turns out to be much more difficult and has been the focus of extensive study.

A distribution consistent with a latent variable DAG must satisfy some independence constraints, but these are not sufficient. In general, some other polynomial equality and

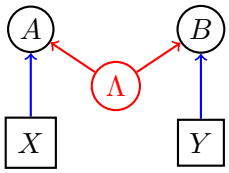


Figure 1: The bipartite Bell Scenario.

inequality constraints must be satisfied as well. The former, known as Verma constraints, were first mentioned in 1990 in [10] and then were also studied by Spirtes in [11]. In 2002 Tian and Pearl [12] proposed an algorithm that finds such constraints, which is now known to be complete by the results in [13].

Inequality constraints however are quite difficult to characterise in general. Therefore, for practical applications they are sometimes omitted and one simply looks at the set of distributions satisfying either the fully observed independence constraints following from d-separation [14] (called the ordinary Markov model) or more strongly all equality constraints [15], [16] (called the nested Markov model). The latter is a subset of the former, since independence constraints are a subset of all equality constraints. In fact, the actual definition of the nested Markov model in [15] is more involved, but the main result in [13] allows us to use this simpler one. They are used in practice as approximations to the classical set, given their nicer properties. Parameterisations exist for the ordinary Markov model [17] and the nested Markov model [18], which makes them convenient to use for model fitting [19].

In parallel, Quantum Theory has developed tremendously in the last century, scientists now generally believing it to be the correct theory to describe Nature, as recognised by the Nobel Prize in Physics in 2022. One central concept in the theory and perhaps one of its most surprising facets is the phenomenon of nonlocality: two spacelike separated systems should intuitively not be able to interact, but there are quantum effects which cannot be explained under this assumption. Having its roots in the EPR paradox [20] published in 1935, Bell inequalities [21] play a crucial role in understanding nonlocality, asserting that quantum correlations cannot be explained via hidden variables. Formalised in a testable way by Clauser, Horne, Shimony, and Holt in 1969, violations of Bell inequalities have been repeatedly confirmed experimentally (see [22] and [23] for example). The conclusion is that indeed quantum predictions seem to model the behaviour of Nature accurately, and hence that reality is nonlocal. These discussions led to other discoveries in Quantum Theory, such as quantum steering [24], the CHSH game, and the GHZ game [25], [26], the physical basis of the latter having been first implemented in [27] and then used in another nonlocality test in [28].

Recently, researchers noticed graphical models are useful in the study of nonlocality. The idea to use graphical reasoning dates back to papers such as [29] and [30], the latter explicitly using algorithms in (classical) Causality literature for the purposes of Quantum research. Multiple mathematical formalisms trying to describe what "Quantum Causality" should look like were proposed, but there is no clear consensus on one as of yet: The Hansen-Lal-Pusey formalism [31] is one of the more popular ones, Fritz [32] developed some theory for correlation scenarios (essentially graphs with no edges between observed variables) and subsequently extended them in [33], and some more inherently-quantum ones are [34], [35], [36]. Conversely, (classical) Causality has seen some progress coming from the Quantum side, some examples including [21], [32], [37], and the general full asymptotic

characterisation of the classical set via the inflation technique in [38], [39][1].

As a concrete example of nonlocality formulated using graphical models, see the Bipartite Bell Scenario in Fig. 1, graph which we will call $\mathcal{B}_2$. A distribution[2] $p(ab|xy)$ is in the classical set $\mathcal{C}(\mathcal{B}_2)$ if it satisfies the factorisation condition

$$p(ab|xy) = \sum_{\lambda \in \mathcal{X}_\Lambda} p(\lambda) \cdot p(a|x,\lambda) \cdot (b|y,\lambda), \tag{1}$$

for some sample space $\mathcal{X}_\Lambda$ and (conditional) distributions $p(a|x,\lambda)$, $p(b|y,\lambda)$, and $p(\lambda)$. Any distribution in the classical set must satisfy the *CHSH inequality* [41]:

$$p(00|00)+p(11|00)+p(00|01)+p(11|01)+p(00|10)+p(11|10)+p(01|11)+p(10|11) \leq 3 \tag{2}$$

Indeed, there are distributions in the quantum set defined in [31] that violate this inequality, which in the interpretation of the CHSH game attests nonlocality.

One of the problems of interest in Quantum Causality is the task of characterising the quantum set, with applications in communication complexity problems [42] and device-independent key distribution [43]. This turns out to be much more difficult than its classical counterpart, mainly because of the unorderly structure of the quantum set: even in the Bell Scenario, whose classical set has been fully characterised as a finite polytope [44], the quantum set is not a finite polytope [45], and in fact not even closed[3] [47].

Given the difficulty of studying the quantum set exactly, it might be worth studying relaxations of it. It is known in any Bell Scenario (i.e. one latent variable and any number of observed variables with separate inputs) that if the latent variable is allowed to follow a quasiprobability distribution (a distribution with possibly negative probabilities), all non-signaling correlations can be achieved [48], [49]. In particular, quantum correlations are therefore special cases of these quasiprobabilistic correlations. Using the framework in [50], we believe this inclusion can be extended to all graphs (although we won't do so this in this paper), in which case we could upper bound the quantum set by this new *quasi set*.

It is hence worth studying what the set of quasicorrelations is in general. It is a folklore belief that quasicorrelations do not have to obey inequality constraints (except for the trivial ones stating the observed distribution is nonnegative). Formally, this is equivalent to the quasi set coinciding with the nested Markov model for any given graph, statement for which there aren't any published proofs as of yet to the best of our knowledge. Validity of this conjecture would imply the quasi set has a very orderly structure (given that the nested Markov model is quite well-understood), and would also give another interpretation to inequality constraints on the classical and quantum sets: inequality constraints come from non-negativity conditions on the former, and from restrictions on which negative distributions can be formulated as quantum states for the latter. In particular, no inequality constraints are inherited from simply using a subtheory of quasiprobability theory.

[1] It is worth mentioning that a quantum counterpart to the inflation technique was introduced in [40]. However, unlike the classical version, the quantum version is not known to be complete.

[2] When discussing Bell Scenarios, it is customary to consider the observed distributions to be conditional over some observed "inputs", in our case X and Y. In the formalism of [15], this is the same as considering the Bell Scenario to be a CDAG with fixed variables X and Y and reasoning about the probability kernels $p(ab|xy)$ consistent with it. In said paper, fixed variables are drawn as squares, convention which we kept in Fig. 1.

[3] In defining the quantum set, there is some subtlety arising from the need to decide whether infinite-dimensional shared quantum states are allowed. If only finite-dimensional resources are allowed, the set is often called C_q, whereas if infinite-dimensional resources are allowed, it is called C_{qs}. It turns out this subtlety makes a difference, as $C_q \neq C_{qs}$. In fact, neither of the two sets is closed. We point the reader to section 1.3 of [46] for a more detailed discussion of the history of this issue and references.

It is certainly not the first time negative probabilities have been mentioned in science, especially not in the context of Quantum Theory. One of the most famous examples dates back to 1932, with Wigner proposing the phase-state representation of quantum states [51], which has seen extensive use in the study of Quantum Mechanics. Feynman also famously vouched for the use of negative probabilities in Physics [52]. In computer science they can be found in the context of affine finite automata [53] and affine Turing machines [54]. We also mention negative probabilities are used in the proof of a classical Causality result in [55].

1.1 Contributions

The main contributions of this paper are generalisations of the results in [48]. In that paper, the authors show that any non-signaling distribution in the Bell Scenario (including non-classical ones) can be written in the form described in (1) if we allow either $p(\lambda)$ or the response functions $p(a|x,\lambda)$ and $p(b|y,\lambda)$ to be negative. Firstly, we will prove that for any graph, it doesn't matter whether we allow latent variables or response functions to be negative; the set of nonnegative observed distributions is the same in both cases. We will define the *quasi set* as the set of distributions which can be written in either of these two equivalent forms. Secondly, we will study whether this quasi set contains the set of all non-signaling distributions in general, i.e. whether it coincides with the nested Markov model for any given graph. We will conjecture that this is indeed true for any DAG and will provide a proof of this result for correlation scenarios with no undirected cycles.

1.2 Paper Structure

We start by introducing graphical models and some results in literature (mainly classical causality results) in Section 2. The next two sections then contain our main results: in Section 3 we talk about defining the quasi set and show that all definitions we mention are equivalent. In Section 4 we conjecture that the quasi set coincides with the nested Markov model for all DAGs. We also introduce Tensor Network Decompositions and show how they are related to directed graphical models in the context of correlation scenarios, and prove the conjecture for tree-structured correlation scenarios. Finally, Section 5 contains some conclusions and future work. Most of the results in the main text are stated without proof, with the proofs deferred to the Appendix.

2 Directed Graphical Models

A directed graph $\mathcal{G}$ consists of a set of vertices V together with a set of edges $E \subseteq V \times V$, where edges from a vertex to itself are not allowed. We follow the convention of using small letters for vertices and big letters for sets of vertices and will denote the edge (v,w) by $v \to w$. A directed path consists of a sequence of edges $v_1 \to v_2, v_2 \to v_3, ..., v_{n-1} \to v_n$. A directed cycle is a directed path together with an additional edge $v_n \to v_1$. A directed acyclic graph (DAG) is a directed graph that has no directed cycles. Often we will write the set of vertices of $\mathcal{G}$ as $V \dot{\cup} L$, to make clear the graph has a set of observed nodes V and a set of latent nodes L.

As usual, given a vertex v, we denote by $\mathrm{pa}_{\mathcal{G}}(v) = \{w \in V : w \to v \in E\}$ and $\mathrm{ch}_{\mathcal{G}}(v) = \{w \in V : v \to w \in E\}$ the sets of parents and children of v. We also define the sets

$$\mathrm{an}^0_{\mathcal{G}}(v) = \mathrm{de}^0_{\mathcal{G}}(v) = \{v\},$$

$$\mathrm{an}_{\mathcal{G}}^{n+1}(v) = \{w \in V : \exists v' \in \mathrm{an}_{\mathcal{G}}^{n}(v).\, w \to v' \in E\},$$
$$\mathrm{de}_{\mathcal{G}}^{n+1}(v) = \{w \in V : \exists v' \in de_{\mathcal{G}}^{n}(v).\, v' \to w \in E\}$$

and finally the sets of ancestors, descendants, and nondescendants of v as $\mathrm{an}_{\mathcal{G}}(v) = \bigcup_{n=0}^{\infty} \mathrm{an}_{\mathcal{G}}^{n}(v)$, $\mathrm{de}_{\mathcal{G}}(v) = \bigcup_{n=0}^{\infty} \mathrm{de}_{\mathcal{G}}^{n}(v)$, and $\mathrm{nd}_{\mathcal{G}}(v) = V \setminus \mathrm{de}_{\mathcal{G}}(v)$. We define all of the above sets for sets of vertices disjunctively, i.e. $\mathrm{pa}_{\mathcal{G}}(A) = \bigcup_{v \in A} \mathrm{pa}_{\mathcal{G}}(v)$. Often we will drop the subscript if the graph is clear from context.

When talking about random variables, X_v will denote the random variable corresponding to vertex v, whereas x_v will denote a value in its sample space $\mathcal{X}_v$. Similarly, X_A and x_A will denote vectors of these quantities indexed by the elements of set A, while $\mathcal{X}_A$ will denote the product space $\prod_{v \in A} \mathcal{X}_v$. For a distribution p over X_V we will write $p(x_V)$ for the probability of X_V taking value x_V under p, and will write $X_A \perp\!\!\!\perp X_B \,|\, X_C \;\; [p]$ to mean the variables X_A and X_B are independent given X_C in p. Sometimes if we want to make explicit that p is defined on some vertex set V, we might refer to the distribution by $p(x_V)$; whether this notation refers to a number denoting an entry of p or the whole distribution itself should hopefully be clear from context.

A DAG $\mathcal{G}$, some sample spaces for its observed vertices $\{\mathcal{X}_v\}_{v \in V}$, and a consistency condition generate a graphical model: a set of probability distributions which are said to be "consistent with the graph", or "explained by the graph", or "generated by the graph". Customarily, the model will contain distributions defined over the set of observed nodes only. If no distinction between observed or latent nodes is made, the convention is the entire graph is observed. Often we will just assume the sample spaces of the observed variables are implicit, so we might omit them when defining models. Also, we will mostly work with finite observed variables, and all product spaces are assumed by default.

The most common consistency condition found in literature is the *factorisation criterion*. We say a distribution p *factorises with respect to* DAG $\mathcal{G}$ with vertices $V \dot{\cup} L$ if the following equation holds

$$p(x_V) = \int_{x_L} \prod_{v \in V \cup L} p(x_v | x_{\mathrm{pa}(v)}),\ \forall x_V \in \mathcal{X}_V, \tag{3}$$

for some sample spaces $\{\mathcal{X}_l\}_{l \in L}$ and (conditional) distributions $\{p(x_v | x_{\mathrm{pa}(v)})\}_{v \in V \cup L}$.

We will call the graphical model given by the factorisation criterion *the classical set* (sometimes found in literature as *the marginal model*) and will denote it by $\mathcal{C}(\mathcal{G})$. The classical set is well-studied and fairly well-understood, although not very well-behaved in general. In the case the graph $\mathcal{G}$ is fully observed, the classical set consists of the distributions which are *Markov with respect to* $\mathcal{G}$, namely the distributions satisfying the three equivalent *Markov properties*. See Chapter 3.2.2 in [7] for details.

A few results in literature allow us to follow some simplifying assumptions in the context of the classical set. Results in [56] (namely lemmas 3.7, 3.8, and 3.9) state that any DAG $\mathcal{G}$ with latent variables has the same classical set as the DAG $\mathcal{G}'$ obtained by:

- exogenising all latent variables, i.e. removing all edges that have arrowheads at that variable and replacing them with arrows going from each parent of the latent to each child of the latent;

- eliminating any latent variables whose children are a subset of the children of another latent;

- eliminating any latent variables that have at most one child.

Additionally, the main result in [37] states that p factorises with respect to a DAG $\mathcal{G}$ if and only if there is a factorisation of p which has finite sample spaces for the latent variables. Using these results, we can assume WLOG that any latent in a DAG $\mathcal{G}$ is finite discrete, has no parents and at least two children, and its set of children is not a subset of the children of another latent. For such a graph, which we will call a *reduced graph*, the factorisation criterion becomes

$$p(x_V) = \sum_{x_L} \prod_{v \in V} p(x_v | x_{\mathrm{pa}(v)}) \prod_{l \in L} p(x_l), \tag{4}$$

for some finite discrete sample spaces $\{\mathcal{X}_l\}_{l \in L}$, (conditional) distributions $\{p(x_v|x_{\mathrm{pa}(v)})\}_{v \in V}$, and distributions $\{p(x_l)\}_{l \in L}$. We will sometimes refer to the conditional distributions $\{p(x_v|x_{\mathrm{pa}(v)})\}_{v \in V}$ as *response functions* (although we are not assuming they are deterministic).

By results in [37], the classical set is known to be semialgebraic in the context of finite observed variables: a union of sets, each constrained by some polynomial equalities and inequalities. By relaxing the inequality constraints we obtain the two models we mentioned before: the *ordinary Markov model* $\mathcal{M}(\mathcal{G})$[4] [57] and the *nested Markov model* [16], [15], which we will denote $\mathcal{N}(\mathcal{G})$. It is shown in [15] (and also in [13] more succintly) that $\mathcal{C}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$. It is worth mentioning that when we talk about $\mathcal{N}(\mathcal{G})$, we are actually referring to the nested Markov model of the *latent projection* of $\mathcal{G}$. We are not putting too much emphasis on this distinction to avoid going into too many details not relevant to the discussion in this paper.

Another example of a graphical model is the one given in [31], corresponding to the set of distributions which can be obtained from a graph under a certain GPT. For example, *the quantum set* $\mathcal{Q}(\mathcal{G})$ is the set of distributions p which are generalised Markov with respect to $\mathcal{G}$ for quantum theory (see [31] for precise definition). The quantum set will contain the classical set for any graph (simply use separable quantum states to simulate the classical latent variables). In [31], the authors also define the model consisting of all distributions p which are generalised Markov for *some* GPT with respect to a graph, which we will denote $\mathcal{GPT}(\mathcal{G})$ (originally called $\mathcal{G}$, which we changed to avoid conflict with our notation of a DAG). This model will contain the quantum set by definition, and is shown in [58] to be contained in the nested Markov model.

In short, some models often found in literature are:

- $\mathcal{C}(\mathcal{G})$: the classical set; note this is sometimes called the marginal model and denoted differently in Statistics literature (commonly $\mathcal{M}(\mathcal{G})$ or a variation, conflicting with our notation of the ordinary Markov model);

- $\mathcal{Q}(\mathcal{G})$: the quantum set in [31];

- $\mathcal{GPT}(\mathcal{G})$: the GPT set, or the set denoted $\mathcal{G}$ in [31];

- $\mathcal{N}(\mathcal{G})$: the nested Markov model [15], [13] consisting of distributions satisfying all polynomial equalities;

[4] This model is denoted by $\mathcal{I}$ in [31]

- $\mathcal{M}(\mathcal{G})$: the ordinary Markov model consisting of all distributions satisfying the d-separation independences, or the set $\mathcal{I}$ in [31].

The inclusion relationships which hold for any DAG $\mathcal{G}$ are

$$\mathcal{C}(\mathcal{G}) \subseteq \mathcal{Q}(\mathcal{G}) \subseteq \mathcal{GPT}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{M}(\mathcal{G}). \tag{5}$$

3 Quasiprobability in Causality

In this section, we will define the quasimarginal model or the quasi set. We will give three definitions of the model and prove they are all equivalent. We want to compare this set with the other models we defined, so we will also assume that observed distributions, even if they are obtained quasiprobabilistically, have entries in $[0, 1]$.

Remark. *We could potentially work with even more general spaces than just* $\mathbb{R}$*. However, given that we expect real probabilities to have maximum possible expressivity, it seems generalising too far wouldn't provide any new interesting insights. Nonetheless, perhaps treating the case of complex numbers would help with some algebraic geometry arguments, given the algebraic closure of* $\mathbb{C}$*. Although we don't prove this in this paper, all of our results should generalise to extensions of* $\mathbb{R}$ *as well.*

We define a *quasiprobability distribution* to be simply a probability distribution with the relaxation that it can take any values in $\mathbb{R}$ instead of being limited to $[0, 1]$. Conditional probability and independence are defined similarly. This will suffice for our purposes, but a more detailed treatment is included in Appendix A. We will only work with the case of finite discrete quasiprobabilistic random variables. Indeed, as we mentioned before, we assume observed variables to be finite, and by extending the result in [37], we can assume latents to be finite even when discussing quasiprobabilistic random variables. The proof of this can be found in Appendix B.1.

Our quasimarginal model will consist of the nonnegative observed distributions which are marginals of a full quasidistribution that factorises with respect to the graph. We distinguish two particular types of such full quasidistributions.

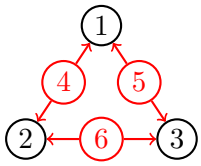
Definition 1 (Distributions with Quasilatents and Quasiresponses)**.** *Consider a DAG* $\mathcal{G}$ *with vertices* $V \dot{\cup} L$ *and quasidistribution* p *over all its vertices. We say* p *has* quasilatents *if* p *can be written as*

$$p(x_{V\cup L}) = \prod_{v\in V\cup L} p(x_v|x_{pa_{\mathcal{G}}(v)}),$$

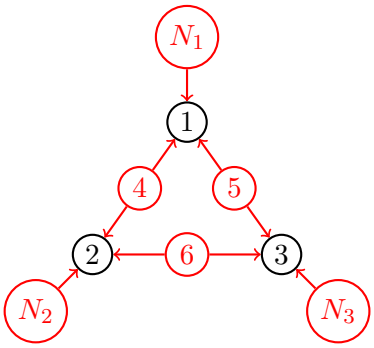
for some nonnegative *distributions* $p(x_v|x_{pa_{\mathcal{G}}(v)})$ *for all* $v \in V$ *and quasidistributions* $p(x_l|x_{pa_{\mathcal{G}}(l)})$ *for all* $l \in L$*. We say* p *has* quasiresponses *if* p *can be written in the same form, but this time with* $p(x_v|x_{pa_{\mathcal{G}}(v)})$ *being general quasidistributions, and* $p(x_l|x_{pa_{\mathcal{G}}(l)})$ *being nonnegative.*

The first definition makes sense from a historical point of view, given that this is the definition often used when analysing negative probabilities in the context of the Bell Scenario. The second one is motivated by the results in [48]. We will show these two definitions are equivalent from the point of view of their marginals over V, meaning there is a sort of duality between latent variables and response functions.

It is worth mentioning that the first definition seemingly follows the same idea as the HLP formalism [31], namely that only latents are non-classical. One might argue however that even in the HLP framework, quantum measurements are inherently a part of Quantum



(a) Triangle Scenario as a DAG with latent variables.



(b) Triangle Scenario with noise variables.

Theory, so if we were to extend the HLP intuition to quasiprobability, we should therefore allow both latents and response functions to be negative. This will turn out to give another model identical to the first two, and we will indeed opt for this HLP-esque definition in the Section 4 discussion.

We define a third model below, which will be useful in proofs.

Definition 2 (Distributions with Quasinoise)**.** *Consider a DAG $\mathcal{G}$ with vertices $V \dot\cup L$ and quasidistribution p over all its vertices. Then we say p has* quasinoise *if there exist random variables N_v for each $v \in V$ (independent of everything in the graph except X_v), called noise variables, such that p can be written as:*

$$p(x_{V\cup L}) = \sum_{n_v} \prod_{v\in V} p(x_v|x_{pa_\mathcal{G}(v)}, n_v) \prod_{l\in L} p(x_l|x_{pa_\mathcal{G}(l)}) \prod_{v\in V} p(n_v),$$

for some deterministic distributions $p(x_v|x_{pa_\mathcal{G}(v)}, n_v)$, nonnegative distributions $p(x_l|x_{pa_\mathcal{G}(l)})$ and quasidistributions $p(n_v)$.

Intuitively, noise variables simply isolate away the internal randomness of the observed nodes. See Figure 2b for an example. They can be considered finite discrete in the case of finite discrete observed variables. To see this, note that one can regard noise variables as simply some additional latent variables. Since the construction in Appendix B.1 does not change the response distributions of the observed variables when creating the finite model, one can replace the noise variables with some finite counterparts.

We are now ready to define the quasimarginal set. We will give three different definitions in terms of the type of full quasidistribution we consider.

Definition 3 (The Three k-Quasimarginal Sets)**.** *Consider DAG $\mathcal{G}$, and nonnegative distribution p over its observed vertices. Then we define the sets*

- *$\mathcal{W}_l(\mathcal{G})$ to be the set of distributions p, such that p is the marginal of a quasidistribution with quasilatents that factorises with respect to $\mathcal{G}$;*

- *$\mathcal{W}_r(\mathcal{G})$ to be the set of distributions p, such that p is the marginal of a quasidistribution with quasiresponses that factorises with respect to $\mathcal{G}$;*

- *$\mathcal{W}_n(\mathcal{G})$ to be the set of distributions p, such that p is the marginal of a quasidistribution with quasinoise that factorises with respect to $\mathcal{G}$.*

The definition above can be extended in the natural way to CADMGs. Also, we can assume the DAG $\mathcal{G}$ is reduced, since the results in [56] extend for quasiprobability distributions as well, which we have shown in Appendix B.2. In the following result we will prove these sets coincide for any DAG $\mathcal{G}$, and hence we can define the quasimarginal set as any of them.

Theorem 4 (Equivalence of Quasimarginal Sets)**.** *For any DAG $\mathcal{G}$, the following holds:*

$$\mathcal{W}_l(\mathcal{G}) = \mathcal{W}_r(\mathcal{G}) = \mathcal{W}_n(\mathcal{G}).$$

We give a full proof of this result in Appendix C. The proof involves showing $\mathcal{W}_n = \mathcal{W}_r$ and $\mathcal{W}_n = \mathcal{W}_l$. The former is fairly simple. For the latter, the $\mathcal{W}_n \subseteq \mathcal{W}_l$ direction involves just encapsulating any noise variable into a latent parent of its observed child. The $\mathcal{W}_l \subseteq \mathcal{W}_n$ direction is the more involved step of the proof, and involves a lemma that shows a quasiprobabilistic latent parent can be replaced by a classical latent parent and quasiprobabilistic noise variables for its children. This process is then applied for each latent.

We mention without proof that Theorem 4 can be extended to also include the set of distributions which are marginals of some general factorising quasidistribution (i.e. full quasidistributions where neither observed nor latent response functions are restricted to $[0, 1]$). Indeed, to see this, one can follow the proof of $\mathcal{W}_r = \mathcal{W}_n$ to show the set is the same as the set of marginals of quasidistributions with both quasiprobabilistic noise and latent variables, which can then be shown to be the same as $\mathcal{W}_l$ by encapsulating the noise variables into the latent variables, i.e. by using the proof of $\mathcal{W}_l = \mathcal{W}_n$.

We can therefore define the quasi-marginal set as follows:

Definition 5 (The Quasimarginal Set)**.** *Given reduced DAG $\mathcal{G}$, we define the quasi-marginal set $\mathcal{W}(\mathcal{G})$ to be any of the sets $\mathcal{W}_l(\mathcal{G})$, $\mathcal{W}_r(\mathcal{G})$, $\mathcal{W}_n(\mathcal{G})$. Equivalently, it is the set of observed distributions p which can be written as*

$$p(x_V) = \sum_{x_L} \prod_{v \in V} p(x_v | x_{pa_{\mathcal{G}}(v)}) \prod_{l \in L} p(x_l),$$

for some (conditional) quasidistributions $p(x_v|x_{pa(v)})$ and $p(x_l)$.

4 Quasiprobability and the Nested Markov Model

We will now discuss the relationship between the set in Definition 5 and the nested Markov model.

Proposition 6. *For any DAG $\mathcal{G}$,*

$$\mathcal{W}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}).$$

Proof. Follows from the fact that the nested Markov model is defined entirely in terms of the factorisation criterion, which is obeyed by distributions in the quasi set (albeit with quasiprobabilistic factors). Indeed, the treatment in [15] is purely symbolic and does not rely on positivity (nor realness, in fact) of the distributions. We include a more detailed discussion in Appendix D. ■

We conjecture that, even more strongly, equality holds for any graph.

Conjecture 1. *For any DAG $\mathcal{G}$,*

$$\mathcal{W}(\mathcal{G}) = \mathcal{N}(\mathcal{G}).$$

We will now study the conjecture in the context of correlation scenarios [32] with a tree structure.

Definition 7 (Correlation Scenario)**.** *A DAG $\mathcal{G}$ is a* correlation scenario *if it is reduced and observed variables have no children.*

Therefore, a correlation scenario $\mathcal{G}$ can be seen as a directed bipartite graph. It has a layer of latent variables and a layer of observed variables, and all edges are from some latent variable to some observed one. We will also assume the graphs we are working with are connected graphs (otherwise just separate the problem into each connected component). The Markov model and the nested Markov model have particularly simple forms in the context of correlation scenarios.

Proposition 8. *Let $\mathcal{G}$ be a correlation scenario. Then a distribution p is in the Markov model of $\mathcal{G}$ iff it satisfies $A \perp B$ $[p]$ whenever A and B are two subsets of V such that no latent variable has children in both sets. Additionally, $\mathcal{N}(\mathcal{G}) = \mathcal{M}(\mathcal{G})$.*

For the first claim, see for example [57] (Theorem 3) for a proof in the case of graphs with only binary latents, which directly generalises. The second claim follows from the definition of the nested Markov property in [13] and an inductive argument.

We will assume from now on that all correlation scenarios we are working with only have latent variables with exactly two children. Less than that would make the latent variable redundant, whereas three children or more can be reduced to the simpler case via the following proposition.

Proposition 9. *Let $\mathcal{G}$ be a correlation scenario with a latent Λ with $k \geq 3$ children. Consider now the correlation scenario $\mathcal{G}'$ which replaces Λ with $\binom{k}{2}$ latents with two children, one for each pair of children of Λ, having that pair as children. Then $\mathcal{M}(\mathcal{G}') = \mathcal{M}(\mathcal{G})$ and $\mathcal{W}(\mathcal{G}') \subseteq \mathcal{W}(\mathcal{G})$.*

The first result follows the fact that the two graphs have the same latent projection, which encodes all observed independences. For the second result, if p can be obtained by a latent assignment in $\mathcal{G}'$, then simply make Λ jointly simulate all of its corresponding latent variables in $\mathcal{G}'$, and make its children just pick out the information they have access to in $\mathcal{G}'$ from it. This shows p has a latent model in $\mathcal{G}$ as well.

Corollary 9.1. *It is enough to prove Conjecture 1 for correlation scenarios where all latent variables have exactly two children.*

Proof. Assume we have proven Conjecture 1 for correlation scenarios where all latent variables have exactly 2 children and let $\mathcal{G}$ be an arbitrary correlation scenario. Since $\mathcal{G}$ is reduced, every latent in $\mathcal{G}$ has at least 2 children. Consider the graph $\mathcal{G}'$ defined in Proposition 9. Then by Propositions 6 and 9 we have

$$\mathcal{M}(\mathcal{G}) \supseteq \mathcal{W}(\mathcal{G}) \supseteq \mathcal{W}(\mathcal{G}') = \mathcal{M}(\mathcal{G}') = \mathcal{M}(\mathcal{G}).$$

■

Now, we will discuss a certain subclass of these correlation scenarios.

Definition 10. *Let $\mathcal{G}$ be a correlation scenario. Consider the undirected graph $\mathcal{G}^t$ that has set of vertices V (i.e. the observed vertices of $\mathcal{G}$), and an edge between two vertices if they have a common latent parent in $\mathcal{G}$. We say $\mathcal{G}$ is a* Tree-structured Correlation Scenario (TCS) *if $\mathcal{G}^t$ is a tree.*

Note that if $\mathcal{G}$ is a TCS, then every latent variable must have exactly 2 children: if Λ has at least 3 children, then any two of them will be connected in $\mathcal{G}^t$, introducing a cycle. As we mentioned before, we will prove Conjecture 1 in the context of TCSs. One important feature that allows this is the very simple structure of the Markov model (and hence the nested Markov model too) of TCSs.

Proposition 11. *Let $\mathcal{G}$ be a TCS. For each vertex v, note that removing v from the graph separates it into a bunch of connected components. We will call the set of such components $\mathcal{B}_v$, meaning*

$$\mathcal{B}_v = \{B \subseteq V \setminus \{v\} : B \text{ is a maximal connected component of the graph } \mathcal{G} \setminus \{v\}\}.$$

Then $p \in \mathcal{M}(\mathcal{G})$ iff for any vertex v and $B \in \mathcal{B}_v$, we have

$$B \perp\!\!\!\perp V \setminus (B \cup \{v\}) \ [p],$$

or equivalently,

$$p(x_{V\setminus\{v\}}) = \prod_{B\in\mathcal{B}_v} p(x_B).$$

In other words, all the (nested) Markov model requires is that after eliminating a vertex v, the distribution with x_v summed out factorises with respect to the branches left.

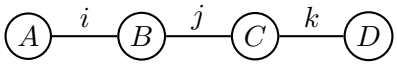
Next, we introduce Tensor Network Decompositions and then we will provide a proof for Conjecture 1 for TCSs.

4.1 Tensor Network Decompositions

Tensor Network Decompositions are tools that allow storing the information of a tensor more compactly. They are used in machine learning, and studying structures of entanglement in multipartite Quantum states, but it was noted they are also related to graphical models [59]. They involve decomposing a tensor into an expression consisting of tensor products and contractions following the structure of a graph, where each vertex is assigned a separate tensor, and edges represent contraction of the neighbouring components over an index of fixed size, called a *bond dimension*. For now we assume the entries of the tensors come from an arbitrary field k, and will replace $k := \mathbb{R}$ later when we get back to graphical models. For an undirected graph $\mathcal{G}$, we will denote by $\text{inc}_{\mathcal{G}}(v) := \{(v_1, v_2) \in E : v_1 = v \text{ or } v_2 = v\}$ the set of incident edges of vertex v, and will drop the index when the graph is clear from context.

Definition 12 (Tensor Network Decomposition)**.** *Let $\mathcal{G}$ be an undirected graph. Assign to each vertex $v \in V$ a finite-dimensional k-vector space $\mathbb{V}_v$ and each edge $e \in E$ a bond dimension $r_e \in \mathbb{Z}_+^*$. We say a tensor $\mathcal{T} \in \prod_{v\in V} \mathbb{V}_v$* decomposes with respect to $\mathcal{G}$ with bond dimensions $\{r_e\}_{e\in E}$ *if there exist k-vector spaces $\{\mathbb{E}_e\}_{e\in E}$ of corresponding dimensions $\{r_e\}_{e\in E}$ and tensors $\left\{T^{(v)} \in \mathbb{V}_v \otimes \left(\bigotimes_{e\in inc(v)} \mathbb{E}_e\right)\right\}_{v\in V}$ such that*

$$\mathcal{T}(x_V) = \sum_{\{1\le i_e\le r_e\}_{e\in E}} \prod_{v\in V} T^{(v)}\left(x_v, \{i_e\}_{e\in inc(v)}\right). \tag{6}$$



(a) A tree $\mathcal{T}$. Each edge is labeled with the index operating on it.

(b) The corresponding TCS $\mathcal{T}^{DAG}$ of $\mathcal{T}$. Edges become latent variables with the same domain size as the corresponding index.

Figure 3: The correspondence between the 4-on-line Scenario and its underlying undirected graph.

From now on, when referring to entries of $T^{(v)}$, we will move the indices i_e of the edges incident to v to the subscript, so equation (6) becomes:

$$\mathcal{T}(x_V) = \sum_{\{1 \le i_e \le r_e\}_{e \in E}} \prod_{v \in V} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v). \tag{7}$$

Example 13. *Let's look at the undirected graph $\mathcal{T}$ in Figure 3a, and the TCS $\mathcal{T}^{DAG}$ in 3b that we obtain by replacing every edge in $\mathcal{T}$ by a latent variable, which we will call the* 4-on-line Scenario. *We associate with $\mathcal{T}$ tensors T in the vector space $\mathbb{R}^{|\mathcal{A}|} \otimes \mathbb{R}^{|\mathcal{B}|} \otimes \mathbb{R}^{|\mathcal{C}|} \otimes \mathbb{R}^{|\mathcal{D}|}$ for some fixed finite sets $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, i.e. for $(a,b,c,d) \in \mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \mathcal{D}$ the tensor T gives an entry $T(a,b,c,d) \in \mathbb{R}$. We say such a tensor T decomposes with respect to $\mathcal{T}$ and bond dimensions $(I,J,K) \in \mathbb{R}^3$ if there exist tensors*

$$A \in \mathbb{R}^{|\mathcal{A}|} \otimes \mathbb{R}^I, \quad B \in \mathbb{R}^{|\mathcal{B}|} \otimes \mathbb{R}^I \otimes \mathbb{R}^J, \quad C \in \mathbb{R}^{|\mathcal{C}|} \otimes \mathbb{R}^J \otimes \mathbb{R}^K,$$
$$D \in \mathbb{R}^{|\mathcal{D}|} \otimes \mathbb{R}^K$$

such that

$$T(a,b,c,d) = \sum_{i,j,k} A_i(a) B_{ij}(b) C_{jk}(c) D_k(d). \tag{8}$$

Now assume we have a distribution $p(a,b,c,d) \in \mathcal{W}\left(\mathcal{T}^{DAG}\right)$, where the random variables X_A, X_B, X_C, X_D take values in the sets $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$. In other words, there exist latent variables Λ_{AB}, Λ_{BC}, Λ_{CD}, such that

$$p(a,b,c,d) = \sum_{\lambda_{AB},\lambda_{BC},\lambda_{CD}} p(a|\lambda_{AB}) \cdot p(b|\lambda_{AB},\lambda_{BC}) \cdot p(c|\lambda_{BC},\lambda_{CD}) \cdot p(d|\lambda_{CD}) \cdot p(\lambda_{AB}) \cdot p(\lambda_{BC}) \cdot p(\lambda_{CD}),$$

for some value assignments to the (conditional) quasidistributions in the expression. However, given this expression, we can equivalently write

$$p(a,b,c,d) = \sum_{\lambda_{AB},\lambda_{BC},\lambda_{CD}} p(a,\lambda_{AB}) \cdot p(b|\lambda_{AB},\lambda_{BC}) \cdot p(c,\lambda_{BC}|\lambda_{CD}) \cdot p(d,\lambda_{CD}),$$

by simply grouping factors, for example $p(a|\lambda_{AB}) \cdot p(\lambda_{AB}) = p(a,\lambda_{AB})$. Therefore, if we now define

$$A_{\lambda_{AB}}(a) := p(a,\lambda_{AB}), \quad B_{\lambda_{AB},\lambda_{BC}}(b) = p(b|\lambda_{AB},\lambda_{BC}),$$
$$C_{\lambda_{BC},\lambda_{CD}}(c) := p(c,\lambda_{BC}|\lambda_{CD}), \quad D_{\lambda_{CD}}(d) := p(d,\lambda_{CD}),$$

and view $p(a, b, c, d)$ as an order-4 tensor, we actually recover equation (8), where the indices now range over the values of the latent variables. Note that we could've instead grouped $p(\lambda_{AB})$ into the factor corresponding to b via $p(b|\lambda_{AB}, \lambda_{BC}) \cdot p(\lambda_{AB}) = p(b, \lambda_{AB}|\lambda_{BC})$ to obtain a different Tensor Network Decomposition. This shows the process we applied is not unique.

In general, given a latent variable representation of a distribution with respect to a correlation scenario $\mathcal{G}$, we can always construct a Tensor Network Decomposition with respect to the corresponding undirected graph $\mathcal{G}^t$ by following the same process, where the bond dimensions are the sizes of the latents. Our proof of Conjecture 1 will explore going in the opposite direction.

A pretty common result in the literature of Tensor Network States is that every tensor can be decomposed with respect to a graph $\mathcal{G}$ as long as we allow the bond dimensions to be large enough (the result is more common for trees [60], [61]; it is proven in the general case in [62]). It is also known that in the case of a tree graph, a *minimal* decomposition exists for every tensor [62], [60].

We now have all the ingredients we need for our proof of Conjecture 1 in the case of TCSs.

Theorem 14. *For any Tree-structured Correlation Scenario $\mathcal{G}$, a distribution is in the quasimarginal model if and only if it satisfies the observed independence constraints enforced by the graph. In other words,*

$$\mathcal{W}(\mathcal{G}) = \mathcal{N}(\mathcal{G}) = \mathcal{M}(\mathcal{G}).$$

Proof. It follows from Propositions 6 and 8 that $\mathcal{W}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) = \mathcal{M}(\mathcal{G})$. For the opposite inclusion, consider an arbitrary element of the ordinary Markov model $p \in \mathcal{M}(\mathcal{G})$. We will build an quasiprobabilistic latent variable representation of p.

We will view $p(x_V)$ as a tensor: it has order $|V|$, one mode for each vertex of the graph (or equivalently observed variable). Now consider the undirected graph $\mathcal{G}^t$ which has as vertices the observed vertices of $\mathcal{G}$ and an edge between v and w iff v and w have a common (latent) parent. By the results in [62] (more precisely Theorem 8.3), there exist minimal bond dimensions $(r_1, \ldots, r_{|E^t|})$ such that $p(x_V)$ decomposes with respect to $\mathcal{G}^t$ (namely the $\mathcal{G}^t$-rank of $p(x_V)$). We will transform this minimal decomposition of $p(x_V)$ to show that $p(x_V) \in \mathcal{W}(\mathcal{G})$. We will illustrate the proof for the case of the 4-on-line Scenario in Fig. 3b and direct the reader to Appendix E for the general case.

In the case of the 4-on-line Scenario, the minimal tensor decomposition corresponds to minimal $I, J, K \in \mathbb{R}$ such that there exist tensors

$$A \in \mathbb{R}^{|\mathcal{A}|} \otimes \mathbb{R}^I, \quad B \in \mathbb{R}^{|\mathcal{B}|} \otimes \mathbb{R}^I \otimes \mathbb{R}^J, \quad C \in \mathbb{R}^{|\mathcal{C}|} \otimes \mathbb{R}^J \otimes \mathbb{R}^K,$$
$$D \in \mathbb{R}^{|\mathcal{D}|} \otimes \mathbb{R}^K$$

that satisfy the equation

$$p(a, b, c, d) = \sum_{i,j,k} A_i(a) B_{ij}(b) C_{jk}(c) D_k(d). \tag{9}$$

The first step is to note that WLOG the following inequations hold:

$$\sum_a A_i(a) \neq 0$$
$$\sum_b B_{ij}(b) \neq 0$$
$$\sum_c C_{jk}(c) \neq 0$$
$$\sum_d D_k(d) \neq 0,$$

for all values $1 \leq i, j, k \leq I, J, K$. If the decomposition we are given doesn't satisfy this, i.e. there exists i such that $\sum_a A_i(a) = 0$ for example, then we can at least say there exists one i for which $\sum_a A_i(a) \neq 0$ does hold: if, for the sake of contradiction, the value is 0 for all i, then by summing equation (9) over all a, b, c, d, we obtain

$$\sum_{a,b,c,d} p(a,b,c,d) = \sum_{i,j,k} \left(\sum_a A_i(a)\right)\left(\sum_b B_{ij}(b)\right)\left(\sum_c C_{jk}(c)\right)\left(\sum_d D_k(d)\right),$$

but the RHS of this equation is 0, which contradicts the fact that p is a probability distribution. If $\sum_a A_q(a) \neq 0$ holds for some $1 \leq q \leq I$, then we can construct a different tensor decomposition of p with the same I, J, K by replacing $A_i(\_)$ by $A_i(\_) + \alpha_i \cdot A_q(\_)$ for some $\alpha_i \in \mathbb{R}^*$ for any i where this is necessary, and then applying the inverse operation to tensor B to maintain the decomposition. If the value of each α_i is chosen carefully, we can make sure $\sum_a A_i(a) \neq 0$ (as well as $\sum_b B_{ij}(b) \neq 0$) holds for all i. See Appendix E.3 for more details and a formal proof of this.

Next, notice that we can write:

$$\begin{aligned} p(a,b,c,d) &= \sum_{i,j,k} A_i(a) B_{ij}(b) C_{jk}(c) D_k(d) \\ &= \sum_{i,j,k} \frac{A_i(a)}{\sum_a A_i(a)} \cdot \frac{B_{ij}(b)}{\sum_b B_{ij}(b)} \cdot \frac{C_{jk}(c)}{\sum_c C_{jk}(c)} \cdot \frac{D_k(d)}{\sum_d D_k(d)} \\ &\quad \cdot \left(\sum_a A_i(a)\right)\left(\sum_b B_{ij}(b)\right)\left(\sum_c C_{jk}(c)\right)\left(\sum_d D_k(d)\right), \end{aligned}$$

and we have $\sum_a \frac{A_i(a)}{\sum_a A_i(a)} = 1, \forall i$, meaning we can consider $\frac{A_i(a)}{\sum_a A_i(a)}$ to be $p(a|\lambda_{AB} = i)$, and the same for the other observed nodes. What we need now is for the factor at the end of the expression to decompose as

$$\left(\sum_a A_i(a)\right)\left(\sum_b B_{ij}(b)\right)\left(\sum_c C_{jk}(c)\right)\left(\sum_d D_k(d)\right) = \phi(i) \cdot \psi(j) \cdot \theta(k).$$

Indeed, if this holds then we can write

$$\begin{aligned}
1 &= \sum_{a,b,c,d} p(a,b,c,d) \\
&= \sum_{i,j,k} \left(\sum_a \frac{A_i(a)}{\sum_a A_i(a)} \right) \cdot \left(\sum_b \frac{B_{ij}(b)}{\sum_b B_{ij}(b)} \right) \cdot \left(\sum_c \frac{C_{jk}(c)}{\sum_c C_{jk}(c)} \right) \cdot \left(\sum_d \frac{D_k(d)}{\sum_d D_k(d)} \right) \\
&\cdot \left(\sum_a A_i(a) \right) \left(\sum_b B_{ij}(b) \right) \left(\sum_c C_{jk}(c) \right) \left(\sum_d D_k(d) \right) \\
&= \sum_{i,j,k} 1 \cdot 1 \cdot 1 \cdot 1 \cdot \phi(i) \cdot \psi(j) \cdot \theta(k) \\
&= \sum_i \phi(i) \cdot \sum_j \psi(j) \cdot \sum_k \theta(k),
\end{aligned}$$

and hence

$$\begin{aligned}
p(a,b,c,d) &= \sum_{i,j,k} A_i(a) B_{ij}(b) C_{jk}(c) D_k(d) \\
&= \sum_{i,j,k} \frac{A_i(a)}{\sum_a A_i(a)} \cdot \frac{B_{ij}(b)}{\sum_b B_{ij}(b)} \cdot \frac{C_{jk}(c)}{\sum_c C_{jk}(c)} \cdot \frac{D_k(d)}{\sum_d D_k(d)} \\
&\cdot \frac{\phi(i)}{\sum_i \phi(i)} \cdot \frac{\psi(j)}{\sum_j \psi(j)} \cdot \frac{\theta(k)}{\sum_k \theta(k)}.
\end{aligned}$$

Now we have $\sum_i \frac{\phi(i)}{\sum_i \phi(i)} = 1$ and the same for j and k, so we can define

$$\begin{aligned}
p(\lambda_{AB} = i) &:= \frac{\phi(i)}{\sum_i \phi(i)}, \\
p(\lambda_{BC} = j) &:= \frac{\psi(j)}{\sum_j \psi(j)}, \\
p(\lambda_{CD} = k) &:= \frac{\theta(k)}{\sum_k \theta(k)},
\end{aligned}$$

and we are done.

It is therefore enough to show that $\sum_b B_{ij}(b)$ and $\sum_c C_{jk}(c)$ factorise into components only depending on i and j, and j and k respectively. This is where we will use the independence constraints of the ordinary Markov model. We will prove that $\sum_b B_{ij}(b)$ decomposes and the same will follow for C as well. Our proof will mainly reason about matrices, which will hopefully be clearer to readers unfamiliar with tensor algebra. The full proof in Appendix E is a tensor generalisation of this argument.

We start by summing (9) over b:

$$p(a,c,d) = \sum_{i,j,k} A_i(a) \cdot \left(\sum_b B_{ij}(b) \right) \cdot C_{jk}(c) \cdot D_k(d). \tag{10}$$

We now rearrange the expression to act on order-two objects, i.e. matrices. The LHS $p(a,c,d)$ is an order-3 tensor, but we can rewrite its entries into a $|\mathcal{A}| \times (|\mathcal{C}| \cdot |\mathcal{D}|)$ matrix with one rows corresponding to a, and columns corresponding to (c,d). We will denote this matrix by $p(a;(c,d))$. Similarly, the tensor $\sum_k C_{jk}(c) \cdot D_k(d)$ has three dimensions: j, c, and d, but we can rewrite it as a matrix with dimensions corresponding to j and (c,d),

matrix which we will denote by M. Hence if we write A as a $|\mathcal{A}| \times I$ matrix, $\sum_b B_{ij}(b)$ as a $I \times J$ matrix, and M is a $J \times (|\mathcal{C}| \cdot |\mathcal{D}|)$ matrix, then equation (10) becomes:

$$p(a;(c,d)) = A\left(\sum_b B_{ij}(b)\right) M. \tag{11}$$

The idea now is to note that A and M have inverses on one side each. Indeed, we claim that the columns of the $|\mathcal{A}| \times I$ matrix A are linearly independent. Assume for the sake of contradiction that there exists $1 \leq q \leq I$ such that

$$A_q(a) = \sum_{i \neq q} \alpha_i A_i(a),$$

for some $\{\alpha_i \in \mathbb{R}\}_{i \neq q}$. Then going back to (9), we get:

$$\begin{aligned}
p(a,b,c,d) &= \sum_{i,j,k} A_i(a)B_{ij}(b)C_{jk}(c)D_k(d) \\
&= \sum_{i \neq q,j,k} A_i(a)B_{ij}(b)C_{jk}(c)D_k(d) + \sum_{j,k} A_q(a)B_{qj}(b)C_{jk}(c)D_k(d) \\
&= \sum_{i \neq q,j,k} A_i(a)B_{ij}(b)C_{jk}(c)D_k(d) + \sum_{i \neq q,j,k} \left(\alpha_i A_i(a)\right) B_{qj}(b)C_{jk}(c)D_k(d) \\
&= \sum_{i \neq q,j,k} A_i(a) \cdot \left(B_{ij}(b) + \alpha_i B_{qj}(b)\right) \cdot C_{jk}(c)D_k(d),
\end{aligned}$$

which is a tensor decomposition of p with bond dimensions $(I-1, J, K)$, contradicting the fact that our initial bond dimensions were minimal. Therefore, the columns or A are linearly independent, meaning A has a left inverse A^L such that $A^L A = \mathcal{I}_I$, where $\mathcal{I}_I$ is the $I \times I$ identity matrix. Similarly (but perhaps less obvious; see Appendix E for details), the matrix M has a right inverse M^R such that $MM^R = \mathcal{I}_J$. Multiplying to the left by A^L and to the right by M^R in equation (11) we get:

$$A^L \cdot p(a;(c,d)) \cdot M^R = \sum_b B_{ij}(b).$$

But $p \in \mathcal{M}(\mathcal{G})$, meaning $p(a,c,d) = p(a)p(c,d)$, or equivalently $p(a;(c,d)) = \mathbf{p}(\mathbf{a}) \cdot \mathbf{p}(\mathbf{c},\mathbf{d})^T$, where we consider $\mathbf{p}(\mathbf{a})$ and $\mathbf{p}(\mathbf{c},\mathbf{d})$ to be column vectors (and hence we write them in bold). Therefore, we get

$$\left(A^L \cdot \mathbf{p}(\mathbf{a})\right)\left(\mathbf{p}(\mathbf{c},\mathbf{d})^T \cdot M^R\right) = \sum_b B_{ij}(b).$$

Finally, if we look at the types in the LHS, A^L is a $I \times |\mathcal{A}|$ matrix, meaning $A^L \cdot p(a)^T$ is a column vector of size I, and similarly $p(c,d)^T \cdot M^R$ is a row vector of size J. Therefore, $\sum_b B_{ij}(b)$ indeed factorises into an i component and a j component. The conclusion follows.

■

Corollary 14.1. *Let $\mathcal{G}$ be a TCS. Then for any distribution $p \in \mathcal{M}(\mathcal{G})$, p has a tensor network decomposition with respect to $\mathcal{G}^t$ and bond dimensions $\{r_e\}_{e \in E^t}$ if and only if it has a quasiprobabilistic latent variable construction where the latent corresponding to edge e in $\mathcal{G}^t$ has domain size equal to r_e.*

Proof. Let $\{r_e^*\}_{e\in E^t}$ be the $\mathcal{G}^t$-rank of p. Then we know by [62] that p has a decomposition with bond dimensions $\{r_e\}_{e\in E^t}$ if and only if $r_e \geq r_e^*$ for all e.

Now assume p has a tensor network decomposition with bond dimensions $\{r_e\}_{e\in E^t}$. Then $r_e \geq r_e^*$ for all e and p has a tensor decomposition with bond dimensions $\{r_e^*\}_{e\in E^t}$. By the discussion in Appendix E, it has a latent variable construction with sizes $\{r_e^*\}_{e\in E^t}$. Since $r_e \geq r_e^*$ for all e, this implies the same must be true for latent variable sizes $\{r_e\}_{e\in E^t}$.

Conversely, if p has a latent variable construction with sizes $\{r_e\}_{e\in E^t}$, then it has a tensor network decomposition with bond dimensions $\{r_e\}_{e\in E^t}$ by the construction we have given in the Example 13. ■

Corollary 14.1 therefore illustrates a sort of equivalence between tensor network decompositions and quasiprobabilistic latent variable constructions in the case of non-signaling distributions and TCSs.

5 Conclusions

In this work, we have formalised the set of quasidistributions consistent with a graph, generalising some folklore results about the Bipartite Bell Scenario. We have shown there is a sort of duality between latent variables and response functions from the point of view of generalised probabilities. Finally, we proved the quasi set is the same as the nested Markov model for any correlation scenario with a tree structure.

A proof of Conjecture 1 would provide an affirmative answer to Proposition 3.4 of [32]: it is known that GPTs are not expressive enough to generate all distributions in the nested Markov model, not even for correlation scenarios, as the perfect correlation in the Triangle Scenario is not in $\mathcal{GPT}$ [31]; however, the theory allowing latents (or equivalently response functions or both) to take all real-valued probabilities is. This theory can be seen as an example of a *non-free generalised probabilistic theory* [63], [54], so it appears dropping the *no-restriction assumption* that GPTs are often studied under is enough to give them the expressivity to achieve all nested Markov model correlations.

In terms of directions for future work, studying Conjecture 1 is a natural direction. We know it is true for tree-structured correlation scenarios, and Appendix F contains a proof for the Triangle Scenario often found in literature, meaning the conjecture might hold in cyclic correlation scenarios as well. What about general DAGs, where the nested Markov model becomes more complex? In lack of a full proof, showing at least that the quasi set contains the $\mathcal{GPT}$ set would confirm our intuition. In parallel, it would be interesting to see what makes the quasi set special, and why it is so much more expressive than other physically motivated theories. One avenue we see towards this is the Inflation Technique [39]: we are aware that the Marginal Problem becomes much simpler in the context of quasiprobability[5], but the completeness proof in [39] requires a version of the de Finetti Theorem, and we are not aware of a quasiprobabilistic version of this theorem that proves completeness of inflation for the quasi set as well.

Acknowledgements: We would like to thank Jonathan Barrett, Robin Evans, Tein van der Lugt, Marina Maciel Ansanelli, Máté Weisz, and Ruiwen Dong for the useful discussions that helped build the results in this paper, as well as Christopher Chang for proofreading.

[5] In fact, we believe we have a proof for the following statement: for any $n \geq 2$ and correlation scenario $\mathcal{G}$, an observed distribution p has a quasiprobabilistic n-th order inflation (defined in [39]) if and only if $p \in \mathcal{N}(\mathcal{G}) = \mathcal{M}(\mathcal{G})$.

References

1. Sangmin Byeon, W.L.: Directed acyclic graphs for clinical research: a tutorial. Journal of Minimally Invasive Surgery **26**(3), 97–107 (2023). DOI: 10.7602/jmis.2023.26.3.97
2. García-Franco, J.D., Díez, F.J., Carrasco, M.Á.: Probabilistic graphical model for the evaluation of the emotional and dramatic personality disorders. Frontiers in Psychology **13** (2022). DOI: 10.3389/fpsyg.2022.996609
3. Imbens, G.W.: Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. Journal of Economic Literature **58**(4), 1129–1179 (2020)
4. Theodoridis, S.: Machine Learning: A Bayesian and Optimization Perspective. Academic Press (2020)
5. Wright, S.: The Method of Path Coefficients. The Annals of Mathematical Statistics **5**(3), 161–215 (1934). `https://www.jstor.org/stable/2957502`
6. Pearl, J.: Causality. Cambridge University Press (2009)
7. Lauritzen, S.L.: Graphical Models. Clarendon Press (1996)
8. Bongers, S., Forré, P., Peters, J., Mooij, J.M.: Foundations of structural causal models with cycles and latent variables. The Annals of Statistics **49**(5) (2021). DOI: 10.1214/21-aos2064
9. Barrett, J., Lorenz, R., Oreshkov, O.: Cyclic quantum causal models. Nature Communications **12**(1) (2021). DOI: 10.1038/s41467-020-20456-x
10. Verma, T., Pearl, J.: Equivalence and synthesis of causal models. In: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. UAI '90, pp. 255–270. Elsevier Science Inc., USA (1990)
11. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. Springer New York, NY (1993)
12. Tian, J., Pearl, J.: On the testable implications of causal models with hidden variables. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. UAI'02, pp. 519–527. Morgan Kaufmann Publishers Inc., Alberta, Canada (2002)
13. Evans, R.J.: Margins of discrete Bayesian networks. The Annals of Statistics **46**(6A) (2018). DOI: 10.1214/17-aos1631
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1988)
15. Richardson, T.S., Evans, R.J., Robins, J.M., Shpitser, I.: Nested Markov Properties for Acyclic Directed Mixed Graphs. Annals of Statistics **51** (2023). DOI: 10.1214/22-AOS2253
16. Shpitser, I., Evans, R.J., Richardson, T.S., Robins, J.M.: Introduction to nested Markov models. Behaviormetrika **41**, 3–39 (2014)
17. Evans, R.J., Richardson, T.S.: Markovian Acyclic Directed Mixed Graphs For Discrete Data. The Annals of Statistics **42**(4), 1452–1482 (2014)
18. Evans, R.J., Richardson, T.S.: Smooth, identifiable supermodels of discrete DAG models with latent variables. Bernoulli **25**(2) (2019). DOI: 10.3150/17-bej1005
19. Shpitser, I., Richardson, T.S., Robins, J.M., Evans, R.: Parameter and Structure Learning in Nested Markov Models, (2012). arXiv: `1207.5058 [stat.ML]`. `https://arxiv.org/abs/1207.5058`.
20. Einstein, A., Podolsky, B., Rosen, N.: Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? Phys. Rev. **47**, 777–780 (1935). DOI: 10.1103/PhysRev.47.777. `https://link.aps.org/doi/10.1103/PhysRev.47.777`
21. Bell, J.S.: On the Einstein Podolsky Rosen paradox. Physics Physique Fizika **1**, 195–200 (1964). DOI: 10.1103/PhysicsPhysiqueFizika.1.195. `https://link.aps.org/doi/10.1103/PhysicsPhysiqueFizika.1.195`
22. Freedman, S.J., Clauser, J.F.: Experimental Test of Local Hidden-Variable Theories. Phys. Rev. Lett. **28**, 938–941 (1972). DOI: 10.1103/PhysRevLett.28.938. `https://link.aps.org/doi/10.1103/PhysRevLett.28.938`
23. Aspect, A., Dalibard, J., Roger, G.: Experimental Test of Bell's Inequalities Using Time-Varying Analyzers. Phys. Rev. Lett. **49**, 1804–1807 (1982). DOI: 10.1103/PhysRevLett.49.1804. `https://link.aps.org/doi/10.1103/PhysRevLett.49.1804`

24. Wiseman, H.M., Jones, S.J., Doherty, A.C.: Steering, Entanglement, Nonlocality, and the Einstein-Podolsky-Rosen Paradox. Phys. Rev. Lett. **98** (2007). DOI: 10.1103/PhysRevLett.98.140402. https://link.aps.org/doi/10.1103/PhysRevLett.98.140402

25. Greenberger, D.M., Horne, M.A., Zeilinger, A.: Going Beyond Bell's Theorem, (2007). arXiv: 0712.0921 [quant-ph]. https://arxiv.org/abs/0712.0921.

26. Greenberger, D.M., Horne, M.A., Shimony, A., Zeilinger, A.: Bell's theorem without inequalities. Americal Journal of Physics **58**, 1131–1143 (1990). DOI: 10.1119/1.16243

27. Bouwmeester, D., Pan, J.-W., Daniell, M., Weinfurter, H., Zeilinger, A.: Observation of Three-Photon Greenberger-Horne-Zeilinger Entanglement. Physical Review Letters **82**(7), 1345–1349 (1999). DOI: 10.1103/physrevlett.82.1345

28. Pan, J.-W., Bouwmeester, D., Daniell, M., Weinfurter, H., Zeilinger, A.: Experimental test of quantum nonlocality in three-photon Greenberger–Horne–Zeilinger entanglement. Nature **403**, 515–519 (2000)

29. Branciard, C., Gisin, N., Pironio, S.: Characterizing the Nonlocal Correlations Created via Entanglement Swapping. Phys. Rev. Lett. **104**, 170401 (2010). DOI: 10.1103/PhysRevLett.104.170401. https://link.aps.org/doi/10.1103/PhysRevLett.104.170401

30. Wood, C.J., Spekkens, R.W.: The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning. New Journal of Physics **17**(3) (2015). DOI: 10.1088/1367-2630/17/3/033002

31. Henson, J., Lal, R., Pusey, M.F.: Theory-independent limits on correlations from generalized Bayesian networks. New Journal of Physics **16**(11) (2014). DOI: 10.1088/1367-2630/16/11/113043

32. Fritz, T.: Beyond Bell's Theorem: Correlation Scenarios. New Journal of Physics **14** (2012). DOI: 10.1088/1367-2630/14/10/103001

33. Fritz, T.: Beyond Bell's Theorem II: Scenarios with Arbitrary Causal Structure. Communications in Mathematical Physics **341**(2), 391–434 (2015). DOI: 10.1007/s00220-015-2495-5

34. Laskey, K.B.: Quantum Causal Networks, (2007). arXiv: 0710.1200 [quant-ph]. https://arxiv.org/abs/0710.1200.

35. Leifer, M.S., Spekkens, R.W.: Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference. Physical Review A **88**(5) (2013). DOI: 10.1103/physreva.88.052130

36. Barrett, J., Lorenz, R., Oreshkov, O.: Quantum Causal Models, (2020). arXiv: 1906.10726 [quant-ph]. https://arxiv.org/abs/1906.10726.

37. Rosset, D., Gisin, N., Wolfe, E.: Universal bound on the cardinality of local hidden variables in networks. Quantum Information and Computation **18**(11 & 12) (2018). DOI: 10.26421/qic18.11-12

38. Wolfe, E., Spekkens, R.W., Fritz, T.: The Inflation Technique for Causal Inference with Latent Variables. Journal of Causal Inference **7**(2) (2019). DOI: 10.1515/jci-2017-0020

39. Navascués, M., Wolfe, E.: The Inflation Technique Completely Solves the Causal Compatibility Problem. Journal of Causal Inference **8**(1), 70–91 (2020). DOI: 10.1515/jci-2018-0008

40. Wolfe, E., Pozas-Kerstjens, A., Grinberg, M., Rosset, D., Acín, A., Navascués, M.: Quantum Inflation: A General Approach to Quantum Causal Compatibility. Physical Review X **11**. DOI: 10.1103/PhysRevX.11.021043

41. Clauser, J.F., Horne, M.A., Shimony, A., Holt, R.A.: Proposed Experiment to Test Local Hidden-Variable Theories. Phys. Rev. Lett. **23**, 880–884 (1969). DOI: 10.1103/PhysRevLett.23.880. https://link.aps.org/doi/10.1103/PhysRevLett.23.880

42. Brassard, G.: Quantum Communication Complexity (A Survey), (2001). arXiv: quant-ph/0101005 [quant-ph]. https://arxiv.org/abs/quant-ph/0101005.

43. Acín, A., Brunner, N., Gisin, N., Massar, S., Pironio, S., Scarani, V.: Device-Independent Security of Quantum Cryptography against Collective Attacks. Physical Review Letters **98** (2007). DOI: 10.1103/PhysRevLett.98.230501

44. Barrett, J., Linden, N., Massar, S., Pironio, S., Popescu, S., Roberts, D.: Non-local correlations as an information theoretic resource. Physical Review A **71** (2005). DOI: 10.1103/PhysRevA.71.022101

45. Tsirelson,
bibinitperiod: Some results and problems on quantum Bell-type inequalities. Hadronic Journal. Supplement **8**(4), 329–345 (1993)

46. Ji, Z., Natarajan, A., Vidick, T., Wright, J., Yuen, H.: MIP*=RE, (2022). arXiv: 2001.04383 [quant-ph]. https://arxiv.org/abs/2001.04383.

47. Slofstra, W.: The set of quantum correlations is not closed, (2017). arXiv: 1703.08618 [quant-ph]. https://arxiv.org/abs/1703.08618.

48. Al-Safi, S.W., Short, A.J.: Simulating all Nonsignaling Correlations via Classical or Quantum Theory with Negative Probabilities. Physical Review Letters **111**(17) (2013). DOI: 10.1103/physrevlett.111.170403

49. Abramsky, S., Brandenburger, A.: The sheaf-theoretic structure of non-locality and contextuality. New Journal of Physics **13**(11), 113036 (2011). DOI: 10.1088/1367-2630/13/11/113036

50. Barrett, J.: Information processing in generalized probabilistic theories, (2006). arXiv: quant-ph/0508211 [quant-ph]. https://arxiv.org/abs/quant-ph/0508211.

51. Wigner, E.: On the Quantum Correction For Thermodynamic Equilibrium. Phys. Rev. **40**, 749–759 (1932). DOI: 10.1103/PhysRev.40.749. https://link.aps.org/doi/10.1103/PhysRev.40.749

52. Feynman, R.P.: Negative Probability. In: Quantum Implications: Essays in Honour of David Bohm. Ed. by B.J. Hiley and D. Peat, pp. 235–248. Methuen (1987)

53. Yakaryılmaz, A.: Real-valued affine automata compute beyond Turing machines, (2022). arXiv: 2212.11834 [cs.FL]. https://arxiv.org/abs/2212.11834.

54. Barrett, J., de Beaudrap, N., Hoban, M.J., Lee, C.M.: The computational landscape of general physical theories. npj Quantum Information **5**(1) (2019). DOI: 10.1038/s41534-019-0156-9

55. Shpitser, I., Pearl, J.: Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models. In: Proceedings of the 21st AAAI Conference on Artificial Intelligence (2006)

56. Evans, R.J.: Graphs for Margins of Bayesian Networks. Scandinavian Journal of Statistics **43**(3), 625–648 (2015). DOI: 10.1111/sjos.12194

57. Richardson, T.: Markov Properties for Acyclic Directed Mixed Graphs. Scandinavian Journal of Statistics **30**(1), 145–157 (2003). DOI: 10.1111/1467-9469.00323

58. Zhang, X., Wang, Y.: On the physics of nested Markov models: a generalized probabilistic theory perspective, (2024). arXiv: 2411.11614 [quant-ph]. https://arxiv.org/abs/2411.11614.

59. Robeva, E., Seigal, A.: Duality of graphical models and tensor networks. Information and Inference: A Journal of the IMA **8**(2), 273–288 (2018). DOI: 10.1093/imaiai/iay009. eprint: https://academic.oup.com/imaiai/article-pdf/8/2/273/28864933/iay009.pdf

60. Bachmayr, M., Schneider, R., Uschmajew, A.: Tensor Networks and Hierarchical Tensors for the Solution of High-Dimensional Partial Differential Equations. Foundations of Computational Mathematics **16**, 1423–1472 (2016). DOI: 10.1007/s10208-016-9317-9

61. Barthel, T., Lu, J., Friesecke, G.: On the closedness and geometry of tensor network state sets. Letters in Mathematical Physics **112**(4) (2022). DOI: 10.1007/s11005-022-01552-z

62. Ye, K., Lim, L.-H.: Tensor network ranks, (2019). arXiv: 1801.02662 [math.NA]. https://arxiv.org/abs/1801.02662.

63. Janotta, P., Lal, R.: Non-locality in theories without the no-restriction hypothesis. Electronic Proceedings in Theoretical Computer Science **171**, 84–89 (2014). DOI: 10.4204/eptcs.171.8

64. Håstad, J.: Tensor rank is NP-complete. Journal of Algorithms **11**, 644–654 (1990). DOI: 10.1016/0196-6774(90)90014-6

A Quasiprobability Definitions

Definition 15 (Quasiprobability Distribution)**.** *Given a sample space Ω and σ-algebra $\mathcal{F}$, a* quasiprobability measure *or* quasimeasure *is a function $p : \mathcal{F} \to \mathbb{R}$ satisfying countable disjoint additivity and that $p(\Omega) = 1$. The triple $(\Omega, \mathcal{F}, p)$ is a* quasiprobability space*. If the space Ω is finite or countable, we call p a* quasiprobability distribution *or* quasidistribution *instead.*

Definition 16 (Conditional Quasidistribution and Independence)**.** *Given quasidistribution p and event B such that $p(B) \neq 0$, we define the conditional distribution $p(A|B)$ as*

$$p(A|B) = \frac{p(A, B)}{p(B)}.$$

We say events A and B are independent given *C with $p(C) \neq 0$ if*

$$p(A, B|C) = p(A|C) \cdot p(B|C).$$

Definition 17 (Quasiprobabilistic Random Variable)**.** *Given quasiprobability space $(\Omega, \mathcal{F}, p)$, a random variable is a function $X : \Omega \to E$, where E is a measurable space.*

B Extensions of Classical Results for Quasiprobability

B.1 Sufficiency of Finite Discrete Latent Variables

In this section we will extend one of the main results in [37] for Quasiprobability. We will use the same measure theoretic notation as the cited paper.

Proposition 18 (Finiteness of Quasilatents)**.** *Given DAG $\mathcal{G}$ with vertices $V \dot{\cup} L$, a distribution $p(x_V)$ over finite discrete variables X_V can be written as*

$$p(x_V) = \int_{\Omega_L} \prod_{v \in V} p(x_v | x_{pa(v)}) \prod_{l \in L} d\rho_l(x_l),$$

for (quasi or not) distributions $p(x_v|x_{pa(v)})$ for $v \in V$ and quasiprobability measures $\rho_l(x_l)$ over sets Ω_l, if and only if it can be written in the same form, but with all latents finite discrete.

Proof. Our proof will follow the one in [37]. We will replace each latent, in turn, with a finite discrete variable over the same space (or equivalently a finite subset of it, i.e. the image of the finite variable), keeping the observed response distributions the same.

Pick a latent variable $l \in L$. We will perform the following experiment: Pick an arbitrary value $x_l^* \in \Omega_l$, fix X_l to deterministically take value x_l^* and then construct p using the factorisation formula as before. The behaviour of the resulting model is:

$$p_{x_l^*}(x_V) = \left[\int_{\Omega_{L \setminus \{l\}}} \prod_{v \in V} p(x_v | x_{\mathrm{pa}(v)}) \prod_{l' \in L \setminus \{l\}} d\rho_{l'}(x_{l'}) \right]_{x_l = x_l^*}.$$

By the square brackets around the right-hand side we mean every occurence of x_l in the expression is replaced by the value x_l^*. Now the vector $\overrightarrow{p_{x_l^*}}$ is a random variable that depends on x_l^* with average

$$\mathbb{E}[\overrightarrow{p_{x_l^*}}] = \int_{\Omega_l} \overrightarrow{p_{x_l^*}} \, d\rho_l(x_l^*) = p(x_V).$$

Note that $p \in \mathbb{R}^d$, where $d = |\mathcal{X}_V|$. We can look at the subset $U = \{p_{x_l^*} | x_l^* \in \Omega_l\}$ of $\mathbb{R}^d$ consisting of the observed distributions resulting from the experiment over the entire range of possible deterministic choices for $x_l^* \in \Omega_l$. By construction, point $p(x_V)$ is an affine mixture with weights in $\mathbb{R}$ of points in U. Then by Lemma 19, $p(x_V)$ is a finite affine combination with weights in $\mathbb{R}$ of points in U, i.e. we can write

$$p(x_V) = \sum_{x_l} \int_{\Omega_{L\setminus\{l\}}} \prod_{v\in V} p(x_v|x_{\mathrm{pa}(v)}) \prod_{l\in L\setminus\{l\}} d\rho_l(x_l).$$

Applying this process to the rest of the latent variables then gives the desired result. ■

Lemma 19. *Given subset $U \subset \mathbb{R}^d$ and $\overrightarrow{x}^* \in \mathbb{R}^d$ a (possibly continuous) affine mixture of points in U, then $\overrightarrow{x}^*$ is a finite affine combination of points in U with weights in $\mathbb{R}$.*

Proof. Consider ρ to be a quasiprobability measure on (the Borel subsets of) $\mathbb{R}^d$ with the support set of the mass function being a subset of U. We need to prove that $\overrightarrow{x}^* = \int_U \overrightarrow{x}\, d\rho(\overrightarrow{x})$ is in the affine hull $\mathrm{aff}(U)$. Note that $\mathrm{aff}(U) = \overrightarrow{u} + \mathrm{span}(U - \overrightarrow{u})$, where we fix some $\overrightarrow{u} \in U$.

Now $\mathrm{span}(U-\overrightarrow{u})$ consists of the solutions to a homogeneous linear system of equations. Consider $\overrightarrow{a} \cdot \overrightarrow{x} = a_1x_1 + \ldots a_dx_d = 0$ to be one such equation, where $x_1, \ldots x_d$ are the entries of an arbitrary vector $\overrightarrow{x} \in \mathbb{R}^d$. Then, substituting $\overrightarrow{x}^* - \overrightarrow{u}$ in the above we get:

$$\begin{aligned}
\overrightarrow{a} \cdot (\overrightarrow{x}^* - \overrightarrow{u}) &= a_1(x_1^* - u_1) + \cdots + a_d(x_d^* - u_d) \\
&= (a_1x_1^* + \cdots + a_dx_d^*) - (a_1u_1 + \cdots + a_du_d) \\
&= \left(a_1 \int_U x_1\, d\rho(\overrightarrow{x}) + \cdots + a_d \int_U x_d\, d\rho(\overrightarrow{x})\right) \\
&\quad - (a_1u_1 + \cdots + a_du_d) \\
&= \int_U (a_1x_1 + \cdots + a_dx_d)\, d\rho(\overrightarrow{x}) - (a_1u_1 + \cdots + a_du_d) \\
&= (a_1u_1 + \cdots + a_du_d) \int_U 1d\, \rho(\overrightarrow{x}) - (a_1u_1 + \cdots + a_du_d) \\
&= 0,
\end{aligned}$$

where the second-to-last equality comes from the fact that

$$a_1x_1 + \cdots + a_dx_d = a_1u_1 + \cdots + a_dx_d,$$

because $\overrightarrow{x} - \overrightarrow{u} \in \mathrm{span}(U)$, and the last equality follows from the fact that ρ is an affine mixture. Indeed, all these calculations will hold for all linear equations in the system, hence $\overrightarrow{x}^* - \overrightarrow{u}$ is a solution to the system, so $\overrightarrow{x}^* \in \mathrm{aff}(U)$. ■

B.2 Soundness of Reducing a Graph

In this section we will show the results in [56] hold for our quasiprobabilistic models as well. However, we cannot assume yet that the three definitions of the quasiset are equivalent (and indeed, we want to use the results in this section to prove the equivalence). Therefore, we would in theory need to prove the three results in [56] for each of the three definitions. For the sake of brevity we will omit the measure theoretic approach in the original paper. If full rigourosity is desired, these proofs could be rewritten in a more formal way by following the original measure theoretical arguments more closely.

Definition 20 (Exogenisation)**.** *Consider a DAG $\mathcal{G}$ and a vertex u. Then the* exogenised graph $\mathfrak{r}(\mathcal{G}, u)$ *is defined as follows: take the parents of u in $\mathcal{G}$ and draw an arrow from each of them to each child of u, and erase the arrows pointing to u.*

Proposition 21 (Quasilatents Exogenisation - Lemma 3.7 in [56])**.** *Consider DAG $\mathcal{G}$ with vertices $V \dot{\cup} \{u\}$. Then $p(x_V)$ is the marginal of a quasidistribution with quasilatents/quasiresponses/quasinoise Markov wrt $\mathcal{G}$ iff the same is true for $\mathfrak{r}(\mathcal{G}, u)$.*

Proof. For the 'if' direction, assume $p(x_V)$ is the marginal of a quasidistribution with quasilatents/quasiresponses/quasinoise Markov wrt $\mathfrak{r}(\mathcal{G}, u)$. Then one can create a quasidistribution of the same type consistent with $\mathcal{G}$ that marginalises to p by simply setting $u_{new}|x_{\mathrm{pa}_{\mathcal{G}}(u)} := (u, x_{\mathrm{pa}_{\mathcal{G}}(u)})$, i.e. make the new u simply behave the same as before, but also send down the values of its parents. If only latents/responses/local noise variables were quasiprobabilistic before, the same is true for the new construction.

For the 'only if' direction, assume $p(x_V)$ is the marginal of a quasidistribution with quasilatents/quasiresponses/quasinoise Markov wrt $\mathcal{G}$. Extract the intrinsic randomness of the variable corresponding to u into a noise variable e_u, i.e. assume $u = f_u(e_u, x_{\mathrm{pa}(u)})$ for some deterministic function f_u. Then, we can create a quasidistribution consistent with $\mathfrak{r}(\mathcal{G}, u)$ that marginalises to p as follows: make $u_{new} := e_u$ and make each child of u apply the function f_u to simulate the value of u. Each such variable has all the ingredients to do so: it is now a child of all the original parents of u, as well as its noise variable. Once again, this constructs the same type of quasidistribution (i.e. with quasilatents/quasiresponses/quasinoise) as before. ■

Proposition 22 (Elimination of Redundant Quasilatents 1 - Lemma 3.8)**.** *Let $\mathcal{G}$ be a DAG with vertices $V \dot{\cup} \{u, w\}$ such that $pa(u) = pa(w) = \emptyset$ and $ch(w) \subseteq ch(u)$. Then $p(x_V)$ is the marginal of a quasidistribution with quasilatents/quasiresponses/quasinoise Markov wrt $\mathcal{G}$ iff the same is true for the graph $\mathcal{G}_{-w}$ obtained by removing w.*

Proof. The 'if' direction is trivial. For the 'only if' direction, make u_{new} simulate both u and w independently and delete w. All nodes in $\mathrm{ch}(u) \setminus \mathrm{ch}(w)$ can simply ignore the w-component of u_{new}. Trivially, the marginal of this construction over V is the same, and the obtained quasidistribution is of the same type. ■

Proposition 23 (Elimination of Latent Variables with One Child - Lemma 3.9)**.** *Let $\mathcal{G}$ be a DAG with vertices $V \dot{\cup} \{u\}$ such that u has no parents and at most one child. Then $p(x_V)$ is the marginal of a quasidistribution with quasilatents/quasireponses/quasinoise Markov with respect to $\mathcal{G}$ iff the same is true for $\mathcal{G}_{-u}$.*

Proof. The 'if' direction is trivial. For the 'only if' direction, we can assume WLOG that u has a child v (otherwise the case is trivial). We can also assume WLOG that v has no other unobserved parents. Otherwise simply apply the proposition just above to eliminate u as a redundant quasilatent. If u is nonnegative, then it can be eliminated using the original version of this result in [56] (alternatively, see the proof of $\mathcal{W}_n \subseteq \mathcal{W}_r$ in Appendix C). If u is quasiprobabilistic, we can actually replace u with a nonnegative variable and keep the same marginal over V (see the beginning of the case $\mathcal{W}_n \subseteq \mathcal{W}_l$ in Appendix C), then eliminate it. In any case, the same type of quasidistribution is obtained. ■

C Proof of Theorem 4

We assume $\mathcal{G}$ is reduced. Before the proof, we start with a few useful definitions.

Definition 24 (Assignments)**.** *Given a DAG $\mathcal{G}$, an* assignment *is a set of fixed quasidistributions $p(x_v|x_{pa_{\mathcal{G}}(v)})$, one for each vertex $v \in V \dot\cup L$ of $\mathcal{G}$ (as well as the necessary probability spaces for the latents). We say the assignment* achieves *distribution $p(x_V)$ over the vertices of $\mathcal{G}$ if*

$$p(x_V) = \sum_{x_L} \prod_{v \in V \cup L} p(x_v|x_{pa_{\mathcal{G}}(v)}).$$

Furthermore, we say the assignment:

- witnesses $p \in \mathcal{W}_l$ *if it achieves p, and $p(x_v|x_{pa_{\mathcal{G}}(v)}) \in [0,1]$ for all $v \in V$.*
- witnesses $p \in \mathcal{W}_r$ *if it achieves p and $p(x_l) \in [0,1]$ for all $l \in L$.*

Definition 25 (Noisy Assignments)**.** *Given a DAG $\mathcal{G}$, a* noisy assignment *is a set of random variables N_v (whose values we will denote by n_v as opposed to x_{N_v}) for each $v \in V$, together with deterministic distributions $p(x_v|x_{pa_{\mathcal{G}}(v)}, n_v)$ for all $v \in V$, quasiprobabilistic distributions $p(x_l)$ for each latent $l \in L$ of $\mathcal{G}$, and quasiprobabilistic distributions $p(n_v)$ for each $v \in V$. We say this noisy assignment* achieves *$p(x_V)$ if*

$$p(x_V) = \sum_{x_L, n_V} \prod_{v \in V} p(x_v|x_{pa_{\mathcal{G}}(v)}, n_v) \prod_{l \in L} p(x_l) \prod_{v \in V} p(n_v).$$

Furthermore, we say the noisy assignment witnesses *$p \in \mathcal{W}_n$ if $p(x_l) \in [0,1]$ for all $l \in L$.*

Note the use of sums instead of integrals. By Appendix B.1 we can assume finiteness of quasiprobabilistic latent variables (which covers the quasinoise and the quasilatents cases) and by [37] we can assume finiteness of nonnegative latent variables (which covers the quasiresponses case).

We are now ready to prove the quasimarginal set definition equivalence.

Theorem (**Equivalence of Quasimarginal Sets**)**.** *For any DAG $\mathcal{G}$, the following holds:*

$$\mathcal{W}_l(\mathcal{G}) = \mathcal{W}_r(\mathcal{G}) = \mathcal{W}_n(\mathcal{G}).$$

Proof. Fix a DAG $\mathcal{G}$. Start by noting that for any noisy assignment, there exist functions $f_v : \mathcal{X}_{\mathrm{pa}_{\mathcal{G}}(v)} \times \mathcal{X}_{N_v} \to \mathcal{X}_v$, one for each $v \in V$, defined as $f_v(x_{\mathrm{pa}_{\mathcal{G}}(v)}, n_v) = x_v^*$, where $x_v^* \in \mathcal{X}_v$ is the only value such that $p(x_v^*|x_{\mathrm{pa}_{\mathcal{G}}(v)}, n_v) \neq 0$, or equivalently $p(x_v^*|x_{\mathrm{pa}_{\mathcal{G}}(v)}, n_v) = 1$ (since for noisy assignments, these conditionals are deterministic). Therefore, a noisy assignment achieves $p(x_V)$ iff the following holds:

$$p(x_V) = \sum_{\substack{x_L, n_V \\ f_v(x_{\mathrm{pa}(v)}, n_v) = x_v \ \forall\, v \in V}} \prod_{l \in L} p(x_l) \prod_{v \in V} p(n_v).$$

The rest of the proof consists of showing that $\mathcal{W}_r = \mathcal{W}_n$ and $\mathcal{W}_l = \mathcal{W}_n$.

• $\mathcal{W}_n \subseteq \mathcal{W}_r$: Assume $p(x_V)$ is in $\mathcal{W}_n$ and consider a noisy assignment p^* that witnesses this. Then,

$$\begin{aligned}
p(x_V) &= \sum_{x_L, n_V} \prod_{v \in V} p^*(x_v|x_{\mathrm{pa}_{\mathcal{G}}(v)}, n_v) \prod_{l \in L} p^*(x_l) \prod_{v \in V} p^*(n_v) \\
&= \sum_{x_L, n_V} \prod_{v \in V} p^*(x_v, n_v|x_{\mathrm{pa}_{\mathcal{G}}(v)}) \prod_{l \in L} p^*(x_l) \\
&= \sum_{x_L} \prod_{v \in V} \left(\sum_{n_v} p^*(x_v, n_v|x_{\mathrm{pa}_{\mathcal{G}}(v)}) \right) \prod_{l \in L} p^*(x_l) \\
&= \sum_{x_L} \prod_{v \in V} p^*(x_v|x_{\mathrm{pa}_{\mathcal{G}}(v)}) \prod_{l \in L} p^*(x_l),
\end{aligned}$$

and hence p^* can be used to construct an assignment that achieves p. Additionally, since we assumed that p^* witnesses $p \in \mathcal{W}_n$, we must have $p^*(x_l) \in [0,1]$ for all $l \in L$, therefore we can indeed say the newly constructed assignment witnesses $p \in \mathcal{W}_r$.

• $\mathcal{W}_r \subseteq \mathcal{W}_n$: Consider $p \in \mathcal{W}_r$ and an assignment p^* that witnesses this. The goal is to create a noise variable for X_v that simulates its distribution conditioned on its parents. This will lead to a noisy assignment witnessing $p \in \mathcal{W}_n$

As such, consider for each X_v the variable N_v given by tuple of independent variables

$$N_v = \left(N_v^{(c_1)}, N_v^{(c_2)}, \ldots, N_v^{(c_s)}\right),$$

$\{c_1, \ldots, c_s\} = \mathcal{X}_{\mathrm{pa}_\mathcal{G}(x)}$, i.e. the c's are the possible values of the parents of v in the assignment p^*. Note that we can assume WLOG that $|\mathcal{X}_{\mathrm{pa}_\mathcal{G}(x)}|$ is a finite quantity by Appendix B.1. We choose

$$N_v^{(c_i)} \sim p^*(x_v | x_{\mathrm{pa}(x_v)} = c_i).$$

Finally, we define the functions f_v of our noisy assignment as:

$$f_v\left(x_{\mathrm{pa}(v)}, n_v = \left(n_v^{(c_1)}, \ldots, n_v^{(c_s)}\right)\right) := n_v^{(x_{\mathrm{pa}(v)})}.$$

Therefore, $N_v^{(c_i)}$ dictates the behaviour of X_v for one value of its parents. If we consider the noisy assignment p^{**} that has $p^{**}(x_l) = p^*(x_l)$ for all $l \in L$ and noise variables described above, then we have the following for every observed node $v \in V$:

$$\begin{aligned}
\sum_{n_v} p^{**}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) p^{**}(n_v) &= \sum_{\substack{n_v \\ f_v(x_{\mathrm{pa}(v)}, n_v) = x_v}} p^{**}(n_v) \\
&= \sum_{\substack{\left(n_v^{(c_1)}, \ldots, n_v^{(c_s)}\right) \\ f_v\left(x_{\mathrm{pa}(v)}, n_v = \left(n_v^{c_1}, \ldots, n_v^{cs}\right)\right) = x_v}} p^{**}\left(n_v = \left(n_v^{(c_1)}, \ldots, n_v^{(c_s)}\right)\right) \\
&= \sum_{\substack{\left(n_v^{(c_1)}, \ldots, n_v^{(c_s)}\right) \\ n_v^{\left(x_{\mathrm{pa}(v)}\right)} = x_v}} p^{**}\left(n_v^{(c_1)}\right) \ldots p^{**}\left(n_v^{(c_s)}\right) \\
&= p^{**}\left(n_v^{\left(x_{\mathrm{pa}_v}\right)} = x_v\right) \\
&= p^*(x_v | x_{\mathrm{pa}(v)}).
\end{aligned}$$

Finally, this implies that:

$$\begin{aligned}
&\sum_{x_L, n_V} \prod_{v \in V} p^{**}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l \in L} p^{**}(x_l) \prod_{v \in V} p^{**}(n_v) \\
&= \sum_{x_L} \prod_{v \in V} \sum_{n_v} p^{**}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) p^{**}(n_v) \prod_{l \in L} p^{**}(x_l) \\
&= \sum_{x_L} \prod_{v \in V} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \prod_{l \in L} p^*(x_l) \\
&= p(x_V),
\end{aligned}$$

so indeed p^{**} achieves $p(x_V)$. Since we haven't changed the latent variables, they are still nonnegative, and the response functions are deterministic, so p^{**} is indeed a noisy

assignment witnessing $p \in \mathcal{W}_n$.

Therefore, we have shown that $\mathcal{W}_r = \mathcal{W}_n$. We move on to $\mathcal{W}_l = \mathcal{W}_n$ next.
• $\mathcal{W}_n \subseteq \mathcal{W}_l$: Consider $p \in \mathcal{W}_n$ and let p^* be a noisy assignment witnessing that. Note that if an observed vertex a has no latent parents in $\mathcal{G}$, then its noise variable can be assumed nonnegative WLOG. To see this, assume N_a is not nonnegative. Then one can do the construction we used to prove $\mathcal{W}_n \subseteq \mathcal{W}_r$ to create an assignment p_1^* witnessing $p \in \mathcal{W}_r$, and then use the construction from $\mathcal{W}_r \subseteq \mathcal{W}_n$ to create another noisy assignment p_2^* witnessing $p \in \mathcal{W}_n$. However, this new noisy assignment has

$$N_a = \left(N_a^{(c_1)}, \ldots, N_a^{(c_s)}\right),$$

where $p_2^*\left(n_a^{(c_i)}\right) = p_1^*(x_a|x_{\mathrm{pa}_\mathcal{G}(a)})$. But since a has no latent parents, the fact that p_1^* achieves $p(x_V)$ implies by Lemma 27 that

$$p_1^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}) = p(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}). \tag{12}$$

Therefore, each of the $N_a^{(c_i)}$'s has nonnegative distributions, so N_a is nonnegative as well.

We will now iteratively eliminate noise variables from the noisy assignment and have them be simulated by latent variables where possible, or by response variables where no latent parents are present.

More precisely, for each observed variable $a \in V$ with no latent parents, make N_a a trivial deterministic variable with a single value 1_a and set

$$p_{new}^*(x_a|x_{\mathrm{pa}_\mathcal{G}(a)}, 1_a) := \sum_{n_a} p^*(x_a|x_{\mathrm{pa}_\mathcal{G}(a)}, n_a)p^*(n_a).$$

Otherwise, p_{new}^* behaves exactly the same as p^*.

Therefore, p_{new}^* achieves $p(x_V)$ as well:

$$\begin{aligned}
p(x_V) &= \sum_{x_L, n_V} \prod_{v\in V} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l\in L} p^*(x_l) \prod_{v\in V} p^*(n_v) \\
&= \sum_{x_L} \left(\sum_{n_a} p^*(x_a|x_{\mathrm{pa}_\mathcal{G}(a)}, n_a)p^*(n_a)\right) \cdot \\
&\quad \left(\sum_{n_{V\setminus\{a\}}} \prod_{v\in V\setminus\{a\}} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l\in L} p^*(x_l) \prod_{v\in V\setminus\{a\}} p^*(n_v)\right) \\
&= \sum_{x_L} p_{new}^*(x_a|x_{\mathrm{pa}_\mathcal{G}(a)}, 1_a) \cdot \\
&\quad \left(\sum_{n_{V\setminus\{a\}}} \prod_{v\in V\setminus\{a\}} p_{new}^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l\in L} p_{new}^*(x_l) \prod_{v\in V\setminus\{a\}} p_{new}^*(n_v)\right) \\
&= \sum_{x_L, n_V^{new}} \prod_{v\in V} p_{new}^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v^{new}) \prod_{l\in L} p_{new}^*(x_l) \prod_{v\in V} p_{new}^*(n_v^{new}),
\end{aligned}$$

and the new response function $p_{new}^*(x_a|x_{\mathrm{pa}(a)})$ we gave a is real nonnegative, as a result of (12). Next, replace p^* by p_{new}^* and repeat the process until we trivialise all of the noise variables associated with variables with no latent parents.

We then apply a similar process for the rest of the observed variables. For these, the noise variables can be from all of $\mathbb{R}$, so we hide them inside a latent parent.

More precisely, for each observed variable a that has at least one latent variable l_a, make N_a be trivial with a single value 1_a, replace l_a by $l_a^{new} = (l_a, N_a)$, and set

$$p^*_{new}\left(x_a | x_{\mathrm{pa}_\mathcal{G}(a)\setminus\{l_a\}}, x_{l_a^{new}} = (x_{l_a}, n_a), 1_a\right) := p^*(x_a | x_{\mathrm{pa}_\mathcal{G}(a)\setminus\{l_a\}}, x_{l_a}, n_a).$$

Otherwise, make p^*_{new} behave the same as p^*. In particular, for any other observed child b of l_a, set

$$p^*_{new}\left(x_b | x_{\mathrm{pa}_\mathcal{G}(b)\setminus\{l_a\}}, x_{l_a^{new}} = (x_{l_a}, n_a), n_b\right) := p^*(x_b | x_{\mathrm{pa}_\mathcal{G}(b)\setminus\{l_a\}}, x_{l_a}, n_b),$$

so given that $x_{l_a^{new}} = (x_{l_a}, n_a)$, we have that

$$p^*_{new}\left(x_b | x_{\mathrm{pa}_\mathcal{G}(b)}, n_b\right) := p^*(x_b | x_{\mathrm{pa}_\mathcal{G}(b)}, n_b).$$

Then, once again, p^*_{new} achieves $p(x_V)$ as well:

$$\begin{aligned}
p(x_V) &= \sum_{x_L, n_V} \prod_{v\in V} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l\in L} p^*(x_l) \prod_{v\in V} p^*(n_v) \\
&= \sum_{x_{L\setminus\{l_a\}}, n_{V\setminus\{a\}}} \sum_{n_a, x_{l_a}} (p^*(n_a)p^*(x_{l_a})) \cdot p^*(x_a | x_{\mathrm{pa}_\mathcal{G}(a)\setminus\{l_a\}}, x_{l_a}, n_a) \cdot \\
&\quad \left(\prod_{v\in V\setminus\{a\}} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l\in L\setminus\{l_a\}} p^*(x_l) \prod_{v\in V\setminus\{a\}} p^*(n_v) \right) \\
&= \sum_{x_{L\setminus\{l_a\}}, n_{V\setminus\{a\}}} \sum_{x^{new}_{l_a} = (n_a, x_{l_a})} p^*_{new}(x^{new}_{l_a}) \cdot p^*_{new}\left(x_a | x_{\mathrm{pa}_\mathcal{G}(a)\setminus\{l_a\}}, x_{l_a^{new}}, 1_a\right) \\
&\quad \left(\prod_{v\in V\setminus\{a\}} p^*_{new}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l\in L\setminus\{l_a\}} p^*_{new}(x_l) \prod_{v\in V\setminus\{a\}} p^*_{new}(n_v) \right) \\
&= \sum_{x^{new}_L, n^{new}_V} \prod_{v\in V} p^*_{new}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n^{new}_v) \prod_{l\in L} p^*_{new}(x^{new}_l) \prod_{v\in V} p^*_{new}(n^{new}_v).
\end{aligned}$$

Note that the response distribution of a is still deterministic and hence real nonnegative in p^*_{new}. Set $p^* := p^*_{new}$ and keep repeating the process until all noise variables have been trivialised.

At the end of this process, every noise variable in the model is trivial, so we can ignore them and obtain an assignment that achieves $p(x_V)$. Additionally, each observed variable has a response distribution that is either deterministic (if it has latent parents), or in any case real nonnegative (if it has no latent parents). Therefore, the constructed assignment witnesses $p \in \mathcal{W}_l$.

• $\mathcal{W}_l \subseteq \mathcal{W}_n$: Given any quasiprobabilistic finite discrete variable Λ, we can split its domain D_Λ into $D_\Lambda = D^+_\Lambda \dot\cup D^-_\Lambda$ as follows:

$$D^+_\Lambda = \{\lambda \in D_\Lambda : p(\lambda) \geq 0\} \qquad\qquad D^-_\Lambda = \{\lambda \in D_\Lambda : p(\lambda) < 0\}.$$

Now let $p^+_\Lambda = \sum_{\lambda\in D^+_\Lambda} p(\lambda)$, $p^-_\Lambda = \sum_{\lambda\in D^-_\Lambda} p(\lambda)$. These are finite quantities satisfying $p^+_\Lambda + p^-_\Lambda = 1$, $p^+_\Lambda \geq 1$, $p^-_\Lambda \leq 0$, with $p^-_\Lambda = 0$ iff Λ is nonnegative.

We can construct the following variables:

- Λ_B having $D_{\Lambda_B} = \{+, -\}$, with $\mathbb{P}(\Lambda_B = +) = p^+_\Lambda$, $\mathbb{P}(\Lambda_B = -) = p^-_\Lambda$;

- Λ^+ having $D_{\Lambda^+} = D^+_\Lambda$, $\mathbb{P}(\Lambda^+ = \lambda^+) = \frac{p(\lambda^+)}{p^+_\Lambda}$;

- Λ^- having $D_{\Lambda^-} = D^-_\Lambda$, $\mathbb{P}(\Lambda^- = \lambda^-) = \frac{p(\lambda^-)}{p^-_\Lambda}$.

Then one can see that sampling Λ is the same as sampling Λ_B and then sampling Λ^+ if the result was $+$ and Λ^- if the result was $-$.

Definition 26 (Ensemble of Components)**.** *We call the random variable* $(\Lambda_B, \Lambda^+, \Lambda^-)$ *the* ensemble of components *of quasiprobabilistic finite discrete random variable* Λ*.*

Now consider $p \in \mathcal{W}_l$ and an assignment witnessing this. Lemma 28 shows that there exists an assignment witnessing $p \in \mathcal{G}_l$ that replaces all latent variables Λ in the previous one by $(\Lambda_B, \lambda^+, \lambda^-)$. We will use this latter assignment (which we denote p^*) to prove that $p \in \mathcal{W}_n$.

The crucial step of the proof is to show that a binary quasiprobabilistic latent can be simulated by its children if they have access to quasiprobabilistic noise and a nonnegative shared latent variable. We will replace each quasiprobabilistic latent variable in p^* by a nonnegative one and will add noise variables to its children which will simulate its original behaviour. After we have done this for each latent, p^* will become a noisy assignment witnessing $p \in \mathcal{W}_n$.

We start with the assignment p^* above, which has each latent variable replaced by its ensemble of components. We can also consider this p^* to have a trivial 'noise variable' for each observed vertex, i.e. each vertex $v \in V$ has a 'noise variable' N_v that can take only one value 1_v with probability 1. Therefore, we can define

$$p^*(x_v | x_{pa_\mathcal{G}(v)}, 1_v) := p^*(x_v | x_{pa_\mathcal{G}(v)}).$$

It is worth mentioning that these are not proper noise variables, and we cannot yet view p^* as a noisy assignment, because the response distributions defined above are not deterministic.

We also have:

$$p(x_V) = \sum_{x_L, n_V} \prod_{v \in V} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l \in L} p^*(x_l) \prod_{v \in V} p^*(n_v), \tag{13}$$

with the reminder that the sum over each n_v is trivial.

Now we will describe a process to replace a latent variable by a nonnegative one while preserving the equality (13). Pick a latent node $l = (l_B, l^+, l^-) \in L$, with $|\mathrm{ch}(l)| = m$. Now, apply Lemma 29 for $n := m$ and $(p^+, p^-) := \Big(p^*(l_B = +),\, p^*(l_B = -)\Big)$, and consider $p^{(l)}$ to be a noisy assignment witnessing $p_{A_1,\dots A_m} \in \mathcal{W}_n(Com_m)$. Note that the lemma uses $\mathcal{W}_r$ instead, but we have already shown that $\mathcal{W}_r = \mathcal{W}_n$.

Note that assignment $p^{(l)}$ has only one latent variable, which we will call $\Lambda^{(l)}$. We will also assign one of the variables $A_1, \dots, A_m$ to each child of l; the one assigned to $v \in \mathrm{ch}(l)$ will be called A_v. The noise variable associated with A_v in $p^{(l)}$ will be called $N^{(l)}_v$. The deterministic value of A_v given $\lambda^{(l)}$ and $n^{(l)}_v$ is $f_v(\lambda^{(l)}, n^{(l)}_v)$.

Now replace in p^* the variable l_B by variable $\Lambda^{(l)}$, and for each $v \in \mathrm{ch}(l)$ replace N_v by $(N_v, N^{(l)}_v)$. Also, set

$$\begin{aligned} &p^*_{new}\left(x_v \middle| x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (\lambda^{(l)}, x^+_l, x^-_l), (n_v, n^{(l)}_v)\right) \\ &\quad := p^*\left(x_v | x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (f_v(\lambda^{(l)}, n^{(l)}_v), x^+, x^-), n_v\right). \end{aligned}$$

Now, note that for all $l \in L, x_V, x_l^+, x_l^-, n_V$, the following holds:

$$\begin{aligned}
&\sum_{x_{l_B}} p^*(x_{l_B}) \prod_{v\in \mathrm{ch}(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (x_{l_B}, x_l^+, x_l^-), n_v) \\
&= \sum_{x_{A_{\mathrm{ch}(l)}}} p_{A_{\mathrm{ch}(l)}}(x_{A_{ch(l)}}) \prod_{v\in \mathrm{ch}(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (x_{A_v}, x_l^+, x_l^-), n_v) \\
= \sum_{x_{A_{\mathrm{ch}(l)}}} &\left(\sum_{\lambda^{(l)}, n^{(l)}_{\mathrm{ch}(l)}} p^{(l)}(\lambda^{(l)}) \prod_{v\in \mathrm{ch}(l)} p^{(l)}(x_{A_v}|\lambda_l, n_v^{(l)}) p^{(l)}(n_v^{(l)}) \right) \prod_{v\in \mathrm{ch}(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (x_{A_v}, x_l^+, x_l^-), n_v) \\
&= \sum_{\lambda^{(l)}, n^{(l)}_{\mathrm{ch}(l)}} p^{(l)}(\lambda^{(l)}) \prod_{v\in \mathrm{ch}(l)} p^{(l)}(n_v^{(l)}) \prod_{v\in \mathrm{ch}(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (f_v(\lambda^{(l)}, n_v^{(l)}), x_l^+, x_l^-), n_v) \\
&= \sum_{\lambda^{(l)}, n^{(l)}_{\mathrm{ch}(l)}} p^{(l)}(\lambda^{(l)}) \prod_{v\in \mathrm{ch}(l)} p^{(l)}(n_v^{(l)}) \prod_{v\in \mathrm{ch}(l)} p^*_{new}\left(x_v \middle| x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (\lambda^{(l)}, x_l^+, x_l^-), (n_v, n_v^{(l)})\right),
\end{aligned} \tag{14}$$

where the first equality comes from the fact that $p_{A_{\mathrm{ch}(l)}}(x_{A_{\mathrm{ch}(l)}})$ is equal to 0 whenever $A_{\mathrm{ch}}(l) \notin \{(+,\ldots,+),(-,\ldots,-)\}$. As a consequence of these calculations we have:

$$\begin{aligned}
p(x_V) &= \sum_{x_L, n_V} \prod_{v\in V} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l'\in L} p^*(x_{l'}) \prod_{v\in V} p^*(n_v) \\
&= \sum_{x_{L\setminus\{l\}}, n_{V\setminus \mathrm{ch}(l)}} \prod_{v\in V\setminus ch(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l'\in L\setminus\{l\}} p^*(x_{l'}) \prod_{v\in V\setminus \mathrm{ch}(l)} p^*(n_v) \\
&\quad \sum_{x_{l_B}, x_l^+, x_l^-, n_{\mathrm{ch}(l)}} p^*(x_{l_B}) p^*(x_l^+) p^*(x_l^-) \prod_{v\in \mathrm{ch}(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (x_{l_B}, x_l^+, x_l^-), n_v) p^*(n_v) \\
&\overset{(14)}{=} \sum_{x_{L\setminus\{l\}}, n_{V\setminus \mathrm{ch}(l)}} \prod_{v\in V\setminus ch(l)} p^*(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v) \prod_{l'\in L\setminus\{l\}} p^*(x_{l'}) \prod_{v\in V\setminus \mathrm{ch}(l)} p^*(n_v) \\
&\quad \sum_{\lambda^{(l)}, x_l^+, x_l^-, n_{\mathrm{ch}(l)}, n^{(l)}_{\mathrm{ch}(l)}} p^{(l)}(\lambda^{(l)}) p^*(x_l^+) p^*(x_l^-) \\
&\quad \prod_{v\in \mathrm{ch}(l)} p^*_{new}\left(x_v \middle| x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, (\lambda^{(l)}, x_l^+, x_l^-), (n_v, n_v^{(l)})\right) p^*(n_v) p^{(l)}(n_v^{(l)}) \\
&= \sum_{x_L^{new}, n_V^{new}} \prod_{v\in V} p^*_{new}(x_v|x_{\mathrm{pa}_\mathcal{G}(v)}, n_v^{new}) \prod_{l\in L} p^*_{new}(x_l^{new}) \prod_{v\in V} p^*_{new}(n_v^{new}).
\end{aligned}$$

Hence, p^*_{new} also achieves p, but has a nonnegative distribution for l. Replace p^* by p^*_{new} and repeat the process for all latents. At the end, we have an almost-noisy-assignment achieving p. The only issue is the response functions are not deterministic, but this can be fixed by extending the noise variables to also contain the response distributions of the observed variables (following the construction we used to prove $\mathcal{W}_r \subseteq \mathcal{W}_n$). Finally, this gives a noisy assignment that achieves p, and hence witnesses $p \in \mathcal{W}_n$.

Therefore, $\mathcal{W}_l = \mathcal{W}_n$ and our proof is done. ■

Lemma 27. *Given DAG $\mathcal{G}$ with vertices $V \dot\cup L$, and p a distribution that can be written in the form*

$$p(x_V) = \sum_{x_L} \prod_{v\in V} p_1(x_v|x_{pa_\mathcal{G}(v)}) \prod_{l\in L} p_1(x_l), \tag{15}$$

for quasidistributions $p_1(x_v|x_{pa_\mathcal{G}(v)})$ and $p_1(x_l)$, then for any vertex $a \in V$ that has no latent parents,

$$p_1(x_a|x_{pa_\mathcal{G}(a)}) = p(x_a|x_{pa_\mathcal{G}(a)}).$$

Proof. We can assume WLOG that a has no children. Otherwise, simply sum out all its descendants (except a) in (15) to obtain that the remaining distribution factorises with respect to the new (ancestral) graph.

Now, denote by $A := V \setminus (\mathrm{pa}_{\mathcal{G}}(a) \cup \{a\})$. Then,

$$\begin{aligned} p(x_a, x_{\mathrm{pa}_{\mathcal{G}}(a)}) &= \sum_{x_L, x_A} \prod_{v \in V} p_1(x_v | x_{\mathrm{pa}_{\mathcal{G}}(v)}) \prod_{l \in L} p_1(x_l) \\ &= p_1(x_a | x_{\mathrm{pa}_{\mathcal{G}}(a)}) \sum_{x_L, x_A} \prod_{\substack{v \in V \\ v \neq a}} p_1(x_v | x_{\mathrm{pa}_{\mathcal{G}}(v)}) \prod_{l \in L} p_1(x_l). \end{aligned}$$

As a result,

$$\begin{aligned} p(x_{\mathrm{pa}_{\mathcal{G}}(a)}) &= \sum_{x_a} p(x_a, x_{\mathrm{pa}_{\mathcal{G}}(a)}) \\ &= \sum_{x_L, x_A} \prod_{\substack{v \in V \\ v \neq a}} p_1(x_v | x_{\mathrm{pa}_{\mathcal{G}}(v)}) \prod_{l \in L} p_1(x_l). \end{aligned}$$

Dividing the above two expressions, we obtain

$$\begin{aligned} p(x_a | x_{\mathrm{pa}_{\mathcal{G}}(a)}) &= \frac{p(x_a, x_{\mathrm{pa}_{\mathcal{G}}(a)})}{p(x_{\mathrm{pa}_{\mathcal{G}}(a)})} \\ &= p_1(x_a | x_{\mathrm{pa}_{\mathcal{G}}(a)}). \end{aligned}$$

■

Lemma 28. *Given DAG $\mathcal{G}$ and assignment p^* witnessing $p \in \mathcal{W}(\mathcal{G})$ for a distribution $p(x_V)$ over $\mathcal{G}$'s vertices. Then there exists assignment p^{**} that also witnesses that, which replaces every latent Λ in p^* by its ensemble of components $(\Lambda_B, \Lambda^+, \Lambda^-)$.*

Proof. We will iteratively replace each latent by its ensemble of components. Pick a latent $l \in L$, replace X_l by its ensemble of components (X_{l_B}, X_l^+, X_l^-) and set

$$p^*_{new}(x_{l_B}, x_l^+, x_l^-) := p_l^{x_{l_B}} \cdot \frac{p^*(x_l^+)}{p_l^+} \cdot \frac{p^*(x_l^-)}{p_l^-},$$

where $p_l^+ = \sum_{x_l;\, p^*(x_l) \geq 0} p^*(x_l)$, $p_l^- = \sum_{x_l;\, p^*(x_l) < 0} p^*(x_l)$, and $p_l^{x_{l_B}} := \begin{cases} p_l^+, \text{ if } x_{l_B} = + \\ p_l^-, \text{ if } x_{l_B} = - \end{cases}$.

For each observed child $v \in \mathrm{ch}(l)$ of l, set

$$p^*_{new}(x_v | x_{\mathrm{pa}_{\mathcal{G}}(v) \setminus \{l\}}, x_{l_B}, x_l^+, x_l^-) := p^*\left(x_v | x_{\mathrm{pa}_{\mathcal{G}}(v) \setminus \{l\}}, x_l = x_l^{x_{l_B}}\right).$$

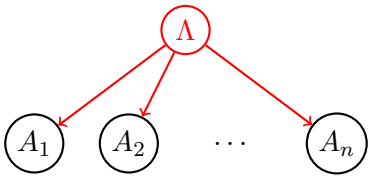


Figure 4: Graph Com_n.

Set everything else in p^*_{new} to behave the same as in p^*. Then p^*_{new} also achieves $p(x_V)$:

$$
\begin{aligned}
p(x_V) &= \sum_{x_L} \prod_{v \in V \cup L} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \\
&= \sum_{x_{L\setminus\{l\}}} \prod_{l' \in L\setminus\{l\}} p^*(x_{l'}) \prod_{v \in V\setminus \mathrm{ch}(l)} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \sum_{x_l} p^*(x_l) \prod_{v \in \mathrm{ch}(l)} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \\
&= \sum_{x_{L\setminus\{l\}}} \prod_{l' \in L\setminus\{l\}} p^*_{new}(x_{l'}) \prod_{v \in V\setminus \mathrm{ch}(l)} p^*_{new}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \\
&\left(\sum_{x_l^+} p^*(x_l^+) \prod_{v \in \mathrm{ch}(l)} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, x_l^+) + \sum_{x_l^-} p^*(x_l^-) \prod_{v \in \mathrm{ch}(l)} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, x_l^-) \right) \\
&= \sum_{x_{L\setminus\{l\}}} \prod_{l' \in L\setminus\{l\}} p^*_{new}(x_{l'}) \prod_{v \in V\setminus \mathrm{ch}(l)} p^*_{new}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \\
&\left(p_l^+ \sum_{x_l^+} \frac{p^*(x_l^+)}{p_l^+} \prod_{v \in \mathrm{ch}(l)} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, x_l^+) \cdot \sum_{x_l^-} \frac{p^*(x_l^-)}{p_l^-} \right. \\
&\left. + p_l^- \sum_{x_l^-} \frac{p^*(x_l^-)}{p_l^-} \prod_{v \in \mathrm{ch}(l)} p^*(x_v | x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, x_l^-) \cdot \sum_{x_l^+} \frac{p^*(x_l^+)}{p_l^+} \right) \\
&= \sum_{x_{L\setminus\{l\}}} \prod_{l' \in L\setminus\{l\}} p^*_{new}(x_{l'}) \prod_{V\setminus \mathrm{ch}(l)} p^*_{new}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)}) \\
&\quad \sum_{x_{l_B}, x_l^+, x_l^-} p^*_{new}(x_{l_B}, x_l^+, x_l^-) \prod_{v \in \mathrm{ch}(l)} p^*_{new}(x_v | x_{\mathrm{pa}_\mathcal{G}(v)\setminus\{l\}}, x_{l_B}, x_l^+, x_l^-) \\
&= \sum_{x_L^{new}} \prod_{v \in V \cup L} p^*_{new}(x_v | x^{new}_{\mathrm{pa}_\mathcal{G}(v)}),
\end{aligned}
$$

and the response distributions of the observed variables are still real nonnegative, therefore p^*_{new} witnesses $p \in \mathcal{W}_l$. Replace p^* by p^*_{new} and repeat the process for the next latent variable, until none are left. We define p^{**} to be the final assignment we obtain, and we are done. ■

Lemma 29. *Consider binary random variables $A_1, \ldots, A_n$ taking values in $\{+, -\}$. We denote $[a_1, \ldots, a_n]$ to mean the deterministic distribution assigning values $a_1, \ldots, a_n$ to the variables. Then the distribution*

$$p_{A_1,\ldots,A_n} = p^+ \cdot [+ + \cdots +] + p^- \cdot [- - \cdots -]$$

is in the set $\mathcal{W}_r$ of the graph Com_n (which stands for 'common ancestor') in Figure 4 for any $p^+, p^- \in \mathbb{R}$ such that $p^+ + p^- = 1$, $p^+ \geq 1$, $p^- \leq 0$.

Proof. We will prove via induction over n that there exists an assignment p^* witnessing $p_{A_1,\dots,A_n} \in \mathcal{W}_r$ that uses a latent variable Λ that is uniform over 2^{n-1} possible outcomes.

Let $D_\Lambda = \{1, \dots, 2^{n-1}\}$ and

$$p^*(\lambda) = \frac{1}{2^{n-1}},\ \forall \lambda \in D_\Lambda.$$

Denote also

$$p^*(a_i = + \,|\, \lambda) =: x_i^{(\lambda)} \qquad\qquad p^*(a_i = - \,|\, \lambda) =: y_i^{(\lambda)},$$

where $x_i^{(\lambda)} + y_i^{(\lambda)} = 1$ for all i and λ.

Therefore, the condition

$$p_{A_1,\dots,A_n}(a_1, \dots, a_n) = \sum_{\lambda \in D_\Lambda} p^*(\lambda) \prod_{i=1}^{n} p^*(a_i|\lambda)$$

is equivalent to the following system:

$$\begin{cases} 2^{n-1}p^+ & = \sum_\lambda \prod_{i=1}^n x_i^{(\lambda)} \\ 0 & = \sum_\lambda \prod_{i \in S^+} x_i^{(\lambda)} \prod_{i \in S^-} y_i^{(\lambda)} \quad , \forall\, S^+ \dot\cup S^- = \{1, \dots, n\}, S^+ \neq 0, S^- \neq 0 \\ 2^{n-1} - 2^{n-1}p^+ & = \sum_\lambda \prod_{i=1}^n y_i^{(\lambda)}. \end{cases} \tag{16}$$

The first equation comes from $p_{A_1,\dots,A_N}(+, \dots, +) = p^+$, the last one comes from $p_{A_1,\dots,A_n}(-, \dots, -) = p^-$, and the ones in the middle make sure that $p_{A_1,\dots,A_n}(a_1, \dots, a_n) = 0$ whenever the a_is are not all equal. Note the important condition $S^+ \neq 0$ and $S^- \neq 0$ in the second row. This ensures these equations do not overlap with the $+ + \cdots +$ and $- - \cdots -$ cases.

It is therefore sufficient to prove that system (16) has solutions $x_i^{(\lambda)}, y_i^{(\lambda)}$ over reals. We can slightly alter the system to bring it to a neater form.

To see this, if we consider $S^- := \{j\}$, the corresponding equation is

$$\begin{aligned} & \sum_\lambda \left(\prod_{i \neq j}^n x_i^{(\lambda)} \right) y_j^{(\lambda)} = 0 \\ \iff & \sum_\lambda \left(\prod_{i \neq j}^n x_i^{(\lambda)} \right) \left(1 - x_j^{(\lambda)}\right) = 0 \\ \iff & \sum_\lambda \prod_{i \in S^+} x_i^{(\lambda)} = 2^{n-1}p^+, \end{aligned}$$

where the last equality comes from using the first equation of the system. Therefore, because we are dealing with a linear system of equations, it is sound to replace the equations that have $|S^+| = n - 1$ with

$$\sum_\lambda \prod_{i \in S^+} x_i^{(\lambda)} = 2^{n-1}p^+.$$

However, we can now proceed inductively to prove that these equations being true for all $S^+ \neq \emptyset$ is enough to conclude we have solved the initial system. To see this, assume we proved it for all $|S^+| \geq k > 1$. Pick S^+ such that $|S^+| = k - 1$. Then the corresponding equation is:

	λ						
	1	2	...	2^{n-2}	$2^{n-2}+1$	...	2^{n-1}
$x_n^{(\lambda)}$	α	α	...	α	β	...	β
$x_{n-1}^{(\lambda)}$ $\vdots$ $x_1^{(\lambda)}$	solution for $n-1$ variables and p_1^+				solution for $n-1$ variables and p_2^+		

Table 1: A table showing the inductive construction of the system solution.

$$\sum_\lambda \prod_{i\in S^+} x_i^{(\lambda)} \prod_{i\in S^-} y_i^{(\lambda)} = 0$$
$$\iff \sum_\lambda \prod_{i\in S^+} x_i^{(\lambda)} \prod_{i\in S^-} \left(1 - x_i^{(\lambda)}\right) = 0$$
$$\iff \sum_\lambda \left(\prod_{i\in S^+} x_i^{(\lambda)} - \sum_{j\in S^-} \prod_{i\in S^+\cup\{j\}} x_i^{(\lambda)} + \sum_{j,k\in S^-} \prod_{i\in S^+\cup\{j,k\}} x_i^{(\lambda)} - \dots \right) = 0$$

But by induction hypothesis, all of the products in the bracket except the first are equal to $2^{n-1}p^+$. By the fact that the number of even subsets of a non-empty set is equal to the number of odd subsets, all but one of the products end up cancelling each other, leaving as desired the equation

$$\sum_\lambda \prod_{i\in S^+} x_i^{(\lambda)} = 2^{n-1}p^+.$$

Therefore, we can replace our system by the equivalent system:

$$\begin{cases} \sum_\lambda \prod_{i\in S^+} x_i^{(\lambda)} = 2^{n-1}p^+ & ,\forall\, S^+ \subseteq \{1,\dots,n\}, S^+ \neq 0 \\ \\ \sum_\lambda \prod_{i=1}^n \left(1 - x_i^{(\lambda)}\right) = 2^{n-1} - 2^{n-1}p^+. \end{cases}$$

In fact, we can completely drop the last equation, since it is a consequence of the other equations and the fact that our conditional distributions are normalised. Therefore, the actual system we will solve over reals $x_i^{(\lambda)}$ is

$$\sum_\lambda \prod_{i\in S} x_i^{(\lambda)} = 2^{n-1}p^+ \qquad ,\forall\, S \subseteq \{1,\dots,n\}, S \neq 0. \tag{17}$$

As mentioned before, we will prove this system has solutions over $\mathbb{R}$ for any $p^+ \geq 1$ using induction over n. For the base case, if $n = 1$ the system is simply $x_1^{(1)} = p^+$.

For the induction step, assume the system has solutions for any $p^+ \geq 1$ for $n-1$ variables and we will prove this is the case for n as well. Now pick $\{x_i^{(\lambda)}\}_{i=\overline{1..n-1},\, \lambda=\overline{1..2^{n-2}}}$ to be a solution for the system with $n-1$ variables with some value $p_1^+ \geq 1$ and similarly $\{x_i^{(\lambda)}\}_{i=\overline{1..n-1},\, \lambda=\overline{(2^{n-2}+1)..2^{n-1}}}$ a solution for the system with $n-1$ variables and another value $p_2^+ \geq 1$; we will pick values for p_1^+ and p_2^+ later.

We also pick $x_n^{(\lambda)} = \begin{cases} \alpha, & \text{if } \lambda \leq 2^{n-2} \\ \beta, & \text{otherwise} \end{cases}$. Once again, we leave α and β unfixed for now and will choose convenient values later. See Table 1 for a visualisation of the construction.

Now for any $\emptyset \neq S \subseteq \{1, \ldots, n-1\}$,

$$\begin{aligned}\sum_{\lambda}^{2^{n-1}} \prod_{i \in S} x_i^{(\lambda)} &= \sum_{\lambda=1}^{2^{n-2}} \prod_{i \in S} x_i^{(\lambda)} + \sum_{\lambda=2^{n-2}+1}^{2^{n-1}} \prod_{i \in S} x_i^{(\lambda)} \\ &= 2^{n-2} p_1^+ + 2^{n-2} p_2^+.\end{aligned}$$

The last equality holds because of the induction hypothesis. Therefore, when $\emptyset \neq S \subseteq \{1, \ldots, n-1\}$,

$$\sum_{\lambda}^{2^{n-1}} \prod_{i \in S} x_i^{(\lambda)} = 2^{n-1} p^+ \iff \boxed{p_1^+ + p_2^+ = 2p^+}.$$

If S contains n, then we need to consider two cases. If $S \setminus \{n\} \neq \emptyset$, we get:

$$\begin{aligned}\sum_{\lambda=1}^{2^{n-1}} \prod_{i \in S} x_i^{(\lambda)} &= \sum_{\lambda=1}^{2^{n-1}} x_n^{(\lambda)} \prod_{i \in S \setminus \{n\}} x_i^{(\lambda)} \\ &= \sum_{\lambda=1}^{2^{n-2}} \alpha \cdot \prod_{i \in S \setminus \{n\}} x_i^{(\lambda)} + \sum_{\lambda=2^{n-2}+1}^{2^{n-1}} \beta \cdot \prod_{i \in S \setminus \{n\}} x_i^{(\lambda)} \\ &= \alpha \cdot 2^{n-2} p_1^+ + \beta \cdot 2^{n-2} p_2^+\end{aligned}$$

The last equality holds because of the induction hypothesis and the fact that $S \setminus \{n\} \neq \emptyset$. Therefore, when $\{n\} \subsetneq S$,

$$\sum_{\lambda}^{2^{n-1}} \prod_{i \in S} x_i^{(\lambda)} = 2^{n-1} p^+ \iff \boxed{\alpha p_1^+ + \beta p_2^+ = 2p^+}.$$

If $S = \{n\}$, we get:

$$\sum_{\lambda=1}^{2^{n-1}} x_n^{(\lambda)} = 2^{n-2}\alpha + 2^{n-2}\beta = 2^{n-1} p^+ \iff \boxed{\alpha + \beta = 2p^+}.$$

Therefore, if we find $p_1^+, p_2^+ \geq 1$ and $\alpha, \beta \in \mathbb{R}$ that satisfy the system

$$\begin{cases} p_1^+ + p_2^+ = 2p^+ \\ \alpha p_1^+ + \beta p_2^+ = 2p^+ \\ \alpha + \beta = 2p^+ \end{cases},$$

we are done. This is actually the system for $n = 2$, and indeed has solution $p_1^+ = 1, p_2^+ = 2p^+ - 1 \geq 1$ and $\alpha = 2p^+, \beta = 0$.

Building the solution as described above therefore attests that the system has solutions for n as well and our induction is complete. Therefore, the system has solutions for any natural number n, which means we can always solve it to construct an assignment that achieves $p_{A_1,\ldots,A_n}$. Note that if $\alpha = 2p^+ > 1$, then this will mean that $y_n^{(\lambda)}$ will be negative for some values of λ, which will translate to some negative response distributions. However, we set the variable Λ to be uniform, and therefore positive, so the assignment we obtain at the end will indeed witness $p_{A_1,\ldots,A_N} \in \mathcal{W}_r$. ■

D Proof of Proposition 6:

In this section we will prove that $\mathcal{W}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$ by following some arguments in [15]. We will not define all the notions used in the proof and direct the reader to the cited paper for more details.

To start off, one can generalise the notion of a kernel to the notion of *quasikernel* in the natural way. We will assume soundness of semigraphoid axioms for quasikernels, given that the proof of this in [15] uses only symbolic manipulations. Similarly, one can extend the (polynomial) fixing operation $\phi_{\mathbf{w}}(\_;\mathcal{G})$ to be defined on quasikernels by simply following the same formula as the one used for kernels.

The nested Markov model of a CADMG $\mathcal{G}$ with observed vertices V and fixed vertices W is defined as the set of kernels $q_{V|W}(x_V|x_W)$ with the property that for any fixing sequence $\mathbf{w}$, the kernel $\phi_{\mathbf{w}}(q;\mathcal{G})$ satisfies the 4 equivalent Markov properties with respect to the fixed graph $\phi_{\mathbf{w}}(\mathcal{G})$, including the local Markov property. The local Markov property enforces some equality conditions on the fixed kernel, which is a polynomial function of the initial kernel, so the set of polynomial equalities enforced by the nested Markov model can be seen as the set of conditions that are obtained by replacing in the local Markov property constraints every occurence of $\phi_{\mathbf{w}}(q;\mathcal{G})$ with the corresponding expression in q.

Now note that the proof of Theorem B.5 in [15] uses only factorisation and the chain rule of probabilities, so it is sound for quasikernels as well, meaning that

$$\mathcal{P}^c_l(\mathcal{G},\prec) = \mathcal{P}^c_f(\mathcal{G}),$$

where we read $\mathcal{P}^c_l(\mathcal{G},\prec)$ to be the set of quasikernels which satisfy the local Markov property with respect to CADMG $\phi_{\mathbf{w}}(\mathcal{G})$, and $\mathcal{P}^c_f(\mathcal{G})$ to be the set of quasikernels which Tian factorise with respect to $\phi_{\mathbf{w}}(\mathcal{G})$, with the caveat that now the factors are allowed to be quasiprobabilistic.

Finally, one can see that all quasikernels that can be obtained from some quasidistribution in $\mathcal{W}(\mathcal{G})$ by applying a fixing sequence must Tian factorise (which can be seen by grouping terms in the original factorisation), meaning they satisfy the local Markov property. Translating these into constraints on the initial quasidistribution, we get that all nested Markov constraints are satisfied.

E Conjecture 1 for Tree-Structured Correlation Scenarios

E.1 Tensor Multilinear Products

In this section of the appendix, we include an introduction to the tensor operations we will be using.

Let k be a field of characteristic 0. A $n_1 \times \cdots \times n_q$ tensor $\mathcal{T}$ with entries in k is called *simple* or *elementary* if it is the tensor product of a bunch of vectors, i.e. there exist vectors $v_1, \ldots, v_q$ such that

$$\mathcal{T} = v_1 \otimes v_2 \otimes \cdots \otimes v_q.$$

Let $\mathcal{T}$ be an arbitrary $n_1 \times \cdots \times n_q$ tensor with entries in k. Then $\mathcal{T}$ has a *tensor rank*, which is a minimal positive integer r such that

$$\mathcal{T} = \sum_{i=1}^{r} v_1^{(i)} \otimes v_2^{(i)} \otimes \cdots \otimes v_q^{(i)},$$

for some k-valued vectors $\{v_j^{(i)}\}_{1\leq i\leq r, 1\leq j\leq q}$, i.e. r is the minimal number of simple tensors that sum to $\mathcal{T}$.

Note the correspondence between the rank of a tensor and the rank of a matrix. Indeed, the rank of a matrix is a particular case of a tensor rank, where the tensor has order 2. Note that in the case of tensors however, the rank can't be computed from the dimension of the fibre vector spaces in the same way as in the case of a matrix (the fibre is the generalisation of the concepts of rows and columns of a matrix). In fact these dimensions need not be equal for tensors for every type of fibre we choose, whereas it is well-known the row rank and column rank are equal for every matrix. Computing the rank over the reals of an arbitrary tensor even as small as order 3 is in fact NP-hard [64].

Given an arbitrary tensor $\mathcal{T}$ as before and its tensor rank representation, we can apply a linear transformation $A : k^{n_i} \to k^{m_i}$ to its ith component. This map sends $v_i \in k^{n_i}$ to $Av_i \in k^{m_i}$, and $v_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes v_q$ to $v_1 \otimes \cdots \otimes (Av_i) \otimes \cdots \otimes v_q$, and so by linearity we can define its application on the arbitrary tensor $\mathcal{T}$ denoted using the symbol " $\cdot_i$" as:

$$A \cdot_i \mathcal{T} = \sum_{j=1}^{r} v_1^{(j)} \otimes \cdots \otimes \left(Av_i^{(j)}\right) \otimes \cdots \otimes v_q^{(j)}.$$

This holds even if the representation is not a tensor rank representation, i.e. even if the value r is not optimal.

It is not too difficult to see that two such linear transformations acting on different components $A_i : k^{n_i} \to k^{m_i}$ and $A_j : k^{n_j} \to k^{m_j}$ commute, i.e.

$$A_i \cdot_i (A_j \cdot_j \mathcal{T}) = A_j \cdot_j (A_i \cdot_i \mathcal{T}).$$

In fact, if we have q of these transformations, $\{A_i : k^{n_i} \to k^{m_i}\}_{1 \leq i \leq q}$, we can define the multilinear multiplication operator $\mathcal{A} = (A_1, \ldots, A_q) : k^{n_1} \otimes \cdots \otimes k^{n_q} \to k^{m_1} \otimes \cdots \otimes k^{m_q}$ as

$$\mathcal{A} \cdot \mathcal{T} = \sum_{i=1}^{r} \left(A_1 v_1^{(i)}\right) \otimes \left(A_2 v_2^{(i)}\right) \otimes \cdots \otimes \left(A_q v_q^{(i)}\right).$$

Additionally, it's easy to check that if A_i has a left-inverse A_i^L for all i, then there is a left-inverse multilinear multiplication operator $\mathcal{A}^L = (A_1^L, \ldots, A_q^L)$ such that

$$\mathcal{A}^L \cdot \mathcal{A} \cdot \mathcal{T} = \mathcal{T}.$$

The right-inverse multilinear multiplication operator and the inverse multilinear multiplication operator are defined similarly in the way one would expect.

E.2 Gauge Freedom of Tensor Network Decompositions

Definition 30 (Gauge Transformations)**.** *Let $\mathcal{G}$ be a tensor network with spaces $\{\mathcal{X}_v\}_{v \in V}$ for its nodes and bond dimensions $\{r_e\}_{e \in E}$ and $\mathcal{T} \in \bigotimes_{v \in V} k^{|\mathcal{X}_v|}$ a tensor that decomposes with respect to this construction. Let such a decomposition be given by tensors $\left\{T^{(v)} \in k^{|\mathcal{X}_v|} \otimes \left(\bigotimes_{e \in inc(v)} k^{r_e}\right)\right\}_{v \in V}$. Now pick an edge $e^* = (a, b) \in E$ and an $r_{e^*} \times r_{e^*}$ invertible matrix X. Then a* gauge transformation *of the decomposition consists of replacing $T^{(a)}$ by $X \times_{i_{e^*}} T^{(a)}$ and $T^{(b)}$ by $\left(X^T\right)^{-1} \times_{i_{e^*}} T^{(b)}$.*

Proposition 31 (Gauge Freedom)**.** *Consider the tensor network $\mathcal{G}$ with bond dimensions $\{r_e\}_{e \in E}$ as before, and a decomposition of $\mathcal{T}$ with respect to $\mathcal{G}$ and $\{r_e\}_{e \in E}$. Then the resulting set of tensors $\{T^{(v)}\}_{v \in V}$ after applying a gauge transformation is still a tensor decomposition of the same tensor $\mathcal{T}$.*

Proof. Let

$$T^{(a)} = \sum_{j_a=1}^{c_a} u^{(j_a)} \otimes \left(\bigotimes_{e \in \mathrm{inc}(a)} u_e^{(j_a)} \right), \tag{18}$$

$$T^{(b)} = \sum_{j_b=1}^{c_b} w^{(j_b)} \otimes \left(\bigotimes_{e \in \mathrm{inc}(b)} w_e^{(j_b)} \right), \tag{19}$$

be rank decompositions of $T^{(a)}$ and $T^{(b)}$. The idea is to show that the RHS of the tensor network decomposition formula, namely

$$\sum_{\{1 \le i_e \le r_e\}_{e \in E}} \prod_{v \in V} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v),$$

only depends on the $\mathbf{u}_{e^*}$ and $\mathbf{w}_{e^*}$ vectors via the dot products $(\mathbf{u}_{e^*})^T \cdot \mathbf{w}_{e^*}$, which remain unchanged after the gauge transformation.

More precisely, if we write the explicit forms of equations (18) and (19) we obtain:

$$T^{(a)}_{\{i_e\}_{e \in \mathrm{inc}(a)}}(x_a) = \sum_{j_a=1}^{c_a} u^{(j_a)}(x_a) \cdot \left(\prod_{e \in \mathrm{inc}(a)} u_e^{(j_a)}(i_e) \right),$$

$$T^{(b)}_{\{i_e\}_{e \in \mathrm{inc}(b)}}(x_b) = \sum_{j_b=1}^{c_b} w^{(j_b)}(x_b) \cdot \left(\prod_{e \in \mathrm{inc}(b)} w_e^{(j_b)}(i_e) \right).$$

Now we have

$$\begin{aligned}
&\sum_{i_{e^*}} T^{(a)}_{\{i_e\}_{e \in \mathrm{inc}(a)}}(x_a) \cdot T^{(b)}_{\{i_e\}_{e \in \mathrm{inc}(b)}}(x_b) = \sum_{j_a=1, j_b=1}^{c_a, c_b} \sum_{i_{e^*}} u^{(j_a)}(x_a) \cdot w^{(j_b)}(x_b) \\
&\cdot \left(\prod_{e \in \mathrm{inc}(a)} u_e^{(j_a)}(i_e) \right) \cdot \left(\prod_{e \in \mathrm{inc}(b)} w_e^{(j_b)}(i_e) \right) \\
&= \sum_{j_a=1, j_b=1}^{c_a, c_b} u^{(j_a)}(x_a) \cdot w^{(j_b)}(x_b) \cdot \left(\prod_{e \in \mathrm{inc}(a) \setminus \{e^*\}} u_e^{(j_a)}(i_e) \right) \cdot \left(\prod_{e \in \mathrm{inc}(b) \setminus \{e^*\}} w_e^{(j_b)}(i_e) \right) \\
&\cdot \sum_{i_{e^*}} \left(u_{e^*}^{(j_a)}(i_{e^*}) \cdot w_{e^*}^{(j_b)}(i_{e^*}) \right) \\
&= \sum_{j_a=1, j_b=1}^{c_a, c_b} u^{(j_a)}(x_a) \cdot w^{(j_b)}(x_b) \cdot \left(\prod_{e \in \mathrm{inc}(a) \setminus \{e^*\}} u_e^{(j_a)}(i_e) \right) \cdot \left(\prod_{e \in \mathrm{inc}(b) \setminus \{e^*\}} w_e^{(j_b)}(i_e) \right) \\
&\left((\mathbf{u}_{e^*}^{(j_a)})^T \cdot \mathbf{w}_{e^*}^{(j_b)} \right),
\end{aligned}$$

where we wrote the vectors at the end of the last expression in bold to emphasise that they are vectors, and part of a dot product.

Applying the gauge transformation only acts on the u_{e^*} and w_{e^*} components, so the only change in the value of this expression will be that the last bracket becomes $\left((X \mathbf{u}_{e^*}^{(j_a)})^T \cdot ((X^T)^{-1} \mathbf{w}_{e^*}^{(j_b)}) \right)$, but this doesn't change its value. Therefore,

$$\sum_{i_{e^*}} T^{(a)}_{\{i_e\}_{e \in \mathrm{inc}(a)}}(x_a) \cdot T^{(b)}_{\{i_e\}_{e \in \mathrm{inc}(b)}}(x_b)$$

doesn't change its value, and so

$$\sum_{\{i_e\}_{e\in E}} \prod_{v\in V} T^{(v)}_{\{i_e\}_{e\in \mathrm{inc}(v)}}(x_v) \;=\; \sum_{\{i_e\}_{e\in E\setminus\{e^*\}}} \prod_{v\in V\setminus\{a,b\}} T^{(v)}_{\{i_e\}_{e\in \mathrm{inc}(v)}}(x_v) \cdot \left(\sum_{i_{e^*}} T^{(a)}_{\{i_e\}_{e\in \mathrm{inc}(a)}}(x_a) \cdot T^{(b)}_{\{i_e\}_{e\in \mathrm{inc}(b)}}(x_b) \right)$$

remains unchanged as well, so it will equal $\mathcal{T}(x_V)$ after the gauge transformation. ■

E.3 Proof of Nonzero Sum Assumption

In this section we will prove the Nonzero Sums Lemma which will be used in the proof of Theorem 14 in the next section of this appendix. The proof makes use of gauge transformations to modify each component tensor in a way that makes these sums nonzero. We start with another result which will be used in the proof of the lemma.

Definition 32 ($\exists\forall$-Partitions)**.** *Let $A \in \mathbb{R}^{|\mathcal{X}_a|} \otimes \left(\bigotimes_{e\in E_a} \mathbb{R}^{r_e}\right)$ and $B \in \mathbb{R}^{|\mathcal{X}_b|} \otimes \left(\bigotimes_{e\in E_b} \mathbb{R}^{r_e}\right)$ be two tensors with exactly one common mode $e^* = E_a \cap E_b$. If we are given partition $E_a = E_a^{\exists} \dot\cup E_a^{\forall}$ for E_a, then we say $(E_a^{\exists}, E_a^{\forall})$ is a* $\exists\forall$-partition *for A if the following property holds:*

$$\exists \{1 \le i_e \le r_e\}_{e\in E_a^{\exists}} \text{ such that } \forall \{1 \le i_e \le r_e\}_{e\in E_a^{\forall}}, \quad \sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a) \neq 0.$$

If we are given partitions $E_a = E_a^{\exists} \dot\cup E_a^{\forall}$ and $E_b = E_b^{\exists} \dot\cup E_b^{\forall}$ for E_a and E_b, then we say the pair of partitions $(E_a^{\exists}, E_a^{\forall}; E_b^{\exists}, E_b^{\forall})$ is a $\exists\forall$-partition pair *if $(E_a^{\exists}, E_a^{\forall})$ is a $\exists\forall$-partition for A and $(E_b^{\exists}, E_b^{\forall})$ is a $\exists\forall$-partition for B.*

Lemma 33. *Let $A \in \mathbb{R}^{|\mathcal{X}_a|} \otimes \left(\bigotimes_{e\in E_a} \mathbb{R}^{r_e}\right)$ and $B \in \mathbb{R}^{|\mathcal{X}_b|} \otimes \left(\bigotimes_{e\in E_b} \mathbb{R}^{r_e}\right)$ be two tensors with exactly one common mode $e^* = E_a \cap E_b$ and a $\exists\forall$-partition pair $(E_a^{\exists}, E_a^{\forall}; E_b^{\exists}, E_b^{\forall})$. Then there exists invertible $r_{e^*} \times r_{e^*}$ matrix X with real entries such that*

$$(E_a^{\exists} \setminus \{e^*\}, E_a^{\forall} \cup \{e^*\}; E_b^{\exists}, E_b^{\forall})$$

is a $\exists\forall$-partition pair for tensors $X \cdot_{e^*} A$ and $(X^T)^{-1} \cdot_{e^*} B$.

Proof. If $e^* \in E_a^{\forall}$ then we are done. Assume therefore that $e^* \in E_a^{\exists}$. Let $\{1 \le i_e \le r_e\}_{e\in E_a^{\exists}}$ be an assignment to the indices of A in $E_a^{\exists}$ such that

$$\forall \{1 \le i_e \le r_e\}_{e\in E_a^{\forall}}, \qquad \sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a) \neq 0.$$

Denote by q the value of i_{e^*} in this assignment.

We define X to have the following structure:

$$X = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix} + \begin{pmatrix} 0 & \dots & \alpha_1 & \dots & 0 \\ 0 & \dots & \alpha_2 & \dots & 0 \\ & & \vdots & & \\ 0 & \dots & \alpha_{r_{e^*}} & \dots & 0 \end{pmatrix},$$

i.e. it has 1s on the main diagonal and some real coefficients α which we will set later on column q. The only coefficient we set for now is $\alpha_q := 0$, so the row corresponding to $i_{e^*} = q$ only contains a single 1 and no additional α entry.

Then one can check that by calculating $X \cdot_{e^*} A$ we get

$$(X \cdot_{e^*} A)_{\{i_e\}_{e\in E_a}}(x_a) = \begin{cases} A_{\{i_e\}_{e\in E_a}}(x_a) & , \text{ if } i_{e^*} = q \\ A_{\{i_e\}_{e\in E_a}}(x_a) + \alpha_{i_e}\left[A_{\{i_e\}_{e\in E_a}}(x_a)\right]_{i_{e^*}=q} & , \text{ if } i_{e^*} \neq q, \end{cases}$$

where by $\left[A_{\{i_e\}_{e\in E_a}}(x_a)\right]_{i_{e^*}=q}$ we denote the expression inside the square brackets with every occurrence of i_{e^*} replaced by the value q.

Therefore, if we fix $\{1 \leq i_e \leq r_e\}_{e\in E_a^{\exists}\setminus\{e^*\}}$ to have the values we considered before, the condition

$$\forall\{1 \leq i_e \leq r_e\}_{e\in E_a^{\forall}\cup\{e^*\}}, \qquad \sum_{x_a}(X \cdot_{e^*} A)_{\{i_e\}_{e\in E_a}}(x_a) \neq 0$$

is equivalent to

$$\forall\{1 \leq i_e \leq r_e\}_{e\in E_a^{\forall}\cup\{e^*\}}, \qquad \sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a) + \alpha_{i_e}\left[\sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a)\right]_{i_{e^*}=q} \neq 0.$$

Note that since $\alpha_q = 0$, this statement does take into consideration the case $i_{e^*} = q$ as well.

Now since we know that

$$\left[\sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a)\right]_{i_{e^*}=q} \neq 0$$

must hold because of the way $\{i_e\}_{e\in E_a^{\exists}\setminus\{e^*\}}$ and q were chosen, we can rewrite the above condition in the equivalent form

$$\forall\{1 \leq i_e \leq r_e\}_{e\in E_a^{\forall}\cup\{e^*\}}, \qquad \alpha_{i_e} \neq -\frac{\sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a)}{\left[\sum_{x_a} A_{\{i_e\}_{e\in E_a}}(x_a)\right]_{i_{e^*}=q}}, \tag{20}$$

so these are the constraints that the α coefficients need to satisfy in order for $E_a^{\exists} \setminus \{e^*\}$ and $E_a^{\forall} \cup \{e^*\}$ to form a $\exists\forall$-partition for $X \cdot_{e^*} A$.

Now let's look at $(X^T)^{-1} \cdot_{e^*} B$. We have that

$$(X^T)^{-1} = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ 0 & 1 & & \dots & & 0 \\ & & & \vdots & & \\ -\alpha_1 & -\alpha_2 & \dots & 1 & \dots & -\alpha_{r_{e^*}} \\ & & & \vdots & & \\ 0 & 0 & & \dots & & 1 \end{pmatrix}.$$

Then one can calculate that

$$\left((X^T)^{-1} \cdot_{e^*} B\right)_{\{i_e\}_{e\in E_b}}(x_b) = \begin{cases} B_{\{i_e\}_{e\in E_b}}(x_b) - \sum_{q'=1}^{r_{e^*}} \alpha_{i_e}\left[B_{\{i_e\}_{e\in E_b}}(x_b)\right]_{i_{e^*}=q'} & , \text{ if } i_{e^*} = q \\ B_{\{i_e\}_{e\in E_b}}(x_b) & , \text{ if } i_{e^*} \neq q. \end{cases}$$

Now pick some values $\{1 \leq i_e \leq r_e\}_{e \in E_b^\exists}$ such that

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall}, \qquad \sum_{x_b} B_{\{i_e\}_{e \in E_b}}(x_b) \neq 0,$$

and distinguish two cases.

- *Case 1:* $e^* \in E_b^\exists$. If there exists $i_{e^*} = q' \neq q$ such that

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall}, \qquad \left[\sum_{x_b} B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q'} \neq 0,$$

then the same is true for $(X^T)^{-1} \cdot_{e^*} B$ for any values for the coefficients α. Pick these coefficients to satisfy (20) and the matrix X obtained in this way is exactly what we need.

Otherwise, $i_{e^*} = q$ is the only value for which

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall}, \qquad \left[\sum_{x_b} B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q} \neq 0$$

holds. In this case, the condition

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall}, \qquad \left[\sum_{x_b} \left((X^T)^{-1} \cdot_{e^*} B\right)_{\{i_e\}_{e \in E_B}}(x_b)\right]_{i_{e^*}=q} \neq 0$$

is equivalent to

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall}, \qquad \left[B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q} - \sum_{q'=1}^{r_{e^*}} \alpha_{i_e} \left[B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q'} \neq 0. \quad (21)$$

We view these as inequations in the α coefficients. Now we know that

$$\left[B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q} \neq 0$$

holds in all of these equations because of the way $\{i_e\}_{e \in E_b^\exists \setminus \{e^*\}}$ and q were chosen. Therefore, for each inequation in (20) and (21), the set of αs which do not satisfy said inequation is a hyperplane in the space of the vectors α. As a consequence, the set of vectors α which satisfy all of these inequations is nonempty (and in fact infinite). We now pick any of these solutions to construct our matrix X and we are done.

- *Case 2:* $e^* \in E_b^\forall$. For any $i_{e^*} = q \neq q$, the condition

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall}, \qquad \left[\sum_{x_b} B_{\{i_e\}_{e \in E_B}}(x_b)\right]_{i_{e^*}=q'} \neq 0$$

remains true after applying $(X^T)^{-1}$ as well, since the values of these sums do not change.

We just need to pick the α vector in such a way that

$$\forall \{1 \leq i_e \leq r_e\}_{e \in E_b^\forall \setminus \{e^*\}}, \qquad \left[B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q} - \sum_{q'=1}^{r_{e^*}} \alpha_{i_e} \left[B_{\{i_e\}_{e \in E_b}}(x_b)\right]_{i_{e^*}=q'} \neq 0$$

holds. However, these equations are the same as (21), so by applying the same reasoning as in Case 1, we can again find a matrix X that satisfies our requirements. ■

Corollary 33.1. *Let $A \in \mathbb{R}^{|\mathcal{X}_a|} \otimes (\bigotimes_{e \in E_a} \mathbb{R}^{r_e})$ and $B \in \mathbb{R}^{|\mathcal{X}_b|} \otimes \left(\bigotimes_{e \in E_b} \mathbb{R}^{r_e}\right)$ be two tensors with exactly one common mode $e^* = E_a \cap E_b$ and a $\exists\forall$-partition pair $(E_a^\exists, E_a^\forall; E_b^\exists, E_b^\forall)$. Then there exists invertible $r_{e^*} \times r_{e^*}$ matrix X with real entries such that*

$$(E_a^\exists \setminus \{e^*\}, E_a^\forall \cup \{e^*\}; E_b^\exists \setminus \{e^*\}, E_b^\forall \cup \{e^*\})$$

is a $\exists\forall$-partition pair for tensors $X \cdot_{e^*} A$ and $(X^T)^{-1} \cdot_{e^*} B$.

Proof. Apply Lemma 33 to find a matrix Y such that

$$(E_a^\exists \setminus \{e^*\}, E_a^\forall \cup \{e^*\}; E_b^\exists, E_b^\forall)$$

is a $\exists\forall$-partition pair for tensors $Y \cdot_{e^*} A$ and $(Y^T)^{-1} \cdot_{e^*} B$. Now apply Lemma 33 again for $(A; B) := \left((Y^T)^{-1} \cdot_{e^*} B; Y \cdot_{e^*} A\right)$ to find Z such that

$$(E_a^\exists \setminus \{e^*\}, E_a^\forall \cup \{e^*\}; E_b^\exists \setminus \{e^*\}, E_b^\forall \cup \{e^*\})$$

is a $\exists\forall$-partition pair for tensors $Z \cdot_{e^*} (Y \cdot_{e^*} A)$ and $(Z^T)^{-1} \cdot_{e^*} \left((Y^T)^{-1} \cdot_{e^*} B\right)$. But $Z \cdot_{e^*} (Y \cdot_{e^*} A) = (ZY) \cdot_{e^*} A$ and $(Z^T)^{-1} \cdot_{e^*} \left((Y^T)^{-1} \cdot_{e^*} B\right) = ((ZY)^T)^{-1} \cdot_{e^*} B$, so the matrix $X := ZY$ satisfies our requirements. ■

We are now ready to prove the Nonzero Sums Lemma.

Lemma 34 (The Nonzero Sums Lemma)**.** *Let $\mathcal{G}$ be a tensor network with vector spaces for its nodes $\{\mathbb{R}^{|\mathcal{X}_v|}\}_{v \in V}$ and bond dimensions $\{r_e\}_{e \in E}$, and let $\mathcal{T} \in \prod_{v \in V} \mathbb{R}^{|\mathcal{X}_v|}$ be a tensor which decomposes with respect to the above construction. Assume $\sum_{x_V} \mathcal{T}(x_V) \neq 0$. Then there exists a decomposition for $\mathcal{T}$ given by tensors $\left\{T^{(v)} \in \mathbb{R}^{|\mathcal{X}_v|} \otimes \left(\bigotimes_{e \in inc(v)} \mathbb{R}^{r_e}\right)\right\}_{v \in V}$ with the property that*

$$\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in inc(v)}}(x_v) \neq 0$$

holds for any $v \in V$ and any values $\{1 \leq i_e \leq r_e\}_{e \in inc(v)}$ for its adjacent indices.

Proof. Let $\left\{T^{(v)} \in \mathbb{R}^{|\mathcal{X}_v|} \otimes \left(\bigotimes_{e \in \mathrm{inc}(v)} \mathbb{R}^{r_e}\right)\right\}_{v \in V}$ be an arbitrary decomposition of $\mathcal{T}$ with respect to $\mathcal{G}$ and the given bond dimensions. In other words, we have the following equation:

$$\mathcal{T}(x_V) = \sum_{\{1 \leq i_e \leq r_e\}_{e \in E}} \prod_{v \in V} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v).$$

Summing over x_V, we get:

$$\sum_{x_V} \mathcal{T}(x_V) = \sum_{\{1 \leq i_e \leq r_e\}_{e \in E}} \prod_{v \in V} \left(\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v) \right).$$

Since $\sum_{x_V} \mathcal{T}(x_V) \neq 0$ by hypothesis, we can conclude that for any vertex $v \in V$, there exist $\{1 \leq i_e \leq r_e\}_{e \in \mathrm{inc}(v)}$ values for its adjacent indices such that

$$\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v) \neq 0.$$

That is to say, $(\mathrm{inc}(v), \emptyset)$ is a $\exists\forall$-partition for $T^{(v)}$ for any vertex v.

Finally our proof of this lemma involves an inductive approach. Start with our tensor decomposition for $\mathcal{T}$ and with the initial $\exists\forall$-partition for every $T^{(v)}$ given by $(\mathrm{inc}(v), \emptyset)$.

Now pick any edge $e \in E$ and apply Corollary 33.1 to find a gauge transformation which makes it so that e will be in the $\forall$-component of both of its adjacent $T^{(v)}$ tensors afterwards. We apply this gauge transformation, replace our decomposition with the resulting decomposition, and modify the $\exists\forall$-partitions of the vertices adjacent to e so that e is in their $\forall$-components. We repeatedly apply this process until every edge in the graph is in the $\forall$-component of their adjacent tensors. At the end, each $T^{(v)}$ will have $\exists\forall$-partition $(\emptyset, \mathrm{inc}(v))$, and this means exactly that

$$\forall v \in V, \{1 \leq i_e \leq r_e\}_{e\in\mathrm{inc}(v)}, \qquad \sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v) \neq 0.$$

■

E.4 Proof of Theorem 14

Theorem. *For any Tree-structured Correlation Scenario $\mathcal{G}$, a distribution is in the quasi-marginal model if and only if it satisfies the observed independence constraints enforced by the graph. In other words,*

$$\mathcal{W}(\mathcal{G}) = \mathcal{M}(\mathcal{G}) = \mathcal{N}(\mathcal{G}).$$

Proof. We recommend the reader familiarises themselves with the simpler proof in the main text first. We will prove the inclusion $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{W}(\mathcal{G})$, since the rest follows from Propositions 6 and 8.

As before, consider a minimal tensor decomposition of $p \in \mathcal{M}(\mathcal{G})$ with respect to $\mathcal{G}^t$. Let r_e be the bond dimension of edge $e \in E^t$ of this decomposition. Recall that the tensor decomposition of p implies the existence of $\left\{T^{(v)} \in \mathbb{R}^{|\mathcal{X}_v|} \otimes \left(\bigotimes_{e\in\mathrm{inc}(v)} \mathbb{R}^{r_e}\right)\right\}_{v\in V}$ such that

$$p(x_V) = \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v). \tag{22}$$

We can assume by Lemma 34 in Appendix E.3 that

$$\sum_{x_V} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v) \neq 0$$

holds for all $v \in V$ and values for indices $\{i_e\}_{e\in\mathrm{inc}(v)}$.

We can therefore write

$$p(x_V) = \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} \frac{T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)} \cdot \prod_{v\in V} \left(\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)\right). \tag{23}$$

Since $\sum_{x_v} \frac{T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)} = 1$, we can create for each edge e of the graph $\mathcal{G}^t$ the corresponding latent variable l_e in $\mathcal{G}$ with size r_e and the response functions of the observed variables to be

$$p\left(x_v | \{x_{l_e} = i_e\}_{e\in\mathrm{inc}(v)}\right) := \frac{T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)}.$$

We just need to set $p(x_{l_e} = i_e)$ for each edge $e \in E^t$ and $1 \leq i_e \leq r_e$. To do so, we will prove that for each vertex $v \in V$, the quantity $\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)$ decomposes into factors depending on only one index as

$$\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v) = \prod_{e\in\mathrm{inc}(v)} \phi^{(v)}_e(i_e),$$

for some functions $\{\phi_e^{(v)}\}_{v\in V, e\in \mathrm{inc}(v)}$.

Let $a \in V$ be an arbitrary vertex of $\mathcal{G}^t$. We sum over x_a in equation (22):

$$p(x_{V\setminus\{a\}}) = \sum_{\{1\le i_e\le r_e\}_{e\in E}} \prod_{v\in V\setminus\{a\}} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v) \cdot \left(\sum_{x_a} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right). \tag{24}$$

As before, we denote the branches left after eliminating a from the graph by

$$\mathcal{B}_a = \{B \subseteq V\setminus\{a\} : B \text{ is a maximal connected component of the graph } \mathcal{G}\setminus\{a\}\}.$$

We now redistribute the entries of the order $|V|-1$ tensor $p(x_{V\setminus\{a\}})$ to obtain another tensor with fewer modes, in particular one mode for each branch of $\mathcal{B}_a$. We do this in the same way we did in the simpler proof in the main text: for each branch $B \in \mathcal{B}_a$, group the modes corresponding to vertices v_i in $B = \{v_1, \dots, v_s\}$ into a single one corresponding to the tuple $(v_1, \dots, v_s)$. In other words, we replace the modes $x_{v_1}, \dots, x_{v_s}$ of sizes $|\mathcal{X}_{v_1}|, \dots, |\mathcal{X}_{v_s}|$ by a single mode x_B of size $\prod_{i=1}^s |\mathcal{X}_{v_i}|$. If we write $\mathcal{B}_a = \{B_1, \dots, B_t\}$, then we will denote the tensor described above by $p(x_{B_1}; x_{B_2}; \dots; x_{B_t})$.

We also do some regrouping in the RHS of equation (24). For each branch $B = \{v_1, \dots, v_s\} \in \mathcal{B}_a$, we group the factors $T^{(v_i)}_{\{i_e\}_{e\in\mathrm{inc}(v_i)}}(x_{v_i})$ together. As such, denote

$$M^{(B)}_{i_{e_B}}(x_B) := \sum_{\{1\le i_e\le r_e\}_{e\in\mathrm{inc}(B)\setminus\{e_B\}}} \prod_{v_i\in B} T^{(v_i)}_{\{i_e\}_{e\in\mathrm{inc}(v_i)}}(x_{v_i}),$$

where $e_B \in E$ is the edge that links the eliminated vertex a and some vertex in the branch B. Now we can rewrite (24) as

$$p(x_{V\setminus\{a\}}) = \sum_{\{1\le i_{e_B}\le r_{e_B}\}_{B\in\mathcal{B}_a}} \left(\prod_{B\in\mathcal{B}_a} M^{(B)}_{i_{e_B}}(x_B)\right) \cdot \left(\sum_{x_a} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right). \tag{25}$$

Let us look at what the object $M^{(B)}$ is in detail for a given branch B. One can see it as a tensor with $|B|+1$ modes: one mode for each $v_i \in B$ and an additional mode corresponding to the value of the index i_{e_B}. However, similar to the proof in the main text, we choose to flatten it by grouping the modes corresponding to x_{v_i} into a single mode corresponding to the value of the tuple x_B. The resulting tensor is a $|\mathcal{X}_B| \times r_{e_B}$ matrix having rows indexed by values of the tuple x_B and columns indexed by i_{e_B}. For brevity we will keep the same notation $M^{(B)}$, but keep in mind this refers to a matrix from now on.

Now if we view the expression $\sum_{x_a} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)$ as an order $|\mathcal{B}_a|$ tensor with a mode i_{e_B} for each branch $B \in \mathcal{B}_a$, then we can rewrite (25) as:

$$p(x_{B_1}; \dots; x_{B_t}) = \left(M^{(B_1)}, M^{(B_2)}, \dots, M^{(B_t)}\right) \left(\sum_{x_a} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right), \tag{26}$$

where $\left(M^{(B_1)}, M^{(B_2)}, \dots, M^{(B_t)}\right)$ is the multilinear transformation operator defined in Appendix E.1.

Recall that $M^{(B)}$ has rows indexed by x_B and columns indexed by i_{e_B}. We now claim that $M^{(B)}$ has linearly independent columns for any $B \in \mathcal{B}_a$. Indeed, assume the contrary holds for some $B^* \in \mathcal{B}_a$. The there exists some $1 \le q \le r_{e_{B^*}}$ and real numbers $\alpha_{i_{e_{B^*}}}$ such that

$$M^{(B^*)}_q(x_{B^*}) = \sum_{i_{e_{B^*}}\neq q} \alpha_{i_{e_{B^*}}} \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*})$$

for all $x_{B^*} \in \mathcal{X}_{B^*}$.

We can replace this in equation (22) to obtain:

$$\begin{aligned}
p(x_V) &= \sum_{\{1 \le i_e \le r_e\}_{e\in E}} \prod_{v\in V} T^{(v)}_{\{i_e\}_{e\in \mathrm{inc}(v)}}(x_v) \\
&= \sum_{\{1 \le i_e \le r_e\}_{e\in E\setminus \mathrm{inc}(B^*)}} \left(\prod_{v\in V\setminus(B^*\cup\{a\})} T^{(v)}_{\{i_e\}_{e\in \mathrm{inc}(v)}}(x_v) \right) \\
&\cdot \sum_{1\le i_{e_{B^*}} \le r_{e_{B^*}}} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) \cdot \left(\sum_{\{1 \le i_e \le r_e\}_{e\in \mathrm{inc}(B^*)\setminus\{e_{B^*}\}}} \prod_{v_i\in B^*} T^{(v_i)}_{\{i_e\}_{e\in\mathrm{inc}(v_i)}}(x_{v_i}) \right) \\
&= \sum_{\{1 \le i_e \le r_e\}_{e\in E\setminus \mathrm{inc}(B^*)}} \left(\prod_{v\in V\setminus(B^*\cup\{a\})} T^{(v)}_{\{i_e\}_{e\in \mathrm{inc}(v)}}(x_v) \right) \\
&\cdot \sum_{1\le i_{e_{B^*}} \le r_{e_{B^*}}} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*})
\end{aligned}$$

However, if we look at the expression on the last line, one can see that for all values of $\{i_e\}_{e\in E\setminus\mathrm{inc}(B^*)}$:

$$\begin{aligned}
&\sum_{1\le i_{e_{B^*}} \le r_{e_{B^*}}} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*}) \\
&= \sum_{i_{e_{B^*}} \ne q} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*}) + \left[T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right]_{i_{e_{B^*}}=q} \cdot M^{(B^*)}_{q}(x_{B^*}) \\
&= \sum_{i_{e_{B^*}} \ne q} T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*}) + \left[T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right]_{i_{e_{B^*}}=q} \cdot \sum_{i_{e_{B^*}} \ne q} \alpha_{i_{e_{B^*}}} \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*}) \\
&= \sum_{i_{e_{B^*}} \ne q} \left(T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) + \alpha_{i_{e_{B^*}}} \cdot \left[T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right]_{i_{e_{B^*}}=q} \right) \cdot M^{(B^*)}_{i_{e_{B^*}}}(x_{B^*}).
\end{aligned}$$

Replacing this in the expression for $p(x_V)$ we obtain:

$$\begin{aligned}
p(x_V) = \sum_{\{1 \le i_e \le r_e\}_{e\in E\setminus\{e_{B^*}\}}} \sum_{\substack{1\le i_{e_{B^*}} \le r_{e_{B^*}} \\ i_{e_{B^*}} \ne q}} &\left(\prod_{v\in V\setminus\{a\}} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v) \right) \\
&\cdot \left(T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) + \alpha_{i_{e_{B^*}}} \cdot \left[T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right]_{i_{e_{B^*}}=q} \right).
\end{aligned}$$

Therefore, we obtain a strictly smaller decomposition for $\mathcal{T}$: by removing the value $i_{e_{B^*}} = q$ and replacing the tensor $T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)$ by the new tensor $T'^{(a)} \in \mathbb{R}^{|\mathcal{X}_a|} \otimes \mathbb{R}^{r_{e_{B^*}}-1} \otimes \left(\bigotimes_{e\in\mathrm{inc}(a)\setminus\{e_{B^*}\}} \mathbb{R}^{r_e}\right)$ defined as

$$T'^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) := T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a) + \alpha_{i_{e_{B^*}}} \cdot \left[T^{(a)}_{\{i_e\}_{e\in\mathrm{inc}(a)}}(x_a)\right]_{i_{e_{B^*}}=q},$$

we obtain a tensor network decomposition for $p(x_V)$ which has bond dimension for e_{B^*} equal to $r_{e_{B^*}} - 1$, contradicting the minimality of the bond dimensions we have picked.

Therefore, indeed the columns of $M^{(B^*)}$ must be linearly independent. As a result, for any $B \in \mathcal{B}_a$, $M^{(B)}$ must have a left inverse $N^{(B)}$ such that

$$N^{(B)} M^{(B)} = \mathcal{I}_{r_{e_B}}.$$

Recall equation (26):

$$p(x_{B_1}; \ldots; x_{B_t}) = \left(M^{(B_1)}, M^{(B_2)}, \ldots, M^{(B_t)}\right) \left(\sum_{x_a} T^{(a)}_{\{i_e\}_{e \in \mathrm{inc}(a)}}(x_a)\right).$$

Applying $\left(N^{(B_1)}, N^{(B_2)}, \ldots, N^{(B_t)}\right)$ to the left of each side of the equation, we get:

$$\left(N^{(B_1)}, N^{(B_2)}, \ldots, N^{(B_t)}\right) \cdot p(x_{B_1}; \ldots; x_{B_t}) = \sum_{x_a} T^{(a)}_{\{i_e\}_{e \in \mathrm{inc}(a)}}(x_a).$$

However, by $p(x_V) \in \mathcal{M}(\mathcal{G})$ and Proposition 11 we get

$$p(x_{V \setminus \{a\}}) = \prod_{B \in \mathcal{B}_a} p(x_B),$$

which we can rewrite in terms of tensor products as:

$$p(x_{B_1}; \ldots; x_{B_t}) = \mathbf{p}(\mathbf{x_{B_1}}) \otimes \cdots \otimes \mathbf{p}(\mathbf{x_{B_t}}),$$

where we write the factors in the RHS in bold to emphasise they are vectors.

Therefore,

$$\begin{aligned}\sum_{x_a} T^{(a)}_{\{i_e\}_{e \in \mathrm{inc}(a)}}(x_a) &= \left(N^{(B_1)}, N^{(B_2)}, \ldots, N^{(B_t)}\right) \cdot (\mathbf{p}(\mathbf{x_{B_1}}) \otimes \cdots \otimes \mathbf{p}(\mathbf{x_{B_t}})) \\ &= \left(N^{(B_1)} \mathbf{p}(\mathbf{x_{B_1}})\right) \otimes \cdots \otimes \left(N^{(B_t)} \mathbf{p}(\mathbf{x_{B_t}})\right).\end{aligned}$$

Note that $N^{(B)}$ is a $r_{e_B} \times |\mathcal{X}_B|$ matrix, meaning $N^{(B)}\mathbf{p}(\mathbf{x_B})$ is a vector of size r_{e_B}, which is what we would expect. Therefore, by setting

$$\phi^{(a)}_{e_B} := N^{(B)}_{i_{e_B}} \mathbf{p}(\mathbf{x_B}),$$

we obtain that indeed

$$\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v) = \prod_{e \in \mathrm{inc}(v)} \phi^{(v)}_e(i_e).$$

To recap, by equation (23), we have:

$$p(x_V) = \sum_{\{1 \le i_e \le r_e\}_{e \in E}} \prod_{v \in V} \frac{T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v)} \cdot \prod_{v \in V} \left(\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v)\right)$$

If we let

$$p\left(x_v | \{x_{l_e} = i_e\}_{e \in \mathrm{inc}(v)}\right) := \frac{T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e\}_{e \in \mathrm{inc}(v)}}(x_v)}$$

as before and define

$$\psi_e(i_e) := \phi^{(v_1)}_e(i_e) \cdot \phi^{(v_2)}_e(i_e),$$

for any edge $e = (v_1, v_2)$, then we obtain:

$$\begin{aligned}
p(x_V) &= \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} \frac{T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)} \cdot \prod_{v\in V}\left(\sum_{x_v} T^{(v)}_{\{i_e\}_{e\in\mathrm{inc}(v)}}(x_v)\right) \\
&= \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} p\left(x_v|\{x_{l_e}=i_e\}_{e\in\mathrm{inc}(v)}\right) \cdot \prod_{v\in V}\prod_{e\in\mathrm{inc}(v)} \phi^{(v)}_e(i_e) \\
&= \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} p\left(x_v|\{x_{l_e}=i_e\}_{e\in\mathrm{inc}(v)}\right) \cdot \prod_{e=(v_1,v_2)\in E} \phi^{(v_1)}_e(i_e)\cdot\phi^{(v_2)}_e(i_e) \\
&= \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} p\left(x_v|\{x_{l_e}=i_e\}_{e\in\mathrm{inc}(v)}\right) \cdot \prod_{e\in E} \psi_e(i_e).
\end{aligned}$$

By summing this equation over x_V we obtain

$$1 = \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{e\in E} \psi_e(i_e) = \prod_{e\in E}\sum_{1\leq i_e\leq r_e} \psi_e(i_e).$$

Therefore,

$$p(x_V) = \sum_{\{1\leq i_e\leq r_e\}_{e\in E}} \prod_{v\in V} p\left(x_v|\{x_{l_e}=i_e\}_{e\in\mathrm{inc}(v)}\right) \cdot \prod_{e\in E} \frac{\psi_e(i_e)}{\sum_{1\leq i_e\leq r_e}\psi_e(i_e)}. \tag{27}$$

Finally, we go back to the original correlation scenario $\mathcal{G}$. Define the latent variable l_e corresponding to edge e in the graph $\mathcal{G}^t$ to have domain $\{1,\ldots,r_e\}$ and quasiprobability distribution

$$p(x_{l_e}=i_e) := \frac{\psi_e(i_e)}{\sum_{1\leq i_e\leq r_e}\psi_e(i_e)}.$$

Define the response function of node $v\in V$ as

$$p(x_v|x_{\mathrm{pa}_{\mathcal{G}}(v)}) := \frac{T^{(v)}_{\{i_e=x_{l_e}\}_{e\in\mathrm{inc}_{\mathcal{G}^t}(v)}}(x_v)}{\sum_{x_v} T^{(v)}_{\{i_e=x_{l_e}\}_{e\in\mathrm{inc}_{\mathcal{G}^t}(v)}}(x_v)},$$

where l_e is the latent node corresponding to edge e in $\mathcal{G}^t$. Then for all nodes we have $\sum_{x_{l_e}} p(x_{l_e}) = 1$ and $\sum_{x_v} p(x_v|x_{\mathrm{pa}_{\mathcal{G}(v)}}) = 1$ as expected, and equation (27) tells us that

$$p(x_V) = \sum_{x_L}\prod_{v\in V} p(x_v|x_{\mathrm{pa}_{\mathcal{G}}(v)})\prod_{l\in L} p(x_l).$$

Therefore, $p(x_V)$ is in the quasi set $\mathcal{W}(\mathcal{G})$, so $\mathcal{M}(\mathcal{G})\subseteq\mathcal{W}(\mathcal{G})$, and so

$$\mathcal{W}(\mathcal{G}) = \mathcal{M}(\mathcal{G}) = \mathcal{N}(\mathcal{G}).$$

■

F Conjecture 1 in the Triangle Scenario

We prove that Conjecture 1 is true for the Triangle Scenario in Fig. 5. We start under the assumption that the observed variables are binary, and we will give two constructions of the perfect correlation $p = \frac{1}{2}([000]+[111])$, where $[abc]$ represents the deterministic distribution with $A=a$, $B=b$, and $C=c$. The first one uses quasiprobabilistic latent

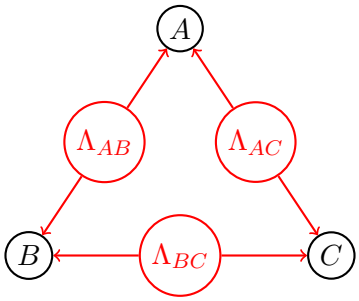


Figure 5: The Triangle Scenario.

variables, and the second one is a symmetric construction using quasiprobabilistic response distributions. In the language of this paper, they are a quasilatents construction and a quasiresponses one.

The quasilatents construction has A, B, and C deterministic given their latent parents. We will write the distributions of the latents as vectors, i.e. the distribution $p_{\Lambda_{BC}}$ below has $p_{\Lambda_{BC}}(0) = p_{\Lambda_{BC}}(1) = \frac{1}{2}$:

- $\Omega_{\Lambda_{BC}} = \{0, 1\}$, $p_{\lambda_{BC}} = (\frac{1}{2}, \frac{1}{2})^T$
- $\Omega_{\Lambda_{AC}} = \{1, 2, 3, 4, 5, 6, 7\}$, $p_{\Lambda_{AC}} = (0.1, 0.9, 0.1, 0.3, -0.4, -0.1, 0.1)^T$
- $\Omega_{\Lambda_{AB}} = \{1, 2, 3, 4, 5\}$, $p_{\Lambda_{AB}} = (1, 1, -1, 1, -1)^T$
- $A = 0 \iff (\Lambda_{AC}, \Lambda_{AB}) \in \{(3, 1), (6, 1), (1, 2), (4, 2), (7, 2), (1, 5), (3, 4), (6, 4), (7, 4)\}$
- $B = 0 \iff (\Lambda_{BC} = 0 \wedge \Lambda_{AB} \in \{1, 2, 3\}) \vee (\Lambda_A = 1 \wedge \Lambda_C \in \{2, 3, 4, 5\})$
- $C = 0 \iff (\Lambda_{BC} = 0 \wedge \Lambda_{AC} \in \{1, 2, 3, 4, 5\}) \vee (\Lambda_A = 1 \wedge \Lambda_B \in \{3, 4, 5, 6, 7\})$.

The quasiresponses construction has all latents behaving as unbiased coinflips and the following response for each observed variable:

$$p(v|\lambda_1, \lambda_2) = \begin{cases} \frac{3}{2} & \text{if } \lambda_1 = \lambda_2 = v; \\ -\frac{1}{2} & \text{if } \lambda_1 = \lambda_2 \neq v; \\ \frac{1}{2} & \text{if } \lambda_1 \neq \lambda_2, \end{cases}$$

where v is a placeholder for a, b, c, with parents λ_1 and λ_2.

Now that the perfect correlation is in $\mathcal{W}$, we can extend the latent variables in the original scenario with a copy of this construction (more precisely, replace each latent Λ by (Λ, Λ'), where Λ' is a copy of its corresponding latent variable in the perfect correlation construction). This results in the observed variables also sharing an unbiased coinflip U (see Fig. 6).

Now simply note there is no reason to stop at one copy of U. Since the latents are allowed to be continuous, we can instead make them generate an infinite sequence of i.i.d. copies of U. Since any finite random variable can be simulated via an infinite number of

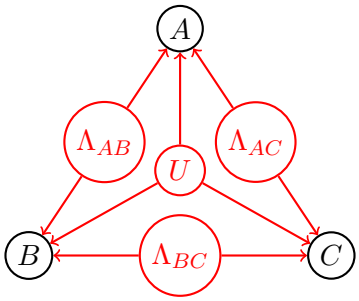


Figure 6: The Triangle Scenario after extending the latents with a copy of the perfect correlation construction.

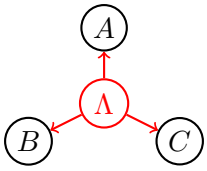


Figure 7: The Common Ancestor Scenario.

coinflips, we conclude that using negative probabilities, the Triangle Scenario can simulate the Common Ancestor Scenario in Fig. 7. It is obvious any distribution on three variables is in the Common Ancestor Scenario, so we are done.