A Deferred Proofs

This section contains proofs which were deferred to it from the body of the article.

A.1 Proof of Theorem 1

We will prove the theorem using induction based on depth of the circuit rooted at a node $v \in V$, i.e. the maximal length of a path connecting a leaf to v. Given that $\Psi_v(\cdot)$ is a non-negative function, it is sufficient to show it is normalized, i.e. that for any fixed values of the variables in $\text{cond}(v)$, denoted by $\mathbf{b} \in \{0,1,*\}^N$ where $b_i = *$ if $i \notin \text{cond}(v)$, the following holds:

$$\sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_v(\mathbf{x}) = 1$$

For the base case of the induction of depth-1 SPQNs, which means $v \in I$ must be an indicator node, i.e. $v = \mathbb{I}[x_i = a]$ for some $i \in [N]$ and $a \in \{0,1\}$, then $\text{eff}(v) = \{i\}$ and $\text{cond}(v) = \emptyset$, and so summing over all possible values of x_i is equal to $\mathbb{I}[x_i = a](0) + \mathbb{I}[x_i = a](1) = 1$ meeting the normalization condition. Let us now assume that our induction assumption holds for all circuit of depth $d \geq 1$, and prove it also holds for $d+1$. Since any SPQNs of depth $d+1$ is greater than 1, then the root node must either be a sum, product or quotient node, and not an indicator node. Additionally, for the root $v \in V$ of such a circuit, because each of its child nodes can be viewed as a depth-d sub-circuit, then according to the induction assumption it represents a normalized probability function over the variables in $\text{eff}(v)$ for any fixed values of the variables in $\text{cond}(v)$. Next we will use this property to show that for any possible node type, v represent a normalized probability function.

if $v \in Q$ is a quotient node, then according to conditional soundness then $Psi_{\text{de}(v)}(\cdot)$ is a strictly positive function and hence the output of the quotient operation is well defined. Additionally, the conditional soundness also entails that $\Psi_{\text{de}(v)}(\cdot)$ is a marginal conditional distribution of $\Psi_{\text{nu}(v)}(\cdot)$, and specifically, that summing $\Psi_{\text{nu}(v)}(\cdot)$ over all the possible values of the variables in $\text{eff}(v)$ equals to $\Psi_{\text{de}(v)}(\cdot)$, and thus:

$$\begin{aligned}
\sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_v(\mathbf{x}) &= \sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \frac{\Psi_{\text{nu}(v)}(\mathbf{x})}{\Psi_{\text{de}(v)}(\mathbf{x})} \\
&= \frac{1}{\Psi_{\text{de}(v)}(\mathbf{b})} \cdot \sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_{\text{nu}(v)}(\mathbf{x}) \\
&= \frac{1}{\Psi_{\text{de}(v)}(\mathbf{b})} \cdot \Psi_{\text{de}(v)}(\mathbf{b}) = 1
\end{aligned}$$

where we have used the fact that changing the values of the coordinates of $\mathbf{x}$ for $i \in \text{eff}(v)$ do not affect the value of $\Psi_{\text{de}(v)}(\mathbf{x})$ as $\text{eff}(\text{de}(v)) \cap \text{eff}(v) = \emptyset$, in combination with the relationship between the sum over $\Psi_{\text{nu}(v)}(\cdot)$ and $\Psi_{\text{de}(v)}(\cdot)$.

If $v \in S$ is a sum node, then according to conditional completeness the effective scopes of its child nodes are identical to its own effective scope. This also entails that $\text{cond}(c) \subset \text{cond}(v)$ for any $c \in \text{ch}(v)$ because that $\text{cond}(c)$ is the complement of $\text{eff}(c)$ with respect to $\text{sc}(c)$. We can also assume without losing our generality that $\text{cond}(c) = \text{cond}(v)$, as variables outside of $\text{cond}(c)$ do not affect the output of $\Psi_c(\cdot)$ regardless of their value. Given the last assumption and the induction assumption, all the children of v represent conditional distributions over the same set of variables, and because the weights of v are normalized to sum to one, then:

$$\begin{aligned}
\sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_v(\mathbf{x}) &= \sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \sum_{c\in\text{ch}(v)} w_c \cdot \Psi_c(\mathbf{x}) \\
&= \sum_{c\in\text{ch}(v)} w_c \cdot \overbrace{\sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_c(\mathbf{x})}^{=1} \\
&= \sum_{c\in\text{ch}(v)} w_c = 1
\end{aligned}$$

where the inner sum equals to 1 due to the normalization of the child nodes.

Finally, we will consider the case that $v \in P$ is a product node. Recall that conditional decomposability means that the effective scopes of each child of v are disjoint sets, and that the directed dependency graph formed by the children of v is an acyclic graph. To prove this case, we will use a secondary induction over the number of children of v. In the base case of v having just a single child $\text{ch}(v) = \{c\}$, it holds that $\Psi_v(\cdot) = \Psi_c(\cdot)$, and thus it is a normalized probability function due to the primary induction assumption. Let us assume that out secondary induction assumption holds for v with t children, and prove it also holds for $t+1$ children. Let $\bar{c} \in \text{ch}(v)$ be child of v that is a sink node in the induced dependency graph, i.e. that none of the variables in its effective scope are part of the conditional scope of another child, hence the following holds:

$$\begin{aligned}
\sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_v(\mathbf{x}) &= \sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v), x_i\in\{0,1\}}} \Psi_{\bar{c}}(\mathbf{x}) \left(\prod_{\substack{c\in\text{ch}(v) \\ c\neq\bar{c}}} \Psi_c(\mathbf{x}) \right) \\
&\overset{(1)}{=} \sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v)\setminus\text{eff}(\bar{c}), x_i\in\{0,1\} \\ \forall i\in\text{eff}(\bar{c}), x_i = *}} \left(\prod_{\substack{c\in\text{ch}(v) \\ c\neq\bar{c}}} \Psi_c(\mathbf{x}) \right) \overbrace{\sum_{\substack{\mathbf{z}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(\bar{c}), z_i = x_i \\ \forall i\in\text{eff}(\bar{c}), z_i\in\{0,1\}}} \Psi_{\bar{c}}(\mathbf{z})}^{=1} \\
&\overset{(2)}{=} \sum_{\substack{\mathbf{x}\in\{0,1,*\}^N \\ \forall i\notin \text{eff}(v), x_i = b_i \\ \forall i\in\text{eff}(v)\setminus\text{eff}(\bar{c}), x_i\in\{0,1\} \\ \forall i\in\text{eff}(\bar{c}), x_i = *}} \left(\prod_{\substack{c\in\text{ch}(v) \\ c\neq\bar{c}}} \Psi_c(\mathbf{x}) \right) = 1
\end{aligned}$$

where the equality marked by (1) is due to decomposing the sum into two nested sums, one where we iterate over the different values of $\mathbf{x}$ just over the coordinates matching the variables in the effective scope of v that are not in the effective scope of $\bar{c}$ and the second nested sum over the remaining coordinates of the effective sum. Because the inner sum affects only the variables in $\Psi_{\bar{c}}(\cdot)$ we can extract all over nodes out of it, this is because of our assumption that $\bar{c}$ is a sink node and hence $\text{eff}(\bar{c})$ is not part of the scopes of the other children, in addition to the fact that

the effective scopes are disjoint sets. The equality marked by (2) is because $\Psi_{\bar{c}})(\cdot)$ is a normalized probability function according to our primary induction assumption, hence the inner sum equals to one. The final equality is due to our secondary induction assumption, as there are only t child nodes left and thus that sum also equals to one. This concludes the proof for both the secondary and the primary induction assumption.

A.2 Proof of Proposition 1

By the second and third conditions in def. 7, all product and sum nodes in an SPQN composed of valid CMOs must be conditionally D&C, and thus, according to theorem. 1, we only need to prove that it is conditionally sound for it to be tractable. We employ induction on the depth of the SPQN rooted at $v \in V$ with the assumption that all SPQNs up to depth d that are composed of valid CMO nodes are strongly conditionally sound, hence also valid distributions, strictly positive functions, and that for all $\mathbf{z} \in \{0,1,*\}^N$ such that $z_i = *$ if $i \in \text{eff}(v)$ it holds that $\Psi_v(\mathbf{z}) = 1$.

We begin with the base case of a CMO node connected to the two indicator leaf nodes $\mathbb{I}[x_i = 0]$ and $\mathbb{I}[x_i = 1]$ for some $i \in [N]$, which according to def. 7 is the only valid CMO node that is connected to the leaves. Under this case the output of the CMO node is equal to a single sum node computing $w_1\mathbb{I}[x_i = 0] + w_2\mathbb{I}[x_i = 1]$, where $\mathbf{w} \in \mathbb{R}^2$ is strictly positive. Since the output is simply a single sum node over indicators of the same variable, it immediately follows that it is conditionally decomposable, complete and sound. Additionally, since $\mathbf{w}$ is strictly positive, then the output of the node is also strictly positive for any value of x_i. Finally, when setting $x_i = *$ the output equals to $w_1 \cdot 1 + w_2 \cdot 1 = 1$.

Let v denote the root CMO node of an SPQN of depth $d+1$. Without losing our generality, we can assume that $\alpha = \beta = 1$ (see def. 6) with children $a_1, \ldots, a_\gamma, b_1, \ldots, b_\gamma \in V$, otherwise we can substitute each of the products, $\prod_{j=1}^{\alpha} A_{ij}$ and $\prod_{j=1}^{\beta} B_{ij}$, with an auxiliary valid CMO node that computes just the product, i.e. with no A-type children, which is trivially conditionally sound. Since we assume all the children of v represents strictly positive functions, and since the output of v is composed of products and weighted sums with positive weights, then the output of v is also strictly positive. According to def. 7, the internal sum and product nodes of v are conditionally D&C, and thus their respective rooted sub-SPQNs are tractable by the induction assumption, which means they represent valid distributions. Additionally, def. 7 also entails that the effective scopes of each of $b_1, \ldots, b_\gamma$ are equal to $\text{eff}(v)$, and do not appear in the conditional scopes of $a_1, \ldots, a_\gamma$. Now, for any $\mathbf{a} \in \{0,1,*\}^N$ the following holds:

$$\sum_{\substack{\mathbf{z}\in\{0,1,*\}^N \\ \forall i\notin\text{eff}(v), z_i=a_i \\ \forall i\in\text{eff}(v), z_i\in\{0,1\}}} \Psi_{\text{nu}(v)}(\mathbf{z}) = \sum_{\substack{\mathbf{z}\in\{0,1,*\}^N \\ \forall i\notin\text{eff}(v), z_i=a_i \\ \forall i\in\text{eff}(v), z_i\in\{0,1\}}} \sum_{i=1}^{\gamma} \Psi_{a_i}(\mathbf{z})\Psi_{b_i}(\mathbf{z})$$

$$\overset{(1)}{=} \sum_{i=1}^{\gamma} \Psi_{a_i}(\mathbf{a}) \overbrace{\sum_{\substack{\mathbf{z}\in\{0,1,*\}^N \\ \forall i\notin\text{eff}(v), z_i=a_i \\ \forall i\in\text{eff}(v), z_i\in\{0,1\}}} \Psi_{b_i}(\mathbf{z})}^{=1} \overset{(2)}{=} \Psi_{\text{de}(v)}(\mathbf{a})$$

where the equality marked by (1) is because the nodes of a_i are not affected by the changing coordinates specified by $\text{eff}(v)$, while the equality marked by (2) follows from our induction assumption that the children $b_1, \ldots, b_\gamma$ already represent normalized probability functions, and thus summing over them equals to one.This proves that the denominator is a marginal of the numerator, which prove that the SPQN rooted at v is conditionally sound. To prove that it is also strongly conditionally sound, we simply notice that for any $\mathbf{z} \in \{0,1,*\}$ such that $z_i = *$ it holds that $\Psi_{b_i}(\mathbf{z}) = 1$ based on our induction assumption as $\text{eff}(b_i) = \text{eff}(v)$, and thus:

$$\Psi_{\text{nu}(v)}(\mathbf{z}) = \sum_{i=1}^{\gamma} \overbrace{\Psi_{a_i}(\mathbf{z})}^{=1} \Psi_{b_i}(\mathbf{z}) = \Psi_{\text{de}(v)}(\mathbf{z})$$

which proves that the SPQN rooted at v is strongly conditionally sound. Additionally from the conditionally sound property we have just proven, it thus follow that

$$\Psi_v(\mathbf{z}) = \frac{\Psi_{\text{nu}(v)}(\mathbf{z})}{\Psi_{\text{de}(v)}(\mathbf{z})} = \frac{\Psi_{\text{de}(v)}(\mathbf{z})}{\Psi_{\text{de}(v)}(\mathbf{z})} = 1$$

proving that all of our induction assumptions hold and completing our proof of the proposition.

A.3 Proof of Theorem 2

We heavily base our proof on Martens and Medabalimi (2014), who have proven a very similar claim on a slightly different distribution on complete graphs, namely, that SPNs cannot approximate the uniform distribution on the spanning trees of a complete graph. Next, we go through the steps of their proof, citing the relevant lemmas, and highlighting the places where our proof diverges.

We begin by citing the following decomposition lemma, paraphrased to match the notations and definition of sec. 2:

Lemma 1 (paraphrase of theorem 39 of Martens and Medabalimi (2014))**.** *Suppose $\{\Psi_j(\mathbf{x})\}_{j=1}^{\infty}$ are the respective outputs of a sequence of D&C SPNs of size at most s over N binary variables, which converges point-wise (considered as functions of $\mathbf{x}$) to some function γ of $\mathbf{x}$. Then we have that γ can be written as:*

$$\gamma = \sum_{i=1}^{k} g_i h_i \tag{2}$$

where $k \leq s^2$ and for all $i \in [k]$ it holds that g_i and h_i are real-valued non-negative functions of $\mathbf{y}_i$ and $\mathbf{z}_i$, respectively, where $\mathbf{y}_i$ and $\mathbf{z}_i$ are sub-sets / tuples of the variables in $\mathbf{x}$ satisfying that $\frac{N}{3} \leq |\mathbf{y}_i|, |\mathbf{z}_i| \leq \frac{2N}{3}$, $\mathbf{y}_i \cap \mathbf{z}_i = \emptyset$, and $\mathbf{y}_i \cup \mathbf{z}_i = \mathbf{x}$.

According to lemma 1, it is sufficient to show that if a function in the form of eq. 2 is equal to a strictly positive distribution of triangle-free graphs of M vertices, denoted by $d(\mathbf{E})$, where $N = \binom{M}{2}$ is the number of variables representing the edges of the graph, then $k = 2^{\Omega(N)}$, because the k is a lower bound on the size of any SPN approximating $d(\mathbf{E})$.

Because the functions that comprise γ are non-negative, then $\gamma = 0$ if and only if for all i it holds that $g_i h_i = 0$. Thus, if $\gamma(\mathbf{E}) = d(\mathbf{E}) > 0$, i.e. $\mathbf{E}$ represents a triangle-free graph, then either $g_i = 0$ or $h_i = 0$ on $\mathbf{E}$. We will prove that $k = 2^{\Omega(N)}$ by showing that each term $g_i h_i$ can be non-zero on at most a small fraction of the triangle-free graphs, and more specifically, that it can be non-zero only on a

small fraction of spanning trees, which are only a sub-set of all triangle-free graphs.

Let g and h be functions as above, such that $\frac{N}{3} \leq |\mathbf{y}|, |\mathbf{z}| \leq \frac{2N}{3}$, $\mathbf{y} \cap \mathbf{z} = \emptyset$, and $\mathbf{y}_i \cup \mathbf{z}_i = \mathbf{E}$, and that $d(\mathbf{E}) = 0$ implies $g(\mathbf{y}) = 0$ or $h(\mathbf{z}) = 0$. Examining the possible triangles of $\mathbf{E}$, we single out all the triangles such that some of the edges are part of $\mathbf{y}$ and some of $\mathbf{z}$. Notice that for such triangles the function $g \cdot h$ must employ a conservative strategy, as each function on its own only see a part of the possible edges of the triangle and hence cannot decide whether all edges are in the graph or not. Martens and Medabalimi (2014) call such triplet of edges *constraint triangles*, and prove the following claims:

Claim 1 (Paraphrase of proposition 42 of Martens and Medabalimi (2014))**.** *Let $E_{i_1 i_2}$, $E_{i_2 i_3}$, and $E_{i_1 i_3}$ be three different edges that form a constraint triangle with respect to g and h as above, for which if all edges are part of the graph then $g \cdot h = 0$. Additionally, suppose that both $E_{i_1 i_2}$ and $E_{i_2 i_3}$ are in the same set of variables with respect to the partition $\mathbf{y} \cup \mathbf{z}$. Then the following properties hold:*

- *$g(\mathbf{y}) \cdot h(\mathbf{z}) = 0$ for all values of $\mathbf{E}$ such that $E_{i_1 i_2} = 1$ and $E_{i_2 i_3} = 1$, i.e. are part of the graph $\mathbf{E}$ represents.*
- *$g(\mathbf{y}) \cdot h(\mathbf{z}) = 0$ for all values of $\mathbf{E}$ such that $E_{i_1 i_3} = 1$, i.e. is part of the graph $\mathbf{E}$ represents.*

Claim 2 (Paraphrase of lemma 43 of Martens and Medabalimi (2014))**.** *Given any partition of the edges of $\mathbf{E}$ into disjoint sets $\mathbf{y} \cup \mathbf{z}$, such that $\frac{N}{3} \leq |\mathbf{y}|, |\mathbf{z}| \leq \frac{2N}{3}$, then the total number of constraint triangles is lower bounded by $\frac{M^3}{60}$.*

Claim 1 means that if $g(\mathbf{y}) \cdot h(\mathbf{z}) > 0$ then either $E_{i_1 i_2}$ and $E_{i_2 i_3}$ are not part of the graph, or $E_{i_1 i_3}$ is not part of it, and thus each constraint limits what graphs it can be non-zero on. Claim 2 finds a lower bound on the number of such constraints, which brings us to the following claim by Martens and Medabalimi (2014), which finds an upper bound on percentage of spanning trees that obey any given C set of distinct constraints:

Claim 3 (Paraphrase of lemma 44 of Martens and Medabalimi (2014))**.** *Suppose we are given C distinct constraints which are each one of the two forms discussed above. Then, of all the spanning trees of the complete graph on M vertices, a proportion of at most:*

$$\left(1 - \frac{C}{M^3}\right)^{C/6M^2}$$

of them obey all of the constraints.

Given that we have $C > \frac{M^3}{60}$, then it holds that $g(\mathbf{y}) \cdot h(\mathbf{z}) > 0$ on at most $\frac{1}{2^{M/15120}}$ of all the possible spanning trees.

To conclude, $\gamma(\mathbf{E})$ can be non-zero on at most $\frac{k}{2^{M/15120}}$ fraction of all spanning trees, and since $d(\mathbf{E})$ should be positive for any $\mathbf{E}$ that represents a triangle free graph, such as any spanning tree, then if $\gamma(\cdot) = d(\cdot)$ it must be that $\frac{k}{2^{M/15120}} \geq 1$, which means $k \geq 2^{M/15120}$, or in other words, $s = O(2^{\Omega(M)})$.

A.4 Proof of Theorem 3

We start by examining all triangles for which the edge $E_{i_2 i_3}$ is the largest edge (according to lexical order). For every $1 < i_2 < i_3 \leq M$, we define the following variables:

$$\varphi^{(1)}_{i_2 i_3} \equiv \prod_{i_1=1}^{i_2-1} \frac{\mathbb{I}\left[E_{i_1 i_2}{=}1\right]\mathbb{I}\left[E_{i_1 i_3}{=}0\right] + \mathbb{I}\left[E_{i_1 i_2}{=}0\right]\mathbb{I}\left[E_{i_1 i_3}{=}1\right] + \mathbb{I}\left[E_{i_1 i_2}{=}0\right]\mathbb{I}\left[E_{i_1 i_3}{=}0\right]}{3}$$

$$\varphi^{(2)}_{i_2 i_3} \equiv \sum_{i_1=1}^{i_2-1} \frac{1}{i_2 - 1}\mathbb{I}\left[E_{i_1 i_2}{=}1\right]\mathbb{I}\left[E_{i_1 i_3}{=}1\right] \cdot \prod_{i' \neq i_1}^{i_2-1} \frac{\mathbb{I}\left[E_{i' i_2}{=}0\right] + \mathbb{I}\left[E_{i' i_2}{=}1\right]}{2} \cdot \frac{\mathbb{I}\left[E_{i' i_3}{=}0\right] + \mathbb{I}\left[E_{i' i_3}{=}1\right]}{2}$$

$$\Phi_{i_2 i_3} \equiv \frac{\frac{1}{2}\varphi^{(1)}_{i_2 i_3} \frac{\mathbb{I}\left[E_{i_2 i_3}=0\right] + \mathbb{I}\left[E_{i_2 i_3}=1\right]}{2} + \frac{1}{2}\varphi^{(2)}_{i_2 i_3}\mathbb{I}\left[E_{i_2 i_3} = 0\right]}{\frac{1}{2}\varphi^{(1)}_{i_2 i_3} + \frac{1}{2}\varphi^{(1)}_{i_2 i_3}}$$

Where $\varphi^{(1)}_{i_2 i_3}$ is a normalized probability over the edges $E_{1,i_2}, \ldots, E_{i_2-1,i_2}$ and $E_{1,i_3}, \ldots, E_{i_2-1,i_3}$, such that $\phi^{(1)}_{i_2 i_3}$ is non-zero if and only if the edge $E_{i_2 i_3}$ cannot complete a triangle, i.e. whether $E_{i_2 i_3} = 0$ or $E_{i_2 i_3} = 1$ the graph can be triangle-free as long as the other triplets of edges not containing $E_{i_2 i_3}$ do not result in a triangle. Similarly, $\varphi^{(2)}_{i_2 i_3}$ is a normalized probability over the same edges, but $\varphi^{(2)}_{i_2 i_3}$ is non-zero if and only if the inclusion of the edge $E_{i_2 i_3}$ will necessarily complete one of the triangles, i.e. for the graph to be triangle-free then it must hold that $E_{i_2 i_3} = 0$. Also notice that both $\varphi^{(1)}_{i_2 i_3}$ and $\varphi^{(2)}_{i_2 i_3}$ can be defined by a D&C SPN. Given the above, either $\varphi^{(1)}_{i_2 i_3} > 0$ or $\varphi^{(2)}_{i_2 i_3} > 0$, hence the denominator of $\Phi_{i_2 i_3}$ is always non-zero. Additionally, if $\varphi^{(2)}_{i_2 i_3}$ is non-zero then $E_{i_2 i_3} = 0$ or else the graph has a triangle, and otherwise either $E_{i_2 i_3} = 0$ or $E_{i_2 i_3} = 1$, hence the numerator of $\Phi_{i_2 i_3}$ is greater than zero if and only if none of the triangles considered are part of the graph. It is also trivial to verify that the numerator is also a D&C SPN, and hence conditionally D&C. Finally, it is clear from the construction that Φ_{i_2, i_3} is strongly conditionally sound, thus it is equivalent to a conditional distribution of $E_{i_2 i_3}$ conditioned on the other edges of the triangles whose $E_{i_2 i_3}$ is their largest edge, where $\text{eff}(\Phi_{i_2 i_3}) = \{E_{i_2 i_3}\}$ and $\text{cond}(\Phi_{i_2 i_3}) = \{E_{i_1 i_2}, E_{i_1 i_3} | 1 \leq i_1 < i_2\}$.

With the above conditional distributions defined for all edges $E_{i_2 i_3}$ such that $1 < i_2 < i_3 \leq M$, we can now construct a strictly positive distribution over triangle-free graphs. First, let us define $\Phi_{1i} \equiv \frac{\mathbb{I}[E_{1i}=0] + \mathbb{I}[E_{1i}=1]}{2}$ for all $1 < i \leq M$, for which $\text{eff}(\Phi_{1i}) = \{E_{1i}\}$ and $\text{cond}(\Phi_{1i}) = \emptyset$. Then, we define the probability as $\Phi \equiv \prod_{1 \leq i < j \leq M} \Phi_{ij}$, and due to the definition of $\Phi_{i_2 i_3}$ it is once more trivial to verify that Ψ is conditionally decomposable, and specifically that the induced dependency graph is indeed cycle-free – this is due to the choice of lexical order which guarantees that $E_{i_2 i_3}$ can only depend on edges which are smaller than it, forbidding the formation of any cycle. In conclusion, Φ is a tractable SPQN, which is non-zero if and only if the edges in $\mathcal{E}$ represent a triangle-free graph – as required. Additionally, since the size of $\varphi^{(1)}_{i_2 i_3}$ is at most $O(M)$, and the size of $\varphi^{(2)}_{i_2 i_3}$ is at most $O(M^2)$, then the size of Φ is at most $O(M^4)$, which proves the main result.

With regards to realizing the same SPQN with valid CMOs, notice that the quotient nodes already follow the structure of valid CMOs, and that the numerator and denominator are simply D&C SPNs which SPQNs composed of valid CMOs can arbitrarily approximate without changing the size of the model. Thus this distribution can be

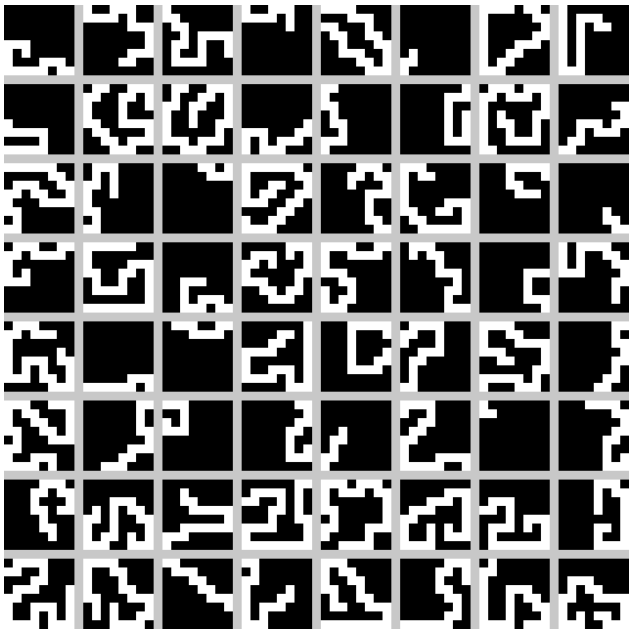

Figure 1: Samples from the synthetic dataset we have designed to showcase the advantages of SPQNs over SPNs.

approximated arbitrarily well with an SPQN composed of valid CMO nodes of size at most $O(M^4)$.

B Experiments

As a preliminary demonstration of the practical advantages of SPQNs over standard SPNs, we have conducted a basic experiment on a synthetic dataset suited to the strengths of SPQNs.

In essence, the difference between the two models is that SPNs have a limited ability to represent intricate correlations between large set of variables – once a sample is drawn for some variable, the result has no further effect on the rest of the sampling process, i.e. the drawing of the children of the remaining sum nodes. In contrast, SPQN can due to the conditional distributions of its nodes. We have came up with a simple synthetic dataset to demonstrate this difference, comprising of $N \times N$ binary images generated as follows: first sample a random location in the image, and then begin drawing a continuous non-overlapping path, where at each step the path can be extended either forward, left, or right, with respect to the direction of movement, with equal probability, and given that the next position is free and not directly adjacent to a previous section of the line (excluding the current position). See fig. 1 for a selection of samples from this distribution for $N = 8$. Since a pixel can be "on" only if there is a free path connecting it to either ends of the drawn path, then it is dependent on all previously sampled pixels, following our initial motivation.

In our experiments we have randomly sampled from the above generative process for the case of $N = 8$: 50000 examples for the training set, 1000 examples for validation and 10000 examples for the test set. We have trained SPNs with the structure learning algorithm proposed by Gens and Domingos (2013), where the hyper-parameters where chosen using grid-search following the same space as in the original article. The best model had in total $189,128$ nodes.

For SPQNs we have first flatten the 8×8 binary image to a 1D array of size 64, and then chosen a simple architecture mimicking a 1D convolutional network. Namely, the graph is composed of a sequence of "convolutional" layers, where each layer d is defined by a stride S_d, receptive field R_d, and number of channels C_d, and is composed of many CMO nodes spatially arranged and stacked according to C_d. For each layer d, spatial position t, and channel c, there is a CMO node whose effective scope is connected to nodes of layer $d-1$ at the spatial locations $t \cdot S - (S-1), \ldots, t \cdot S$ via intermediate sum nodes that are each connected the channels of a given spatial location, and similarly for the conditional scope at the spatial locations $t \cdot S - R + 1, t \cdot S - S$. Essentially, the output $O_{d,t,c}$ of layer d at location t and channel c is equivalent to the following:

$$O_{d,t,c} = \frac{\sum_{i=1}^{C_d} W_{c,i}^{\text{Out}} \left(\prod_{j=1}^{S} \sum_{k=1}^{C_{d-1}} W_{i,j,k}^{\text{In}} O_{d-1,t\cdot S-j+1,k} \right) \left(\prod_{j=S+1}^{R} \sum_{k=1}^{C_{d-1}} W_{i,j,k}^{\text{In}} O_{d-1,t\cdot S-j+1,k} \right)}{\sum_{i=1}^{C_d} W_{c,i}^{\text{Out}} \prod_{j=1}^{S} \sum_{k=1}^{C_{d-1}} W_{i,j,k}^{\text{In}} O_{d-1,t\cdot S-j+1,k}}.$$

The above architecture was trained with the Adam (Kingma and Ba, 2015) variant of SGD, using $\beta_1 = \beta_2 = 0.9$, a learning rate of $5e-2$, mini-batches of 100 samples each, and for 20 epochs. The other hyper-parameters where chosen using cross-validation, where the best performing model was composed of 4 layers, with receptive fields equal to $R_1 = R_2 = 32, R_3 = 16, R_4 = 1$, strides equal to $S_1 = S_2 = 2, S_3 = 16, S_4 = 1$, and number of channels equal to $R_1 = R_2 = R_3 = 64, R_4 = 1$. In terms of sum, product, and quotient nodes involved in the computation, it amounts to just $108,817$ nodes, on par with the SPN model found via structure learning.

In the final results, the best SPN model attained a log-likelihood score of -22.68 on the training set, -24.35 on the validation set, and -24.71 on the test set. In contrast, the best SPQN model attained -15.94 on the training set, -16.36 on the validation set, and -16.51 on the test set, which amounts to a 35% improvement over SPNs. Given both models are of similar size, and despite the fact no structure learning was used for SPQN model, then the SPQN model clearly outperform by a large margin the standard SPN model. Nevertheless, it is important to stress that further empirical evaluations are required to completely validate the advantages of SPQNs over SPNs, even more so on real-world tasks.