

# On the Relationship Between Monotone and Squared Probabilistic Circuits

Benjie Wang, Guy Van den Broeck

University of California, Los Angeles  
benjiewang@cs.ucla.edu, guyvdb@cs.ucla.edu

## Abstract

Probabilistic circuits are a unifying representation of functions as computation graphs of weighted sums and products. Their primary application is in probabilistic modeling, where circuits with non-negative weights (*monotone* circuits) can be used to represent and learn density/mass functions, with tractable marginal inference. Recently, it was proposed to instead represent densities as the *square* of the circuit function (*squared* circuits); this allows the use of negative weights while retaining tractability, and can be exponentially more expressive efficient than monotone circuits. Unfortunately, we show the reverse also holds, meaning that monotone circuits and squared circuits are incomparable in general. This raises the question of whether we can reconcile, and indeed improve upon the two modeling approaches. We answer in the positive by proposing Inception PCs, a novel type of circuit that naturally encompasses both monotone circuits and squared circuits as special cases, and employs complex parameters. Empirically, we validate that Inception PCs can outperform both monotone and squared circuits on a range of tabular and image datasets.

**Code** — <https://github.com/wangben88/InceptionPCs>

**Extended version** — <https://arxiv.org/abs/2408.00876>

## 1 Introduction

The philosophy of tractable probabilistic modeling advocates for modeling high-dimensional probability distributions using model architectures that support exact and efficient inference for various probabilistic queries *by-design*. This property makes them extremely useful for probabilistic reasoning; for example, tractable models have found wide-ranging applications from enhancing and controlling intractable generative models (Zhang et al. 2023; Liu, Niepert, and Van den Broeck 2024), to neuro-symbolic AI (Ahmed et al. 2022; Ahmed, Chang, and Van den Broeck 2023), to causal discovery and inference (Wang, Wicker, and Kwiatkowska 2022; Wang and Kwiatkowska 2023).

The *lingua franca* for tractable models is the probabilistic circuits framework (PC) (Choi, Vergari, and Van den Broeck 2020), which specify functions using *computation graphs* of sums and products, unifying many previous tractable models

such as arithmetic circuits (Darwiche 2003), sum-product networks (Poon and Domingos 2011) and cutset networks (Rahman, Kothalkar, and Gogate 2014). The standard approach to modeling is to directly represent the probability distribution (density/mass function) using a PC. In order to enforce non-negativity of the PC output, one typically restricts to the PC to have non-negative parameters; these are known as *monotone* PCs (Darwiche 2003; Poon and Domingos 2011). However, recent works have also shown that there exist many tractable classes of probability distributions that provably cannot be expressed in this way (Zhang, Holtzen, and Broeck 2020; Yu, Trapp, and Kersting 2023; Broadrick, Zhang, and Van den Broeck 2024).

This motivates the development of new approaches for *practically constructing* and *learning* generalized tractable models. To this end, Loconte et al. (2024b) recently proposed *squared circuits*, where the probability distribution is defined to be (proportional to) the *square* of the circuit function. In this formulation, a PC can employ real (possibly negative parameters) while still defining a valid probability distribution. This was shown theoretically to lead to an exponential advantage in expressive efficiency compared to monotone PCs, for certain classes of distributions.

In this work, we reexamine monotone and squared (structured-decomposable) PCs, and show that they are incomparable in general: either can be exponentially more expressive efficient than the other. Motivated by this observation, we draw on the latent variable interpretation of PCs (Peharz et al. 2016) and show an elegant connection between the two types of circuits; namely, that they simply correspond to summing the latent variables *outside or inside the square*, i.e. a deep sum-of-squares or square-of-sums. By combining these two types of latent variables, we introduce a novel class of tractable models representing deep *sum-of-square-of-sums*, which we call Inception PCs (IncPC), strictly generalizing and extending (structured-decomposable) monotone and squared PCs.

We further investigate learning Inception PCs effectively at scale. Following recent trends in PC learning (Peharz et al. 2020b,a; Liu and Van den Broeck 2021; Mari, Vesio, and Vergari 2023), we design an efficient tensorized implementation of Inception PCs that subsumes tensorized (structured-decomposable) monotone and squared PCs. Empirical results validate the improved expressive efficiency of

Inception PCs on density estimation benchmarks.

Our contributions can be summarized as follows:

- Theoretically, we show that monotone circuits can be *exponentially* smaller than squared circuits for a simple class of distributions, meaning that monotone and squared circuits are incomparable in general in terms of expressive efficiency (Section 3);
- We analyze both monotone and squared circuits from a latent variable perspective, showing that they can be interpreted as deep *sums-of-squares* and *squares-of-sums* respectively. We then propose a novel class of tractable circuits, Inception PCs (IncPC), which generalize and extend both types of circuits as a deep *sum-of-square-of-sums* and can employ complex parameters (Section 4).
- We propose an efficient tensorized architecture for Inception PCs, enabling their application to large-scale datasets (Section 5);
- Empirically, we demonstrate that (i) combining both types of latents, together with complex parameters, can improve performance; and (ii) Inception PCs can outperform monotone PCs and squared PCs on challenging benchmarks including downscaled versions of ImageNet, even when normalized for computation time (Section 7).

In concurrent work, Loconte, Mengel, and Vergari (2024) proved an exponential separation between monotone and squared circuits (i.e., our Theorem 2) using a similar family of separating functions. They also proposed two new model classes, SOCS and  $\mu$ SOCS, which generalize squared circuits. In the extended version, we discuss in detail connections with our Inception PCs.

## 2 Preliminaries

**Notation** We use capital letters to denote variables and lowercase to denote their assignments/values (e.g.  $X, x$ ). We use boldface (e.g.  $\mathbf{X}, \mathbf{x}$ ) to denote sets of variables/assignments.

Probabilistic circuits are *computation graphs* representing functions constructed by hierarchical compositions of weighted sums and products.

**Definition 1** (Probabilistic Circuit). *A probabilistic circuit  $\mathcal{C}$  over a set of variables  $\mathbf{V}$  is a rooted DAG consisting of three types of nodes  $n$ : input, product and sum nodes. Each input node  $n$  is a leaf encoding a function  $f_n : \mathbf{W} \rightarrow \mathbb{R}$  for some  $\mathbf{W} \subseteq \mathbf{V}$ , and for each internal (product or sum) node  $n$ , denoting the set of children (i.e. nodes  $n'$  for which  $n \rightarrow n'$ ) by  $\text{in}(n)$ , we define:*

$$f_n = \begin{cases} \prod_{n_i \in \text{in}(n)} f_{n_i} & \text{if } n \text{ is product;} \\ \sum_{n_i \in \text{in}(n)} \theta_{n, n_i} f_{n_i} & \text{if } n \text{ is sum.} \end{cases} \quad (1)$$

where each sum node has a set of weights  $\{\theta_{n, n_i}\}_{n_i \in \text{in}(n)}$  with  $\theta_{n, n_i} \in \mathbb{R}$ . Each node  $n$  thus encodes a function over a set of variables  $\text{sc}(n)$ , which we call its *scope*; this is given by  $\text{sc}(n) = \bigcup_{n_i \in \text{in}(n)} \text{sc}(n_i)$  for internal nodes. The function encoded by the circuit  $f_{\mathcal{C}}$  is the function encoded by its root node. The size of a probabilistic circuit  $|\mathcal{C}|$  is defined to be the number of edges in its DAG.

In this paper, we will assume that sum and product nodes alternate; this is without loss of generality as this property can be enforced on a PC at most a linear increase in size. A key feature of the sum-product structure of probabilistic circuits is that they allow for efficient (linear-time) computation of marginals, for example the partition function  $Z = \sum_{\mathbf{v}} f(\mathbf{v})^1$ , if they are *smooth* and *decomposable*:

**Definition 2** (Smoothness, Decomposability). *A probabilistic circuit is smooth if for every sum node  $n$ , its inputs  $n_i$  have the same scope. A probabilistic circuit is decomposable if for every product node  $n$ , its inputs have disjoint scope.*

We will also need a stronger version of decomposability that enables circuits to be multiplied together efficiently (Pitarsrisawat and Darwiche 2008; Vergari et al. 2021):

**Definition 3** (Structured Decomposability). *A smooth and decomposable probabilistic circuit is structured-decomposable if any two product nodes  $n, n'$  with the same scope decompose in the same way.*

Structured-decomposability is commonly enforced when learning PCs from data (Liu and Van den Broeck 2021) as well as when compiling a circuit from another model such as a Bayesian network (Choi, Kisa, and Darwiche 2013).

## 3 Expressive Efficiency of Monotone and Squared Structured-Decomposable Circuits

One of the primary applications of probabilistic circuits is as a tractable representation of probability distributions. As such, we typically require the function output of the circuit to be a non-negative real. The usual way to achieve this is to enforce non-negativity of the weights and input functions:

**Definition 4** (Monotone PC). *A probabilistic circuit is monotone if all weights are non-negative reals, and all input functions map to the non-negative reals.*

Given a *monotone* PC  $\mathcal{C}$ , one can define a probability distribution  $p_1(\mathbf{V}) := \frac{f_{\mathcal{C}}(\mathbf{V})}{Z_{\mathcal{C}}}$  where  $Z_{\mathcal{C}}$  is the partition function of the PC. However, this is not the only way to construct a non-negative function. In Loconte et al. (2024b), it was proposed to instead use  $f_{\mathcal{C}}$  to represent a *real* (i.e. possibly negative) function, by allowing for real weights/input functions; this can then be squared to obtain a non-negative function.

That is, we define  $p_2(\mathbf{V}) := \frac{f_{\mathcal{C}}(\mathbf{V})^2}{\sum_{\mathbf{v}} f_{\mathcal{C}}(\mathbf{v})^2}$ .

In order for  $\sum_{\mathbf{v}} f_{\mathcal{C}}(\mathbf{v})^2$  to be tractable to compute, a sufficient condition is for the circuit  $\mathcal{C}$  to be structured-decomposable; one can then explicitly construct a smooth and (structured-)decomposable circuit  $\mathcal{C}^2$  such that  $f_{\mathcal{C}^2}(\mathbf{V}) = f_{\mathcal{C}}(\mathbf{V})^2$  of size and in time  $O(|\mathcal{C}|^2)$  (Vergari et al. 2021). Then we have that  $p_2(\mathbf{V}) = \frac{f_{\mathcal{C}^2}(\mathbf{V})}{Z_{\mathcal{C}^2}}$ , i.e. the distribution induced by the PC  $\mathcal{C}^2$ . Crucially, the circuit  $\mathcal{C}^2$  is not necessarily monotone; squaring thus provides an alternative means of constructing PCs that represent non-negative functions. In fact, it is known that squared real PCs can

<sup>1</sup>alternatively,  $\int$  in the case of continuous variables

be exponentially more expressive efficient than structured-decomposable monotone PCs for representing probability distributions:

**Theorem 1.** (Loconte et al. 2024b) *There exists a class of non-negative functions  $p(\mathbf{V})$  such that there exist structured-decomposable PCs  $\mathcal{C}$  with  $p(\mathbf{V}) = f_{\mathcal{C}}(\mathbf{V})^2$  of size polynomial in  $|\mathbf{V}|$ , but the smallest structured-decomposable monotone PC  $\mathcal{C}'$  such that  $p(\mathbf{V}) = f_{\mathcal{C}'}(\mathbf{V})$  has size  $2^{\Omega(|\mathbf{V}|)}$ .*

However, we now show that, in fact, the other direction also holds: monotone PCs can also be exponentially more expressive efficient than squared (real) PCs.

**Theorem 2.** *There exists a class of non-negative functions  $p(\mathbf{V})$ , such that there exist monotone structured-decomposable PCs  $\mathcal{C}$  with  $p(\mathbf{V}) = f_{\mathcal{C}}(\mathbf{V})$  of size polynomial in  $|\mathbf{V}|$ , but the smallest structured-decomposable PC  $\mathcal{C}'$  such that  $p(\mathbf{V}) = f_{\mathcal{C}'}(\mathbf{V})^2$  has size  $2^{\Omega(|\mathbf{V}|)}$ .*

*Proof.* (Sketch) The function class we use to separate the circuit classes is  $p(\mathbf{V}) = \sum_{i=0}^{d-1} 2^i \mathbb{1}_{V_i=1} + 1$ , i.e. a function that outputs the integer encoded in binary by the variables (+1). This can easily be represented as a structured-decomposable monotone circuit of linear size. We show a lower bound on the size of structured-decomposable circuits computing any square root  $F(\mathbf{V})$  of this function, by (i) reducing to bounding the rank of a matrix representation of  $F$  w.r.t. balanced partitions  $(\mathbf{X}, \mathbf{Y})$  of  $\mathbf{V}$  (Martens and Medabalimi 2014), using a standard technique from communication complexity; (ii) lower bounding the number of distinct prime square roots in the matrix; (iii) showing the existence of a sufficiently large submatrix with full rank, thus lower bounding the rank of the original matrix (Fawzi et al. 2015).  $\square$

This is perhaps surprising, as squaring PCs generate structured PCs with possibly negative weights, suggesting that they should be more general than monotone structured PCs. The key point is that not all circuits that represent a positive function (not even all monotone structured ones) can be generated by squaring. Taken together, these results are somewhat unsatisfying, as we know that there are some distributions better represented by an unsquared monotone PC, and some by a squared real PC. In the next section, we will investigate how to reconcile these different approaches to specifying probability distributions.

## 4 Towards a Unified Model for Deep Sums-of-Squares-of-Sums

In this section, we investigate the relationship between monotone and squared circuits in depth. Firstly, we show how to introduce complex parameterizations to squared circuits (Prop 1). Secondly, we provide a new interpretation of monotone and squared circuits as different ways of marginalizing out latent variables (Figure 1); and introduce a new tractable model, Inception PCs, that combines the two approaches (Theorem 3).

### 4.1 Complex Parameters

We begin by noting that, beyond simply negative parameters, one can also allow for *complex* weights and input functions<sup>2</sup>, i.e. take values in the field  $\mathbb{C}$ . Then, to ensure the non-negativity of the squared circuit, we multiply a circuit with its *complex conjugate*. That is:

$$p_2(\mathbf{V}) = \frac{|f_{\mathcal{C}}(\mathbf{V})|^2}{\sum_{\mathbf{v}} |f_{\mathcal{C}}(\mathbf{v})|^2} = \frac{\overline{f_{\mathcal{C}}(\mathbf{V})} f_{\mathcal{C}}(\mathbf{V})}{\sum_{\mathbf{v}} \overline{f_{\mathcal{C}}(\mathbf{v})} f_{\mathcal{C}}(\mathbf{v})}$$

As complex conjugation is a field isomorphism of  $\mathbb{C}$ , taking a complex conjugate of a circuit is as straightforward as taking the complex conjugate of each weight and input function, retaining the same DAG as the original circuit. This allows us to efficiently compute  $p_2(\mathbf{V})$  as a (smooth and structured decomposable) circuit:

**Proposition 1** (Tractability of Complex Conjugation). *Given a smooth and decomposable circuit  $\mathcal{C}$ , it is possible to compute a smooth and decomposable circuit  $\bar{\mathcal{C}}$  such that  $f_{\bar{\mathcal{C}}}(\mathbf{V}) = \overline{f_{\mathcal{C}}(\mathbf{V})}$  of size and in time  $O(|\mathcal{C}|)$ . Further, if  $\mathcal{C}$  is structured decomposable, then it is possible to compute a smooth and structured decomposable  $\mathcal{C}^2$  s.t.  $f_{\mathcal{C}^2}(\mathbf{V}) = \overline{f_{\mathcal{C}}(\mathbf{V})} f_{\mathcal{C}}(\mathbf{V})$  of size and in time  $O(|\mathcal{C}|^2)$ .*

### 4.2 Deep Sums-of-Squares-of-Sums: A Latent Variable Interpretation

In the latent variable interpretation (LVI) of probabilistic circuits (Peharz et al. 2016), for every sum node, one assigns a categorical latent variable, where each state of the latent variable is associated with one of the inputs to the sum node; we show an example in Figure 1a. In this interpretation, when performing inference in the probabilistic circuit, we marginalize over all of the latent variables beforehand.

In the case when the circuit is structured decomposable, one can assign latent variables with variable scopes (sets) appearing in the circuit, such that two sum nodes  $n, n'$  with the same scope  $sc(n) = sc(n')$  are associated with the same latent  $Z_{sc(n)}$ . Then, writing  $\mathbf{Z}$  for the set of all latents, the function represented by the PC can be expressed as:

$$f_{\mathcal{C}}(\mathbf{V}) = \sum_{\mathbf{z}} f_{\mathcal{C}}(\mathbf{V}, \mathbf{z}) \quad (2)$$

where  $f_{\mathcal{C}}(\mathbf{V}, \mathbf{z})$  is a *product of input functions* for any value of  $\mathbf{z}$ .<sup>3</sup> In other words, the distribution represented by the PC is fully factorized conditioned on a latent value  $\mathbf{z}$ .

However, interpreting these latent variables becomes tricky when we consider probability distributions defined by squaring circuits. The key question is, does one marginalize out the latent variables before or after squaring? We show

<sup>2</sup>Circuits with complex-valued input functions have previously been proposed by Yu, Trapp, and Kersting (2023) in the context of representing characteristic functions (an alternative representation of probability distributions).

<sup>3</sup>Concretely,  $f_{\mathcal{C}}(\mathbf{V}, \mathbf{z})$  is a function obtained by traversing the circuit top-down, selecting one child of every sum node according to the value  $\mathbf{z}$ , and all children of a product node; cf. also the notion of induced subcircuits/trees (Chan and Darwiche 2006; Zhao, Poupart, and Gordon 2016).

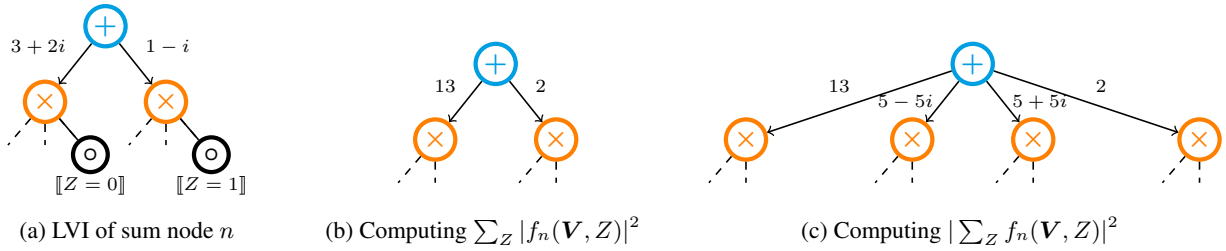


Figure 1: Latent variable interpretation for squaring PCs. The sum node in Figure 1a has two children with complex weights and associated with different values of the latent  $Z$ . A sum-of-squares (Figure 1b) gives a monotone PC, where the parameters necessarily become non-negative. A square-of-sums (Figure 1c) leads to a squared PC, with four children each corresponding to the product of any two children from the original circuit.

both options in Figures 1b and 1c. In Figure 1b, we square before marginalizing  $Z$ . In this case, each sum node weight is multiplied by its conjugate, and we are left with a sum node with non-negative real parameters. On the other hand, if we marginalize before squaring, we have a sum node with four children and complex parameters. Interestingly, the former case is very similar to directly constructing a monotone PC, while the latter is more like an explicit squaring without latent variables. This suggests that we can switch between monotone and squared PCs simply by *deciding whether to sum the latent variables inside or outside the square*. Using this perspective, we propose the following model, which explicitly introduces both type of latent variables into the circuit, producing a deep *sum-of-square-of-sums*:

**Definition 5** (Inception PC<sup>4</sup>). *An Inception PC (IncPC)  $\mathcal{C}_{\text{Inception}}$  is a smooth and structured-decomposable probabilistic circuit over observed variables  $\mathbf{V} \cup \mathbf{U} \cup \mathbf{W}$ . The probability distribution of an Inception PC is defined by:*

$$p_{\text{Inception}}(\mathbf{V}) = \frac{\sum_{\mathbf{u}} \left| \sum_{\mathbf{w}} f_{\mathcal{C}_{\text{Inception}}}(\mathbf{V}, \mathbf{u}, \mathbf{w}) \right|^2}{\sum_{\mathbf{v}} \sum_{\mathbf{u}} \left| \sum_{\mathbf{w}} f_{\mathcal{C}_{\text{Inception}}}(\mathbf{v}, \mathbf{u}, \mathbf{w}) \right|^2} \quad (3)$$

As  $U$ -latents are outside the square, and  $W$ -latents are inside the square, we will refer to them as 1-norm and 2-norm latents respectively. The next Theorem shows that, given an Inception PC, we can efficiently “materialize” it into a tractable PC over just the observed variables  $\mathbf{V}$  representing the Inception PC’s distribution  $p_{\text{Inception}}(\mathbf{V})$ :

**Theorem 3** (Tractability of InceptionPC). *Given an Inception PC  $\mathcal{C}_{\text{Inception}}$ , it is possible to compute a smooth and structured decomposable circuit  $\mathcal{C}_{\text{mat}}$  such that  $f_{\mathcal{C}_{\text{mat}}}(\mathbf{V}) = \sum_{\mathbf{u}} \left| \sum_{\mathbf{w}} f_{\mathcal{C}_{\text{Inception}}}(\mathbf{V}, \mathbf{u}, \mathbf{w}) \right|^2$  of size and in time  $O(|\mathcal{C}|^2)$ .*

*Proof.* The most systematic way to see this is to take advantage of the compositional inference framework of Vergari et al. (2021). We can marginalize out  $\mathbf{W}$  from  $\mathcal{C}_{\text{Inception}}$  to obtain a PC  $\mathcal{C}_1$  such that  $f_{\mathcal{C}_1}(\mathbf{V}, \mathbf{U}) = \sum_{\mathbf{w}} f_{\mathcal{C}_{\text{Inception}}}(\mathbf{V}, \mathbf{U}, \mathbf{w})$ , retaining smoothness and structured decomposability. Then the computation of the square is possible by Proposition 1, returning a smooth and structured-decomposable circuit  $\mathcal{C}_2$  such that  $f_{\mathcal{C}_2}(\mathbf{V}, \mathbf{U}) = \left| \sum_{\mathbf{w}} f_{\mathcal{C}_{\text{Inception}}}(\mathbf{V}, \mathbf{U}, \mathbf{w}) \right|^2$ . Finally,

<sup>4</sup>The nomenclature is inspired by the deep(er) layering of summation and squaring in Inception PCs.

we can marginalize out  $\mathbf{U}$  from this circuit to obtain the structured decomposable and smooth circuit  $\mathcal{C}_{\text{mat}}$ .  $\square$

In Figure 2, we show a (fragment of) an example Inception PC, and its materialization via Theorem 3. In the rest of the paper, we will refer to this as the *materialized* Inception PC. The point is that we can efficiently perform inference on the distribution of an Inception PC through standard PC inference procedures on the materialized IncPC. Though this is not strictly necessary, we will assume (in accordance with the LVI for monotone and squared PCs) that  $\mathbf{W}, \mathbf{U}$  are categorical, and the input nodes with scope over these variables are indicators (e.g.  $\llbracket W = 1 \rrbracket$ ).

This provides an elegant resolution to the tension between monotone and squared (real/complex) PCs. To retrieve a monotone PC, we need only set  $\mathbf{W} = \emptyset$ ; then there is no summation inside the square, and  $\mathcal{C}_{\text{mat}}$  has the same structure as  $\mathcal{C}_{\text{Inception}}$  but with the parameters and input functions squared (and so non-negative real).<sup>5</sup> To retrieve a squared PC, we simply set  $\mathbf{U} = \emptyset$ ; then there is no summation outside the square. However, by using both types of latents, we obtain a generalized PC model that is strictly more expressive efficient than either individually<sup>6</sup>.

**Corollary 1** (Expressive Efficiency of Inception PCs). *Inception PCs are strictly more expressive efficient than either monotone or squared PCs.*

## 5 Tensorized Inception PCs

Thus far, we have described a general formulation of Inception PCs that simply requires smoothness and structured decomposability of the circuit. In practice, we will want to design specific circuit architectures for learning. The purpose of this section is to (i) propose such an architecture suitable for learning and inference on GPUs; and (ii) present an alternative view of Inception PCs as performing hierarchical tensor contractions.

<sup>5</sup>Interestingly enough, in this case we can relax the conditions of Theorem 3 to require decomposability rather than structured decomposability, assuming that children of sum nodes correspond to different indicators and thus  $\mathcal{C}_{\text{Inception}}$  is deterministic. Multiplying a *deterministic* circuit with itself (or its conjugate) is tractable in linear time (Vergari et al. 2021).

<sup>6</sup>The special cases can also be *learned* in the general case by making  $f_{\mathcal{C}_{\text{Inception}}}$  constant w.r.t.  $\mathbf{U}$  or  $\mathbf{W}$ .

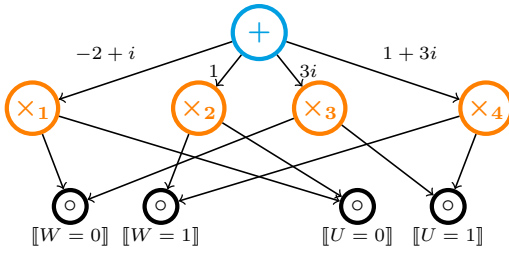
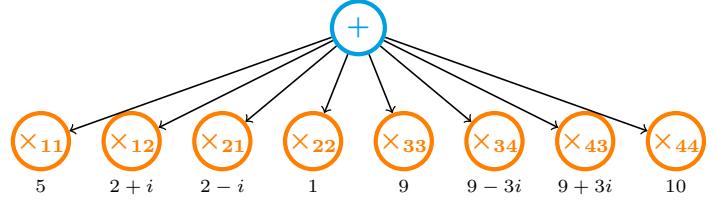
(a) Inception PC  $\mathcal{C}_{\text{Inception}}$ (b) Materialized Inception PC  $\mathcal{C}_{\text{mat}}$ 

Figure 2: Diagrams showing an Inception PC, and the corresponding materialized IncPC. Each product node is labelled with an index, such that e.g.  $\times_{34}$  is the product of the product nodes  $\times_3, \times_4$ . For clarity, the children of product nodes have been omitted (except the latent indicators), and edge weights for the materialized Inception PC are displayed below the corresponding child.

We follow recent trends in probabilistic circuit learning (Peharz et al. 2020a; Mari, Vessio, and Vergari 2023) and consider tensorized architectures for  $\mathcal{C}_{\text{Inception}}$ , where sum and product nodes are grouped into regions by scope. In particular, for each region (scope), we construct  $N_1 \times N_2$  sum and product nodes in the Inception PC, where the sum nodes are connected in a dense fashion to the product nodes via a weight tensor  $\theta_{ii'kk'} \in \mathbb{C}^{N_1^2 \times N_2^2}$  with  $1 \leq k, k' \leq N_1$ ,  $1 \leq i, i' \leq N_2$ ; and the product nodes are connected to sum nodes in subsequent regions via a Hadamard product (Loconte et al. 2024a). Each region is associated with a categorical latent  $W$  and  $U$ ; and each product node corresponds explicitly to a value of  $W$  and  $U$ , having two indicators  $\llbracket W = i' \rrbracket$  and  $\llbracket U = k' \rrbracket$  as children where  $1 \leq k' \leq N_1$ ,  $1 \leq i' \leq N_2$ . For example, the product nodes in the Inception PC in Figure 2a correspond to a tensorized Inception PC with  $N_1 = 2, N_2 = 2$ .

Once materialized using Theorem 3, we obtain a materialized IncPC  $\mathcal{C}_{\text{mat}}$  with  $N_1 \times N_2^2$  sum and product nodes; the increase in size occurs from the “expansion” from two-norm latents  $\mathbf{W}$  as seen in Figure 1c. For a given region, let us write  $n_{ijk}^{(S)}$  for the sum nodes for  $1 \leq k \leq N_1, 1 \leq i, j \leq N_2$ , and  $n_{ijk}^{(P)}$  for the product nodes for  $1 \leq k \leq N_1, 1 \leq i, j \leq N_2$  in  $\mathcal{C}_{\text{mat}}$ . The product nodes are still connected to their  $M$  child sum regions  $n^{(S_1)}, \dots, n^{(S_M)}$  via a Hadamard product:

$$f_{n_{ijk}^{(P)}} = \prod_{m=1}^M f_{n_{ijk}^{(S_m)}} \quad (4)$$

However, the sum-to-product connection is no longer dense. In particular, writing we have the following relationship between the sum and product node tensors:

$$f_{n_{ijk}^{(S)}} = \sum_{i'j'k'} \theta_{ii'kk'} \overline{\theta_{jj'kk'}} f_{n_{i'j'k'}^{(P)}} \quad (5)$$

We illustrate the structure of this sum region in Figure 3. On a high level, the connections between the groups of nodes in Figure 3 can be viewed as a standard, monotone PC; in particular, the value of each sum group is simply an weighted sum of its children. However, each weighted edge is between 2D groups of “squared” nodes, rather than between scalar nodes.

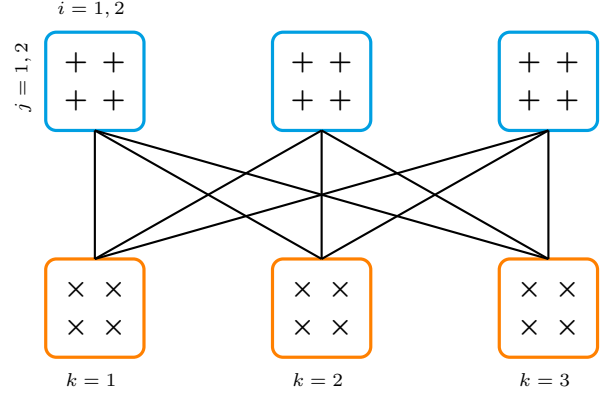


Figure 3: Illustration of tensorized Inception PC sum region, where  $N_1 = 3, N_2 = 2$ .

The complexity of a forward pass can be deduced from the tensor contraction in Equation 5. In Table 1, we compare the parameter counts (i.e., memory cost) and time complexity of a forward pass for each region, between monotone, squared and Inception PCs; where  $B$  is the size of the data batch. It can be seen that Inception PCs have the same number of parameters/complexity as monotone and squared PCs when setting  $N_2 = 1$  and  $N_1 = 1$  respectively, except with the complexity differing for squared PCs: this is because one can perform the squaring only once-per-batch in this special case (Loconte et al. 2024b).

As with previous works (Peharz et al. 2020a), we organize the circuit into layers (sets of regions that can be computed simultaneously) to further take advantage of GPU parallelization. For training, we use gradient descent on the negative log-likelihood of the training set. In the case of using complex parameters, we use Wirtinger derivatives (Kreutz-Delgado 2009) in order to optimize the complex weights and input functions. To achieve numerical stability, we use a variant of the log-sum-exp trick for complex numbers, described in the extended version of this paper.

	Monotone	Squared	Inception
Complexity	$BN_1^2$	$BN_2^2 + N_2^3$	$BN_1^2N_2^3$
Parameter Count	$N_1^2$	$N_2^2$	$N_1^2N_2^2$

Table 1: Complexity of batched forward pass and parameter count for circuit variants, per region.

## 6 Related Work

Our work builds upon a long line of work in the circuit literature examining the effect of relaxing the monotonicity condition in circuits. Theoretically, it is known that allowing negative parameters in arithmetic circuits can result in exponential gains in succinctness (Valiant 1979), though this not true of all circuit subclasses (de Colnet and Mengel 2021). Practically, recent works have aimed to exploit negative parameters in probabilistic modeling (Zhang, Holtzen, and Broeck 2020; Zhang, Juba, and Van den Broeck 2021; Sladek, Trapp, and Solin 2023; Loconte et al. 2024b). Yu, Trapp, and Kersting (2023) design circuits to represent the characteristic function of a distribution, which can take complex values (in particular, using complex leaves). Concurrently with our work, Loconte, Mengel, and Vergari (2024) proposed to employ a sum of squared circuits to overcome similar limitations to those we observe in this paper.

There also exist a range of other probabilistic models which employ negative parameterizations and/or squaring. Tensor networks, such as the popular matrix-product states (MPS) (Perez-Garcia et al. 2007), compactly encode functions through sparse tensor contractions. They are often used to model quantum states, whereby the probability of observations is given by squaring via the Born rule (Dirac 1981); recently, tensor networks have been used for probabilistic modeling (Cheng et al. 2019; Glasser et al. 2019). Positive semidefinite models (Rudi and Ciliberto 2021) in the kernel methods literature utilize a shallow sum of squares to define unnormalized distributions. Tsuchida, Ong, and Sejdi-novic (2023, 2024) recently introduced squared neural families, which employ the squared 2-norm of a neural network in the density function and strictly generalize exponential families. We discuss technical connections with these models in the extended version of this paper.

## 7 Experiments

In our experiments, we aim to answer the following research questions: (1) do Inception PCs improve upon the expressivity and modeling capabilities of monotone and squared PCs?; (2) what is the tradeoff between the number of one-norm latents and two-norm latents in terms of modeling performance and computational cost?; (3) how do complex parameterizations of Inception PCs compare to non-negative and real parameterizations?

We use hidden Chow-Liu trees (HCLT) in all experiments as the PC vtree (Liu and Van den Broeck 2021), as it satisfies the required structured-decomposability property, and has been shown to provide state-of-the-art likelihoods for PCs. Further experimental details can be found in the extended version.

## 7.1 Binary Datasets

We begin by evaluating on the 20 binary datasets, which has been extensively used as a density estimation benchmark (Rooshenas and Lowd 2014; Peharz et al. 2020b). We aim to investigate across this range of datasets how (1) non-negative, real, and complex parameterizations differ; and (2) whether Inception PCs can effectively make use of both types of latents, by comparing to its monotone and squared PC counterparts. In particular, we set  $(N_1, N_2) = (8, 1)$  for monotone PCs,  $(N_1, N_2) = (1, 8)$  for squared PCs, and  $(N_1, N_2) = (8, 8)$  for Inception PCs. We train each PC on negative log-likelihood for 100 epochs, using the Adam optimizer (Kingma and Ba 2015) with learning rate 0.01. We average over 5 runs for each configuration.

The results are shown in Table 2. It can be seen that Inception PCs achieve the best likelihoods on 14 of the 20 datasets, confirming that they provide more modeling capacity compared to squaring alone. They also perform very favorably compared to existing models, including (tractable) RAT-SPNs (Peharz et al. 2020b) and (intractable) importance weighed autoencoders (Burda, Grosse, and Salakhutdinov 2016). Squared PCs, even with non-negative parameters, outperform their monotone counterparts that have the same number of parameters, as also observed in Figure 3 of Loconte et al. (2024b). Very interestingly, despite the fact that complex parameterizations of edge weights strictly generalizes real parameterizations and non-negative parameterizations, there is no clear trend in performance, with the non-negative parameterization often achieving comparable or better likelihoods than real or complex parameterizations. We observed that this is (at least partially) due to the complex and negative parameterizations being more prone to overfitting (cf. training LLs in the extended version).

## 7.2 Scaling to Large Image Datasets

In this section, we conduct experiments on large-scale image datasets, in particular, downscaled versions of ImageNet to  $32 \times 32$  and  $64 \times 64$  (Deng et al. 2009). Following recent work on PC modeling for these datasets (Liu, Zhang, and Van den Broeck 2023; Liu et al. 2023; Liu, Ahmed, and Van den Broeck 2024), we transform the data from RGB using the lossy YCoCg transform. Note that likelihoods on on YCoCg transformed data are thus not comparable to likelihoods on the original RGB dataset. Additionally, to improve training efficiency, we train the circuits on  $16 \times 16$  patches of the original image<sup>7</sup>, such that the PC models  $16 \times 16 \times 3 = 768$  variables each with 256 categories. We train the models on negative log-likelihood, using the Adam optimizer with learning rate 0.01. We evaluate models using test-set bits-per-dimension (bpd).

We begin by examining the tradeoff between the two types of latents. To this end, in Figure 4 we plot the test bpd for a range of configurations of  $(N_1, N_2)$  on ImageNet32. These configurations were chosen based on a search in the

<sup>7</sup>The entire image is then modelled using the same PC for each of the patches; the performance could potentially be improved further by modelling correlation using a PC over patch-level PCs (Liu et al. 2023).

Dataset	Model								
	Monotone PC	Squared PC			Inception PC			RAT-SPN	VAE
		Non-negative	Real	Complex	Non-negative	Real	Complex		
nltcs	6.04	6.02	6.09	6.03	<b>6.00</b>	6.01	6.02	6.01	5.99
msnbc	6.25	6.23	6.10	6.07	6.05	6.06	<b>6.04</b>	6.04	6.09
kdd-2k	2.13	<b>2.10</b>	2.23	2.12	2.12	2.13	2.12	2.13	2.12
plants	13.70	13.38	15.06	13.13	12.81	13.06	<b>12.76</b>	13.44	12.34
jester	52.77	52.56	54.51	52.97	<b>52.51</b>	52.60	52.75	52.97	51.54
audio	40.27	40.08	41.49	40.01	<b>39.88</b>	39.91	40.05	39.96	38.67
netflix	57.09	56.85	57.68	56.70	<b>56.52</b>	56.57	56.74	56.85	54.73
accidents	29.57	27.93	28.15	27.05	26.70	27.30	<b>26.61</b>	35.49	29.11
retail	10.99	10.82	11.00	<b>10.95</b>	11.00	<b>10.95</b>	<b>10.95</b>	10.91	10.83
pumsb-star	27.98	24.95	25.69	23.98	23.69	24.85	<b>23.03</b>	32.53	25.16
dna	80.21	79.95	80.15	80.17	79.85	80.11	<b>79.77</b>	97.23	94.56
kosarek	10.77	<b>10.54</b>	12.03	10.59	10.69	10.83	10.60	10.89	10.64
msweb	10.44	<b>9.92</b>	10.41	10.17	10.84	10.34	10.10	10.12	9.727
book	33.70	<b>33.32</b>	37.02	33.95	33.51	34.18	33.67	34.68	33.19
eachmovie	52.83	51.28	62.03	52.33	<b>50.76</b>	51.22	51.41	53.63	47.43
web-kb	155.34	151.84	162.03	155.00	<b>151.74</b>	153.32	153.98	157.53	146.9
reuters-52	95.22	92.63	96.25	93.90	93.17	<b>89.67</b>	93.80	87.37	81.33
20ng	155.05	<b>152.98</b>	164.19	154.79	154.15	155.47	155.18	152.06	146.9
bbc	253.98	<b>250.88</b>	259.04	255.13	251.28	253.28	253.37	252.14	240.94
ad	16.93	15.54	16.32	15.35	16.02	15.81	<b>15.32</b>	48.47	18.81

Table 2: Test negative log-likelihoods on 20 binary datasets. Lower is better.

	MPC	SPC	IncPC	EM	LVD
ImageNet32	4.35	4.26	<b>4.19</b>	4.82	4.38
ImageNet64	4.12	3.98	<b>3.90</b>	4.67	4.12

Table 3: Test bpd on large-scale image datasets (lower is better). All PCs use the HCLT vtrees.

range  $N_1, N_2 \in \{1, 2, 4, 8, 16, 32, 64\}$  according to a maximum budget of  $2^{18} = 262144$  floating-point operations per second (FLOPS) per region (c.f. Table 1). It can be seen that the optimal configuration ( $N_1 = 32, N_2 = 4$  with bpd 4.19) given these constraints uses a combination of 1-norm latents and 2-norm latents, as opposed to a pure monotone ( $N_2 = 1$ ) or pure squared circuit ( $N_1 = 1$ ).

In Table 3, we summarize results from the ImageNet32 and ImageNet64 datasets, where MPC refers to a monotonic PC ( $(N_1, N_2) = (64, 1)$ ), SQC a squared PC ( $(N_1, N_2) = (1, 64)$ ), and IncPC a tensorized Inception PC with the optimal configuration of  $(N_1, N_2)$  under the computation time constraint. For comparison, we also show results for monotone HCLT PCs trained using expectation-maximization (EM), and using latent variable distillation (LVD) in which guidance for the PC latent space is provided by distilling information from existing deep generative models (Liu, Zhang, and Van den Broeck 2023). The results show that it is possible to effectively train tractable PC models with Adam with bpd's competitive with the state-of-the-art.

## 8 Conclusion

To conclude, we have shown that two important classes of tractable probabilistic models, namely monotone and squared real structured-decomposable PCs are incompa-

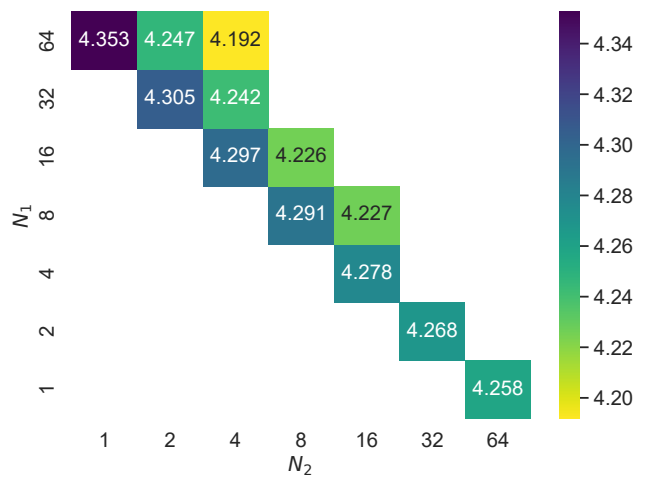


Figure 4: Test bpd for a range of configurations of  $(N_1, N_2)$  on ImageNet32 (lower is better); configurations are limited to  $2^{18} = 262K$  FLOPS per region. The Inception PC with  $N_1 = 32, N_2 = 4$  achieves the best performance.

table in terms of expressive efficiency in general. Thus, we propose a new class of probabilistic circuits, Inception PCs, based on *deep sums-of-squares-of-sums* that generalizes these approaches, and can employ complex parameters. We further show empirically that Inception PCs can offer better performance on a range of tabular and image datasets. Promising avenues to investigate in future work would be improving the optimization of complex parameters in Inception PCs; as well as reducing the computational cost of training by designing more efficient architectures.

## Acknowledgements

BW and GVdB gratefully acknowledge support from the National Artificial Intelligence Research Resource Pilot under project NAIRR240143. This work was funded in part by the DARPA ANSR program under award FA8750-23-2-0004, the DARPA PTG Program under award HR00112220005, and NSF grant #IIS-1943641.

## References

- Ahmed, K.; Chang, K.-W.; and Van den Broeck, G. 2023. A Pseudo-Semantic Loss for Deep Generative Models with Logical Constraints. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- Ahmed, K.; Teso, S.; Chang, K.-W.; Van den Broeck, G.; and Vergari, A. 2022. Semantic Probabilistic Layers for Neuro-Symbolic Learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*.
- Broadrick, O.; Zhang, H.; and Van den Broeck, G. 2024. Polynomial Semantics of Tractable Probabilistic Circuits. In *Conference on Uncertainty in Artificial Intelligence*. PMLR.
- Burda, Y.; Grosse, R.; and Salakhutdinov, R. 2016. Importance Weighted Autoencoders. In *International Conference on Learning Representations (ICLR)*.
- Chan, H.; and Darwiche, A. 2006. On the robustness of most probable explanations. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 63–71.
- Cheng, S.; Wang, L.; Xiang, T.; and Zhang, P. 2019. Tree tensor networks for generative modeling. *Physical Review B*, 99(15): 155131.
- Choi, A.; Kisa, D.; and Darwiche, A. 2013. Compiling probabilistic graphical models using sentential decision diagrams. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 12th European Conference, EC-SQARU 2013, Utrecht, The Netherlands, July 8-10, 2013. Proceedings 12*, 121–132. Springer.
- Choi, Y.; Vergari, A.; and Van den Broeck, G. 2020. Probabilistic Circuits: A Unifying Framework for Tractable Probabilistic Models.
- Darwiche, A. 2003. A differential approach to inference in Bayesian networks. *Journal of the ACM (JACM)*, 50(3): 280–305.
- de Colnet, A.; and Mengel, S. 2021. A compilation of succinctness results for arithmetic circuits. In *18th International Conference on Principles of Knowledge Representation and Reasoning (KR)*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Dirac, P. A. M. 1981. *The principles of quantum mechanics*. 27. Oxford university press.
- Fawzi, H.; Gouveia, J.; Parrilo, P. A.; Robinson, R. Z.; and Thomas, R. R. 2015. Positive semidefinite rank. *Mathematical Programming*, 153: 133–177.
- Glasser, I.; Sweke, R.; Pancotti, N.; Eisert, J.; and Cirac, I. 2019. Expressive power of tensor-network factorizations for probabilistic modeling. *Advances in neural information processing systems*, 32.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Kreutz-Delgado, K. 2009. The complex gradient operator and the CR-calculus. *arXiv preprint arXiv:0906.4835*.
- Liu, A.; Ahmed, K.; and Van den Broeck, G. 2024. Scaling Tractable Probabilistic Circuits: A Systems Perspective. In *Forty-first International Conference on Machine Learning*.
- Liu, A.; Niepert, M.; and Van den Broeck, G. 2024. Image Inpainting via Tractable Steering of Diffusion Models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.
- Liu, A.; and Van den Broeck, G. 2021. Tractable regularization of probabilistic circuits. *Advances in Neural Information Processing Systems*, 34: 3558–3570.
- Liu, A.; Zhang, H.; and Van den Broeck, G. 2023. Scaling Up Probabilistic Circuits by Latent Variable Distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Liu, X.; Liu, A.; Van den Broeck, G.; and Liang, Y. 2023. Understanding the distillation process from deep generative models to tractable probabilistic circuits. In *International Conference on Machine Learning*, 21825–21838. PMLR.
- Loconte, L.; Mari, A.; Gala, G.; Peharz, R.; de Campos, C.; Quaeghebeur, E.; Vessio, G.; and Vergari, A. 2024a. What is the Relationship between Tensor Factorizations and Circuits (and How Can We Exploit it)? *arXiv preprint arXiv:2409.07953*.
- Loconte, L.; Mengel, S.; and Vergari, A. 2024. Sum of Squares Circuits. *arXiv:2408.11778*.
- Loconte, L.; Sladek, A. M.; Mengel, S.; Trapp, M.; Solin, A.; Gillis, N.; and Vergari, A. 2024b. Subtractive Mixture Models via Squaring: Representation and Learning. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.
- Mari, A.; Vessio, G.; and Vergari, A. 2023. Unifying and Understanding Overparameterized Circuit Representations via Low-Rank Tensor Decompositions. In *The 6th Workshop on Tractable Probabilistic Modeling*.
- Martens, J.; and Medabalimi, V. 2014. On the expressive efficiency of sum product networks. *arXiv preprint arXiv:1411.7717*.
- Peharz, R.; Gens, R.; Pernkopf, F.; and Domingos, P. 2016. On the latent variable interpretation in sum-product networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(10): 2030–2044.
- Peharz, R.; Lang, S.; Vergari, A.; Stelzner, K.; Molina, A.; Trapp, M.; Van den Broeck, G.; Kersting, K.; and Ghahramani, Z. 2020a. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *International Conference on Machine Learning*, 7563–7574. PMLR.

- Peharz, R.; Vergari, A.; Stelzner, K.; Molina, A.; Shao, X.; Trapp, M.; Kersting, K.; and Ghahramani, Z. 2020b. Random sum-product networks: A simple and effective approach to probabilistic deep learning. In *Uncertainty in Artificial Intelligence*, 334–344. PMLR.
- Perez-Garcia, D.; Verstraete, F.; Wolf, M.; and Cirac, J. 2007. Matrix product state representations. *Quantum Information & Computation*, 7(5): 401–430.
- Pipatsrisawat, K.; and Darwiche, A. 2008. New Compilation Languages Based on Structured Decomposability. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 517–522.
- Poon, H.; and Domingos, P. 2011. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 689–690. IEEE.
- Rahman, T.; Kothalkar, P.; and Gogate, V. 2014. Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of chow-liu trees. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, 630–645. Springer.
- Rooshenas, A.; and Lowd, D. 2014. Learning sum-product networks with direct and indirect variable interactions. In *International Conference on Machine Learning*, 710–718. PMLR.
- Rudi, A.; and Ciliberto, C. 2021. PSD representations for effective probability models. *Advances in Neural Information Processing Systems*, 34: 19411–19422.
- Sladek, A. M.; Trapp, M.; and Solin, A. 2023. Encoding Negative Dependencies in Probabilistic Circuits. In *The 6th Workshop on Tractable Probabilistic Modeling*.
- Tsuchida, R.; Ong, C. S.; and Sejdinovic, D. 2023. Squared Neural Families: A New Class of Tractable Density Models. *Advances in Neural Information Processing Systems*, 36.
- Tsuchida, R.; Ong, C. S.; and Sejdinovic, D. 2024. Exact, Fast and Expressive Poisson Point Processes via Squared Neural Families. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20559–20566.
- Valiant, L. G. 1979. Negation can be exponentially powerful. In *Proceedings of the eleventh annual ACM symposium on theory of computing*, 189–196.
- Vergari, A.; Choi, Y.; Liu, A.; Teso, S.; and Van den Broeck, G. 2021. A compositional atlas of tractable circuit operations for probabilistic inference. *Advances in Neural Information Processing Systems*, 34: 13189–13201.
- Wang, B.; and Kwiatkowska, M. 2023. Compositional Probabilistic and Causal Inference using Tractable Circuit Models. In Ruiz, F.; Dy, J.; and van de Meent, J.-W., eds., *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 9488–9498. PMLR.
- Wang, B.; Wicker, M. R.; and Kwiatkowska, M. 2022. Tractable Uncertainty for Structure Learning. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 23131–23150. PMLR.
- Yu, Z.; Trapp, M.; and Kersting, K. 2023. Characteristic Circuits. *Advances in Neural Information Processing Systems*, 36.
- Zhang, H.; Dang, M.; Peng, N.; and Van den Broeck, G. 2023. Tractable Control for Autoregressive Language Generation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
- Zhang, H.; Holtzen, S.; and Broeck, G. 2020. On the relationship between probabilistic circuits and determinantal point processes. In *Conference on Uncertainty in Artificial Intelligence*, 1188–1197. PMLR.
- Zhang, H.; Juba, B.; and Van den Broeck, G. 2021. Probabilistic generating circuits. In *International Conference on Machine Learning*, 12447–12457. PMLR.
- Zhao, H.; Poupart, P.; and Gordon, G. J. 2016. A unified approach for learning the parameters of sum-product networks. *Advances in neural information processing systems*, 29.