Negative Mixture Models via Squaring: Representation and Learning

Lorenzo Loconte
University of Edinburgh, UK

Stefan Mengel
Univ. Artois, CNRS, FR

Nicolas Gillis
Université de Mons, BE

Antonio Vergari
University of Edinburgh, UK

Finite Mixture Models

A finite mixture model (MM) with K components p_i encodes a distribution p over variables $\mathbf{X}$ as

$$p(\mathbf{X}) = \sum_{i=1}^{K} \theta_i \, p_i(\mathbf{X}), \quad \sum_{i=1}^{K} \theta_i = 1, \quad \theta_i \geq 0.$$

- Relaxing the convex combination constraint on θ_i can potentially yield more expressive MMs with fewer parameters.
- Learning by allowing $\theta_i < 0$ while modeling a valid distribution supporting tractable integration is hard: *e.g., requires component-tailored constraints [1, 4, 2]*

Squaring Mixture Models

A squared negative MM (NMM^2) encodes a (possibly unnormalized) distribution over variables $\mathbf{X}$ as

$$c(\mathbf{X})^2 = \left(\sum_{i=1}^{K} \theta_i c_i(\mathbf{X})\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{K} \theta_i\theta_j c_i(\mathbf{X}) c_j(\mathbf{X})$$

with $p(\mathbf{X}) = c(\mathbf{X})^2/Z$ and

$$Z = \int c(\mathbf{x})^2 d\mathbf{X} = \sum_{i=1}^{K}\sum_{j=1}^{K} \theta_i\theta_j \int c_i(\mathbf{x}) c_j(\mathbf{x}) d\mathbf{X},$$

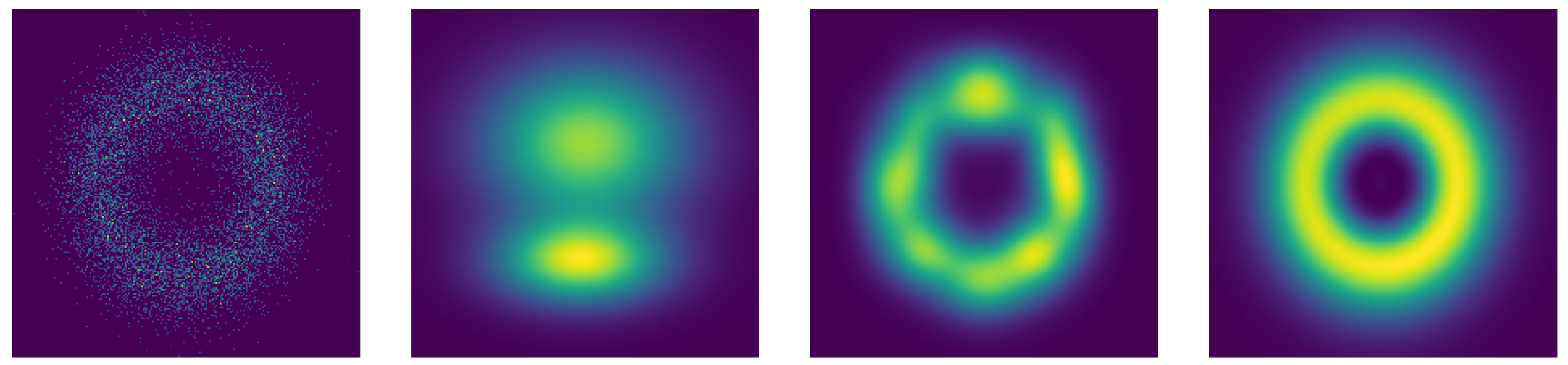
requiring evaluating $\binom{K}{2}$ integrals.

(Hierarchical) mixture models can be formalized under the framework of **probabilistic circuits** (PCs) [3].

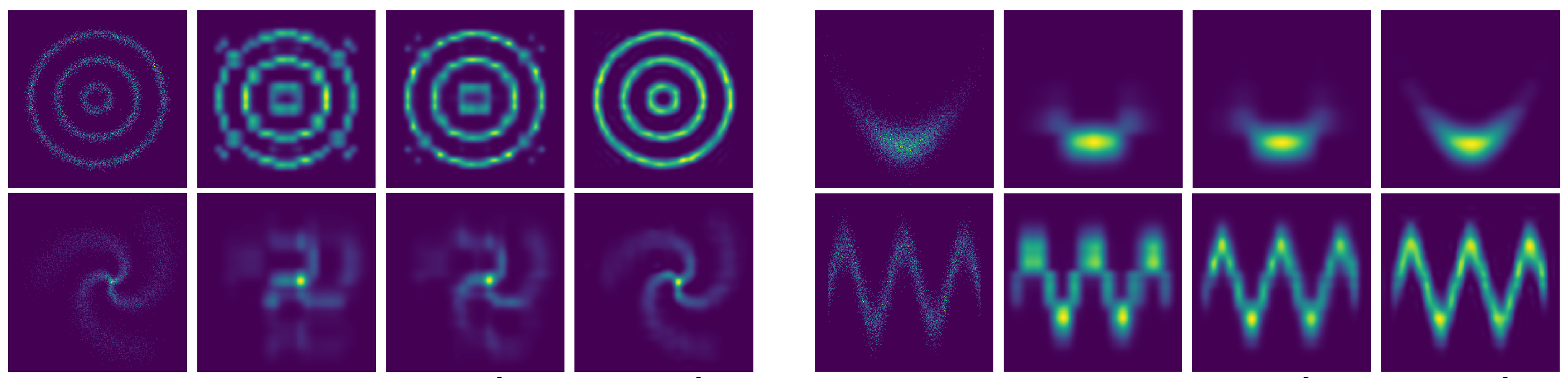
Theorem (Expressive efficiency)

There is a class of non-negative functions $\mathcal{F}$ over variables $\mathbf{X}$ that can be compactly represented as shallow squared NMMs (and hence squared non-monotonic PCs) but for which the smallest structured-decomposable monotonic PC computing any $F \in \mathcal{F}$ has size $2^{\Omega(|\mathbf{X}|)}$.

TL;DR: *"We build a framework for **deep mixture models with negative parameters**, and prove their increased expressiveness both theoretically and empirically."*

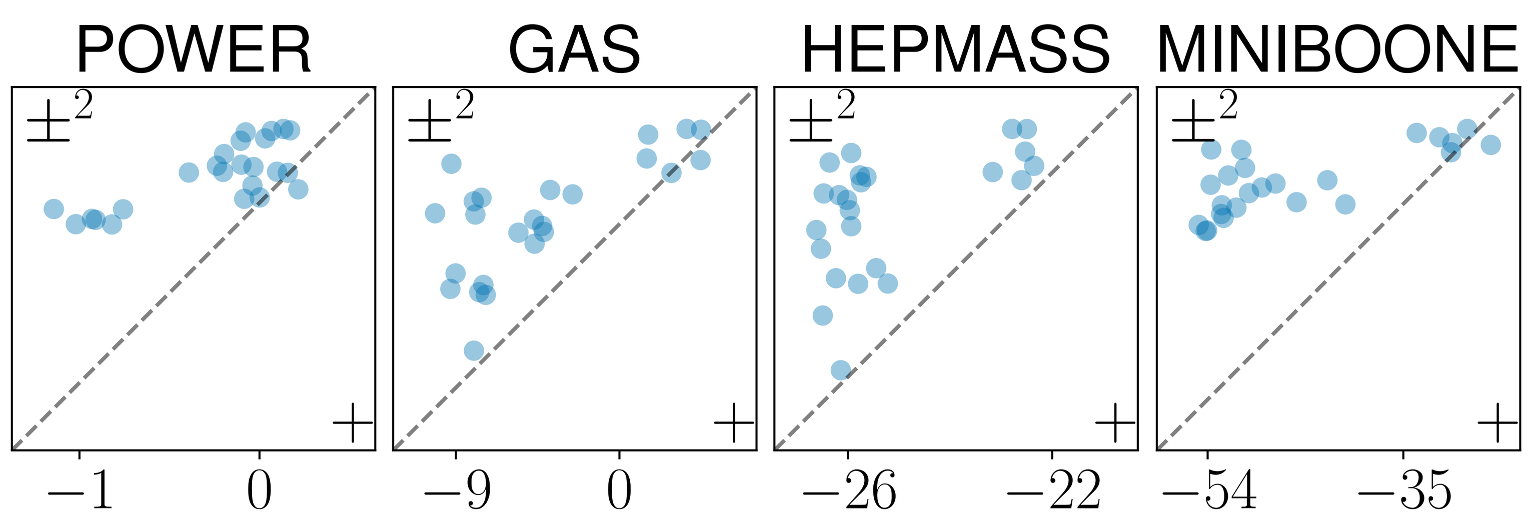


Samples | GMM $(K=2)$ | GMM $(K=16)$ | NGMM^2 $(K=2)$



Ground Truth | MM | MM^2 | NMM^2 | Ground Truth | MM | MM^2 | NMM^2

Negative parameters increase MMs expressiveness.



Better density estimation with the same model size.

Squaring Hierarchical Mixtures

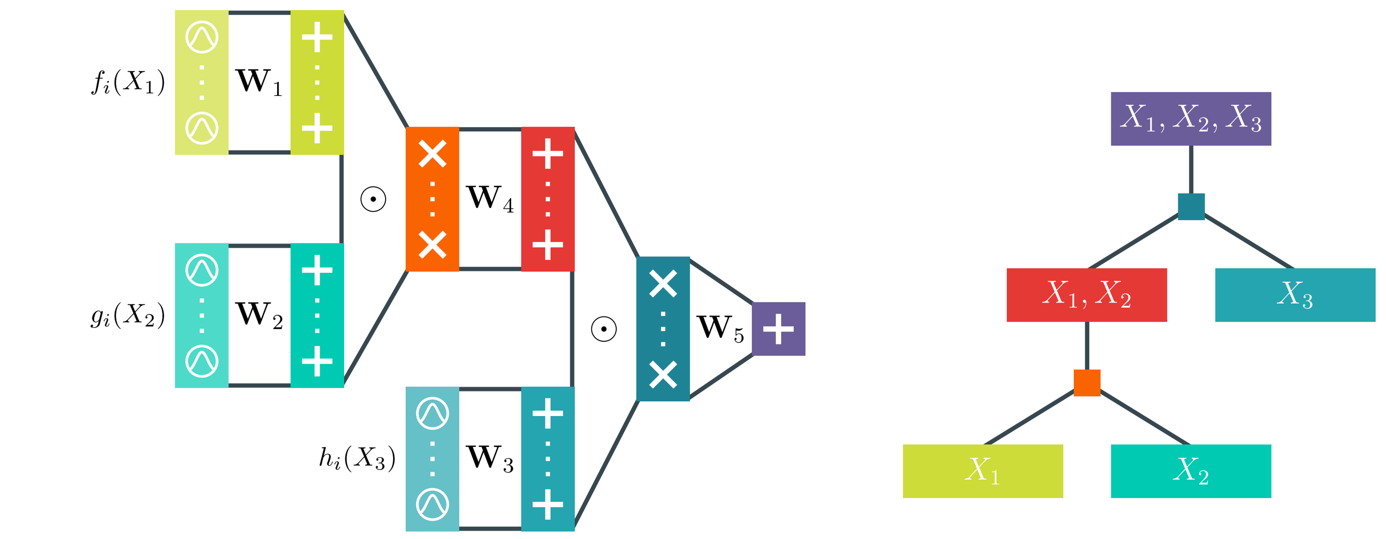
Squaring circuits such that marginalizing variables would still be efficient is possible if

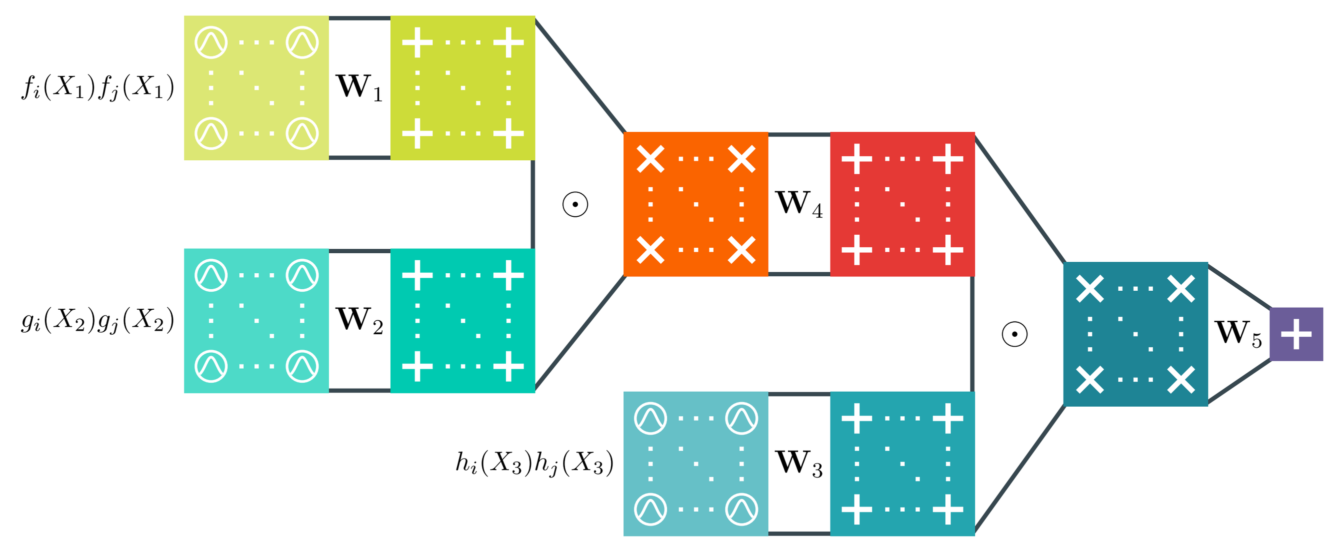
1. they induce a tree-shaped variable partitioning
2. products of input units are efficiently integrable: *e.g., Gaussian, polynomials, splines*

Squaring *tensorized circuits* is easy

e.g., a sum-product layer computes $\mathbf{c}_o = \mathbf{W}_o \left(\mathbf{c}_l \odot \mathbf{c}_r\right)$

$\implies$ *its squaring computes* $\mathbf{c}_o^2 = \mathbf{W}_o \left(\mathbf{c}_l^2 \odot \mathbf{c}_r^2\right) \mathbf{W}_o^T$





References

[1] R Jiang, MJ Zuo, and H-X Li. "Weibull and inverse Weibull mixture models allowing negative weights". In: *Reliability Engineering & System Safety* 66.3 (1999), pp. 227–234.

[2] Guillaume Rabusseau and François Denis. "Learning negative mixture models by tensor decompositions". In: *arXiv preprint arXiv:1403.4224* (2014).

[3] Antonio Vergari et al. "A Compositional Atlas of Tractable Circuit Operations for Probabilistic Inference". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 13189–13201.

[4] Baibo Zhang and Changshui Zhang. "Finite mixture models with negative components". In: *Machine Learning and Data Mining in Pattern Recognition: 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005. Proceedings 4*. Springer. 2005, pp. 31–41.