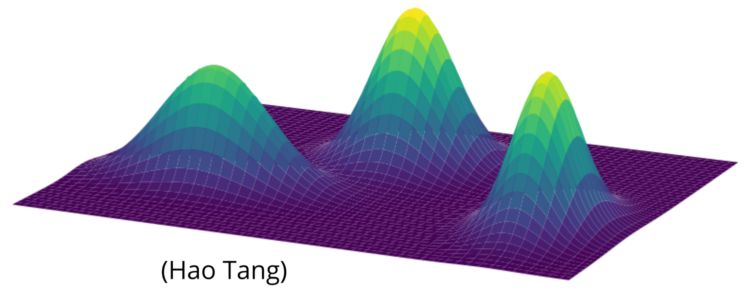
Subtractive Mixture Models via Squaring: Representation and Learning

Paper Code

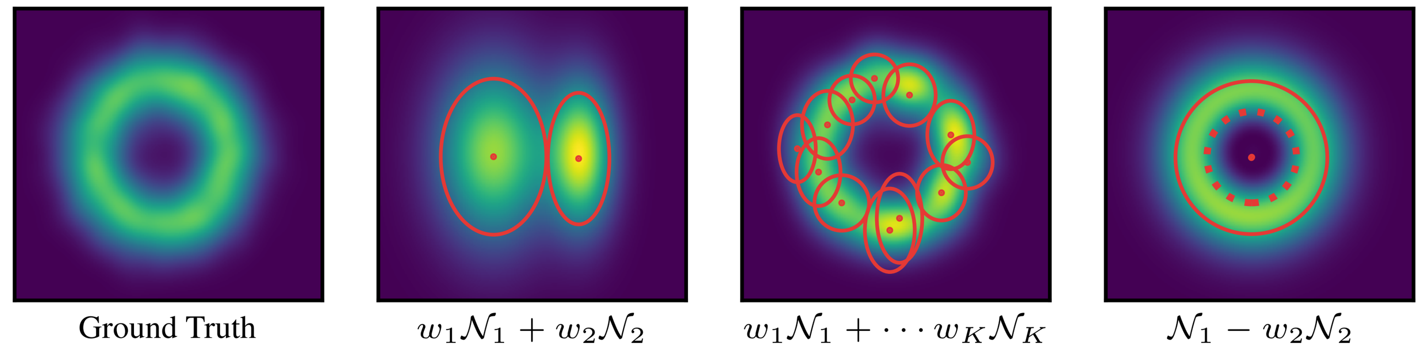
Lorenzo Loconte, University of Edinburgh, UK
Aleksanteri M. Sladek, Aalto University, FI
Stefan Mengel, University of Artois, CNRS, CRIL, FR
Martin Trapp, Aalto University, FI
Arno Solin, Aalto University, FI
Nicolas Gillis, Université de Mons, BE
Antonio Vergari, University of Edinburgh, UK

0 Mixture models

$$p(\mathbf{X}) = \sum_{i=1}^{K} w_i\, p_i(\mathbf{X}) \quad \text{subject to} \quad w_i \geq 0 \quad \sum_{i=1}^{K} w_i = 1$$



✗ components can only be added together!



Fewer components with subtractions

Questions? ...**Contributions!** 1 2 3

1 How to learn subtractive mixture models?

$$p(\mathbf{X}) = \sum_{i=1}^{K} w_i\, p_i(\mathbf{X}) \qquad w_i \in \mathbb{R}$$

How to ensure $p(\mathbf{X})$ is non-negative?

$\Longrightarrow$ Impose ad-hoc constraints over the parameters

✗ challenging to derive in closed-form [1][2][3]

2 How much more expressive are they?

with respect to traditional additive-only mixtures

3 What is their relationship with other models?

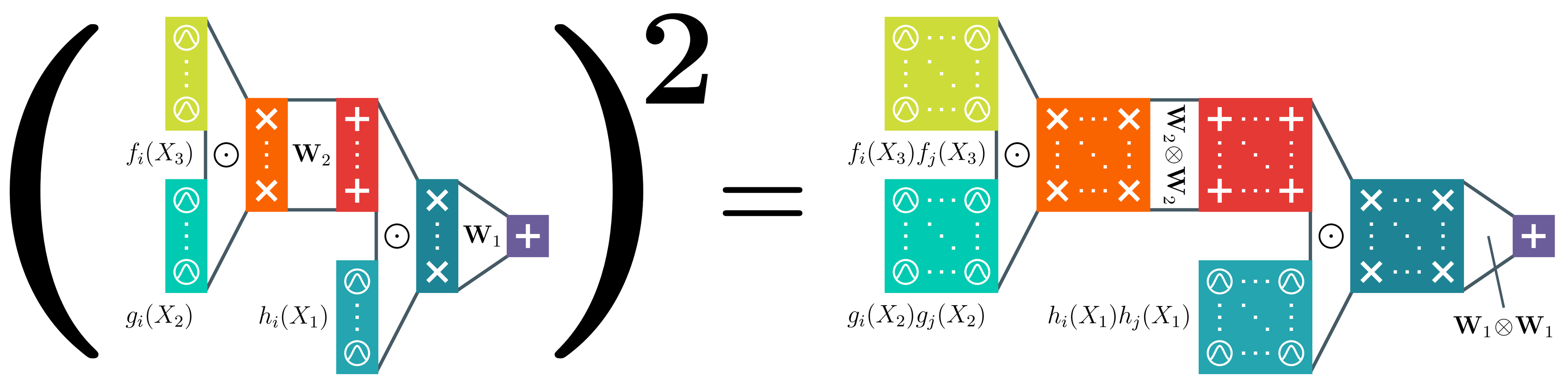
understanding why they are expressive ...

... and why they support tractable inference

TL; DR

"We learn exponentially more expressive mixture models with subtractions, by squaring deep tensorized mixtures"

Learning deep subtractive mixtures by squaring layers of a deep circuit

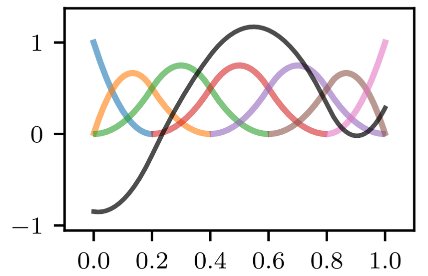


1 Squaring mixtures ...

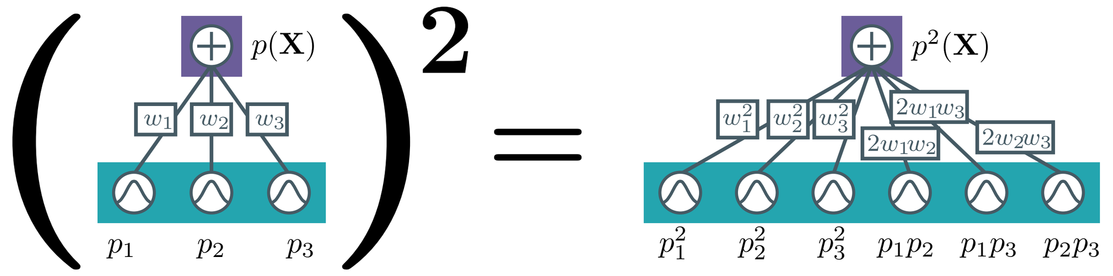
$$p(\mathbf{X}) \propto \left(\sum_{i=1}^{K} w_i\, p_i(\mathbf{X})\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j\, p_i(\mathbf{X}) p_j(\mathbf{X})$$

Renormalization:

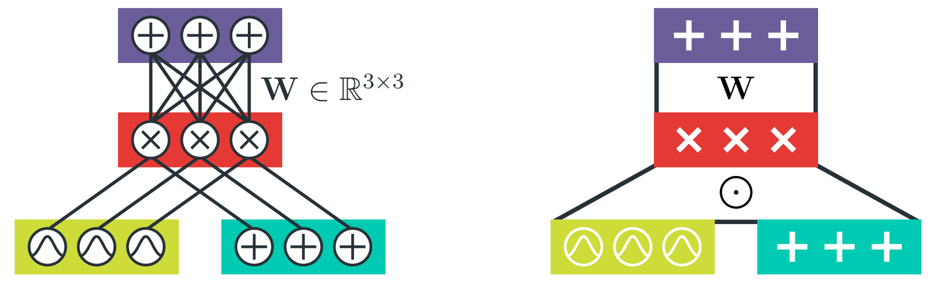
$$Z = \sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j \int p_i(\mathbf{X}) p_j(\mathbf{X}) \mathrm{d}\mathbf{X}$$



Tractable marginalization is supported by exponential families [2] and splines components



... by squaring circuits



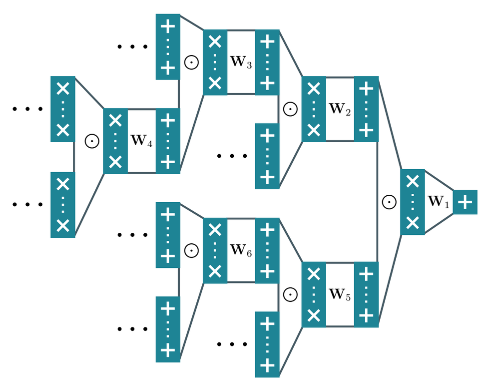
Build deep mixtures with layers as "Lego blocks"

2

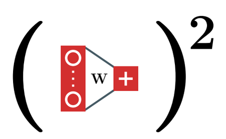
Theorem. exponential separation [4] [5]

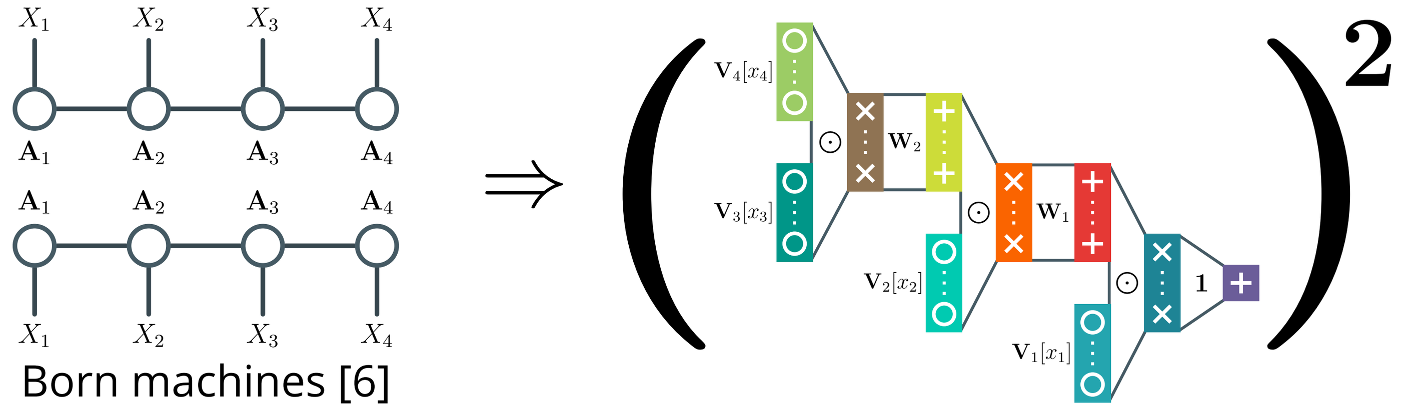
There is a class of distributions $\mathcal{F}$ over variables $\mathbf{X}$ that can be compactly represented as a shallow squared mixture with negative weights, but the smallest structured decomposable additive-only mixture of any depth computing any $F \in \mathcal{F}$ has size $2^{\Omega(|\mathbf{X}|)}$.

Deep additive-only mixtures



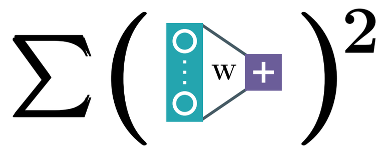
Squared subtractive mixture model





Born machines [6]

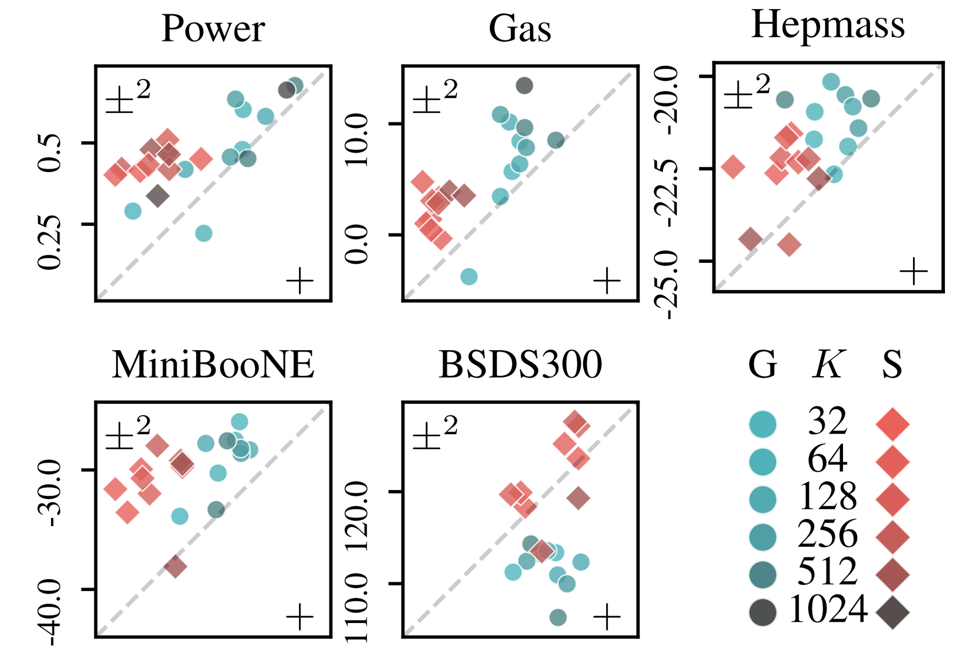
$$p(\mathbf{x}) \propto \kappa(\mathbf{x})^\top \mathbf{A}\, \kappa(\mathbf{x}) \quad \Rightarrow \quad \sum\left(\ \right)^2$$



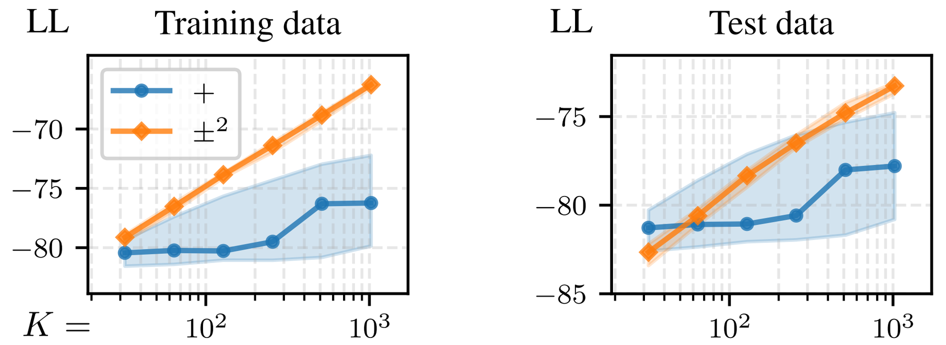
PSD models [7]

3 Understanding the expressiveness of other models in a unifying framework

Density estimation



GPT2 distillation [8]



References

[1] B. Zhang and C. Zhang. "Finite mixture models with negative components". In: *MLDM*. Springer. 2005, pp. 31–41.
[2] G. Rabusseau and F. Denis. "Learning negative mixture models by tensor decompositions". In: *arXiv preprint arXiv:1403.4224* (2014).
[3] R. Jiang, M. J. Zuo, and H. Li. "Weibull and inverse Weibull mixture models allowing negative weights". In: *Reliability Engineering & System Safety* 66.3 (1999), pp. 227–234.
[4] J. Martens and V. Medabalimi. "On the expressive efficiency of sum product networks". In: *arXiv preprint arXiv:1411.7717* (2014).
[5] A. de Colnet and S. Mengel. "A Compilation of Succinctness Results for Arithmetic Circuits". In: *KR*. 2021, pp. 205–215.
[6] I. Glasser et al. "Expressive power of tensor-network factorizations for probabilistic modeling". In: *NeurIPS*. Curran Associates, Inc., 2019, pp. 1498–1510.
[7] A. Rudi and C. Ciliberto. "PSD Representations for Effective Probability Models". In: *NeurIPS*. Curran Associates, Inc., 2021, pp. 19411–19422.
[8] H. Zhang et al. "Tractable Control for Autoregressive Language Generation". In: *ICML*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 40932–40945.