Sparse Probabilistic Circuits via Pruning and Growing

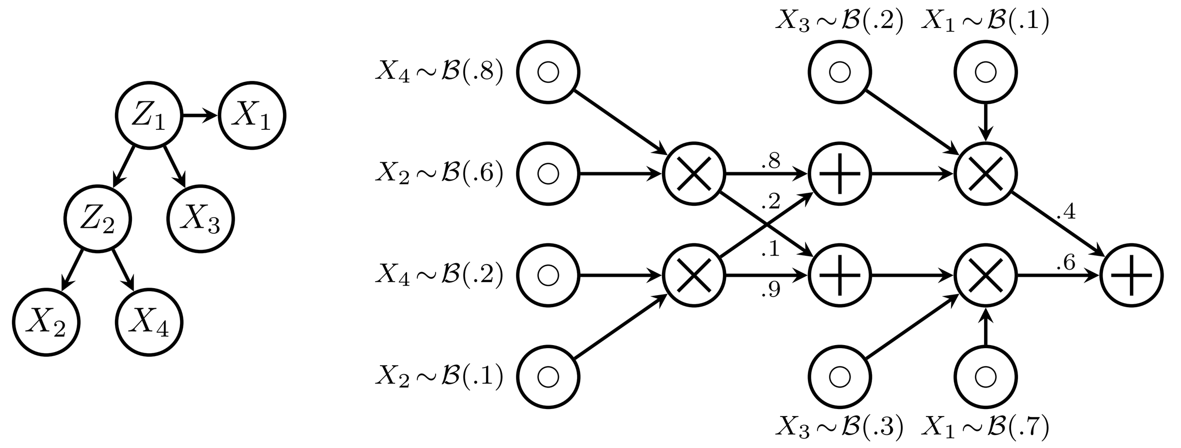
Selected as Oral

Meihua Dang Anji Liu Guy Van den Broeck
University of California, Los Angeles

Probabilistic Circuits

Probabilistic circuits (PCs) encode a probability distribution $p_n(\mathrm{x})$ defined recursively as follows.

$$p_n(\boldsymbol{x}) := \begin{cases} f_n(\boldsymbol{x}) & \text{if } n \text{ is an input unit,} \\ \prod_{c \in \mathsf{in}(n)} p_c(\boldsymbol{x}) & \text{if } n \text{ is a product unit,} \\ \sum_{c \in \mathsf{in}(n)} \theta_{c|n} \cdot p_c(\boldsymbol{x}) & \text{if } n \text{ is a sum unit,} \end{cases}$$



An example PC defined over 4 random variables and an equivalent Bayesian network

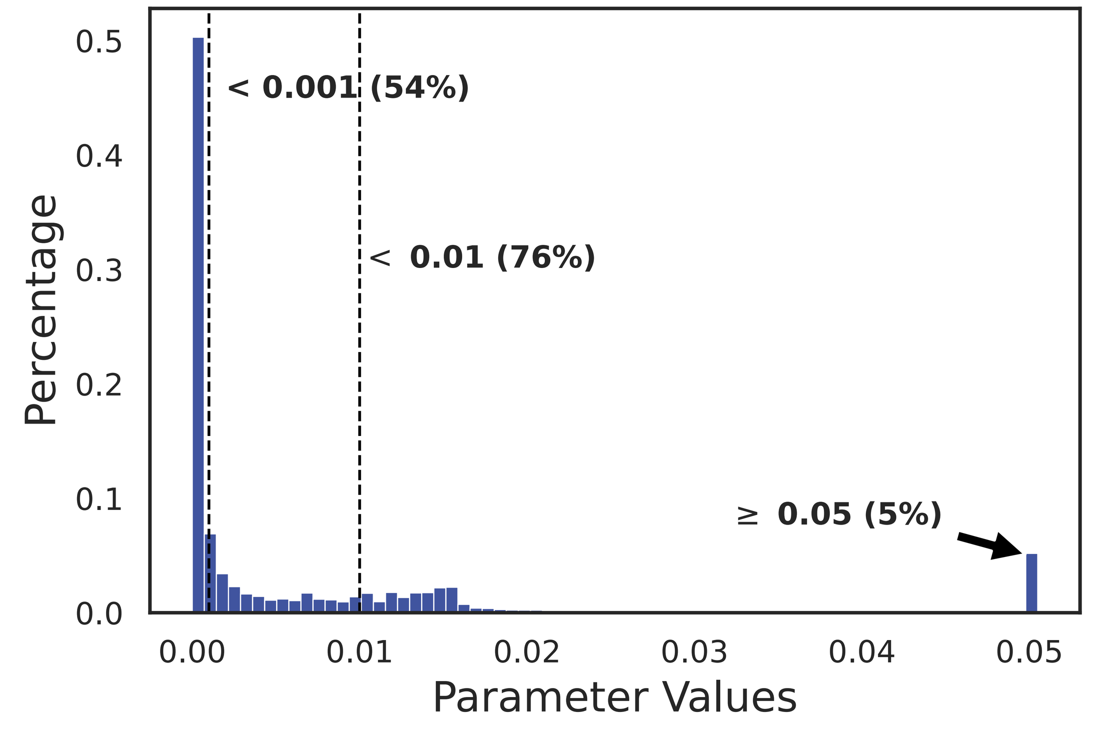
We study *smooth* and *decomposable* PCs.

Two perspectives:

1. *Computational graph*: inference as forward propagation.
2. *Probability semantics*: parameter value represents local conditional probability.

Motivation: Fully-connected Layers are Sparsely Used

As we *scale up learning PCs*, the performance of PCs plateaus as model size increases. Thus, we need to better utilize the available capacity.



Histograms of a SoTA PC on MNIST, 95% of the parameters have close-to-zero values

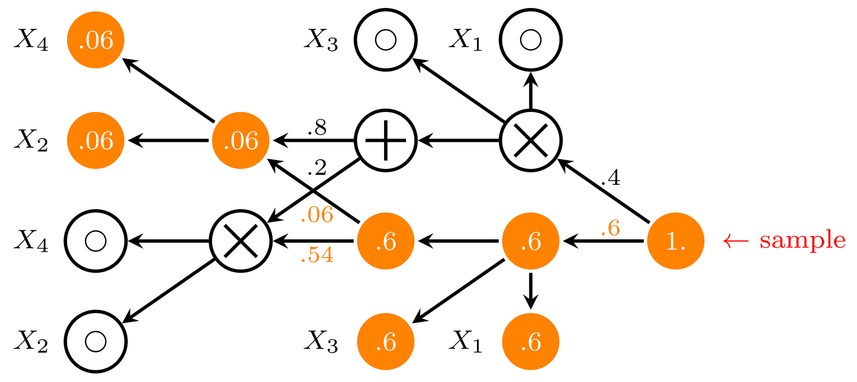
Though PC structures have fully-connected parameter layers, the parameter values are only sparsely used.

Methods: Pruning Parameters by Probability Semantics

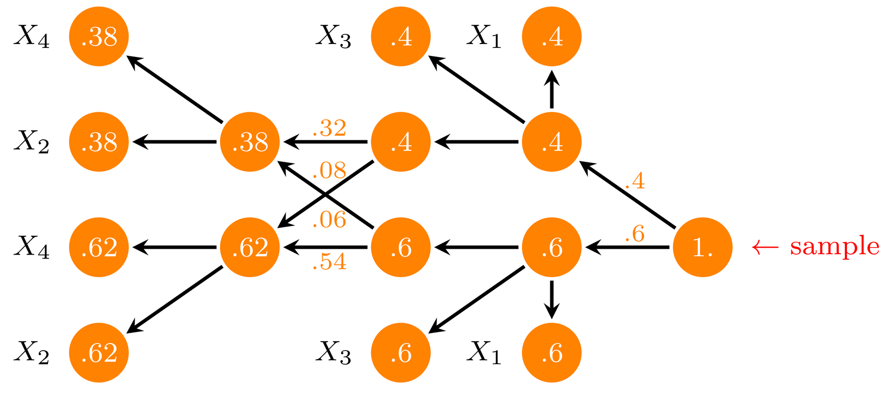
By pruning away "unimportant" parameters, it is possible to significantly reduce model size while maximally retaining model expressiveness.

Intuition: when drawing a sample from PC, if a parameter is seldom visited in the generative sampling process, removing it will not significantly affect the PC's distribution.

Sampling from PC



Sampling as a backward propagation



The probability of each unit being sampled

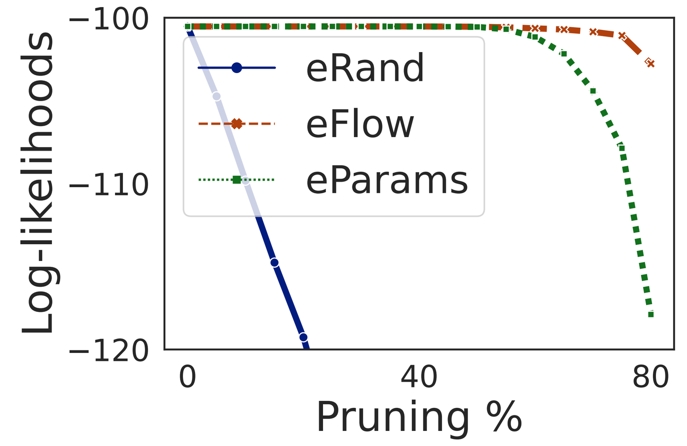
Circuit Flows

The circuit flow of unit n on example x is the probability that n will be visited during the sampling procedure *conditioned on x being sampled.*

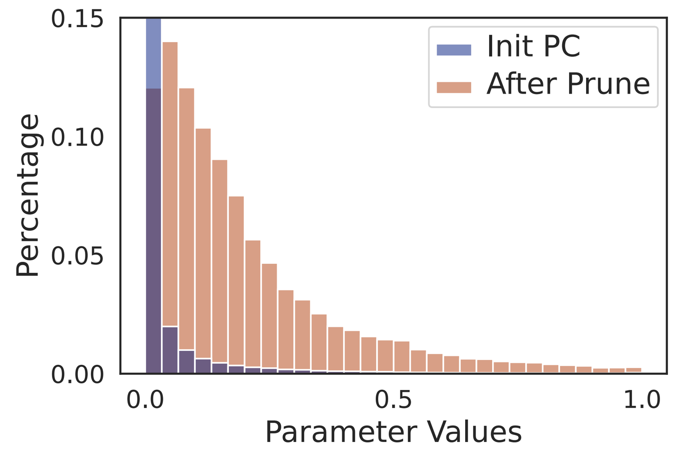
$$\mathrm{F}_n(\boldsymbol{x}) = \begin{cases} 1 & \text{if } n \text{ is the root unit,} \\ \sum_{m \in \mathsf{out}(n)} \mathrm{F}_m(\boldsymbol{x}) & \text{if } n \text{ is a sum unit,} \\ \sum_{m \in \mathsf{out}(n)} \frac{\theta_{n|m} \cdot p_n(\boldsymbol{x})}{p_m(\boldsymbol{x})} \cdot \mathrm{F}_m(\boldsymbol{x}) & \text{if } n \text{ is a product unit.} \end{cases}$$

Therefore, we prune edges with the smallest aggregate circuit flow.

Pruning by Circuit Flows



Pruning by circuit flows can prune up to 80% of the parameters without much log-likelihoods decrease.



The parameter values take higher significance after pruning.

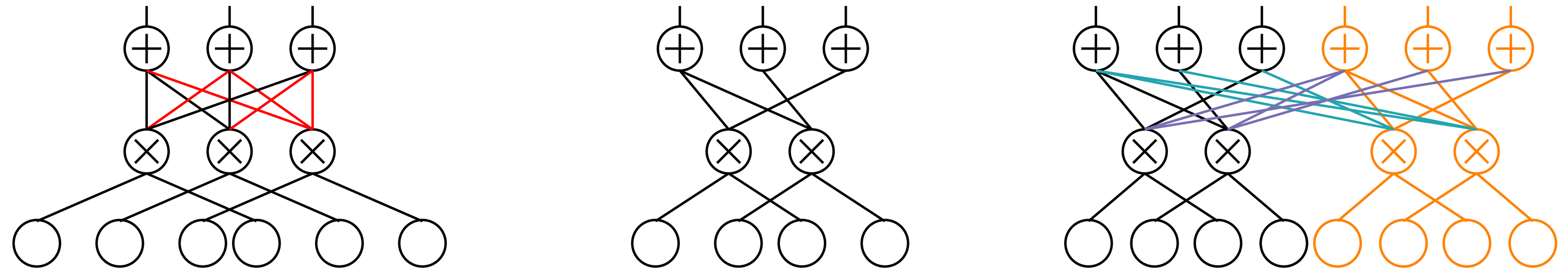
(Theorem) The log-likelihood drop by pruning away multiple edges is bounded and approximated by

$$\Delta\mathcal{LL}(\mathcal{D},\mathcal{C},\mathcal{E}) \le -\frac{1}{|\mathcal{D}|}\sum_{\boldsymbol{x}} \log\Big(1 - \sum_{(n,c)\in\mathcal{E}} \mathrm{F}_{n,c}(\boldsymbol{x})\Big) \approx \frac{1}{|\mathcal{D}|}\sum_{(n,c)\in\mathcal{E}} \mathrm{F}_{n,c}(\mathcal{D}).$$

Learning Sparse PCs Structures

Learning Sparse PCs

1. Growing operation copies parameters and injects noise
2. Apply pruning, growing, EM iteratively to learn structures



(a) PC with fully connected layers (b) PC after pruning operation (c) PC after growing operation

Density Estimation Benchmarks

- Image datasets: MNIST, EMNIST, FashionMNIST
- Character-level language modeling task: PTB

Baselines: PC learners (HCLT, RatSPN), VAEs, flow-based models

Table 1: Density estimation performance on MNIST-family datasets in test set bpd.

Dataset	Sparse PC (ours)	HCLT	RatSPN	IDF	BitSwap	BB-ANS	McBits
MNIST	**1.14**	1.20	1.67	1.90	1.27	1.39	1.98
EMNIST(MNIST)	**1.52**	1.77	2.56	2.07	1.88	2.04	2.19
EMNIST(Letters)	**1.58**	1.80	2.73	1.95	1.84	2.26	3.12
EMNIST(Balanced)	**1.60**	1.82	2.78	2.15	1.96	2.23	2.88
EMNIST(ByClass)	**1.54**	1.85	2.72	1.98	1.87	2.23	3.14
FashionMNIST	**3.27**	3.34	4.29	3.47	3.28	3.66	3.72

Table 2: Character-level language modeling results on Penn Tree Bank in test set bpd.

Dataset	Sparse PC (ours)	Bipartite flow [42]	AF/SCF [48]	IAF/SCF [48]
Penn Tree Bank	**1.35**	1.38	1.46	1.63

Main Message

1. Most capacity in existing large PC structures is wasted: **fully-connected parameter layers are only sparsely used**.
2. We prune the unimportant parameters based on their probability semantics: **pruning away low probability sub-structures**.
3. The main idea can be generalized to compress other deep generative models or neural networks.

Scan this code →

Acknowledgements: This work was funded in part by the DARPA Perceptually-enabled Task Guidance (PTG) Program under contract number HR00112220005, NSF grants #IIS-1943641, #IIS-1956441, #CCF-1837129, Samsung, CISCO, a Sloan Fellowship, a UCLA Samueli Fellowship. We thank Honghua Zhang for proofreading and insightful comments on this paper's final version.