Scalable Learning of Probabilistic Circuits

Renato Lui Geh

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Master of Science

Program: Computer Science
Advisor: Prof. Denis Deratani Mauá, PhD

This work was supported by CNPq grant #133787/2019-2,
CAPES grant #88887.339583/2019-00 and EPECLIN FM-USP.

São Paulo
November 1, 2021

Scalable Learning of Probabilistic Circuits

Renato Lui Geh

This is the original version of the thesis prepared by candidate Renato Lui Geh, as submitted to the Examining Committee.

Acknowledgements

Resumo

Renato Lui Geh. **Aprendizado Escalável de Circuitos Probabilísticos**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Palavras-chave: Circuitos probabilísticos. Aprendizado de máquina. Modelos probabilísticos.

Abstract

Renato Lui Geh. **Scalable Learning of Probabilistic Circuits**. Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Keywords: Probabilistic circuits. Machine learning. Probabilistic models.

Nomenclature

List of Symbols

σ Sigmoid function

List of Figures

List of Tables

List of Algorithms

Notation

We use the following notation throughout the work. Random variables are written in upper case (e.g. X, Y) and their values in lower case (e.g. x, y). We identify propositional variables with 0/1-valued random variables, and use them interchangeably. Sets of variables and their joint values are written in boldface (e.g. $\mathbf{X}$, $\mathbf{x}$). Given a Boolean formula f, we write $\langle f \rangle$ to denote its semantics, i.e. the Boolean function represented by f. For Boolean formulas f and g, we write $f \equiv g$ if they are logically equivalent, that is, if $\langle f \rangle = \langle g \rangle$; we abuse notation and write $\phi \equiv f$ to indicate that $\phi = \langle f \rangle$ for a Boolean function ϕ.

Contents

Chapter 1

Introduction

When reasoning about the world, rarely can we find a realistic model that perfectly subsumes all of the needed relationships for flawless prediction. As such, the presence of a reliable uncertainty quantifier in intelligent systems is essential in developing performant yet diagnostible agents. This is made explicitly clear in the case of high-risk settings, such as autonomous vehicles or automated power plant systems, where a wrong prediction could cause disastrous consequences. A well known example in the case of the former is obstacle avoidance: while the agent should be capable of accurately identifying obstructions in its way during normal conditions, so should it be able to identify its own lack of confidence in high uncertainty situations like ones brought by environmental factors, such as severe blizzard or heavy rain. In these situations where predictions are highly unreliable, the safest option might be for the agent to first identify its uncertainty, and second to reach out for human help. Other interesting applications of uncertainty quantification include out-of-distribution detection, which in our previous example could be visualized as the agent identifying a human's driving as abnormally irregular (due to inebriation, infirmity, etc.) and appropriately taking control of the vehicle before a potential accident takes place.

One popular approach to quantifying uncertainty is through probability theory. By abstracting the world as a probability distribution with a finite number of observable random variables that encode a possibly incomplete knowledge of the environment, we are, in theory, able to answer a diverse set of complex queries as long as we have access to the (approximate) true data distribution. Practice is far different from theory, however, as most machine learning models either lie within a very limited range in the tractability spectrum in terms of inference (Vergari, Choi, *et al.*, 2020) or are too simplistic for complex real-world problems. Besides, although the majority of recent advances in deep learning claim some probabilistic meaning from the model's output, they are often uncalibrated distributions, a result of focusing on maximizing predictive accuracy at the expense of predictive uncertainty (Guo *et al.*, 2017; Ovadia *et al.*, 2019; Chernikova *et al.*, 2019), ultimately producing overconfident and peculiar results (Szegedy *et al.*, 2013; Wei *et al.*, 2018; Su *et al.*, 2019; Chernikova *et al.*, 2019). Crucially, mainstream deep models (i.e. standard neural networks) usually optimize a conditional distribution over the to-be-predicted random variables – and are thus often called *discriminative* models – and do

not model the actual joint distribution of the data, limiting inference capabilities and uncertainty estimation.

In contrast, *generative* models seek to extract information from the joint (in varying capacities), and have lately seen a sharp rise in interest within deep learning. Despite this, most popular models do not admit either exact or tractable querying of key inference scenarios. For instance, although Generative Adversarial Networks (GANs) allow for efficient sampling (Goodfellow *et al.*, 2014), basic queries such as likelihood or marginals are outside of their capabilities. Similarly, Normalizing Flows (NF) also permit access to efficient sampling, with the added feature of computing likelihoods, but are severely limited by the base distribution when it comes to discrete data (Rezende and Mohamed, 2015; Papamakarios *et al.*, 2021) albeit recent works on discretizing NFs have shown empirically good results (Lippe and Gavves, 2021; Ziegler and Rush, 2019). Variational Auto-Encoders (VAEs) are (under certain conditions) a generalization of NFs (Yu, 2020; Gritsenko *et al.*, 2019) with known extensions for categorical data (Rolfe, 2017; Vahdat, Macready, *et al.*, 2018; Vahdat, Andriyash, *et al.*, 2018), but only permit access to sampling and an upper-bound on the likelihood, with the latter available only after solving a complex optimization task (Kingma and Welling, 2014).

Despite the impressive achievements of the aforementioned generative models on realistically producing samples consistent with evidence, in none of the previous models are complex queries like structured prediction under partial observations, *maximum a posteriori* (MAP), conditional or marginal probabilities tractable. An obvious alternative would be Probabilistic Graphical Models (PGMs), although they too suffer from intractability when dealing with high treewidth networks (R. Dechter, 1998; Koller and Friedman, 2009), severely limiting expressivity. Instead, we draw our attention to an expressive class of models that subsumes several families of probabilistic models with tractable inference capabilities.

Probabilistic Circuits (PCs) define a superclass of probabilistic models distincly specified by recursive compositions of distributions through graphical formalisms. Vaguely speaking, PCs are computational graphs akin to neural networks, but whose network structure and computational units abide by special constraints. Within these specific conditions span a wide range of subclasses, each establishing a distinct set of restrictions on their structure in order to enable different segments within the tractability spectrum. As an example, Sum-Product Networks (SPNs, Poon and P. Domingos, 2011) are usually loosely defined over a couple of constraints: namely *smoothness* and *decomposability*, which in turn enables likelihood, marginal and conditional computations. Arithmetic Circuits (ACs, Darwiche, 2003) add *determinism* to the mix, allowing for tractable computation of MAP probabilities. Similarly, Cutset Networks (CNets, Rahman *et al.*, 2014) employ the same constraints as ACs, but accept more expressive distributions as part of their computational units. Probabilistic Sentential Decision Diagrams (PSDDs, Kisa *et al.*, 2014), Probabilistic Decision Graphs (PDGs, Jaeger, 2004) and And/Or-Graphs (AOGs, Rina Dechter and Mateescu, 2007) all require *smoothness* and *determinism*, but also call for a stronger version of *decomposability*, permitting all queries previously mentioned as well as computation of the Kullback-Leibler divergence and expectation between two circuits (Choi *et al.*, 2020). Usually, PCs represent the joint distribution of the data, although they are sufficiently expressive for generative *and* discriminative modeling (Khosravi *et al.*, 2019; Rashwan

et al., 2018; Rooshenas and Lowd, 2016; Gens and P. Domingos, 2012; Shao *et al.*, 2020). In this dissertation though, we shall focus on the generative side of PCs.

While inference is usually straightforward, as we shall see in ??, learning the structure of PCs so that they obey the needed structural restrictions requires either careful handcrafted architectures (Poon and P. Domingos, 2011; Cheng *et al.*, 2014; Nath and P. M. Domingos, 2016) or usually involves running costly (in)dependence tests over most (if not all) variables (Gens and P. Domingos, 2013; Jaini *et al.*, 2018; Vergari, Mauro, *et al.*, 2015; Di Mauro *et al.*, 2017), which can become prohibitive in higher dimensions. Alternatively, some learning algorithms resort to structure preserving iterative methods to grow a PC that already initially satisfies desired constraints, adding complexity to the underlying distribution at each iteration (Liang *et al.*, 2017; Dang *et al.*, 2020). However, these can take several iterations until visible improvement and often take several minutes for each iteration when the circuit is big. Common techniques used in deep learning for generating scalable architectures for neural network also pose a problem, as the nature of the needed structural constraints make for sparse computational graphs. To circumvent these issues, work on scaling PCs to higher dimensions has focused mainly on random architectures, with competitive results (Peharz, Vergari, *et al.*, 2020; Mauro *et al.*, 2021; Geh and Mauá, 2021; Peharz, Lang, *et al.*, 2020).

1.1 Contributions and Dissertation Outline

referenced structural constraints in passing

References

[Cheng *et al.* 2014] Wei-Chen Cheng, Stanley Kok, Hoai Vu Pham, Hai Leong Chieu, and Kian Ming A. Chai. "Language modeling with sum-product networks". In: *Fifteenth Annual Conference of the International Speech Communication Association*. 2014 (cit. on p. 3).

[Chernikova *et al.* 2019] A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim. "Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction". In: *2019 IEEE Security and Privacy Workshops*. 2019, pp. 132–137 (cit. on p. 1).

[Choi *et al.* 2020] YooJung Choi, Antonio Vergari, and Guy Van den Broeck. "Probabilistic circuits: a unifying framework for tractable probabilistic models". In: (2020). In preparation (cit. on p. 2).

[Dang *et al.* 2020] Meihua Dang, Antonio Vergari, and Guy Van den Broeck. "Strudel: learning structured-decomposable probabilistic circuits". In: *Proceedings of the 10th International Conference on Probabilistic Graphical Models*. PGM. 2020 (cit. on p. 3).

[Darwiche 2003] Adnan Darwiche. "A differential approach to inference in bayesian networks". In: *Journal of the ACM* 50.3 (2003), pp. 280–305 (cit. on p. 2).

[R. Dechter 1998] R. Dechter. "Bucket elimination: a unifying framework for probabilistic inference". In: *Learning in Graphical Models*. Ed. by Michael I. Jordan. Dordrecht: Springer Netherlands, 1998, pp. 75–104. isbn: 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_4. url: https://doi.org/10.1007/978-94-011-5014-9_4 (cit. on p. 2).

[Rina Dechter and Mateescu 2007] Rina Dechter and Robert Mateescu. "And/or search spaces for graphical models". In: *Artificial Intelligence* 171.2 (2007), pp. 73–106. issn: 0004-3702. doi: https://doi.org/10.1016/j.artint.2006.11.003. url: https://www.sciencedirect.com/science/article/pii/S000437020600138X (cit. on p. 2).

[Di Mauro *et al.* 2017] Nicola Di Mauro, Floriana Esposito, Fabrizio G. Ventola, and Antonio Vergari. "Alternative variable splitting methods to learn sum-product networks". In: *AI*IA 2017 Advances in Artificial Intelligence*. Ed. by Floriana Esposito, Roberto Basili, Stefano Ferilli, and Francesca A. Lisi. Cham: Springer International Publishing, 2017, pp. 334–346. isbn: 978-3-319-70169-1 (cit. on p. 3).

[Geh and Mauá 2021] Renato Lui Geh and Denis Deratani Mauá. "Learning probabilistic sentential decision diagrams under logic constraints by sampling and averaging". In: *Proceedings of The 37th Uncertainty in Artificial Intelligence Conference*. Proceedings of Machine Learning Research. PMLR, 2021 (cit. on p. 3).

[Gens and P. Domingos 2012] Robert Gens and Pedro Domingos. "Discriminative learning of sum-product networks". In: *Advances in Neural Information Processing Systems 25*. NIPS, 2012, pp. 3239–3247 (cit. on p. 3).

[Gens and P. Domingos 2013] Robert Gens and Pedro Domingos. "Learning the structure of sum-product networks". In: *Proceedings of the 30th International Conference on Machine Learning*. ICML. 2013, pp. 873–880 (cit. on p. 3).

[Goodfellow *et al.* 2014] Ian Goodfellow *et al.* "Generative adversarial nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680. url: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (cit. on p. 2).

[Gritsenko *et al.* 2019] Alexey A. Gritsenko, Jasper Snoek, and Tim Salimans. "On the relationship between normalising flows and variational- and denoising autoencoders". In: *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. url: https://openreview.net/forum?id=HklKEUUY%5C_E (cit. on p. 2).

[Guo *et al.* 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. "On calibration of modern neural networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1321–1330 (cit. on p. 1).

[Jaeger 2004] Manfred Jaeger. "Probabilistic decision graphs-combining verification and ai techniques for probabilistic inference". In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 12.1 supp (Jan. 2004), pp. 19–42. issn: 0218-4885. doi: 10.1142/S0218488504002564. url: https://doi.org/10.1142/S0218488504002564 (cit. on p. 2).

[Jaini *et al.* 2018] Priyank Jaini, Amur Ghose, and Pascal Poupart. "Prometheus: directly learning acyclic directed graph structures for sum-product networks". In: *International Conference on Probabilistic Graphical Models*. PGM. 2018, pp. 181–192 (cit. on p. 3).

[Khosravi *et al.* 2019] Pasha Khosravi, Yitao Liang, YooJung Choi, and Guy Van den Broeck. "What to expect of classifiers? reasoning about logistic regression with missing features". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 2716–2724. doi: 10.24963/ijcai.2019/377. url: https://doi.org/10.24963/ijcai.2019/377 (cit. on p. 2).

[Kingma and Welling 2014] Diederik P. Kingma and Max Welling. "Auto-encoding variational bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014. url: http://arxiv.org/abs/1312.6114 (cit. on p. 2).

[Kisa *et al.* 2014] Doga Kisa, Guy Van den Broeck, Arthur Choi, and Adnan Darwiche. "Probabilistic sentential decision diagrams". In: *Knowledge Representation and Reasoning Conference* (2014) (cit. on p. 2).

[Koller and Friedman 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. isbn: 0262013193 (cit. on p. 2).

[Liang *et al.* 2017] Yitao Liang, Jessa Bekker, and Guy Van den Broeck. "Learning the structure of probabilistic sentential decision diagrams". In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*. 2017 (cit. on p. 3).

[Lippe and Gavves 2021] Phillip Lippe and Efstratios Gavves. "Categorical normalizing flows via continuous transformations". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. url: https://openreview.net/forum?id=-GLNZeVDuik (cit. on p. 2).

[Mauro *et al.* 2021] Nicola Di Mauro, Gennaro Gala, Marco Iannotta, and Teresa M. A. Basile. "Random probabilistic circuits". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. 2021 (cit. on p. 3).

[Nath and P. M. Domingos 2016] Aniruddh Nath and Pedro M Domingos. "Learning tractable probabilistic models for fault localization". In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016 (cit. on p. 3).

[Ovadia *et al.* 2019] Yaniv Ovadia *et al.* "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019 (cit. on p. 1).

[Papamakarios *et al.* 2021] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64. url: http://jmlr.org/papers/v22/19-1028.html (cit. on p. 2).

[Peharz, Lang, *et al.* 2020] Robert Peharz, Steven Lang, *et al.* "Einsum networks: fast and scalable learning of tractable probabilistic circuits". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 7563–7574. url: https://proceedings.mlr.press/v119/peharz20a.html (cit. on p. 3).

[Peharz, Vergari, *et al.* 2020] Robert Peharz, Antonio Vergari, *et al.* "Random sum-product networks: a simple and effective approach to probabilistic deep learning". In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 115. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 334–344. url: https://proceedings.mlr.press/v115/peharz20a.html (cit. on p. 3).

[Poon and P. Domingos 2011] Hoifung Poon and Pedro Domingos. "Sum-product networks: a new deep architecture". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. 2011, pp. 337–346 (cit. on pp. 2, 3).

[Rahman *et al.* 2014] Tahrima Rahman, Prasanna Kothalkar, and Vibhav Gogate. "Cutset networks: a simple, tractable, and scalable approach for improving the accuracy of chow-liu trees". In: *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases*. 2014, pp. 630–645 (cit. on p. 2).

[Rashwan *et al.* 2018] Abdullah Rashwan, Pascal Poupart, and Chen Zhitang. "Discriminative training of sum-product networks by extended baum-welch". In: *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*. Vol. 72. Proceedings of Machine Learning Research. 2018, pp. 356–367 (cit. on p. 2).

[Rezende and Mohamed 2015] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1530–1538. url: https://proceedings.mlr.press/v37/rezende15.html (cit. on p. 2).

[Rolfe 2017] Jason Tyler Rolfe. "Discrete variational autoencoders". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. url: https://openreview.net/forum?id=ryMxXPFex (cit. on p. 2).

[Rooshenas and Lowd 2016] Amirmohammad Rooshenas and Daniel Lowd. "Discriminative structure learning of arithmetic circuits". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, Sept. 2016, pp. 1506–1514. url: https://proceedings.mlr.press/v51/rooshenas16.html (cit. on p. 3).

[Shao *et al.* 2020] Xiaoting Shao *et al.* "Conditional sum-product networks: imposing structure on deep probabilistic architectures". In: *Proceedings of the 10th International Conference on Probabilistic Graphical Models*. Ed. by Manfred Jaeger and Thomas Dyhre Nielsen. Vol. 138. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 401–412. URL: https://proceedings.mlr.press/v138/shao20a.html (cit. on p. 3).

[Su *et al.* 2019] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks". In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 828–841 (cit. on p. 1).

[Szegedy *et al.* 2013] Christian Szegedy *et al.* "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013) (cit. on p. 1).

[Vahdat, Andriyash, *et al.* 2018] Arash Vahdat, Evgeny Andriyash, and William Macready. "Dvae#: discrete variational autoencoders with relaxed boltzmann priors". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio *et al.* Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/9f53d83ec0691550f7d2507d57f4f5a2-Paper.pdf (cit. on p. 2).

[Vahdat, Macready, *et al.* 2018] Arash Vahdat, William G. Macready, Zhengbing Bian, Amir Khoshaman, and Evgeny Andriyash. "DVAE++: discrete variational autoencoders with overlapping transformations". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 5042–5051. URL: http://proceedings.mlr.press/v80/vahdat18a.html (cit. on p. 2).

[Vergari, Choi, *et al.* 2020] Antonio Vergari, YooJung Choi, Robert Peharz, and Guy Van den Broeck. *Probabilistic Circuits: Representations, Inference, Learning and Applications*. AAAI Tutorial. 2020 (cit. on p. 1).

[Vergari, Mauro, *et al.* 2015] Antonio Vergari, Nicola Di Mauro, and Floriana Esposito. "Simplifying, regularizing and strengthening sum-product network structure learning". In: *ECML/PKDD*. 2015 (cit. on p. 3).

[Wei *et al.* 2018] Wenqi Wei *et al.* "Adversarial examples in deep learning: characterization and divergence". In: *arXiv preprint arXiv:1807.00051* (2018) (cit. on p. 1).

[Yu 2020] Ronald Yu. *A Tutorial on VAEs: From Bayes' Rule to Lossless Compression*. 2020. arXiv: 2006.10273 [cs.LG] (cit. on p. 2).

[Ziegler and Rush 2019] Zachary M. Ziegler and Alexander M. Rush. "Latent normalizing flows for discrete sequences". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7673–7682. URL: http://proceedings.mlr.press/v97/ziegler19a.html (cit. on p. 2).