On the Computational Tractability of Probabilistic Global and Local Sufficient Explanation

Xiliang Yang[1]

[1]College of Computer and Data Science,
Nanyang Technological University
50 Nanyang Avenue, Singapore
xiliang001@e.ntu.edu.sg

Abstract

Explainable AI (XAI) seeks to answer the question: *which features of the data led a model to make its decision?* Sufficient reasons are an important concept for understanding the behaviour of machine learning models, as they identify the key characteristics responsible for the prediction of an individual instance. Recent work introduced probabilistic global sufficient reasons, extending sufficient reasons from the single-instance level to all instances in the feature domain, thereby providing a global understanding of the classifier. However, prior work on this notion has been purely theoretical, without empirical evaluation. In this paper, we aim to fill this gap by developing practical methods for computing probabilistic global sufficient reasons and evaluating them on decision trees and circuit-based models.

Introduction

We revisit the well-established sufficiency criterion for this selection, known as a sufficient reason, which aligns with many explainability methods. Model-agnostic methods such as LIME (Ribeiro, Singh, and Guestrin 2018) and SHAP (Lundberg and Lee 2017) produce local explanations through perturbation but often suffer from instability. Logical reasoning approaches, including sufficient reasons and knowledge compilation (Darwiche and Marquis 2002), provide principled guarantees but frequently yield explanations requiring many features.

A common criticism of the logical definitions of sufficient and contrastive reasons is their rigidity and lack of flexibility, as they apply in an absolute sense over whole domains, often leading to excessively large or uninformative explanations. Recent work in tractable probabilistic reasoning (Khosravi et al. 2019b; Wang, Khosravi, and Van den Broeck 2021) demonstrates promising directions. Building on these threads, we propose a probabilistic generalization of sufficient reasons. Recent work (Bassan, Amir, and Katz 2024) extends local sufficient explanations to the global setting, which, instead of explaining a single instance, aims to provide an overall understanding of the classifier's behavior. However, their studies primarily focus on theoretical investigation rather than on the practical effectiveness of computing global explanations, which is the goal of this work.

Research Framework

We define a *local δ-sufficient reason* as a feature subset S such that, with probability at least δ, the classifier's prediction remains unchanged when S is fixed. Similarly, a *global δ-sufficient reason* ensures this property across the distribution of all inputs. This framework generalizes classical sufficient reasons, enabling explanations that are both flexible and probabilistically reliable. We further define cardinal minimal and subset minimal δ-explanations to formalize the optimality of S with respect to the local/global value functions. Preliminary results were obtained by X Huang, a postdoc in our group, and his collaborators, which includes: (1) While the local value function for sufficient reasons is non-monotonic, its global counterpart is monotonic and non-decreasing. (2) The global sufficient value function is supermodular, whereas the local one is neither supermodular nor submodular. (3) Global probabilistic sufficient explanations can be computed in polynomial time, even for complex models. In contrast, local explanations remain NP-hard even for simpler models such as decision trees. (4) Stronger approximation results for computing cardinally minimal global explanations for complex models, along with strong inapproximability results for the local setting.

Research Plan and Anticipated Contributions

Building on prior work in feature selection (Choi, Darwiche, and Van den Broeck 2017; Choi and Van den Broeck 2018; Khosravi et al. 2019b,a; Wang, Khosravi, and Van den Broeck 2021, 2020), we plan to extend tractability results to a broader class of models. Beyond traditional probabilistic circuits (PCs (Choi, Vergari, and Van den Broeck 2020)), recent advances in learning PCs involve constructing over-parameterized circuits with millions or even billions of parameters (Liu, Zhang, and Van den Broeck 2022), trained via SGD or EM (Peharz et al. 2016, 2020b). This trend has been driven by architectures such as RAT-SPNs (Peharz et al. 2020b), Einsum networks (EiNets) (Peharz et al. 2020a), and hidden Chow-Liu trees (HCLTs) (Liu, Zhang, and Van den Broeck 2022). We aim to analyze the tractability of our local/global value functions in these settings.

My planned contributions are: (1) Complexity results, identification, and algorithms for tractable classes. (2) Tractability analyses for discriminative models such as decision trees (Khosravi et al. 2020) and discriminative cir-

cuits (Liang and Van den Broeck 2019), as well as generative models such as probabilistic circuits. Recent interest in overparameterized circuits motivates tractability analysis where both discriminative and generative models are compiled from the same tree-shaped region graph (Loconte et al. 2024). (3) Approximation methods for intractable cases. (4) Applications to benchmark classifiers to evaluate explanation fidelity and usability.

This research aims to provide both theoretical insights and practical tools for XAI, advancing trustworthy AI in uncertain settings.

Progress and Timeline

Completed as of Sept 30, 2025:

- Defined probabilistic sufficient reasons and formalized their local/global versions.
- Analyzed complexity connections to weighted model counting and expectation queries.
- Identified tractable cases where generative and discriminative models are compatible (Khosravi et al. 2019a), including tensorized circuits that share the same tree-shaped region graph (Loconte et al. 2024), as well as probabilistic and discriminative circuits compiled with the same vtree. Tractability also holds when the discriminative model is a decision tree or forest.

In progress (Oct 2025–Jan 2026):

- Develop more effective Monte Carlo methods for rare-event sampling in intractable cases, where plain Monte Carlo performs poorly.
- Implement algorithms for tractable subclasses.

Future (2026–2027):

- Implement advanced Monte Carlo approximation methods.
- Design efficient algorithms for tractable subclasses.
- Explore continuous relaxations.
- Apply the proposed methods to challenging real-world classification tasks with large-scale datasets.

References

Bassan, S.; Amir, G.; and Katz, G. 2024. Local vs. global interpretability: A computational complexity perspective. *arXiv preprint arXiv:2406.02981*.

Choi, Y.; Darwiche, A.; and Van den Broeck, G. 2017. Optimal feature selection for decision robustness in bayesian networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

Choi, Y.; and Van den Broeck, G. 2018. On robust trimming of bayesian network classifiers. *arXiv preprint arXiv:1805.11243*.

Choi, Y.; Vergari, A.; and Van den Broeck, G. 2020. Probabilistic circuits: A unifying framework for tractable probabilistic models. *UCLA. URL: http://starai. cs. ucla. edu/papers/ProbCirc20. pdf*, 6.

Darwiche, A.; and Marquis, P. 2002. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17: 229–264.

Khosravi, P.; Choi, Y.; Liang, Y.; Vergari, A.; and Van den Broeck, G. 2019a. On tractable computation of expected predictions. *Advances in Neural Information Processing Systems*, 32.

Khosravi, P.; Liang, Y.; Choi, Y.; and Van den Broeck, G. 2019b. What to expect of classifiers? reasoning about logistic regression with missing features. *arXiv preprint arXiv:1903.01620*.

Khosravi, P.; Vergari, A.; Choi, Y.; Liang, Y.; and Van den Broeck, G. 2020. Handling missing data in decision trees: A probabilistic approach. *arXiv preprint arXiv:2006.16341*.

Liang, Y.; and Van den Broeck, G. 2019. Learning Logistic Circuits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 4277–4286.

Liu, A.; Zhang, H.; and Van den Broeck, G. 2022. Scaling up probabilistic circuits by latent variable distillation. *arXiv preprint arXiv:2210.04398*.

Loconte, L.; Mari, A.; Gala, G.; Peharz, R.; de Campos, C.; Quaeghebeur, E.; Vessio, G.; and Vergari, A. 2024. What is the Relationship between Tensor Factorizations and Circuits (and How Can We Exploit it)? *arXiv preprint arXiv:2409.07953*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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 III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 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.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wang, E.; Khosravi, P.; and Van den Broeck, G. 2020. Towards probabilistic sufficient explanations. In *Extending Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI)*, 68–69.

Wang, E.; Khosravi, P.; and Van den Broeck, G. 2021. Probabilistic sufficient explanations. *arXiv preprint arXiv:2105.10118*.