Regular Tree Priors for Scientific Symbolic Reasoning

Tim Schneider[1], Amin Totounferoush, Wolfgang Nowak, Steffen Staab

[1]) University of Stuttgart, Department for Analytic Computing,
Universitätsstraße 32, 70563 Stuttgart, Germany
timphillip.schneider@ipvs.uni-stuttgart.de

A significant aspect of the scientific process in natural science is to formulate new research hypotheses based on experimental data. Artificial Intelligence (AI) has been proposed as a *new tool* [8] for scientists to support hypotheses formulation. Recently, Bayesian methods [5, 2] for the task of symbolic regression have become key candidates to aid scientists, because of their well-defined posterior of symbolic solutions given priors and evidence in the form of experimental data. In general, it has been beneficial to exploit the vast amount of knowledge available in natural sciences as a yield of centuries of research. While Bayesian frameworks lend themselves naturally to encode it as prior distributions, existing encodings lacked expressiveness or closure properties to adequately capture prior knowledge about equations. Standard approaches to encode priors about the structure of the unknown equation would leverage formal languages, i.e. (weighted) subsets, of arbitrary *strings*. This formalism has several weaknesses: (i) The set of syntactically correct arithmetic expressions (compare *Dyck's language*) is not *regular*, rendering it not expressive enough for languages of practical interest. Thus Probabilistic Context-Free Grammar (PCFG) are used. But (ii) PCFG can also generate strings of mathematical symbols that are not syntactically valid equations, and (iii) *context-free* grammars are not closed under boolean operations [4], making arbitrary combinations of prior knowledge from different sources infeasible.

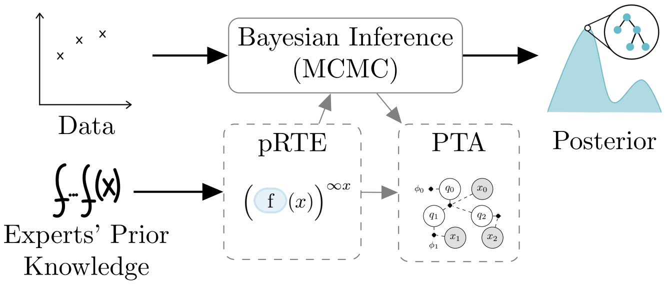

Figure 1: Scientists have data and prior knowledge. Our *Bayesian Inference* requires *samples* and *density evaluations* of the latter and yields a posterior distribution over expressions that fit the data consistently to the prior knowledge.

Method We propose to enhance existing encodings of priors by using formal tree languages $L \subseteq T_\Sigma$ [1] and suggest an inferencing method (fig. 1) that allows for the inclusion of these priors into symbolic regression. Since syntactical correctness of equations is ensured with a *ranked alphabet* Σ in a tree language, it is possible to define *expressive* regular languages L with: (i) compact Probabilistic Regular Tree Expression (pRTE) [7], and (ii) closure properties [6] under Boolean set operations that allow to easily combine the prior knowledge. A scientist expresses *prior knowledge* through pRTEs. The pRTEs are automatically translated to a joint Probabilistic Tree Automaton (PTA). Our inference algorithm works on the data interacting with the pRTE to sample proposal expressions $t \sim p_L$, as well as with the PTA to evaluate probability densities $p_L(t)$ to decide on the acceptance of proposals $t \in T_\Sigma$. This Markov Chain Monte Carlo (MCMC) inference yields a *posterior distribution* over arithmetic expressions fitting the data evidence and respecting the experts' prior knowledge.

Application *Sorption* is the transition of ions or molecules from a solution to a solid phase. All kinds of sorption isotherm equations (e.g. *Freundlich* or *Langmuir*) relate the *equilibrium sorptive concentration* c $(\frac{\text{mg}}{L})$ in the solution phase with the *sorbate concentration* s $(\frac{\text{mg}}{\text{Kg}})$ in the soil to characterize contaminants and its retention in soils. This knowledge has been *unified* [3] as

$$s = s_T \sum_{i=1}^{n} f_i \prod_{j=1}^{m_i} \left(\frac{q_{ij} c^{\alpha_{ij}}}{1 + p_{ij} c^{\beta_{ij}}} \right)^{\gamma_{ij}}, \tag{1}$$

with each choice for n, m_i and assignment to parameters $s_T, f_i, q_{ij}, \ldots$ being a model for $c \mapsto s$ that all share some regular pattern which we express as

$$\begin{aligned} E_{\text{iso}} = {} & \boxed{\cdot}\left(\boxed{s_T}, (\boxed{+}(y,x) + y)^{\infty x}\right) \circ_y \boxed{\cdot}\left(\boxed{f_i}, (\boxed{\cdot}(z,x) + z)^{\infty x}\right) \\ & \circ_z \boxed{\text{pow}}\left(\boxed{\div}(u,v), \boxed{\gamma_{ij}}\right) \circ_u \boxed{\cdot}\left(\boxed{q_{ij}}, \boxed{\text{pow}}\left(\boxed{c}, \boxed{\alpha_{ij}}\right)\right) \\ & \circ_v \boxed{+}\left(\boxed{1}, \boxed{\cdot}\left(\boxed{p_{ij}}, \boxed{\text{pow}}\left(\boxed{c}, \boxed{\beta_{ij}}\right)\right)\right), \end{aligned}$$

and exploit it for inference in our framework. Our experiments show that this approach outperforms standard symbolic regression algorithms in the majority of our scenarios and extrapolation datasets.

References

[1] H. Comon, M. Dauchet, R. Gilleron, F. Jacquemard, D. Lugiez, C. Löding, S. Tison, and M. Tommasi. Tree automata techniques and applications, 2008.

[2] R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F. A. Massucci, M. Miranda, J. Pallarès, and M. Sales-Pardo. A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science advances*, 6(5):eaav6971, 2020.

[3] C. Hinz. Description of sorption data with isotherm equations. *Geoderma*, 99(3-4):225–243, 2001.

[4] J. E. Hopcroft, R. Motwani, and J. D. Ullman. Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65, 2001.

[5] Y. Jin, W. Fu, J. Kang, J. Guo, and J. Guo. Bayesian symbolic regression. *arXiv preprint arXiv:1910.08892*, 2019.

[6] G. Rozenberg and A. Salomaa. *Handbook of Formal Languages: Volume 3 Beyond Words*. Springer Science & Business Media, 2012.

[7] T. Weidner. Probabilistic regular expressions and mso logic on finite trees. In *35th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

[8] L. Zdeborová. New tool in the box. *Nature Physics*, 13(5):420–421, 2017.