Neurosymbolic Reinforcement Learning With Sequential Guarantees

Lennert De Smet[*1], **Gabriele Venturato**[*1], **Luc De Raedt**[1,2], **Giuseppe Marra**[1]

[1]KU Leuven, Belgium
[2]Örebro University, Sweden
firstname.lastname@kuleuven.be

Reinforcement learning (RL) is successfully applied in various domains (Juang and Rabiner 1991; Khiatani and Ghose 2017; Schrittwieser et al. 2020; Van Roy et al. 2023; Mor, Garhwal, and Kumar 2020), yet it struggles to provide safety and behavioural guarantees (Garcıa and Fernández 2015; Yang et al. 2023). Neurosymbolic AI (NeSy), with its ability to combine logical reasoning and neural perception, has been explored as a potential solution (Yang et al. 2023; Zhang et al. 2023; Reichstein et al. 2019). However, existing NeSy methods, such as probabilistic logic shields (Yang et al. 2023), focus on single-step guarantees, limiting their effectiveness where multistep reasoning is required. To extend NeSy to efficient sequential reasoning, we introduced *relational* neurosymbolic Markov models (NeSy-MMs) that have been shown promising results on generative tasks (De Smet et al. 2024).

We propose a new framework for neurosymbolic reinforcement learning that incorporates relational NeSy-MMs as internal models for an RL agent. NeSy-MMs allow the agent to reason over multiple time steps and provide safety guarantees throughout the training process. We expect that this integration will provide policies that are resilient to test-time perturbations and adhere to given constraints over time, e.g. safety constraints.

Relational Neurosymbolic Markov Models

Relational NeSy-MMs are sequential probabilistic models over neurally-parametrised discrete-continuous random variables (Figure 1). They are probabilistic reasoning models that use random variables to model symbols, relations, and logical constraints. Neural predicates φ and φ_g map raw inputs (e.g. images) to symbols and vice versa, for *discriminative* and *generative* tasks. For instance, consider a MiniHack (Samvelyan et al. 2021) game (Figure 2), where the monsters can attack the player. With NeSy-MM we can model the sequences of interactions as well as a safety constraint for the player not being attacked.

Because of the sequential structure of NeSy-MMs, part of the world model can be specified by replacing unknown transition functions by neural networks. Finally, NeSy-MMs are relational models, a popular and very expressive representation for representing states in, for instance, databases and planning (Russell and Norvig 2020). Moreover, relational representations facilitate strong generalisation behaviour (Hummel and Holyoak 2003).

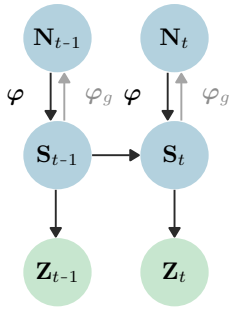

Figure 1: NeSy-MMs sequentially factorise neural ($\mathbf{N}_t$) and symbolic states ($\mathbf{S}_t$) over time. They can be conditioned on evidence ($\mathbf{Z}_t$).

Inference and Learning in NeSy-MMs

To bridge the gap between planning (sequential inference) and reinforcement learning, we propose a new, differentiable inference technique that combines non-parametric approximate Bayesian inference with exact NeSy inference.

We address the differentiability limitations of traditional particle filters by leveraging a novel approach rooted in neurosymbolic reasoning. Resampling, which hampers differentiability, is circumvented using a Rao-Blackwellised particle filter (RBPF) (Murphy and Russell 2001). The RBPF recursively computes $p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{Z}_{0:t+1})$ as

$$\int p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{Z}_{t+1}) p_{\boldsymbol{\varphi}}(\mathbf{x}_t \mid \mathbf{Z}_{0:t}) \, \mathrm{d}\mathbf{x}_t, \qquad (1)$$

where $p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{Z}_{t+1})$ can be computed exactly in NeSy settings by leveraging advancements in exact inference (Kisa et al. 2014; Darwiche 2020).

By removing resampling and having access to the exact transition probabilities, we can exploit an up-until-now unexplored synergy with gradient estimation methods (Kool, van Hoof, and Welling 2019; De Smet, Sansone, and Zuidberg Dos Martires 2023), which approximate gradients for $p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{Z}_{0:t+1})$ recursively. For example, using the Log-

[*]These authors contributed equally.

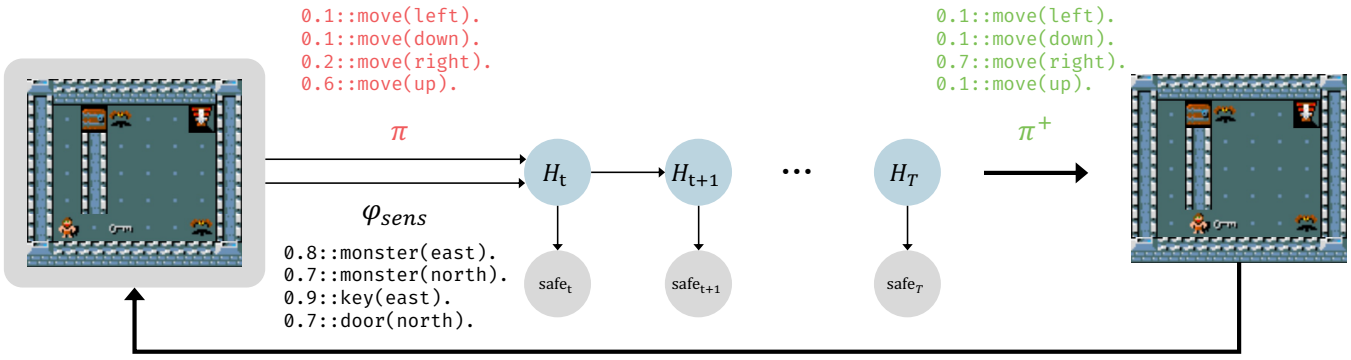

Figure 2: NeSy-MMs used as neurosymbolic policies that provide safety guarantees. As in a classic RL algorithms, ($\rightarrow$) executes an action in the environment, and ($\leftarrow$) provides a new observation to the policy. The agent (bottom left) has to reach the staircases (top right). Each NeSy state H_i can contain raw data, or relational symbols. The transition from H_i to H_{i+1} can be fully logical, neural, or a mixture of both. Each state is also conditioned on a safety property, such that the agent is not killed by the monsters.

Derivative trick (Williams 1992), we compute:

$$\begin{aligned}&\nabla_{\boldsymbol{\varphi}} p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{Z}_{0:t+1}) \\ &= \mathbb{E}_{\mathbf{X}_t}\left[\nabla_{\boldsymbol{\varphi}} p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{X}_t, \mathbf{Z}_{t+1})\right] \\ &+ \mathbb{E}_{\mathbf{X}_t}\left[p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{X}_t, \mathbf{Z}_{t+1}) \nabla_{\boldsymbol{\varphi}} \log p_{\boldsymbol{\varphi}}(\mathbf{X}_t \mid \mathbf{Z}_{0:t})\right].\end{aligned} \tag{2}$$

To ensure scalability, we employ cluster factorisation to decompose $\mathbf{X}$ into clusters $\{\mathbf{X}^i\}_{i=1}^B$ that become conditionally independent given $\mathbf{Z}$. This factorisation reduces the computational burden by allowing exact computation of $p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{Z}_{t+1})$ for each cluster:

$$p_{\boldsymbol{\varphi}}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{Z}_{t+1}) = \prod_{i=1}^{B} p_{\boldsymbol{\varphi}}(\mathbf{X}^i_{t+1} \mid \mathbf{x}_t, \mathbf{Z}_{t+1}). \tag{3}$$

For clusters containing infinite variables $\mathbf{I}^i_t$, i.e. both countably infinite and continuous (uncountable) domains, we first obtain samples using a traditional particle filter. This leaves a purely finite distribution for the remaining variables $\mathbf{F}^i_t$, which is computed exactly:

$$\prod_{i=1}^{B} p_{\boldsymbol{\varphi}}(\mathbf{F}^i_{t+1} \mid \mathbf{I}^i_{t+1}, \mathbf{x}_t, \mathbf{Z}_{t+1}) p_{\boldsymbol{\varphi}}(\mathbf{I}^i_{t+1} \mid \mathbf{x}_t, \mathbf{Z}_{t+1}). \tag{4}$$

This hybrid approach unites local exact inference, cluster factorisation, and tailored gradient estimation methods to enable optimisation across finite, infinite, and logical variables in hybrid domains. Our resulting differentiable particle filter effectively exploits the conditional dependency structure of the NeSy states $\mathbf{X}_t$, providing a scalable and generalisable solution. Intuitively, one can view NeSy-MMs as differentiable planning models that can specify only part of the underlying environment, while the rest is learned while interacting with it.

Neurosymbolic Reinforcement Learning

The goal of using NeSy-MMs as RL policies is to obtain formal guarantees within a given time horizon. Previous efforts (Yang et al. 2023) have focused on providing single-step guarantees by shielding (Jansen et al. 2020) a neural policy with a probabilistic logic program (De Raedt, Kimmig, and Toivonen 2007). While effective, this approach does not scale to multistep guarantees because of the #P-hardness of its inference procedure. NeSy-MMs resolve this problem by using unbiased approximate inference techniques instead.

Consider again a MiniHack level where the agent is in a room with two monsters and has to reach a goal (Figure 2). The optimal strategy in this case is to take the key and wait to lure the two monsters away from the goal. Only once the monsters are close enough and the agent has the key, it can move through the corridor, open the door, and move safely to the goal before the monsters can catch up. Hence, safely reaching the goal is not something that can be decided by single-step reasoning. Concretely, if the agent is governed by a policy π and a sensor φ_{sens} gives an estimate of the current state of the game, then these will form the input to a NeSy-MM. The NeSy-MM then updates the policy to $\pi^+(a|$  $) = \pi(a \mid \text{safe}_{t:T},$ $)$ that incorporates the safety constraints via approximate Bayesian inference. Finally, we want to obtain a policy such that,

$$P_{\pi^+}(\text{safe}_{t:T} \mid \text{}) \geq P_{\pi^+}(\text{safe}_t \mid \text{}) \tag{5}$$

$$\geq P_{\pi}(\text{safe}_{t:T} \mid \text{}) \tag{6}$$

This means our NeSy policy is going to be safer than the single time-step shielded policy (5) from Yang et al. (2023), that is in turn safer than the unshielded policy (6), for any time horizon. In the future, we aim to empirically verify this idea and more closely integrate NeSy-MMs into the RL framework by analysing the behaviour of the expected reward in the presence of neurosymbolic policies.

References

Darwiche, A. 2020. An Advance on Variable Elimination with Applications to Tensor-Based Computation. In *ECAI 2020*, 2559–2568. IOS Press.

De Raedt, L.; Kimmig, A.; and Toivonen, H. 2007. ProbLog: A Probabilistic Prolog and Its Application in Link Discovery. In *IJCAI*. Hyderabad.

De Smet, L.; Sansone, E.; and Zuidberg Dos Martires, P. 2023. Differentiable Sampling of Categorical Distributions Using the CatLog-Derivative Trick. In *NeurIPS*.

De Smet, L.; Venturato, G.; De Raedt, L.; and Marra, G. 2024. Neurosymbolic Markov Models. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1): 1437–1480.

Hummel, J. E.; and Holyoak, K. J. 2003. A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, 110(2): 220.

Jansen, N.; Könighofer, B.; Junges, S.; Serban, A.; and Bloem, R. 2020. Safe reinforcement learning using probabilistic shields. In *31st International Conference on Concurrency Theory (CONCUR 2020)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Juang, B. H.; and Rabiner, L. R. 1991. Hidden Markov models for speech recognition. *Technometrics*, 33(3): 251–272.

Khiatani, D.; and Ghose, U. 2017. Weather forecasting using hidden Markov model. In *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 220–225. IEEE.

Kisa, D.; Van den Broeck, G.; Choi, A.; and Darwiche, A. 2014. Probabilistic sentential decision diagrams. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Kool, W.; van Hoof, H.; and Welling, M. 2019. Buy 4 reinforce samples, get a baseline for free! *ICLR Deep RL Meets Structured Prediction Workshop*.

Mor, B.; Garhwal, S.; and Kumar, A. 2020. A Systematic Review of Hidden Markov Models and Their Applications. *Archives of Computational Methods in Engineering*, 28: 1429 – 1448.

Murphy, K.; and Russell, S. 2001. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Sequential Monte Carlo methods in practice*, 499–515. Springer.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; and Prabhat, F. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204.

Russell, S.; and Norvig, P. 2020. *Artificial Intelligence: A Modern Approach*. Hoboken: Pearson, 4th edition edition. ISBN 978-0-13-461099-3.

Samvelyan, M.; Kirk, R.; Kurin, V.; Parker-Holder, J.; Jiang, M.; Hambro, E.; Petroni, F.; Kuttler, H.; Grefenstette, E.; and Rocktäschel, T. 2021. MiniHack the Planet: A Sandbox for Open-Ended Reinforcement Learning Research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609.

Van Roy, M.; Robberechts, P.; Yang, W.-C.; De Raedt, L.; and Davis, J. 2023. A Markov framework for learning and reasoning about strategies in professional soccer. *Journal of Artificial Intelligence Research*, 77: 517–562.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Yang, W.-C.; Marra, G.; Rens, G.; and De Raedt, L. 2023. Safe Reinforcement Learning via Probabilistic Logic Shields. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 5739–5749. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhang, H.; Dang, M.; Peng, N.; and Van Den Broeck, G. 2023. Tractable Control for Autoregressive Language Generation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 40932–40945. PMLR.