Predictive information in reinforcement learning of embodied agents

Keyan Zahedi[1], Georg Martius[1] and Nihat Ay[1,2]

[1]MPI Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany, {zahedi,martius,nay}@mis.mpg.de
[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 8501, USA

Information-driven self-organisation, by the means of maximizing the one-step approximation of the predictive information (PI) has proven to produce a coordinated behaviour among physically coupled but otherwise independent agents [1, 2]. The reason is that the PI inherently addresses two important issues for self-organised adaptation, as the following equation shows: $I(S_t; S_{t+1}) = H(S_{t+1}) - H(S_{t+1}|S_t)$, where S_t are the sensor values, intrinsically accessible by the agent. The first term leads to a diversity of the behaviour, as every possible sensor state must be visited with equal probability. The second term ensures that the behaviour is compliant with the constraints given by the environment and morphology, as the behaviour, measured by the sensor stream, must be predictable. The PI maximization is also related to other self-organisation principles, such as the Homeokinses [3], and therefore, is a good candidate for a general first principle for embodied artificial intelligence.

In general it is not desirable that an embodied agent senses every possible stimulus with an equal probability, as tasks have to be solved and dangerous situations need to be avoided. Hence, guidance is required. We chose the framework of reinforcement learning for this purpose, where the reinforcement signal is understood as an external guidance, and the PI maximization is understood as an internal drive. This idea is not new, as combining extrinsic and intrinsic rewards in the context of reinforcement learning goes back to the pioneering work of [4], but is also in the focus of more recent work by [5–7]. The significant difference here is that the PI, measured on the sensor values, accompanies (or may even replace) the exploration of the reinforcement learning such that it is conducted compliant to the morphology and environment. The actual embodiment is now taken into account, without modelling it explicitly in the learning process.

A different approach used the PI, estimated on the spatio-temporal phase-space of an embodied system, as part of fitness function in an artificial evolution setting [8]. It was shown that the resulting locomotion behaviour of a snake-bot was more robust, compare to the setting, in which only the travelled distance determined the fitness. This is a good indication, that more than just a speed up can be expected from combining information-driven self-organisation and reinforcement learning. In first experiments, we apply the PI as intrinsic reward to the problem of the cart pole swing up experiment. Although maximizing the PI seems contradicting here, as the goal is to minimize the entropy over the pole deviation, we will show that the learning process benefits from the PI.

References

[1] K. Zahedi, N. Ay, and R. Der. Higher coordination with less control – a result of information maximization in the sensori-motor loop. *Adaptive Behavior*, 18(3–4):338–355, 2010.

[2] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich. Predictive information and explorative behavior of autonomous robots. *European Physical Journal B*, 63(3):329–339, 2008.

[3] R. Der and G. Martius. *The Playful Machine: Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Cognitive Systems Monographs. Springer, 2012.

[4] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of SAB'90*, pages 222–227, 1990.

[5] F. Kaplan and P.-Y. Oudeyer. Maximizing learning progress: An internal reward system for development. *Embodied Artificial Intelligence*, pages 259–270, 2004.

[6] P.-Y. Oudeyer, F. Kaplan, and V.V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Trans. on Evo. Computation*, 11(2):265–286, 2007.

[7] A.G. Barto, S. Singh, and N. Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proc. of 3rd Int. Conf. Development Learn.*, pages 112–119, 2004.

[8] M. Prokopenko, V. Gerasimov, and I. Tanev. Evolving spatiotemporal coordination in a modular robotic system. In *Proc. SAB'06*, volume 4095, pages 558–569, 2006.