Exploration using meta-reinforcement learning

Manav Choudhary

April 4, 2019

Abstract

We present a brief review of the literature in exploration in reinforcement learning.

1 Introduction

Exploration is a fundamental aspect of a reinforcement learning(RL) algorithm. Exploration can be achieved while dealing with the 'atomic action space', 'skill space' and 'navigation routes'. A lot of research work has been done to address this. A good summary can be found in Oudeyer & Kaplan (2007), Oudeyer et al. (2007). Based on this review, we believe that an episodic memory is fundamentally required to implement structured exploration in an episode. The memory is used to keep track of already explored state space. The existing techniques can be classified in the following categories:

1. Intrinsic motivation / 'Prediction error', error in predicting the environment dynamics:
Schmidhuber (1990), Singh et al. (2005), Sun et al. (2011), Stadie et al. (2015), Pathak et al. (2017), Al-Shedivat et al. (2017), Burda, Edwards, Pathak, Storkey, Darrell & Efros (2018), Burda, Edwards, Storkey & Klimov (2018), Stanton & Clune (2018), Savinov et al. (2018). In psychology: Ryan & Deci (2000), Smith & Gasser (2005), Ryan & Silvia (2012)
2. State visitation counts :
Lopes et al. (2012), Strehl & Littman (2008), Bellemare et al. (2016), Poupart et al. (2006), Tang et al. (2016), Oh et al. (2015), Fu et al. (2017), Ostrovski et al. (2017), Machado et al. (2018), Choshen et al. (2018)
3. Curiosity/ Information gain/ Prediction uncertainty :
Storck et al. (1995), Schmidhuber (2010), Graziano et al. (2011), Ngo et al. (2012), Still & Precup (2012), Ying-Jeh Little & Sommer (2013), Schmidhuber (2015), Houthooft et al. (2016), Chen et al. (2017), Achiam & Sastry (2017).
4. Exploration through meta-reinforcement learning :
Schmidhuber et al. (1998), Schmidhuber (2003), Gupta et al. (2018), Rothfuss et al. (2018), Stadie et al. (2018)
5. Improvement in environment model in Model based RL:
Schmidhuber (1991), Lopes et al. (2012), Shyam et al. (2018)
6. Using adversarial models :
Schmidhuber (1997, 1999), Sukhbaatar et al. (2017)
7. Novelty search and quality diversity :
Lehman & Stanley (2008, 2011), Stanley & Lehman (2015), Eysenbach et al. (2018), Ecoffet et al. (2018)
8. Empowerment : Information gain based on entropy of actions & policy entropy
Mohamed & Jimenez Rezende (2015), Klyubin et al. (2005), Gregor et al. (2016), Eysenbach et al. (2018)
9. Thompson sampling and bootstrapped models :
Chapelle & Li (2011), Osband et al. (2016)
10. Optimism in the face of uncertainty :
Brafman & Tennenholtz (2003), Kearns & Singh (2002), Kolter & Ng (2009), Kompella et al. (2017), propose exploration algorithms polynomial in the number of state space parameters.
11. Parameter space exploration :
Plappert et al. (2018), Fortunato et al. (2017)
12. Noise from a learnt latent space:
Hausman et al. (2018), Florensa et al. (2017)
13. PowerPlay : Schmidhuber (2011)

2 Research work in exploration

We will now focus on 3 types of methods, using prediction error, meta-RL and novelty search methods.

2.1 Prediction Error

Schmidhuber (1990) introduced intrinsic reward based on the error in the predictions of a environment model, which in turn is learnt from the experience gathered by the agent. Pathak et al. (2017) introduced the Intrinsic Curiosity Module (ICM) to calculate the intrinsic reward. ICM uses a inverse dynamics model to learn the state features. Burda, Edwards, Pathak, Storkey, Darrell & Efros (2018) did a large scale study and compared different ways of learning the state features. RND [Burda, Edwards, Storkey & Klimov (2018)] takes a slightly different approach, in that, it calculates the error based on the predicted features of the observation by the target network (a fixed random network) and the predictor network. This predictor network is not a forward dynamics model. Also they use separate value function heads for the intrinsic and extrinsic reward and then add these values functions.

While designing a prediction error based intrinsic reward system following are the key design choices:

- Feature space
- Combination of intrinsic and extrinsic rewards
- CNN or RNN for the dynamics model
- Reward, observation and features normalization
- Distributed training

Following qualities (as identified in Burda, Edwards, Pathak, Storkey, Darrell & Efros (2018)) make a good feature space:

1. Compact: Features should be low dimensional and should filter out irrelevant parts of the observation space.
2. Sufficient: Features should have enough capacity to contain all important information.
3. Stable: Intrinsic reward, by design, is non-stationary, since it decreases as the state space is better explored. Now, since the features change over time as they are learnt, this introduces a second non-stationarity in the method. The first non-stationarity is inherent to the method, however, second should be minimized.

Burda, Edwards, Pathak, Storkey, Darrell & Efros (2018) investigated the use of 'Pixels', 'Random features', 'VAE features' and 'Inverse dynamics features' and found that for complex environments inverse dynamics features performed best. Also, they found that such agents get stuck in the 'noisy-TV problem', since they are using the 'prediction error' of the forward dynamics model as reward, instead of the 'improvement in the prediction error'.

RND Burda, Edwards, Storkey & Klimov (2018) used the error in prediction of features of an observation when compared to those of a fixed random network as the intrinsic reward. Since RND did not use a forward dynamics model, therefore it is not susceptible to the 'noisy-TV problem'.

Episodic curiosity through reachability [Savinov et al. (2018)] uses the concept of reachability to define novelty and uses an episodic buffer to store previous interesting states.

2.2 Exploration in Meta Reinforcement Learning

Finn et al. (2017) proposed a model agnostic meta learner (MAML) which can be used to learn the optimal 'initial weights' of an RL agent w.r.t a task distribution. However, the gradient update used in MAML doesn't account for the impact of the pre-update sampling distribution on the post-update rewards. Stadie et al. (2018) then proposed E-MAML based on this insight and introduce an extra term to account for the impact of pre-update sampling distribution. This extra-term encourages explorative trajectories in the pre-update sampling, if the updated agent learns well. This term allows the policy to attempt to deliver the maximal amount of information useful for the future rewards, without worrying

about it own rewards. They also propose E-RL^2, which is the exploratory version of RL^2. In E-RL^2, we perform k trials for each MDP, where we sample p exploratory rollouts and $k - p$ non-exploratory rollouts. The rewards of the exploratory rollouts are zeroed out, therefore during the forward pass the exploratory episodes contribute to RNN hidden weights perform better system identification and during the backward pass the rewards of the exploratory episodes do not contribute to the gradient directly. It should be noted that none of these methods use intrinsic reward.

Rothfuss et al. (2018) proposed ProMP (proximal meta policy search), which uses low variance curvature objective for the gradient update. ProMP outpeforms E-MAML in their experiments.

Gupta et al. (2018) proposed MAESN (model agnostic exploration with structured noise). ProMP, E-MAML, E-RL^2, do not employ structured exploration, i.e their exploration is time-invariant, however it is better to use a temporally correlated exploration strategy. For MAML and E-MAML the stochasticity of the policy is limited to time-invariant noise, this fundamentally limits the exploratory behaviour it can represent. The distribution $\pi_\theta(a|s)$ is typically represented with simple parametric distributions, such as unimodal Gaussians, which restricts its ability to model task-dependent covariances. As pointed out in Gupta et al. (2018), for RL^2, E-RL^2 and Wang et al. (2016), the policy is limited in its adaptation ability by the forward pass of the RNN. If this forward pass does not produce a good policy, then there is no way for further improvement. Gradient descent based meta-learning methods such as MAML can revert to the standard policy gradient and make slow but steady improvement. MAESN uses noise sampled in a learned latent space for structure noise. It conditions the policy on random variables which are sampled only once in the beginning of the episode and are drawn from a learned latent distribution. Also, MAESN then learns the policy network parameters and the parameters of the latent space distribution for a given task distribution in MAML fashion. However, different latent noise space parameters are learnt for different tasks and a single common policy network parameters. Also there is a penalty term for the KL-divergence between the gaussian distribution of pre-update variational parameters and the latent variable prior, a unit Gaussian. At test time the latent noise distribution is initialized to a unit Gaussian. Again note, this work doesn't use intrinsic reward, but the authors suggest such rewards can be used.

DIYAN [Eysenbach et al. (2018)], learns diverse "skills", which can then be used by a meta-controller to act in sparse reward environment and to perform exploration. Here the mutual information between "skills" and states is maximized and the policy entropy is maximized. This leads to skills which visit different parts of the state space.

Schmidhuber et al. (1998), Schmidhuber (2003) use a SMP with SSA to perform exploration with meta-RL.

2.3 Novelty search methods

Ecoffet et al. (2018) proposed 'Go-Explore' that performs greatly on Montezuma's revenge and Pitfall. They hypothesize that a major weakness for intrinsic motivation algorithms is *detachment*, wherein the algorithms forget about promising areas they have visited, i.e they do not return to them to see if they lead to new states. The phase 1 of the algorithm is to choose a cell from an 'archive', **Go** back to the chosen cell, **Explore** from the cell, if the new trajectory is better update the archive with it. Phase 2 is to robustify the found solutions by imitation learning. They used simple cell representations, domain specific knowledge and simple exploration techniques. Go-explore decomposes exploration into three elements : *Accumulate* stepping-stones, *return* to promising stepping-stones, *explore* from them for additional stepping-stones. The idea of preserving and exploring from stepping-stones in an archive comes from the Quality diversity (QD) [Pugh et al. (2016)] family of algorithms, MAP-elites and novelty search with local competition.

2.4 Questions addressed by these papers

How to define the intrinsic reward?
How to combine intrinsic reard with the extrinsic reward?
How to create Long term structured exploration?
How to define the policy of the agent?

How to address *detachment*?

3 Conclusion

We provided a brief summary of existing literature on exploration in reinforcement learning.

References

Achiam, J. & Sastry, S. (2017), 'Surprise-based intrinsic motivation for deep reinforcement learning', *CoRR* **abs/1703.01732**.
URL: *http://arxiv.org/abs/1703.01732*

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I. & Abbeel, P. (2017), 'Continuous adaptation via meta-learning in nonstationary and competitive environments', *CoRR* **abs/1710.03641**.
URL: *http://arxiv.org/abs/1710.03641*

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D. & Munos, R. (2016), 'Unifying count-based exploration and intrinsic motivation', *CoRR* **abs/1606.01868**.
URL: *http://arxiv.org/abs/1606.01868*

Brafman, R. I. & Tennenholtz, M. (2003), 'R-max - a general polynomial time algorithm for near-optimal reinforcement learning', *J. Mach. Learn. Res.* **3**, 213–231.
URL: *https://doi.org/10.1162/153244303765208377*

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T. & Efros, A. A. (2018), 'Large-Scale Study of Curiosity-Driven Learning', *ArXiv e-prints* p. arXiv:1808.04355.

Burda, Y., Edwards, H., Storkey, A. & Klimov, O. (2018), 'Exploration by Random Network Distillation', *ArXiv e-prints* p. arXiv:1810.12894.

Chapelle, O. & Li, L. (2011), An empirical evaluation of thompson sampling, *in* J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems 24', Curran Associates, Inc., pp. 2249–2257.
URL: *http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf*

Chen, R. Y., Sidor, S., Abbeel, P. & Schulman, J. (2017), 'UCB and infogain exploration via $\boldsymbol{Q}$-ensembles', *CoRR* **abs/1706.01502**.
URL: *http://arxiv.org/abs/1706.01502*

Choshen, L., Fox, L. & Loewenstein, Y. (2018), 'DORA The Explorer: Directed Outreaching Reinforcement Action-Selection', *ArXiv e-prints* p. arXiv:1804.04012.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O. & Jeff, C. (2018), 'Montezuma's revenge solved by go-explore, a new algorithm for hard-exploration problems'.
URL: *https://eng.uber.com/go-explore/*

Eysenbach, B., Gupta, A., Ibarz, J. & Levine, S. (2018), 'Diversity is all you need: Learning skills without a reward function', *CoRR* **abs/1802.06070**.
URL: *http://arxiv.org/abs/1802.06070*

Finn, C., Abbeel, P. & Levine, S. (2017), 'Model-agnostic meta-learning for fast adaptation of deep networks', *CoRR* **abs/1703.03400**.
URL: *http://arxiv.org/abs/1703.03400*

Florensa, C., Duan, Y. & Abbeel, P. (2017), 'Stochastic neural networks for hierarchical reinforcement learning', *CoRR* **abs/1704.03012**.
URL: *http://arxiv.org/abs/1704.03012*

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C. & Legg, S. (2017), 'Noisy networks for exploration', *CoRR* **abs/1706.10295**.
URL: *http://arxiv.org/abs/1706.10295*

Fu, J., Co-Reyes, J. D. & Levine, S. (2017), 'EX2: exploration with exemplar models for deep reinforcement learning', *CoRR* **abs/1703.01260**.
URL: *http://arxiv.org/abs/1703.01260*

Graziano, V., Glasmachers, T., Schaul, T., Pape, L., Cuccu, G., Leitner, J. & Schmidhuber, J. (2011), Artificial curiosity for autonomous space exploration.

Gregor, K., Rezende, D. J. & Wierstra, D. (2016), 'Variational intrinsic control', *CoRR* **abs/1611.07507**.
URL: *http://arxiv.org/abs/1611.07507*

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P. & Levine, S. (2018), 'Meta-reinforcement learning of structured exploration strategies', *CoRR* **abs/1802.07245**.
URL: *http://arxiv.org/abs/1802.07245*

Hausman, K., Springenberg, J. T., Wang, Z., Heess, N. & Riedmiller, M. (2018), Learning an embedding space for transferable robot skills, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=rk07ZXZRb*

Houthooft, R., Chen, X., Duan, Y., Schulman, J., Turck, F. D. & Abbeel, P. (2016), 'Curiosity-driven exploration in deep reinforcement learning via bayesian neural networks', *CoRR* **abs/1605.09674**.
URL: *http://arxiv.org/abs/1605.09674*

Kearns, M. & Singh, S. (2002), 'Near-optimal reinforcement learning in polynomial time', *Machine Learning* **49**(2), 209–232.
URL: *https://doi.org/10.1023/A:1017984413808*

Klyubin, A. S., Polani, D. & Nehaniv, C. L. (2005), Empowerment: a universal agent-centric measure of control, *in* '2005 IEEE Congress on Evolutionary Computation', Vol. 1, pp. 128–135 Vol.1.

Kolter, J. Z. & Ng, A. Y. (2009), Near-bayesian exploration in polynomial time, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ICML '09, ACM, New York, NY, USA, pp. 513–520.
URL: *http://doi.acm.org/10.1145/1553374.1553441*

Kompella, V. R., Stollenga, M., Luciw, M. & Schmidhuber, J. (2017), 'Continual curiosity-driven skill acquisition from high-dimensional video inputs for humanoid robots', *Artificial Intelligence* **247**, 313 – 335. Special Issue on AI and Robotics.
URL: *http://www.sciencedirect.com/science/article/pii/S000437021500017X*

Lehman, J. & Stanley, K. O. (2008), Exploiting open-endedness to solve problems through the search for novelty, *in* 'Proceedings of the Eleventh International Conference on Artificial Life (Alife XI', MIT Press.

Lehman, J. & Stanley, K. O. (2011), 'Abandoning objectives: Evolution through the search for novelty alone', *Evolutionary Computation* **19**(2), 189–223.

Lopes, M., Lang, T., Toussaint, M. & yves Oudeyer, P. (2012), Exploration in model-based reinforcement learning by empirically estimating learning progress, *in* F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems 25', Curran Associates, Inc., pp. 206–214.
URL: *http://papers.nips.cc/paper/4642-exploration-in-model-based-reinforcement-learning-by-empirically-estimating-learning-progress.pdf*

Machado, M. C., Bellemare, M. G. & Bowling, M. (2018), 'Count-Based Exploration with the Successor Representation', *ArXiv e-prints* p. arXiv:1807.11622.

Mohamed, S. & Jimenez Rezende, D. (2015), 'Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning', *ArXiv e-prints* p. arXiv:1509.08731.

Ngo, H., Luciw, M., Forster, A. & Schmidhuber, J. (2012), Learning skills from play: Artificial curiosity on a katana robot arm, *in* 'The 2012 International Joint Conference on Neural Networks (IJCNN)', pp. 1–8.

Oh, J., Guo, X., Lee, H., Lewis, R. L. & Singh, S. P. (2015), 'Action-conditional video prediction using deep networks in atari games', *CoRR* **abs/1507.08750**.
URL: *http://arxiv.org/abs/1507.08750*

Osband, I., Blundell, C., Pritzel, A. & Van Roy, B. (2016), Deep exploration via bootstrapped dqn, *in* D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, eds, 'Advances in Neural Information Processing Systems 29', Curran Associates, Inc., pp. 4026–4034.
URL: *http://papers.nips.cc/paper/6501-deep-exploration-via-bootstrapped-dqn.pdf*

Ostrovski, G., Bellemare, M. G., van den Oord, A. & Munos, R. (2017), 'Count-based exploration with neural density models', *CoRR* **abs/1703.01310**.
URL: *http://arxiv.org/abs/1703.01310*

Oudeyer, P., Kaplan, F. & Hafner, V. V. (2007), 'Intrinsic motivation systems for autonomous mental development', *IEEE Transactions on Evolutionary Computation* **11**(2), 265–286.

Oudeyer, P.-Y. & Kaplan, F. (2007), 'What is intrinsic motivation? a typology of computational approaches', *Frontiers in Neurorobotics* **1**, 245 – 270.

Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. (2017), 'Curiosity-driven exploration by self-supervised prediction', *CoRR* **abs/1705.05363**.
URL: *http://arxiv.org/abs/1705.05363*

Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen, R. Y., Chen, X., Asfour, T., Abbeel, P. & Andrychowicz, M. (2018), Parameter space noise for exploration, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=ByBAl2eAZ*

Poupart, P., Vlassis, N., Hoey, J. & Regan, K. (2006), An analytic solution to discrete bayesian reinforcement learning, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ICML '06, ACM, New York, NY, USA, pp. 697–704.
URL: *http://doi.acm.org/10.1145/1143844.1143932*

Pugh, J. K., Soros, L. B. & Stanley, K. O. (2016), 'Quality diversity: A new frontier for evolutionary computation', *Frontiers in Robotics and AI* **3**, 40.
URL: *https://www.frontiersin.org/article/10.3389/frobt.2016.00040*

Rothfuss, J., Lee, D., Clavera, I., Asfour, T. & Abbeel, P. (2018), 'ProMP: Proximal Meta-Policy Search', *ArXiv e-prints* p. arXiv:1810.06784.

Ryan, R. M. & Deci, E. L. (2000), 'Intrinsic and extrinsic motivations: Classic definitions and new directions', *Contemporary Educational Psychology* **25**(1), 54 – 67.
URL: *http://www.sciencedirect.com/science/article/pii/S0361476X99910202*

Ryan, R. M. & Silvia, P. J. (2012), 'Curiosity and motivation'.
URL: *http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195399820.001.0001/oxfordhb-9780195399820-e-10*

Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T. P. & Gelly, S. (2018), 'Episodic curiosity through reachability', *CoRR* **abs/1810.02274**.
URL: *http://arxiv.org/abs/1810.02274*

Schmidhuber, J. (1990), A possibility for implementing curiosity and boredom in model-building neural controllers, *in* 'Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats', MIT Press, Cambridge, MA, USA, pp. 222–227.
URL: *http://dl.acm.org/citation.cfm?id=116517.116542*

Schmidhuber, J. (1991), Curious model-building control systems, *in* '[Proceedings] 1991 IEEE International Joint Conference on Neural Networks', pp. 1458–1463 vol.2.

Schmidhuber, J. (1997), 'What's interesting?'.
URL: *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.6362rep=rep1type=pdf*

Schmidhuber, J. (1999), Artificial curiosity based on discovering novel algorithmic predictability through coevolution, *in* 'Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)', Vol. 3, pp. 1612–1618 Vol. 3.

Schmidhuber, J. (2003), Advances in evolutionary computing, Springer-Verlag, Berlin, Heidelberg, chapter Exploring the Predictable, pp. 579–612.
URL: *http://dl.acm.org/citation.cfm?id=903758.903782*

Schmidhuber, J. (2010), 'Formal theory of creativity, fun, and intrinsic motivation (1990–2010)', *IEEE Transactions on Autonomous Mental Development* **2**(3), 230–247.

Schmidhuber, J. (2011), 'POWERPLAY: training an increasingly general problem solver by continually searching for the simplest still unsolvable problem', *CoRR* **abs/1112.5309**.
URL: *http://arxiv.org/abs/1112.5309*

Schmidhuber, J. (2015), 'On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models', *CoRR* **abs/1511.09249**.
URL: *http://arxiv.org/abs/1511.09249*

Schmidhuber, J., Zhao, J. & Schraudolph, N. N. (1998), *Reinforcement Learning with Self-Modifying Policies*, Springer US, Boston, MA, pp. 293–309.
URL: *https://doi.org/10.1007/978-1-4615-5529-2*$_1$2

Shyam, P., Jaśkowski, W. & Gomez, F. (2018), 'Model-Based Active Exploration', *ArXiv e-prints* p. arXiv:1810.12162.

Singh, S. P., Barto, A. G. & Chentanez, N. (2005), Intrinsically motivated reinforcement learning, *in* L. K. Saul, Y. Weiss & L. Bottou, eds, 'Advances in Neural Information Processing Systems 17', MIT Press, pp. 1281–1288.
URL: *http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf*

Smith, L. & Gasser, M. (2005), 'The development of embodied cognition: six lessons from babies', *Artif. Life* **11**(1-2), 13–29.

Stadie, B. C., Levine, S. & Abbeel, P. (2015), 'Incentivizing exploration in reinforcement learning with deep predictive models', *CoRR* **abs/1507.00814**.
URL: *http://arxiv.org/abs/1507.00814*

Stadie, B. C., Yang, G., Houthooft, R., Chen, X., Duan, Y., Wu, Y., Abbeel, P. & Sutskever, I. (2018), 'Some considerations on learning to explore via meta-reinforcement learning', *CoRR* **abs/1803.01118**.
URL: *http://arxiv.org/abs/1803.01118*

Stanley, K. O. & Lehman, J. (2015), *Why Greatness Cannot Be Planned: The Myth of the Objective*, Springer Publishing Company, Incorporated.

Stanton, C. & Clune, J. (2018), 'Deep Curiosity Search: Intra-Life Exploration Can Improve Performance on Challenging Deep Reinforcement Learning Problems', *ArXiv e-prints* p. arXiv:1806.00553.

Still, S. & Precup, D. (2012), 'An information-theoretic approach to curiosity-driven reinforcement learning', *Theory in Biosciences* **131**(3), 139–148.
URL: *https://doi.org/10.1007/s12064-011-0142-z*

Storck, J., Hochreiter, S. & Schmidhuber, J. (1995), 'Reinforcement driven information acquisition in non-deterministic environments'.

Strehl, A. L. & Littman, M. L. (2008), 'An analysis of model-based interval estimation for markov decision processes', *J. Comput. Syst. Sci.* **74**, 1309–1331.

Sukhbaatar, S., Kostrikov, I., Szlam, A. & Fergus, R. (2017), 'Intrinsic motivation and automatic curricula via asymmetric self-play', *CoRR* **abs/1703.05407**.
URL: *http://arxiv.org/abs/1703.05407*

Sun, Y., Gomez, F. J. & Schmidhuber, J. (2011), 'Planning to be surprised: Optimal bayesian exploration in dynamic environments', *CoRR* **abs/1103.5708**.
URL: *http://arxiv.org/abs/1103.5708*

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D. & Abbeel, P. (2016), '#exploration: A study of count-based exploration for deep reinforcement learning', *CoRR* **abs/1611.04717**.
URL: *http://arxiv.org/abs/1611.04717*

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D. & Botvinick, M. (2016), 'Learning to reinforcement learn', *CoRR* **abs/1611.05763**.
URL: *http://arxiv.org/abs/1611.05763*

Ying-Jeh Little, D. & Sommer, F. (2013), 'Learning and exploration in action-perception loops', *Frontiers in neural circuits* **7**, 37.