Model-free and Model-based Reinforcement Learning, the Intersection of Learning and Planning

Doctoral Consortium

Piotr Januszewski
Gdańsk University of Technology
Gdańsk, Poland
piotr.januszewski@pg.edu.pl

ABSTRACT

My doctoral dissertation is intended as the compound of four publications considering: structure and randomness in planning and reinforcement learning, continuous control with ensemble deep deterministic policy gradients, toddler-inspired active representation learning, and large-scale deep reinforcement learning costs.

KEYWORDS

deep reinforcement learning; planning and learning; representation learning; energy considerations

ACM Reference Format:
Piotr Januszewski. 2022. Model-free and Model-based Reinforcement Learning, the Intersection of Learning and Planning: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 STRUCTURE AND RANDOMNESS IN PLANNING AND REINFORCEMENT LEARNING

Planning in large state spaces inevitably needs to balance the depth and breadth of the search. It has a crucial impact on the performance of a planner and most manage this interplay implicitly. We[1] present a novel method *Shoot Tree Search (STS)*, which makes it possible to control this trade-off more explicitly by redesigning the expansion phase of the Monte-Carlo Tree Search (MCTS) [4]. Given a leaf and a planning horizon H, the method expands H consecutive vertices according to the in-tree policy and *add them to the search tree*. Our algorithm can be understood as an interpolation between two celebrated search mechanisms: MCTS and random shooting [1].

We tested the STS algorithm in the Google Research Football (GRF) [18] and Sokoban domains. GRF is an advanced, physics-based simulator of the game of football that facilitates the Football Academy consisting of 11 scenarios highlighting various tactical and strategical difficulties. Table 1 compares STS to two baselines: model-free PPO [21] and model-based AlphaZero [23] with a minor environment-specific modification. STS Conv. completely solves 8 out of 11 academies and is the best or close to the best on the remaining 3. Sokoban is a well-known combinatorial puzzle where the agent's goal is to push all boxes to the designed spots and deciding whether a level of Sokoban is solvable or not is PSPACE-complete [9]. The STS learning curve dominates the MCTS learning curve throughout training and, since the difficulty of Sokoban levels increases progressively, it achieves a significant improvement in the final solved rate from 89.5% to 91%.

Our experiments presented here and in the paper support the hypothesis that STS builds a more efficient search tree. Having empirically verified the efficiency of multi-step expansion in many challenging scenarios, we argue that it could be included in a standard MCTS toolbox.

2 CONTINUOUS CONTROL WITH ENSEMBLE DEEP DETERMINISTIC POLICY GRADIENTS

The growth of deep reinforcement learning (RL) has brought multiple exciting tools and methods to the field of decision-making and control [12–14, 20, 26]. This rapid expansion makes it important to understand the interplay between individual elements of the RL toolbox. We[2] approach this task from an empirical perspective by conducting a study in the continuous control setting. We present multiple insights including:

(1) an average of multiple actors trained from the same data boosts performance;
(2) the existing methods are unstable across training runs, epochs of training, and evaluation runs;
(3) the critics' initialization plays a major role in ensemble-based actor-critic exploration.
(4) a commonly used additive action noise is not required for effective training;
(5) a strategy based on posterior sampling explores better than the approximated UCB;
(6) the weighted Bellman backup can neither augment nor replace the clipped double Q-Learning;

We show how existing RL tools can be brought together in a novel way, giving rise to the Ensemble Deep Deterministic Policy Gradients (ED2) method, to yield state-of-the-art results on continuous control tasks from OpenAI Gym MuJoCo. ED2 is an off-policy algorithm for continuous control, which constructs an ensemble of streamlined versions of TD3 [12] agents. Figure 1 shows the results of ED2 contrasted with three strong baselines: SUNRISE [19] (an ensemble-based method), SOP [26], and SAC [13]. In both Hopper and Walker environments, ED2 achieves state-of-the-art performance exceeding all the baseline results but SUNRISE on Walker. ED2 substantially improves the results on the two hardest tasks,

[1] [8] is the joint work with K. Czechowski, P. Kozakowski, Ł. Kuciński and P. Miłoś

[2] [16] is the joint work with M. Olko, M. Królikowski, J. Świątkowski, M. Andrychowicz (Google Brain), Ł. Kuciński and P. Miłoś

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

Table 1: Comparison of selected algorithms on GRF. Entries are rounded solved rates. PPO results come from [18].

		3 vs. 1 with keeper	Corner	Counterattack easy	Counterattack hard	Empty goal	Empty goal close	Pass and shoot with keeper	Run pass and shoot with keeper	Run to score	Run to score with keeper	Single goal versus lazy
	PPO	0.90	0.10	0.70	0.65	0.90	**1.00**	0.65	0.90	0.90	**1.00**	0.90
	AlphaZero	0.81	0.50	0.31	0.31	0.99	**1.00**	0.45	0.89	0.70	0.00	0.00
STS	MLP	**1.00**	0.78	**1.00**	0.97	**1.00**	**1.00**	0.94	0.97	**1.00**	0.94	0.94
STS	Conv.	**1.00**	**0.81**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.97	**0.97**

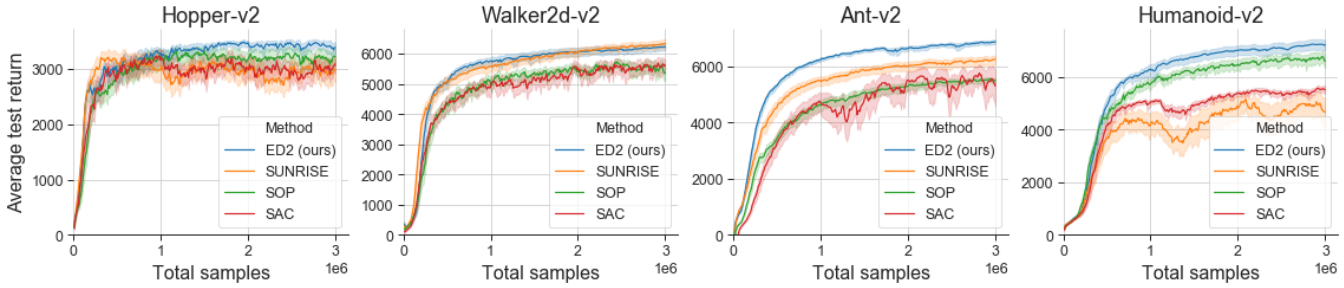

Figure 1: The average test returns (solid lines) across training and the 95% bootstrap confidence interval (shaded regions) over 30 seeds.

Ant and Humanoid, significantly outperforming the baselines. ED2 is also more sample-efficient, achieving the same performance as the next best method with up to 3x fewer environment interactions on Ant. Moreover, experiments in the paper show that ED2 is more stable than the baselines across training runs, epochs of training, and evaluation runs.

We believe that our findings can be useful to both RL researchers and practitioners and allow them to build on top of our work while avoiding pitfalls that we described and tested empirically. From the practical side, ED2 is conceptually straightforward, easy to code, and does not require knowledge outside of the existing RL toolbox.

3 TODDLER-INSPIRED ACTIVE REPRESENTATION LEARNING

Human toddlers are incredibly adept at learning to associate objects to names, and to recognise object instances despite variations in point of view, distractors, and occlusion. As these are the first steps that the human visual system goes through in the process of acquiring the capabilities of adulthood, they are seen as providing critical clues for developing artificial vision systems. Recent research [24] has shown that a toddler's ability to learn is supported by a number of innate strategies, such as the way in which objects are held or how head motions are used to robustly attend to objects of interest. We[1] aim to investigate whether such strategies also result in improved visual learning in artificial agents and if these strategies can be recovered by the agent trained without supervision to learn good visual representation, by developing methods that facilitate deep reinforcement learning.

This is a work in progress. We specified the toy task in the Gym-MiniWorld [7], tuned the PPO [21] agent to the environment, and run the preliminary experiments with rewards being an improvement of the SimCLR [6] objective trained in the inner-loop (the outer-loop being the RL agent training). The next steps include adopting the [5] findings on how to train from a loss as a reward signal and tuning SimCLR to the training from sequential observations stream, in contrast to i.i.d. samples from an offline dataset.

4 LARGE-SCALE DEEP REINFORCEMENT LEARNING COSTS

Recent progress in high-performance computing hardware and deep learning has ushered in a new generation of deep RL algorithms supported by large networks trained on abundant data. These have achieved multiple breakthroughs in a range of challenging domains [2, 3, 22, 25]. However, these depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result, these methods are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware. In this paper, we[2] will measure how much it costs, both financially and environmentally, to reproduce the results of state-of-the-art large-scale deep RL methods [10, 11, 15, 17] and propose actionable strategies to reduce the costs. This work is yet to begin.

ACKNOWLEDGMENTS

The work was or is supported by the Polish National Science Center grant UMO-2017/26/E/ST6/00622 and the PL-Grid, VGG at Oxford, and KASK at GranskTech infrastructure. The experiments were managed using https://neptune.ai.

[1] Joint work with J. F. Henriques and W. Xie from VGG, University of Oxford.

[2] Joint work with P. Czarnul. from Gdańsk University of Technology

REFERENCES

[1] Bruce Abramson. 1990. Expected-Outcome: A General Model of Static Evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1990). https://doi.org/10.1109/34.44404

[2] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research* 39, 1 (2020), 3–20.

[3] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[4] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games* 4, 1 (2012), 1–43.

[5] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. 2019. Exploration by Random Network Distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607.

[7] Maxime Chevalier-Boisvert. 2018. gym-miniworld environment for OpenAI Gym. https://github.com/maximecb/gym-miniworld.

[8] Konrad Czechowski, Piotr Januszewski, Piotr Kozakowski, Łukasz Kuciński, and Piotr Miłoś. 2021. Structure and Randomness in Planning and Reinforcement Learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533317

[9] Dorit Dor and Uri Zwick. 1999. SOKOBAN and other motion planning problems. *Computational Geometry* 13, 4 (1999), 215–228.

[10] Lasse Espeholt, Raphaël Marinier, Piotr Stanczyk, Ke Wang, and Marcin Michalski. 2020. SEED RL: Scalable and Efficient Deep-RL with Accelerated Central Inference. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[11] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. 2018. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1406–1415.

[12] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1582–1591. http://proceedings.mlr.press/v80/fujimoto18a.html

[13] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic Algorithms and Applications. *CoRR* abs/1812.05905 (2018). arXiv:1812.05905 http://arxiv.org/abs/1812.05905

[14] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. 2018. Rainbow: Combining improvements in deep reinforcement learning. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. 3215–3222. arXiv:1710.02298 http://arxiv.org/abs/1710.02298

[15] Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado van Hasselt, and David Silver. 2018. Distributed Prioritized Experience Replay. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[16] Piotr Januszewski, Mateusz Olko, Michał Królikowski, Jakub Świątkowski, Marcin Andrychowicz, Łukasz Kuciński, and Piotr Miłoś. 2021. Continuous Control With Ensemble Deep Deterministic Policy Gradients. In *2021 NeurIPS DeepRL Workshop*.

[17] Steven Kapturowski, Georg Ostrovski, John Quan, Rémi Munos, and Will Dabney. 2019. Recurrent Experience Replay in Distributed Reinforcement Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[18] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. 2019. Google Research Football: A Novel Reinforcement Learning Environment. *CoRR* abs/1907.11180 (2019). http://arxiv.org/abs/1907.11180

[19] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2020. SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. arXiv:2007.04938

[20] Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. 2019. Deep Exploration via Randomized Value Functions. *J. Mach. Learn. Res.* 20 (2019), 124:1–124:62. http://jmlr.org/papers/v20/18-339.html

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347

[22] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[23] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 1144 (2018), 1140–1144.

[24] Lauren K. Slone, Linda B. Smith, and Chen Yu. 2019. Self-Generated Variability in Object Images Predicts Vocabulary Growth. *Developmental Science* 22, 6 (Nov. 2019), e12816. https://doi.org/10.1111/desc.12816

[25] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[26] Che Wang, Yanqiu Wu, Quan Vuong, and Keith Ross. 2020. Striving for Simplicity and Performance in Off-Policy DRL: Output Normalization and Non-Uniform Sampling. *Proceedings of the 37th International Conference on Machine Learning* 119 (13–18 Jul 2020), 10070–10080.