DYNAMIC NOISES OF MULTI-AGENT ENVIRONMENTS CAN IMPROVE GENERALIZATION: AGENT-BASED MODELS MEETS REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the benefits of reinforcement learning (RL) environments based on agent-based models (ABM). While ABMs are known to offer microfoundational simulations at the cost of computational complexity, we empirically show in this work that their non-deterministic dynamics can improve the generalization of RL agents. To this end, we examine the control of an epidemic SIR environments based on either differential equations or ABMs. Numerical simulations demonstrate that the intrinsic noise in the ABM-based dynamics of the SIR model not only improve the average reward but also allow the RL agent to generalize on a wider ranges of epidemic parameters.

1 INTRODUCTION

1.1 BACKGROUND AND RELATED WORK

Most dominant learning algorithms rely on single-agent and optimization-based paradigms. However, many intelligent systems in various domains such as economy, biology, gaming, ecology and sociology are multi-agent and/or decentralized in nature (Ponomarev & Voronkov (2017)). This calls for the need to $i)$ design systems accounting for the intrinsic interactions between its multiple components, and $ii)$ investigate the impact of this consideration in the overall performance and the generalization of the existing algorithms.

Generalization has become one of the major challenges of real-world Reinforcement Learning (RL). Despite the tremendous success of RL in a myriad of applications, the learned policies have difficulty generalizing to new test environments, even when these environments are slightly different from the training ones (Farebrother et al. (2018); Zhang et al. (2018); Cobbe et al. (2019); Song et al. (2020)). Indeed, RL agents are commonly trained and evaluated on the same environment, thereby leading to overfitting and biased evaluation (Whiteson et al. (2011)). This practice is fundamentally different from supervised learning settings where the optimization and the evaluation are conducted on two separate sets (i.e., train and test sets). Empirical risk minimization guarantees a good generalization for supervised learning in absence of data distributional shift. This motivated the construction of similar settings in RL to improve generalization: The agent learns using a training set of contexts or tasks and is evaluated on new unseen test contexts (Cobbe et al. (2019)).

The existing RL suites for generalization such as OpenAI Procgen (Cobbe et al. (2020)) and 2D CoinRun (Cobbe et al. (2019)) are solely devoted to quantify generalization in vision-based RL tasks. To do so, a set of game levels are procedurally generated by varying visual styles and level layouts without changing the environment dynamics. However, little attention has been given for non-vision RL tasks where the agent is supposed to generalize across different dynamics (Packer et al. (2018); Peng et al. (2018)). This is because artificially perturbing the dynamics of complex non-linear dynamical systems (e.g., fluid dynamics, electromagnetism, etc.) requires a careful design of the dynamics randomization procedure to avoid unrealistic scenarios (e.g. deterministic chaos, strange attractors). Such problems are challenging for traditional and deep control methods to handle (Bucci et al. (2019)).

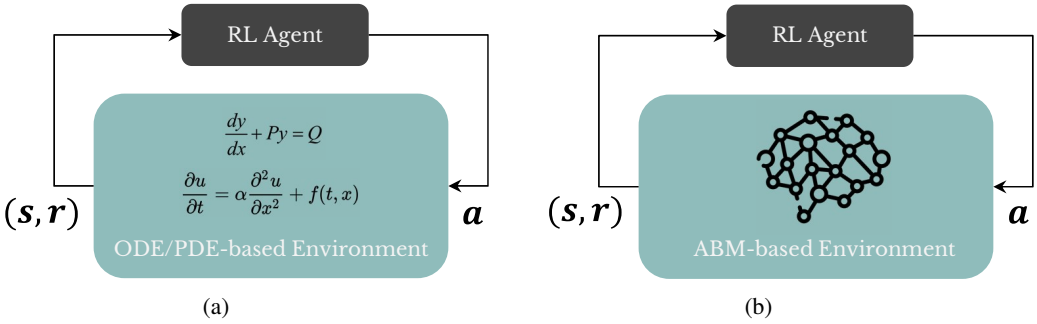

Figure 1: The RL setting using RL environments whose dynamics are simulated using (a) ODEs and PDEs, and (b) ABMs.

1.2 Motivation

As depicted in Fig. 1, we advocate in this work the use of agent-based models (ABM) to design a class of environments for dynamics generalization without the injection of artificial noise. The motivation behind this work stems from the fact that several real-world applications in social and network sciences rely on complex interactions between the components of the overall simulated system. The dynamics of the latter are out of reach for mathematical methods (Epstein & Axtell (1996); Colman (1998)), i.e., they cannot be explicitly governed by ordinary differential equations (ODE) or partial differential equations (PDE). This is in contrast to many real-word applications such as robotics (Kober et al. (2013)), ventilating (Farahmand et al. (2016)), and fluid dynamics Pirmorad et al. (2021) whose complex dynamical systems can be captured using ODEs and PDEs. Moreover, it has been shown that RL agents are prone to exploiting idiosyncrasies of specific implementations of RL environments from which they learn infeasible behaviors in the real world (Heess et al. (2017)). This fact justifies the need for more realistic simulators for RL environments based on ABMs despite their computational complexity.

ABMs come into play as a computational modelling approach for several dynamical systems based on the microfoundational simulation of heterogeneous interactions across predefined agents with heterogeneous preferences (Bianchi & Squazzoni (2015)). ABMs capture emergent self-organizing phenomena induced by the interactions of the individuals, leading to a system behavior whose properties are decoupled of the agents' properties. This aspect commonly summarized by "the whole is more than the sum of its parts" is a modelling bottleneck because its explicit mathematical description is unknown and/or not easy to obtain.

Furthermore, several ODE/PDE-based RL environments tend to smooth out fluctuations by modelling the average behavior across all the agents. However, under certain conditions, fluctuations can be amplified when the system is linearly stable but unstable to larger perturbations. Therefore, simulating a population of agents using dynamic parameters computed from averaging over the behavior of all agents is an oversimplification (Bonabeau (2002)). Indeed, the behavior of each individual is nonlinear and highly dependent on the dynamics of its neighbors. For example, some behavioral scenarios can follow a hard-thresholding decision making strategies. The latter induce a discontinuity in the behavior that cannot be captured by differential equations.

Last but not least, the dynamics of ABMs are non-deterministic due to the randomness in the behavior of each agent. Hence, they induce intrinsic noises that are not artificially injected to examine the effect of the dynamic randomization on the generalization of the RL agent.

1.3 Contributions

The use of ABMs for the microfoundational simulations of the dynamics of RL environments is not new a idea (Rand (2006); Sert et al. (2020)). However, to our best knowledge, the effect of the intrinsic noise in the ABM dynamics on the generalization of RL agents has never been investigated. In this paper, we benchmark the generalization of an RL agent interacting with a standard epidemic SIR environment implemented with either ODEs (Kermack & McKendrick (1927)) or ABMs (Perez & Dragicevic (2009)). We first compare the evolution of the SIR dynamics using ODEs and ABMs

to seize the intrinsic noise in the dynamics of the ABM-based SIR environment. We then randomize the epidemic parameters of the SIR model during both training and inference. By doing so, we show that an RL agent obtains a significant improvement in the average reward and an ability to generalize across different ABM-based environments whose epidemic parameters correspond to distinct pandemic regimes. Our work present an empirical evidence about the relationship between ABM-based environments and RL generalization, thereby calling for the need to further investigate this relationship in future work.

2 THE ABM-BASED RL ENVIRONMENT: THE SIR MODEL AS A STUDY CASE

2.1 THE ODE-BASED SIR DYNAMICS

The spread of a disease can be modelled using the well-known SIR model Kermack & McKendrick (1927). The latter divides the (fixed) population of N individuals into the following three varying group of individuals over the time t (measured in days):

- $S(t)$ is the number of susceptible but not yet infected individuals,
- $I(t)$ is the number of infected individuals,
- $R(t)$ is the number of recovered individual from the disease, and hence are naturally immunized.

Using the finite state machine representation, these three groups are associated to three states as shown in Fig. 2 The transition between them is governed by two key parameters:

- β represents the effective contact rate of the disease. At every time step t, an infected individual in the group $I(t)$ comes into contact with a sub-population of size $\beta\, N$ corresponding to the fraction $S(t)/N(t)$ of individuals that are susceptible to contract the disease.
- γ is the mean recovery rate. This corresponds to a mean period of time, γ^{-1}, during which an infected individual among the group $I(t)$ can spread the virus to its contacts.

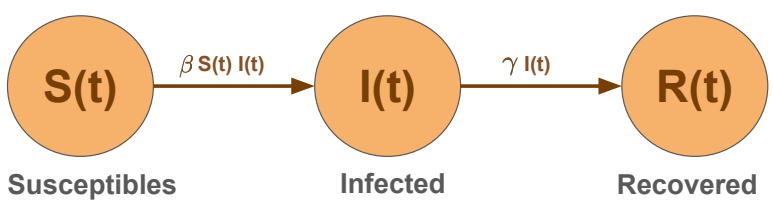

Figure 2: The finite state machine representation of the SIR model.

The dynamics of this system can also be formulated using the following set of differential equations (Kermack & McKendrick (1927)):

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\beta\, \frac{S(t)\, I(t)}{N}, \\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \beta\, \frac{S(t)\, I(t)}{N} - \gamma\, I(t), \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma\, I(t). \end{aligned} \tag{1}$$

2.2 THE ABM-BASED SIR DYNAMICS

We consider a complete undirected graph between N homogeneous agents. Each agent can be in one of the three possible SIR states in Fig 2. At each time step, if the agent is infected, it transmit the disease with a probability $p_{\text{trans}} = \beta/N$ to each neighboring agent in the susceptible state. Moreover, each infected agent transits to the recovered state after γ^{-1} time steps (in days).

The choice of the complete graph topology stems from the fact that it captures more accurately the idea of the aggregate population. In other words, the more interactions between agents are allowed, the closer the average dynamics of the ABM becomes as compared to the ODE-based one.

2.3 THE INTRINSIC NOISE IN THE ABM-BASED SIR DYNAMICS

To verify our implementation of the SIR model with an ABM, we compare the ABM-based dynamics vs. the ODE-based one. Fig. 3 depicts the time evolution of the number of individuals belonging to the susceptible, infected, and recovered populations, namely, $S(t)$, $I(t)$ and $R(t)$, respectively. There, it is seen that the values produced by the ODE model (in solid lines) and the ABM (in dashed lines) converge to approximately the same values. This validates the asymptotic behavior of the ABM model against the ODE model. We also observe a discrepancy between the time evolution of the ABM-based and ODE-based values. This is due to the realistic discrete behavior of the agents as opposed to the averaged one from the ODE. The "microscopic modelling" of ABMs lead to significant deviations from predicted aggregate behavior based on the coarse-grained models (Moein et al. (2021)).

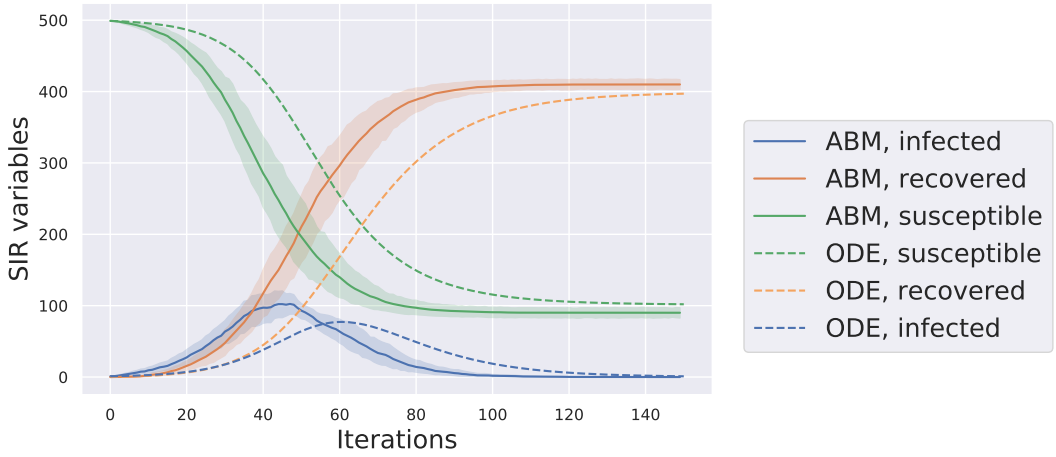

Figure 3: The state evolution of the SIR model as a function of iterations for a population size $N = 500$, $\beta = 0.2$ and $\gamma = 0.1$ when simulated with ODEs (in dashed line) and 100 Monte Carlo realizations of ABMs (in solid line).

The observed deviations from one single simulation to another can be seen as an intrinsic noise in the ABM-based dynamics shown in Fig.3. This non-deterministic aspect of ABM simulations are different from the artificial noise incorporation to deterministic ODE-based dynamics in order to study generalization in RL.

3 SIMULATION RESULTS

In this section, we empirically show how the intrinsic noise in ABM-based dynamics can be beneficial for RL generalization. To this end, we consider two pandemic propagation regimes: a low-spreading one with $\beta = 0.2$ and a wide-spreading one with $\beta = 0.8$.

We train an RL agent using proximal policy optimization (PPO) (Schulman et al. (2017)) to choose an action from the set $\mathcal{A} = \{\text{lockdown, social distancing, open}\}$ to control three SIR models:

- **ODE** : an ODE-based SIR model as described in subsection 2.1;
- **Randomized ODE**: an ODE-based SIR model with dynamics randomisation at each time-step of each episode, where the SIR state ($S(t)$, $I(t)$, $R(t)$) is perturbed with a noise sampled from a uniform probability mass distribution between 1 and 10 individuals.

- **ABM**: an ABM-based SIR model as described in subsection 2.2.

We train and evaluate the RL agent interacting with each one of the aforementioned environments with the default parameters of the PPO implementation of the stable baselines (Raffin et al. (2021)). We evaluate the RL agent on 5000 episodes. The number of evaluation episodes was chosen such as the evaluation time of ABMs does not exceed 12 hours.

To investigate the learned policies of the RL agents, we present in Tables 1 and 2 the distribution of the actions for wide- and low- spreading pandemics, respectively. We also report the average reward during the evaluation in Table 3.

Table 1: The percentage of the taken actions for RL environments simulated based on ODEs, randamized ODEs and ABM for wide-spreading pandemic $\beta = 0.8$.

	Lockdown	**Social distancing**	**Open**
ODE	4.2%	0.9%	94.9%
Randomized ODE	99.3%	0.2%	0.5%
ABM	0.7%	20.1%	79.1%

Table 2: The percentage of the taken actions for RL environments simulated based on ODEs, randamized ODEs and ABM for low-spreading pandemic $\beta = 0.2$.

	Lockdown	**Social distancing**	**Open**
ODE	0.8%	42.1%	57.1%
Randomized ODE	99.2%	0.4%	0.4%
ABM	0.4%	0.3%	99.3%

Table 3: The average reward obtained with different RL environments for both low-spreading ($\beta = 0.2$) and wide-speading ($\beta = 0.8$) pandemics.

	ODE	**Randomized ODE**	**ABM**
$\beta = 0.2$	0.77 $\pm$0.12	-0.16 $\pm$ 0.15	**0.84** $\pm$ **0.14**
$\beta = 0.8$	0.78 $\pm$0.8	-0.15 $\pm$ 0.13	**0.85** $\pm$ **0.11**

We observe that the uniform randomisation of the ODE dynamics has drastically shifted the action distribution of the RL agent and led to a poor average reward in both pandemic regimes. This confirms that the blind randomization of the dynamics cannot be beneficial without domain knowledge, unlike visual-based randomization (Lee et al. (2019)).

Moreover, we observe that the RL agent trained on an ABM-based environment achieves a high average reward in the evaluation. Its policy has shifted from being ignorant of the pandemic dynamics (i.e., almost picks the action "open") in the low-spreading pandemic to a policy that consider social distancing with a frequency of 20% in the wide-spreading regime. This suggests that the noise introduced by the non-deterministic dynamics of ABMs from one episode to another has improved the overall averaged reward. Further experiments should be conducted over a higher number of Monte Carlo simulations to validate this fact.

Finally, we evaluate the generalization of the trained RL agents on test pandemic regimes with different β values. To this end, we randomly sample the values of the parameter β between 0.2 and 0.8 and run the learned policies on the train regimes (i.e., low and wide spreading pandemic regimes). Figure 4 illustrates the obtained results. It shows that the agent trained with ABM-based environments is able to generalize to different pandemic parameters and maintain similar performance (2.3% decrease) as in the train regimes. However, using ODE-based environments, we observe that there is a decrease in the performance (approximately 8%). This confirms that ABMs can offer a better design

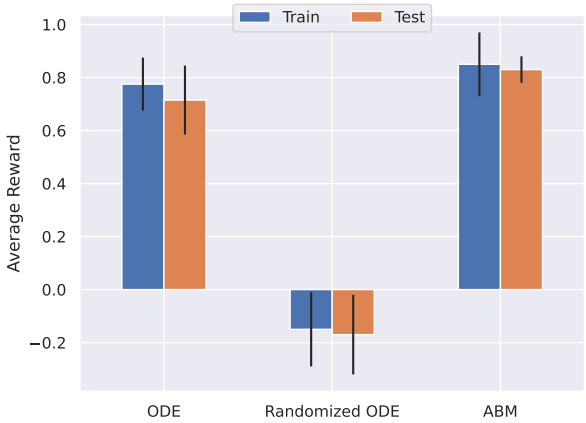

Figure 4: The mean and standard deviation of the average reward over pandemic regimes with β sampled uniformly in $[0.2,\ 0.8]$, while $\gamma = 0.1$.

of RL environments not only for the sake of realistic modelling but also for RL generalization because of their non-deterministic dynamics. Moreover, our results highlight that artificially injecting noise in the agent state did not improve generalization (i.e., negative average reward) on the contrary to the common practice in generalization and sim-to-real works, e.g., (Peng et al. (2018)).

4 Conclusion

In this work, we empirically demonstrated that ABM-based modeling improves the performance and generalization of RL agents due to the intrinsic noise of their dynamics. Our experiments showed that dynamics randomization through random perturbation leads to deterioration of the agent's performance, thereby calling for a more careful design of simulated environments for better generalization. We therefore advocated ABM-based modeling as an approach to tackle dynamics randomization without artificial noise injection during training.

References

Federico Bianchi and Flaminio Squazzoni. Agent-based models in sociology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(4):284–306, 2015.

Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl 3):7280–7287, 2002.

Michele Alessandro Bucci, Onofrio Semeraro, Alexandre Allauzen, Guillaume Wisniewski, Laurent Cordier, and Lionel Mathelin. Control of chaotic systems by deep reinforcement learning. *Proceedings of the Royal Society A*, 475(2231):20190351, 2019.

Karl Cobbe, Oleg Klimov, Christopher Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 2019. URL `http://proceedings.mlr.press/v97/cobbe19a.html`.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2048–2056. PMLR, 2020. URL `http://proceedings.mlr.press/v119/cobbe20a.html`.

Andrew M Colman. The complexity of cooperation: Agent-based models of competition and collaboration, 1998.

Joshua M Epstein and Robert Axtell. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996.

Amir-massoud Farahmand, Saleh Nabi, Piyush Grover, and Daniel N Nikovski. Learning to control partial differential equations: Regularized fitted q-iteration approach. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 4578–4585. IEEE, 2016.

Jesse Farebrother, Marlos C. Machado, and Michael Bowling. Generalization and regularization in DQN. *CoRR*, abs/1810.00123, 2018. URL http://arxiv.org/abs/1810.00123.

Nicolas Heess, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, SM Eslami, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396*, 2019.

Shiva Moein, Niloofar Nickaeen, Amir Roointan, Niloofar Borhani, Zarifeh Heidary, Shaghayegh Haghjooy Javanmard, Jafar Ghaisari, and Yousof Gheisari. Inefficiency of sir models in forecasting covid-19 epidemic: a case study of isfahan. *Scientific Reports*, 11(1):1–9, 2021.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pp. 1–8. IEEE, 2018. doi: 10.1109/ICRA.2018.8460528. URL https://doi.org/10.1109/ICRA.2018.8460528.

Liliana Perez and Suzana Dragicevic. An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics*, 8(1):1–17, 2009.

Erfan Pirmorad, Faraz Khoshbakhtian, Farnam Mansouri, and Amir-massoud Farahmand. Deep reinforcement learning for online control of stochastic partial differential equations. *arXiv preprint arXiv:2110.11265*, 2021.

S Ponomarev and AE Voronkov. Multi-agent systems and decentralized artificial superintelligence. *arXiv preprint arXiv:1702.08529*, 2017.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

William Rand. Machine learning meets agent-based modeling: When not to go to a bar. In *Conference on Social Agents: Results and Prospects*, 2006.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Egemen Sert, Yaneer Bar-Yam, and Alfredo J Morales. Segregation dynamics with reinforcement learning and agent based modeling. *Scientific reports*, 10(1):1–12, 2020.

Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, and Behnam Neyshabur. Observational overfitting in reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJli2hNKDH.

Shimon Whiteson, Brian Tanner, Matthew E. Taylor, and Peter Stone. Protecting against evaluation overfitting in empirical reinforcement learning. In *2011 IEEE Symposium on Adaptive Dynamic Programming And Reinforcement Learning, ADPRL 2011, Paris, France, April 12-14, 2011*, pp. 120–127. IEEE, 2011. doi: 10.1109/ADPRL.2011.5967363. URL https://doi.org/10.1109/ADPRL.2011.5967363.

Chiyuan Zhang, Oriol Vinyals, Rémi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *CoRR*, abs/1804.06893, 2018. URL http://arxiv.org/abs/1804.06893.