
Revisiting Rainbow: Promoting more Insightful and Inclusive Deep Reinforcement Learning Research

Johan S. Obando-Ceron Pablo Samuel Castro¹

Abstract

Since the introduction of DQN, a vast majority of reinforcement learning research has focused on reinforcement learning with deep neural networks as function approximators. New methods are typically evaluated on a set of environments that have now become standard, such as Atari 2600 games. While these benchmarks help standardize evaluation, their computational cost has the unfortunate side effect of widening the gap between those with ample access to computational resources, and those without. In this work we argue that, despite the community’s emphasis on large-scale environments, the traditional small-scale environments can still yield valuable scientific insights and can help reduce the barriers to entry for underprivileged communities. To substantiate our claims, we empirically revisit the paper which introduced the Rainbow algorithm (Hessel et al., 2018) and present some new insights into the algorithms used by Rainbow.

1. Introduction

Since the introduction of DQN (Mnih et al., 2015) reinforcement learning has witnessed a dramatic increase in research papers (Henderson et al., 2018). A large portion of these papers propose new methods that build on the original DQN algorithm and network architecture, often adapting methods introduced before DQN to work well with deep networks (e.g., (van Hasselt et al., 2016; Bacon et al., 2017; Castro, 2020)). New methods are typically evaluated on a set of environments that have now become standard, such as the Atari 2600 games made available in the Arcade Learning Environment (ALE) (Bellemare et al., 2012) and the control tasks available in MuJoCo and DM control suites (Todorov et al., 2012; Tassa et al., 2020).

¹Google Research, Brain Team. Correspondence to: Johan S. Obando-Ceron <jobando0730@gmail.com>, Pablo Samuel Castro <psc@google.com>.

While these benchmarks have helped to evaluate new methods in a standardized manner, they have also implicitly established a minimum amount of computing power in order to be recognized as valid scientific contributions. Although classic reinforcement learning tasks such as MountainCar, CartPole, Acrobot, and grid worlds have not gone away, they are now used mostly for evaluating theoretical contributions (e.g., (Nachum et al., 2019; Lyle et al., 2019)); indeed, in our experience it is quite difficult to have a paper proposing a new reinforcement learning method accepted at one of the major machine learning conferences unless it includes experiments with one of the benchmarks just mentioned. This is unfortunate, as the low computational cost and speed at which one can train on small-scale environments enables broad hyper-parameter sweeps and more thorough investigations into the nuances of the methods being considered, as well as the reevaluation of certain empirical choices that have become “folk wisdom” in these types of experiments.

Furthermore, at a time when efforts such as Black in AI and LatinX in AI are helping bring people from underrepresented (and typically underprivileged) segments of society into the research community, these newcomers are faced with enormous computational hurdles to overcome if they wish to be an integral part of said community.

It thus behooves the reinforcement learning research community to incorporate a certain degree of flexibility and creativity when proposing and evaluating new research; of course, this should not be at the expense of scientific rigour. This paper is partly a position paper, partly an empirical evaluation. We argue for a need to change the status-quo in evaluating and proposing new research to avoid exacerbating the barriers to entry for newcomers from underprivileged communities. In section 4 we complement this argument by revisiting the Rainbow algorithm (Hessel et al., 2018), which proposed a new state of the art algorithm by combining a number of recent advances, on a set of small- and medium-sized tasks. This allows us to conduct a “counterfactual” analysis: would Hessel et al. (2018) have reached the same conclusions if they had run on the smaller-scale experiments we investigate here? In section 5 we extend this analysis by investigating the interaction between the different algorithms considered and the network architecture used,

varying the distribution parameterization and bootstrapping methodology, and the base loss used by DQN. Finally, in section 6 we compare the Rainbow variants we considered, as well as provide insights into the properties of the different environments used in our study.

2. Preliminaries

Reinforcement learning methods are used for learning how to act (near) optimally in sequential decision making problems in uncertain environments. In the most common scenario, an *agent* transitions between *states* in an environment by making *action* choices at discrete time steps; upon performing an action, the environment produces a numerical *reward* and transitions the agent to a new state. This is formalized as a Markov decision process (Puterman, 1994) $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$, where \mathcal{S} is a finite set of states, \mathcal{A} is a finite set of actions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow [R_{min}, R_{max}]$ is the reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is the transition function, where $\Delta(X)$ denotes the set of probability distributions over a set X , and $\gamma \in [0, 1]$ is a horizon discount factor.

An agent’s behaviour is formalized as a policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ which induces a *value function* $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$:

$$\begin{aligned} V^\pi(s) &:= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) | s_0 = s, s_t, a_t \right] \\ &= \mathbb{E}_{a \sim \pi(s)} [\mathcal{R}(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s, a)} V^\pi(s')] , \end{aligned}$$

where $s_t \sim \mathcal{P}(s_{t-1}, a_{t-1})$ and $a_t \sim \pi(s_t)$. The second line is the well-known Bellman recurrence. It is also convenient to consider the value of actions that differ from those encoded in π via the function $Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: $Q^\pi(s, a) := \mathcal{R}(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s, a)} V^\pi(s')$.

It is well known that there exist policies π^* that are optimal in the sense that $V^*(s) := V^{\pi^*}(s) \geq V^\pi(s)$ for all policies π and states $s \in \mathcal{S}$. In reinforcement learning we are typically interested in having agents find these policies by interacting with the environment. One of the most popular ways to do so is via *Q*-learning, where an agent maintains a function Q_θ , parameterized by θ (e.g. the weights in a neural network), and updates it after observing the transition $s \xrightarrow{a, r} s'$ using the method of *temporal differences* (Sutton, 1988):

$$Q_\theta(s, a) \leftarrow Q_\theta(s, a) + \alpha (Q^*(s', a') - Q_\theta(s, a)) \quad (1)$$

where α is a learning rate and the optimal target values $Q^*(s', a') = r + \gamma \max_{a' \in \mathcal{A}} Q_\theta(s', a')$

2.1. DQN

Mnih et al. (2015) introduced DQN, which combined *Q*-learning with deep networks. Some of the most salient design choices are:

- The *Q* function is represented using a feed forward neural network consisting of three convolutional layers followed by two fully connected layers. Two copies of the *Q*-network are maintained: an *online* network (parameterized by θ) and a *target* network (parameterized by $\bar{\theta}$). The online network is updated via the learning process described below, while the target network remains fixed and is synced with the online weights at less frequent (but regular) intervals.

- A large *replay buffer* D is maintained to store experienced transitions (s, a, r, s') (Lin, 1992).
- The update in Equation 1 is implemented using the following *loss function* to update the *online* network:

$$L(\theta) = \mathbb{E}_{(s, a, r, s') \sim U(D)} [(Y^{DQN} - Q_\theta(s, a))^2] \quad (2)$$

using $Y^{DQN} = (r + \gamma \max_{a' \in \mathcal{A}} Q_{\bar{\theta}}(s', a'))$ and *mini-batches* of size 32.

2.2. Rainbow

In this section we briefly present the enhancements to DQN that were combined by Hessel et al. (2018) for the Rainbow agent.

Double Q-learning: van Hasselt et al. (2016) added double Q-learning (Hasselt, 2010) to mitigate overestimation bias in the *Q*-estimates by decoupling the maximization of the action from its selection in the target bootstrap.

Prioritized experience replay: Instead of sampling uniformly from the replay buffer ($U(D)$), prioritized experience replay (Schaul et al., 2016) proposed to sample a trajectory $t = (s, a, r, s')$ with probability p_t proportional to the temporal difference error.

Dueling networks: Wang et al. (2016) introduced dueling networks by modifying the DQN network architecture. Specifically, two streams share the initial convolutional layers, separately estimating $V^*(s)$, and the advantages for each action: $A(s, a) := Q^*(s, a) - V^*(s)$. The output of the network is a combination of these two streams.

Multi-step learning: Instead of computing the temporal difference error using a single-step transition, one can use multi-step targets instead (Sutton, 1988), where for a trajectory $(s_0, a_0, r_0, s_1, a_1, \dots)$ and update horizon n : $R_t^{(n)} := \sum_{k=0}^{n-1} \gamma^k r_{t+k+1}$, yielding the multi-step temporal difference: $R_t^{(n)} + \gamma^n \max_{a' \in \mathcal{A}} Q_{\bar{\theta}}(s_{t+n}, a') - Q_\theta(s_t, a_t)$.

Distributional RL: Bellemare et al. (2017) demonstrated that the Bellman recurrence also holds for *value distributions*: $Z(x, a) \stackrel{D}{=} R(s, a) + \gamma Z(X', A')$, where Z , R , and (X', A') are random variables representing the return, immediate reward, and next state-action, respectively. The authors present an algorithm (C51) to maintain an estimate

Z_θ of the return distribution Z by use of a parameterized categorical distribution with 51 atoms.

Noisy nets: Fortunato et al. (2018) propose replacing the simple ϵ -greedy exploration strategy used by DQN with noisy linear layers that include a noisy stream.

3. The cost of Rainbow

Although the *value* of the hybrid agent uncovered by Hessel et al. (2018) is undeniable, this result could have only come from a large research laboratory with ample access to compute. Indeed, each Atari 2600 game from the ALE (there are 57 in total) takes roughly 5 days to fully train using specialized hardware (for example, an NVIDIA Tesla P100 GPU).¹ Additionally, in order to be able to report performance with confidence bounds it is common to use at least five independent runs (Henderson et al., 2018; Machado et al., 2018).

Thus, to provide the convincing empirical evidence for Rainbow, Hessel et al. (2018) required approximately 34,200 GPU hours (or 1425 days). Note that this cost does not include the hyper-parameter tuning that was necessary to optimize the various components. Considering that the cost of a Tesla P100 GPU is around US\$6,000, providing this evidence will take an unreasonably long time as it is prohibitively expensive to have multiple GPUs in a typical academic lab so they can be used in parallel. As a reference point, the average minimum monthly wage in South America (excluding Venezuela) is approximately US\$313²; in other words, one GPU is the equivalent of approximately 20 minimum wages. Needless to say, this expectation is far from inclusive.

In light of this, we wish to investigate three questions:

1. Would Hessel et al. (2018) have arrived at the same qualitative conclusions, had they run their experiments on a set of smaller-scale experiments?
2. Do the results of Hessel et al. (2018) generalize well to non-ALE environments, or are their results overly-specific to the chosen benchmark?
3. Is there scientific value in conducting empirical research in reinforcement learning when restricting oneself to small- to mid-scale environments?

We investigate the first two in Section 4, and the last in Sections 5 and 6.

¹The computational expense is not limited to the ALE: MuJoCo tasks from the DM control suite take about 2 days with the same hardware.

²Taken from <https://www.statista.com/statistics/953880/latin-america-minimum-monthly-wages/>

4. Revisiting Rainbow

4.1. Methodology

We follow a similar process as Hessel et al. (2018) in evaluating the various algorithmic variants mentioned above: we investigate the effect of adding each on top of the original DQN agent as well as the effect of dropping each from the final Rainbow agent, sweeping over learning rates for each. Our implementation is based on the Dopamine framework (Castro et al., 2018). Note that Dopamine includes a “lite” version of Rainbow, which does not include noisy networks, double DQN, nor dueling networks, but we have added all these components in our implementation³.

We perform our empirical evaluation on small-scale environments (CartPole, Acrobot, LunarLander, and MountainCar) which are all available as part of the OpenAI Gym library (Brockman et al., 2016) (see Appendix A for a detailed description of each environment). We used multilayer perceptrons (MLPs) with 2 layers of 512 units each for these experiments. The agents were all trained on a CPU; it is worth noting that of these environments the one that took longest to train (LunarLander) is still able to finish in under two hours.

In order to strengthen the Rainbow Connection, we also ran a set of experiments on the Minatar environment (Young and Tian, 2019), which is a set of miniaturized versions of five ALE games (Asterix, Breakout, Freeway, Seaquest, and SpaceInvaders). These environments are considerably larger than the four classic control environments previously explored, but they are significantly faster to train than regular ALE environments. Specifically, training one of these agents takes approximately 12-14 hours on a P100 GPU. For these experiments, we followed the network architecture used by Young and Tian (2019) consisting of a single convolutional layer followed by a linear layer.

4.2. Empirical evaluation

Under constant hyper-parameter settings (see appendix for details), we evaluate both the *addition* of each algorithmic component to DQN, as well as their *removal* from the full Rainbow agent on the classic control environments (Figure 1) and on Minatar (Figure 2).

We analyze our results in the context of the first two questions posed in Section 3: *Would Hessel et al. (2018) have arrived at the same qualitative conclusions, had they run their experiments on a set of smaller-scale experiments? Do the results of Hessel et al. (2018) generalize well to non-ALE environments, or are their results overly-specific to the chosen benchmark?*

³Source code available at https://github.com/JohanSamir/revisiting_rainbow

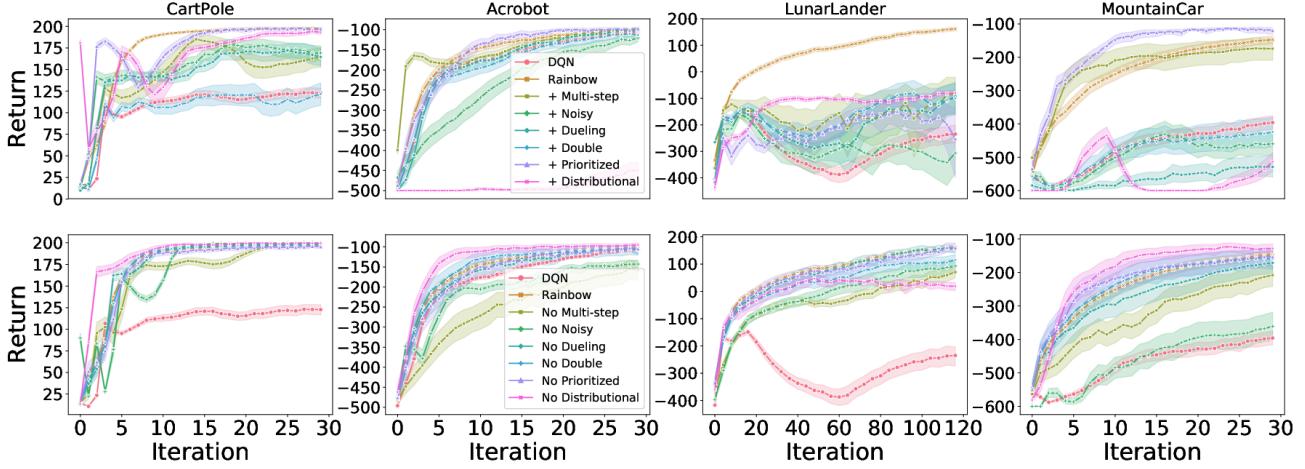


Figure 1. Comparison of the different algorithmic components on the four small environments using the optimal hyper-parameters from Table 1 for each, averaged over 100 independent runs (shaded areas show 95% confidence intervals). Top row explores adding on top of DQN, bottom row explores removing from Rainbow.

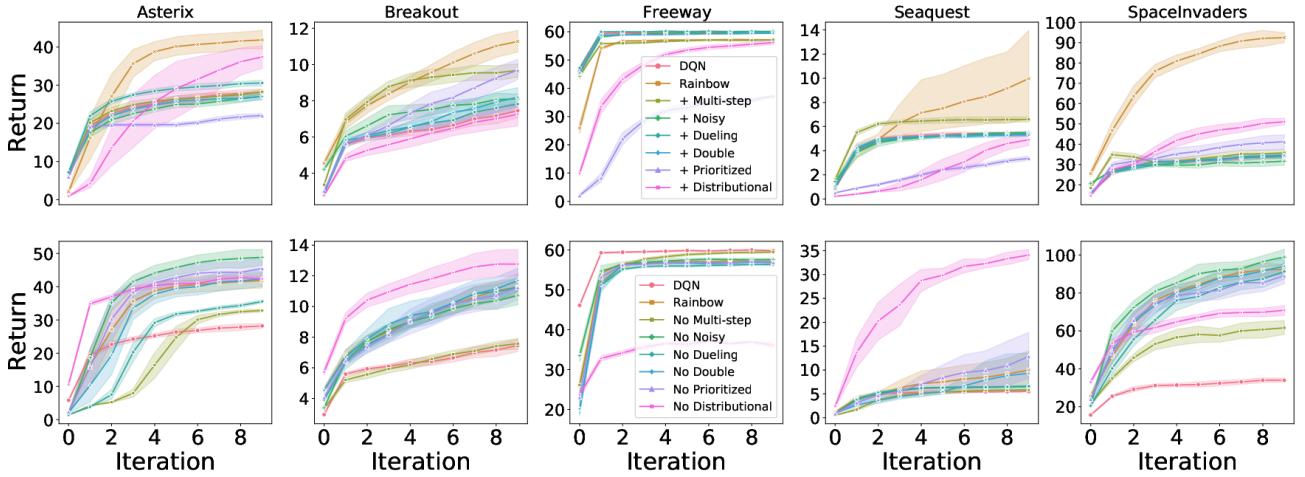


Figure 2. Comparison of the different algorithmic components on the five MinAtar games, averaged over 10 independent runs (shaded areas show 90% confidence intervals). Top row explores adding on top of DQN, bottom row explores removing from Rainbow.

What we find is that the performance of the different components is not uniform throughout all environments; a finding which is consistent with the results observed by Hessel et al. (2018). However, if we were to suggest a single agent that balances the tradeoffs of the different algorithmic components, our analysis would be consistent with Hessel et al. (2018): combining all components produces a better overall agent.

Nevertheless, there are important details in the variations of the different algorithmic components that merit a more thorough investigation. An important finding of our work is that distributional RL, when added on its own to DQN, may actually hurt performance (e.g. Acrobot and Freeway); similarly, performance can sometimes increase when distributional RL is removed from Rainbow (e.g. MountainCar

and Seaquest); this is in contrast to what was found by Hessel et al. (2018) on the ALE experiments and warrants further investigation. As Lyle et al. (2019) noted, under non-linear function approximators (as we are using in these experiments), using distributional RL generally produces *different* outcomes than the non-distributional variant, but these differences are not always positive.

5. Beyond the Rainbow

In this section we seek to answer the third question posed in section 3: *Is there scientific value in conducting empirical research in small- to mid-scale environments?* We leverage the low cost of the small-scale environments to conduct a thorough investigation into some of the algorithmic compo-

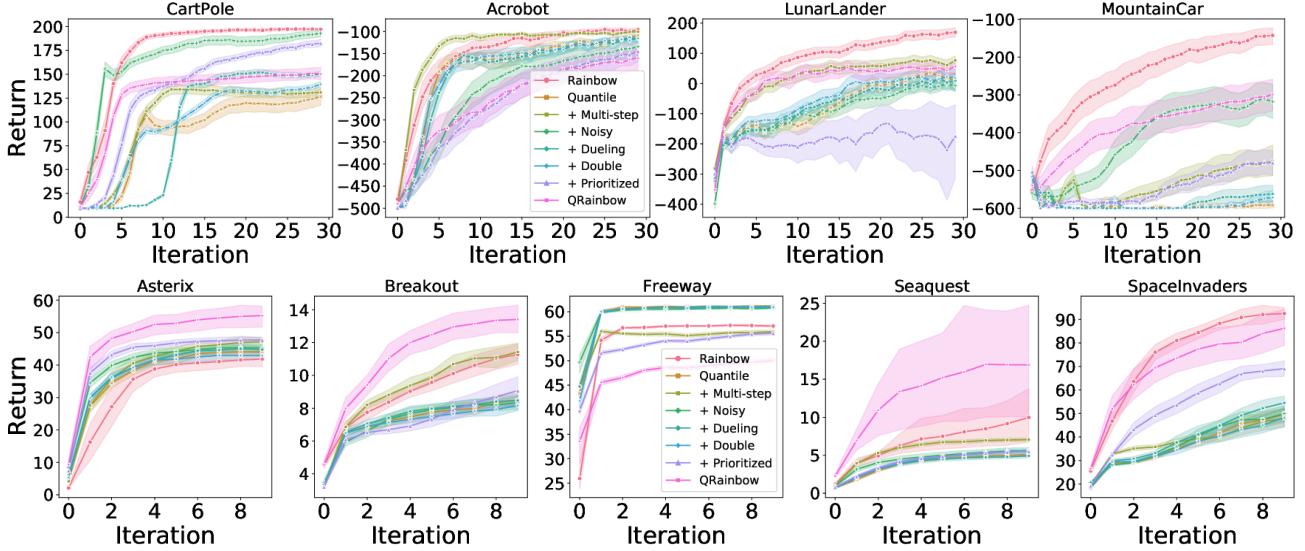


Figure 3. Investigating the combination of the different algorithmic components with QR-DQN (shaded areas show 95% (top) and 90% (bottom) confidence intervals).

nents studied. Unless otherwise specified, the classic control and MinAtar environments are averaged over 100 and 10 independent runs, respectively; in both cases shaded areas report 95% confidence intervals.

5.1. Examining network architectures and batch sizes

We investigated the interaction of the best per-game hyperparameters with the number of layers and units per layer. Due to space constraints we omit the figures from the main paper, but include them in the appendix. We found that in general using 2-3 layers with at least 256 units each yielded the best performance. Further, aside from Cartpole, the algorithms were generally robust to varying network architecture dimensions.

Another often overlooked hyper-parameter in training RL agents is the batch size. We investigated the sensitivity of DQN and Rainbow to varying batch sizes and found that while for DQN it is sub-optimal to use a batch size below 64, Rainbow seems fairly robust to both small and large batch sizes.

5.2. Examining distribution parameterizations

Although distributional RL is an important component of the Rainbow agent, at the time of its development Rainbow was only evaluated with the C51 parameterization of the distribution, as originally proposed by [Bellemare et al. \(2017\)](#). Since then there have been a few new proposals for parameterizing the return distribution, notably quantile regression ([Dabney et al., 2017; 2018a](#)) and implicit quantile networks ([Dabney et al., 2018b](#)). In this section we investigate the in-

teraction of these parameterizations with the other Rainbow components.

Quantile Regression for Distributional RL

In contrast to C51, QR-DQN ([Dabney et al., 2017; 2018a](#)) computes the return quantile values for N fixed, uniform probabilities. Compared to C51, QR-DQN has no restrictions or bound for value, as the distribution of the random return is approximated by a uniform mixture of N Diracs: $Z_\theta(x, a) := \frac{1}{N} \sum_{i=1}^N \delta_{\theta_i(x, a)}$, with each θ_i assigned a quantile value trained with quantile regression. In [Figure 3](#) we evaluate the interaction of the different Rainbow components with Quantile and find that, in general, QR-DQN responds favourably when augmented with each of the components. We also evaluate a new agent, *QRainbow*, which is the same as Rainbow but with the QR-DQN parameterization. It is interesting to observe that in the classic control environments Rainbow outperforms QRainbow, but QRainbow tends to perform better than Rainbow on Minatar (with the notable exception of Freeway), suggesting that perhaps the quantile parameterization of the return distribution has greater benefits when used with networks that include convolutional layers.

Implicit quantile networks

We continue by investigating using implicit quantile networks (IQN) as the parameterization of the return distribution [Dabney et al. \(2018b\)](#). IQN learns to transform a base distribution (typically a uniform distribution in $[0, 1]$) to the quantile values of the return distribution. This can result in greater representation power in comparison to QR-DQN, as well as the ability to incorporate *distortion risk measures*.

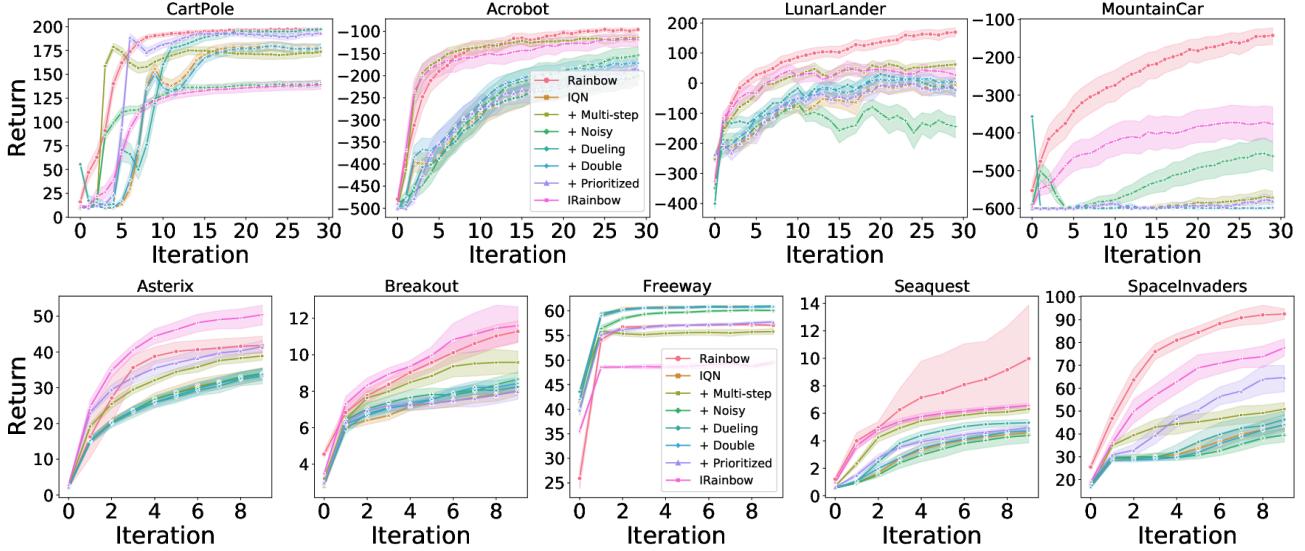


Figure 4. Investigating the combination of the different algorithmic components with IQN (shaded areas show 95% (top) and 90% (bottom) confidence intervals).

We repeat the “Rainbow experiment” with IQN and report the results in Figure 4. In contrast to QR-DQN, in the classic control environments the effect on performance of various Rainbow components is rather mixed and, as with QR-DQN *IRainbow* underperforms Rainbow. In Minatar we observe a similar trend as with QR-DQN: *IRainbow* outperforms Rainbow on all the games except Freeway.

5.3. Munchausen Reinforcement Learning

Vieillard et al. (2020) introduced *Munchausen RL* as a simple variant to any temporal difference learning agent consisting of two main components: the use of stochastic policies and augmenting the reward with the scaled log-policy. Integrating their proposal to DQN yields M-DQN with performance superior to that of C51; the integration of Munchausen-RL to IQN produced M-IQN, a new state-of-the art agent on the ALE.

In Figure 5 we report the results when repeating the Rainbow experiment on M-DQN and M-IQN. In the classic control environments neither of the Munchausen variants seem to yield much of an improvement over their base agents. In Minatar, while M-DQN does seem to improve over DQN, the same cannot be said of M-IQN. We explored combining all the Rainbow components⁴ with the Munchausen agents and found that, in the classic control environments, while M-Rainbow underperforms relative to its non-Munchausen counterpart, M-IRainbow can provide gains. In Minatar, the results vary from game to game, but it appears that the

⁴We were unable to successfully integrate M-DQN with C51 nor double-DQN, so our M-Rainbow agent is compared against Rainbow without distributional RL and without double-DQN.

Munchausen agents yield an advantage on the same games (Asterix, Breakout, and SpaceInvaders).

5.4. Reevaluating the Huber loss

The Huber loss is what is usually used to train DQN agents as it is meant to be less sensitive to outliers. Based on recent anecdotal evidence, we decided to evaluate training DQN using the mean-squared error (MSE) loss and found the surprising result that on *all* environments considered using the MSE loss yielded much better results than using the Huber loss, sometimes even surpassing the performance of the full Rainbow agent (full classic control and Minatar results are provided in the appendix). This begs the question as to whether the Huber loss is truly the best loss to use for DQN, especially considering that reward clipping is typically used for most ALE experiments, mitigating the occurrence of outlier observations. Given that we used Adam (Kingma and Ba, 2015) for all our experiments while the original DQN algorithm used RMSProp, it is important to consider the choice of optimizer in answering this question.

To obtain an answer, we compared the performance of the Huber versus the MSE loss when used with both the Adam and RMSProp optimizers on all 60 Atari 2600 games. In Figure 6 we present the improvement obtained when using the Adam optimizer with the MSE loss over using the RMSProp optimizer with the Huber loss and find that, overwhelmingly, Adam+MSE is a superior combination than RMSProp+Huber. In the appendix we provide complete comparisons of the various optimizer-loss combinations that confirm our finding. Our analyses also show that, when using RMSProp, the Huber loss tends to perform better

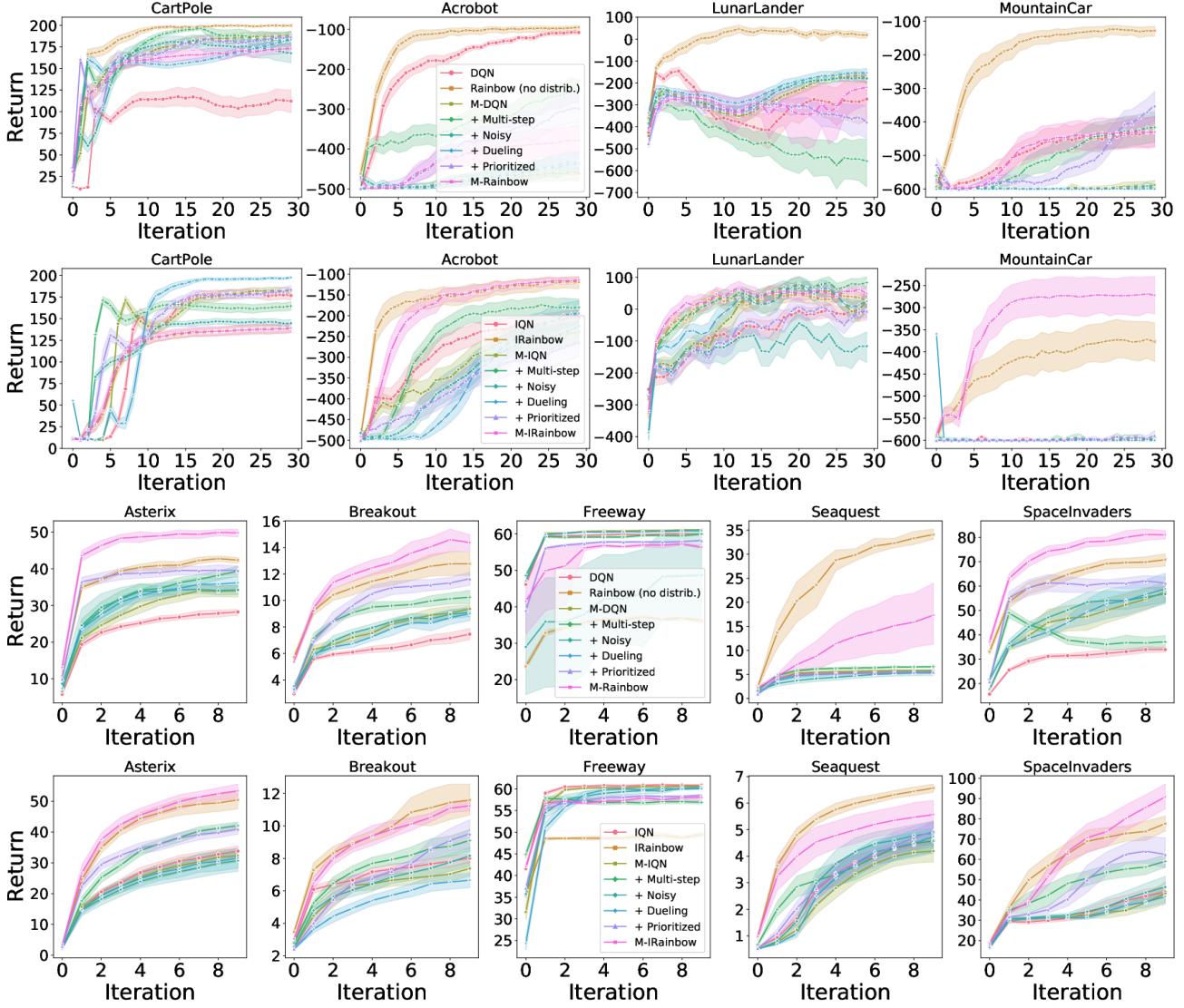


Figure 5. Investigating the combination of the different algorithmic components with M-DQN and M-IQN (shaded areas show 95% (top 2) and 90% (bottom 2) confidence intervals).

than MSE, which in retrospect explains why (Mnih et al., 2015) chose the Huber over the simpler MSE loss when introducing DQN.

Our findings highlight the importance in properly evaluating the interaction of the various components used when training RL agents, as was also argued by Fujimoto et al. (2020) with regards to loss functions and non-uniform sampling from the replay buffer; as well as by Hessel et al. (2019) with regards to inductive biases used in training RL agents.

6. Putting it all together

6.1. Rainbow flavours

We compare the performance of DQN against all of the Rainbow variants and show the results of two environments in Figure 7 (full comparisons in the appendix). These two environments highlight the fact that, although Rainbow does outperform DQN, there are important differences amongst the various flavours that invite further investigation.

6.2. Environment properties

Our exhaustive experimentation on the four classic control and five MinAtar environments grant us some insight into

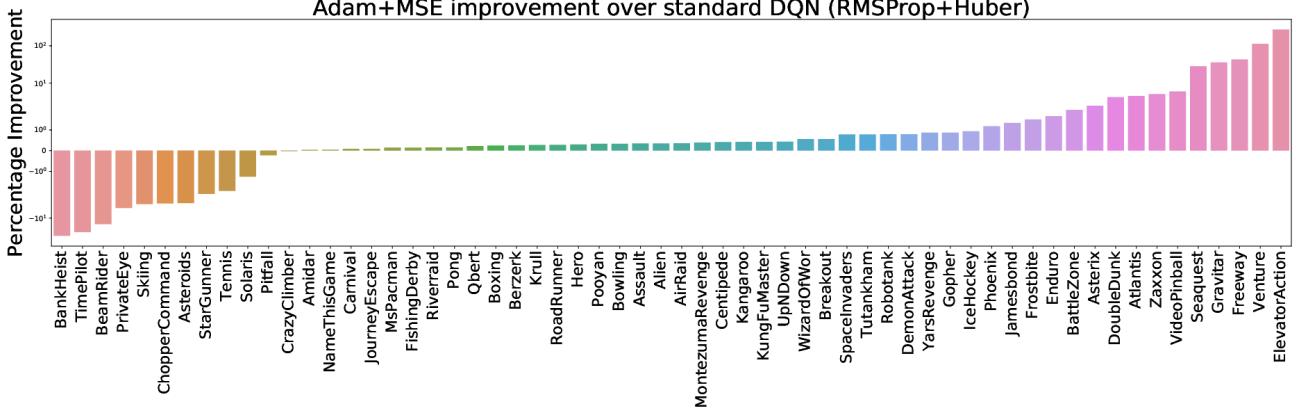


Figure 6. Comparison of training DQN using the Adam optimizer with the MSE loss, versus the Huber loss with the RMSProp. All results report the average of 5 independent runs.

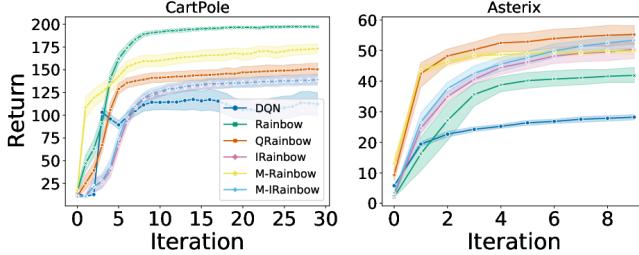


Figure 7. Comparing DQN against the various Rainbow flavours.

their differing properties. We believe these environments pose a variety of interesting challenges for RL research and present a summary of our insights here, with a more thorough analysis in the appendix, in the hope that it may stimulate future investigations.

Classic control

Although **CartPole** is a relatively simple task, DQN can be quite sensitive to learning rates and as such, can prove an interesting testbed for optimizer stability. We found the use of Noisy networks and MSE loss to dramatically help with this sensitivity. It seems that distributional RL is required for obtaining good performance on **LunarLander** (see the poor results with DQN, M-DQN, and M-Rainbow in comparison with the others), suggesting this would be a good environment to investigate the differences in expectation and distributional RL as proposed by Lyle et al. (2019). Both **Acrobot** and **MountainCar** are sparse reward environments, which are typically good environments for investigating exploration strategies; indeed, we observe that Noisy networks tend to give a performance gain in both these environments. MountainCar appears to be the more difficult of the two, a fact also observed by Tang et al. (2017), Colas et al. (2018a), and Declan et al. (2020).

MinAtar

The value of MinAtar environments has already been argued by Young and Tian (2019) and recently exemplified by Ghassian et al. (2020) and Buckman et al. (2021). It is worth highlighting that both **Seaquest** and **Freeway** appear to lend themselves well for research on exploration methods, due to their partial observability and reward sparsity. We would like to stress that these environments enable researchers to investigate the importance and effect of using convolutional networks in RL without the prohibitive expense of the ALE benchmark.

7. Conclusion

On a limited computational budget we were able to reproduce, at a high-level, the findings of Hessel et al. (2018) and uncover new and interesting phenomena. Evidently it is much easier to revisit something than to discover it in the first place; however, our intent with this work was to argue for the relevance and significance of empirical research on small- and medium-scale environments. We believe that these less computationally intensive environments lend themselves well to a more critical and thorough analysis of the performance, behaviours, and intricacies of new algorithms (a point also argued by Osband et al. (2020)). It is worth remarking that when we initially ran 10 independent trials for the classic control environments, the confidence intervals were very wide and inconclusive; boosting the independent trials to 100 gave us tighter confidence intervals with small amounts of extra compute. This would be impractical for most large-scale environments. Ensuring statistical significance when comparing algorithms, a point made by Colas et al. (2018b) and Jordan et al. (2020), is facilitated with the ability to run a large number of independent trials.

We are by no means calling for less emphasis to be placed on

large-scale benchmarks. We are simply urging researchers to consider smaller-scale environments as a valuable tool in their investigations, and reviewers to avoid dismissing empirical work that focuses on these smaller problems. By doing so, we believe, we will get both a clearer picture of the research landscape and will reduce the barriers for newcomers from diverse, and often underprivileged, communities. These two points can only help make our community and our scientific advances stronger.

8. Acknowledgements

The authors would like to thank Marlos C. Machado, Sara Hooker, Matthieu Geist, Nino Vieillard, Hado van Hasselt, Eleni Triantafillou, and Brian Tanner for their insightful comments on our work.

References

- Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI’17, page 1726–1734, 2017.
- Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, Vol. 47:253–279, 2012. cite arxiv:1207.4708.
- Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 449–458, 2017.
- Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.
- Jacob Buckman, Carles Gelada, and Marc G. Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *Proceedings of the Ninth International Conference on Learning Representations*, ICLR’21, 2021.
- Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
- Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning. 2018. URL <http://arxiv.org/abs/1812.06110>.
- Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1039–1048. PMLR, 2018a.
- Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *CoRR*, abs/1806.08295, 2018b.
- W. Dabney, M. Rowland, Marc G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *AAAI*, 2018a.
- Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression, 2017.
- Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1096–1105. PMLR, 2018b.
- Oller Declan, Glasmachers Tobias, and Cuccu Giuseppe. Analyzing reinforcement learning benchmarks with random weight guessing. *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, AAMAS 2020, pages 975–982, 2020.
- Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alexander Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *Proceedings of the International Conference on Representation Learning (ICLR 2018)*, Vancouver (Canada), 2018.
- Scott Fujimoto, David Meger, and Doina Precup. An equivalence between loss functions and non-uniform sampling in experience replay. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.
- Sina Ghassian, Andrew Patterson, Shivam Garg, Dhawal Gupta, Adam White, and Martha White. Gradient temporal-difference learning with regularized corrections. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3524–3534. PMLR, 13–18 Jul 2020.
- Hado V. Hasselt. Double q-learning. In *Advances in Neural Information Processing Systems 23*, pages 2613–2621. Curran Associates, Inc., 2010.

- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirhy-Second AAAI Conference On Artificial Intelligence (AAAI), 2018*, 2018.
- Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining Improvements in Deep Reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- Matteo Hessel, Hado van Hasselt, Joseph Modayil, and David Silver. On inductive biases in deep reinforcement learning. *CoRR*, abs/1907.02908, 2019.
- Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. doi: 10.1214/aoms/1177703732. URL <https://doi.org/10.1214/aoms/1177703732>.
- Scott Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip Thomas. Evaluating the performance of reinforcement learning algorithms. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4962–4973. PMLR, 13–18 Jul 2020.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learn.*, 8(3–4):293–321, May 1992.
- Clare Lyle, Marc G. Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *AAAI*, 2019.
- Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the Arcade Learning Environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 2018.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.
- Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *NeurIPS*, 2019.
- Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvári, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado van Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rygf-kSYwh>.
- Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *Proceedings of ICLR*, 2016, 2016.
- Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, August 1988.
- Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 2753–2762. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3a20f62a0af1aa152670bab3c602feed-Paper.pdf>.
- Yuval Tassa, Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, and Nicolas Heess. dm_control: Software and tasks for continuous control, 2020.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, pages 5026–5033. IEEE, 2012.
- Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirthieth AAAI Conference On Artificial Intelligence (AAAI), 2016*, 2016. cite arxiv:1509.06461Comment: AAAI 2016.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen reinforcement learning, 2020.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1995–2003, 2016.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.