

A LARGE-SCALE ANALYSIS ON METHODOLOGICAL CHOICES IN DEEP REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

1 SAMPLE COMPLEXITY WITH UNKNOWN SUPPORT

Proposition 1.1 (*Sample Complexity with Unknown Support*). *Let $\mathcal{N} > \mathcal{M} \geq 2$, $\epsilon > \frac{\mathcal{M}}{4\mathcal{N}}$, and $\theta_i \in \mathbb{R}$ for $i \in [\mathcal{N}]$. The number of samples required to learn a distribution of the form $\mathcal{Z} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta_{\theta_i}$ to within total variation distance ϵ is $\Omega\left(\frac{\mathcal{M}}{\epsilon^2}\right)$.*

Proof. Let $\mathcal{M} \geq 2$ and $\mathcal{D} = \{1, 2, \dots, \mathcal{M}\} \subseteq \mathbb{R}$. First we will show that any distribution $p(z)$ supported on $z \in \mathcal{D}$ is within total-variation distance $\frac{k}{4\mathcal{N}}$ of a distribution of a random variable of the form $\mathcal{Z} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta_{\theta_i}$ for numbers $\theta_i \in \mathcal{D}$. Indeed we can construct such a distribution as follows. First let $\tilde{p}(z)$ be the rounded distribution obtained by rounding each probability $p(z)$ to the nearest integer multiple of $\frac{1}{\mathcal{N}}$. The total variation distance between $p(z)$ and $\tilde{p}(z)$ is given by

$$\frac{1}{2} \sum_{z=1}^{\mathcal{M}} |p(z) - \tilde{p}(z)| \leq \frac{1}{2} \sum_{z=1}^{\mathcal{M}} \frac{1}{2\mathcal{N}} \leq \frac{\mathcal{M}}{4\mathcal{N}}. \quad (1)$$

Next partition the set of θ_i into \mathcal{M} groups $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_{\mathcal{M}}$, where group \mathcal{G}_z has size $\mathcal{N}\tilde{p}(z)$ (this size is an integer by construction of \tilde{p}). Finally, for each $\theta_i \in \mathcal{G}_z$ assign $\theta_i = z$. Thus for $\mathcal{Z} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta_{\theta_i}$ we have for each $z \in \mathcal{D}$

$$\mathbb{P}[\mathcal{Z} = z] = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbb{1}[\theta_i = z] = \frac{1}{\mathcal{N}} |\mathcal{G}_z| = \tilde{p}(z). \quad (2)$$

Therefore, any distribution $p(z)$ can be approximated to within total variation distance $\frac{\mathcal{M}}{4\mathcal{N}}$ by a distribution \mathcal{Z} of the prescribed form. Thus, by the sample complexity lower bounds for learning a discrete distribution with known support, for any $\epsilon > \frac{\mathcal{M}}{4\mathcal{N}}$ at least $\frac{\mathcal{M}}{\epsilon^2}$ samples are required to learn a distribution of the form $\mathcal{Z} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta_{\theta_i}$. \square

2 MEAN ESTIMATION VERSUS LEARNING THE DISTRIBUTION

To get a fundamental understanding of the additional cost of learning the state-action value distribution, we compare the sample complexity of learning the distribution of a finitely supported random variable with that of estimating the mean.

Proposition 2.1 (Canonne (2020)). *Let \mathcal{X} be a real-valued random variable with support on exactly k known values. Further, assume $|\mathcal{X}| < 1$ and let $\epsilon > 0$. Any algorithm that learns the distribution $\mathbb{P}(\mathcal{X})$ within total variation distance ϵ requires $\Omega(k/\epsilon^2)$ samples, while there exists an algorithm to estimate $\mathbb{E}[\mathcal{X}]$ to within error ϵ using only $O(1/\epsilon^2)$ samples.*

Proof. Learning a distribution with known discrete support of size k requires $\Omega(k/\epsilon^2)$ samples to achieve total variation distance at most ϵ with constant probability (Canonne, 2020). On the other hand, let $\mathcal{X}_1, \dots, \mathcal{X}_n$ be independent samples of the random variable \mathcal{X} and consider the sample mean $\bar{\mathcal{X}} = \frac{1}{n} \sum_{i=1}^n \mathcal{X}_i$. The expectation is given by $\mathbb{E}[\bar{\mathcal{X}}] = \mathbb{E}[\mathcal{X}]$ and the variance is $\sigma^2(\bar{\mathcal{X}}) = \frac{1}{n} \sigma^2(\mathcal{X})$. Further, since $|\mathcal{X}| < 1$ we have that $\sigma^2(\mathcal{X}) < 1$ and so $\sigma^2(\bar{\mathcal{X}}) \leq \frac{1}{n}$. Hence, by Chebyshev's inequality

$$\mathbb{P}[|\bar{\mathcal{X}} - \mathbb{E}[\mathcal{X}]| > \epsilon] \leq \frac{1}{\epsilon^2 n}. \quad (3)$$

Thus with $n = O(\frac{1}{\epsilon^2})$ samples, $\bar{\mathcal{X}}$ is within ϵ of $\mathbb{E}[\mathcal{X}]$ with constant probability. \square

3 REPRODUCIBILITY AND CONFIGURATION DETAILS

The hyperparameter settings of all of the algorithms in our paper, double- Q , dueling, $QRDQN$, and IQN for the high-data regime are exactly the same with the original papers that proposed these algorithms in the high-data regime. See the hyperparameter settings in Hasselt et al. (2016) for double- Q , Wang et al. (2016) for dueling architecture, Bellemare et al. (2017) for C51, Dabney et al. (2018a) for $QRDQN$, and Dabney et al. (2018b) for IQN .

Table 1: Hyperparameter settings and architectural details for the dueling algorithm, double- Q learning, C51, $QRDQN$, and IQN in the low-data regime of the Arcade Learning Environment.

| Hyperparameters | Settings |
|--|--------------------------------------|
| Grey-scaling | True |
| Observation down-sampling | (84, 84) |
| Action repetitions | 4 |
| Frames stacked | 4 |
| Batch Size | 32 |
| Update | Double-Q |
| Max Frames per episode | 108000 |
| Evaluation exploration epsilon | 0.01 |
| Min replay size for sampling | 1600 |
| Max gradient norm | 10 |
| Discount factor | 0.99 |
| Maximum absolute rewards | 1 |
| Training steps | 100000 |
| Evaluation steps | 125000 |
| Exploration epsilon decay frame fraction | 0.0125 |
| Gradient error bound | 0.03125 |
| Optimizer | Adam |
| Replay period every | 1 |
| n-step length | 10 |
| Exploration | ϵ -greedy |
| ϵ -decay | 5000 |
| Number of atoms | 51 |
| Number of quantiles | 201 |
| v_{\max} | 10 |
| Q -Network channels | 32,64,64 |
| Q -Network filter size | $8 \times 8, 4 \times 4, 3 \times 3$ |
| Q -Network stride | (4, 4), (2, 2), (1, 1) |
| Q -Network hidden units | 512 |

For a fair and transparent comparison, we kept the hyperparameters exactly the same with the DRQ^{ICLR} paper for all of the baseline Q algorithms in the low-data region. Note that DRQ is an observation regularization study; hence the hyperparameters in the DRQ paper are specifically tuned for the purpose of the paper besides tuning for the Arcade Learning Environment 100K low-data regime. We did not tune any of the hyperparameters for the baseline algorithms (i.e. dueling architecture). Hence, it is even further possible to conduct hyperparameter tuning and get better performance profile results with the simple baseline dueling architecture. For the purpose of this paper we kept the hyperparameters exactly the same with the DRQ^{ICLR} paper. However, we would strongly encourage further research to conduct hyperparameter optimization to obtain better results from the baseline dueling architecture in the low-data regime.

We have also tried the hyperparameter settings reported in the data efficient Rainbow (DER) paper for C51, IQN and $QRDQN$ in the low-data regime. The performance results are provided in Table 2 for the hyperparameter settings of DER. As can be seen, the hyperparameter settings of DRQ^{ICLR} gave better performance results also for C51, IQN and $QRDQN$ in the low-data region. The results in Table 2 also align with the claims of the DER paper in which there has not been extensive hyperparameter tuning conducted to achieve the results provided, and it is possible to obtain better results by further hyperparameter tuning.

Table 2: Human normalized mean, human normalized median, and human normalized 20th percentile results for the C51 algorithm, *QRDQN*, and *IQN* in the low-data regime of the Arcade Learning Environment with the hyperparameter settings reported in the DER paper.

| Algorithms | Human Normalized Median | Human Normalized Mean | Human Normalized 20 th Percentile |
|--------------|-------------------------|-----------------------|--|
| C51 | 0.0490±0.0038 | 0.1352±0.0057 | 0.0163±0.0029 |
| <i>QRDQN</i> | 0.0203±0.0033 | 0.0778±0.0101 | -0.0012±0.0053 |
| <i>IQN</i> | 0.0202±0.0020 | 0.0590±0.0139 | -0.0035±0.0031 |

4 RESULTS ON THE COMPLETE LIST OF GAMES FROM THE ARCADE LEARNING ENVIRONMENT 100K BASELINE

Table 3: Average returns for human, random, dueling Wang et al. (2016), C51, *QRDQN* and *IQN* across all the games in the Arcade Learning Environment 100K benchmark.

| Games | Human | Random | C51 | <i>QRDQN</i> | <i>IQN</i> | Dueling |
|----------------|---------|---------|-----------------|-----------------|----------------|----------------|
| Alien | 7127.7 | 227.8 | 547.16 | 509.57 | 330.81 | 705.58 |
| Amidar | 1719.5 | 5.8 | 78.41 | 55.70 | 74.98 | 199.31 |
| Assault | 742.0 | 222.4 | 465.30 | 314.58 | 488.55 | 503.82 |
| Asterix | 8503.3 | 210.0 | 475.90 | 367.32 | 286.26 | 705.16 |
| BankHeist | 753.1 | 14.2 | 22.81 | 21.53 | 18.17 | 243.19 |
| BattleZone | 37187.5 | 2360.0 | 2728.52 | 6238.27 | 3105.70 | 6880.37 |
| Boxing | 12.1 | 0.1 | 9.60 | 2.03 | 12.41 | 1.68 |
| Breakout | 30.5 | 1.7 | 11.35 | 16.50 | 15.09 | 8.28 |
| ChopperCommand | 7387.8 | 811.0 | 831.83 | 752.51 | 629.04 | 1313.90 |
| CrazyClimber | 35829.4 | 10780.5 | 71776.14 | 21366.42 | 22649.44 | 17039.44 |
| DemonAttack | 1971.0 | 152.1 | 789.09 | 198.01 | 1035.17 | 694.42 |
| Freeway | 29.6 | 0.0 | 20.42 | 5.98 | 19.37 | 5.93 |
| FrostBite | 4334.7 | 65.2 | 215.25 | 218.11 | 192.33 | 259.18 |
| Gopher | 2412.5 | 257.6 | 791.83 | 576.19 | 466.81 | 429.85 |
| Hero | 30826.4 | 1027.0 | 7097.42 | 1108.44 | 1322.63 | 8210.53 |
| Jamesbond | 302.8 | 29.0 | 43.85 | 108.71 | 26.23 | 296.46 |
| Kangaroo | 3035.0 | 52.0 | 301.01 | 120.60 | 294.46 | 1914.86 |
| Krull | 2665.5 | 1598.0 | 3744.04 | 2040.50 | 2319.74 | 2867.78 |
| KungFuMaster | 22736.3 | 258.5 | 6877.62 | 11574.02 | 1526.76 | 5367.90 |
| Mspacman | 6951.6 | 307.3 | 917.78 | 749.29 | 533.98 | 1355.21 |
| Pong | 14.6 | -20.7 | 11.17 | -7.49 | -10.86 | -4.20 |
| PrivateEye | 69571.3 | 24.9 | -103.30 | -6.32 | 33.83 | 100.00 |
| Qbert | 13455.0 | 163.9 | 528.30 | 590.05 | 582.72 | 1710.23 |
| RoadRunner | 7845.0 | 11.5 | 3993.34 | 400.59 | 1202.20 | 6031.80 |
| Seaquest | 42054.7 | 68.4 | 163.69 | 183.25 | 213.87 | 351.10 |
| UpNdown | 11693.2 | 533.4 | 1970.28 | 1622.67 | 1552.27 | 3553.12 |

Table 3 reports the average scores obtained by the human player, random player, baseline *Q*-based algorithm dueling architecture, baseline algorithm C51 that focuses on learning the distribution, *QRDQN* and *IQN* across all the games in the Arcade Learning Environment 100K baseline. These results once more demonstrate that the baseline *Q*-based algorithm performs significantly better than any algorithm that aims to learn the distribution as has also been explained in detail in Section 5 in the main body of the paper.

Figure 1 reports the learning curves of the complete list of the games in the Arcade Learning Environment 100K benchmark; in particular, for Alien, Amidar, Asterix, BankHeist, BattleZone, Boxing, Breakout, ChopperCommand, Hero, CrazyClimber, JamesBond, Kangaroo, PrivateEye, MsPacman, FrostBite, Qbert, RoadRunner, Seaquest, Pong, Gopher, DemonAttack, Krull, and UpNdown with dueling architecture Wang et al. (2016), C51 (Bellemare et al., 2017), *IQN* (Dabney et al., 2018a) and *QRDQN* (Dabney et al., 2018b; Bellemare et al., 2023) algorithms with 100K environment interaction training. Note that the results for deep double-*Q* learning (Hasselt et al.,

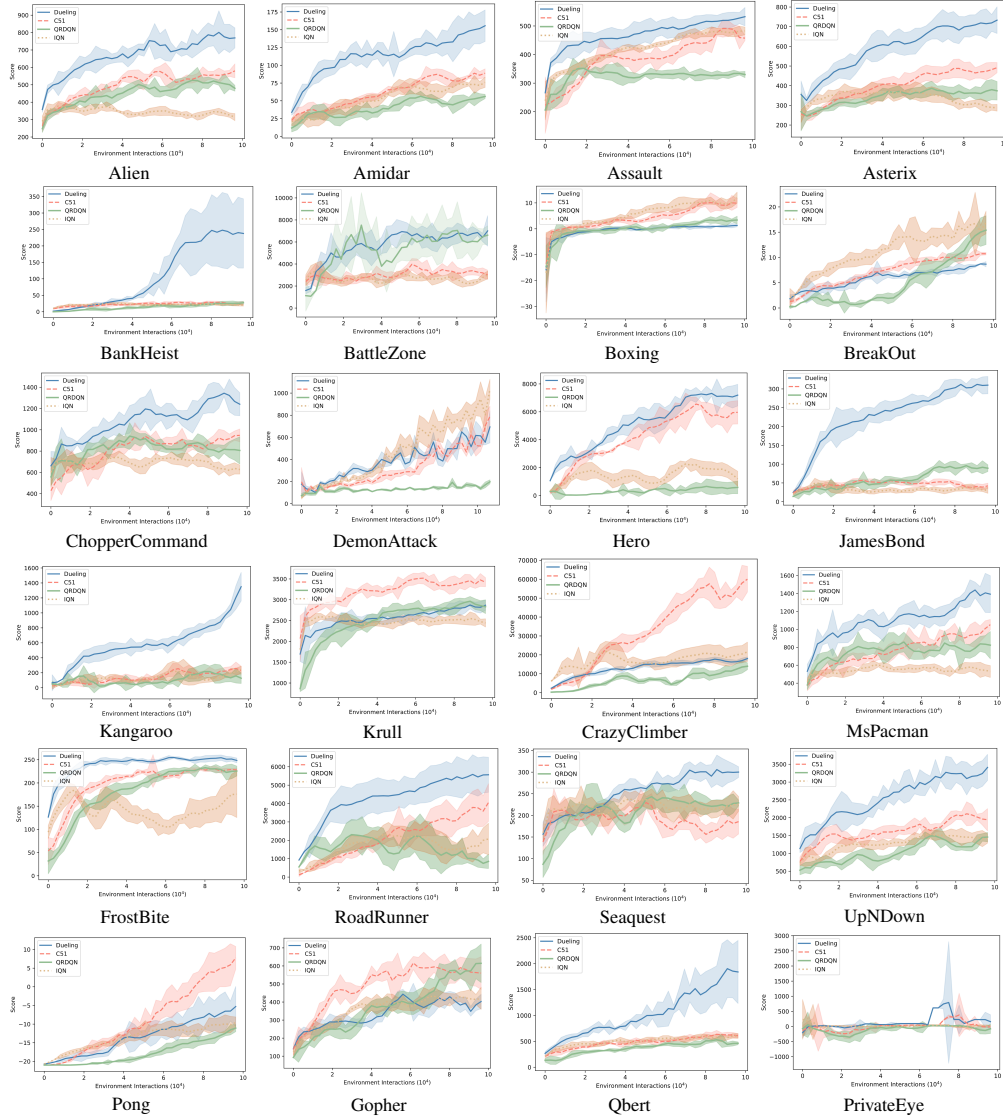


Figure 1: The learning curves of Alien, Amidar, Asterix, BankHeist, BattleZone, Boxing, Breakout, ChopperCommand, Hero, CrazyClimber, JamesBond, Kangaroo, PrivateEye, MsPacman, FrostBite, Qbert, RoadRunner, Seaquest, Pong, Gopher, DemonAttack, Krull, and UpNDown with dueling architecture Wang et al. (2016), C51, IQN and QRDQN algorithms in the Arcade Learning Environment with 100K environment interaction training.

2016), prior (Schaul et al., 2016) and DQN (Mnih et al., 2015) are reported in the main body of the paper.

The learning curves reported in Figure 1 demonstrate that the number of samples required to obtain the performance level achieved via the simple base dueling architecture is significantly higher for any reinforcement learning algorithm that learns the distribution. Note that the baseline reinforcement learning algorithm C51 focusing on learning the distribution represents the state-action value distribution as a discrete probability distribution supported on 51 fixed atoms evenly spaced between a pre-specified minimum and maximum value. In contrast, QR-DQN represents the value distribution as the uniform distribution over a larger number of atoms with variable positions on the real line. Thus, QR-DQN is able to more accurately approximate a broader class of state-action value distributions. Finally, IQN parameterizes the quantile function of the state-action value distribution via a deep neural network, leading to a yet more flexible representation of the state-action value distribution. As

discussed in Section 4, more complex representations for broader classes of distributions come at the cost of a higher sample complexity required for learning.

REFERENCES

- Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017.
- Bellemare, M. G., Dabney, W., and Rowland, M. Distributional reinforcement learning. *MIT Press*, 2023.
- Canonne, C. L. A short note on learning discrete distributions. 2020. URL <http://arxiv.org/abs/2002.11457>. cite arxiv:2002.11457Comment: This is a review article; its intent is not to provide new results, but instead to gather known (and useful) ones, along with their proofs, in a single convenient location.
- Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1104–1113. PMLR, 2018a.
- Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. Distributional reinforcement learning with quantile regression. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2892–2901. AAAI Press, 2018b.
- Hasselt, H. v., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, a. G., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.
- Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*, 2016.
- Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., and De Freitas, N. Dueling network architectures for deep reinforcement learning. *International Conference on Machine Learning ICML*, pp. 1995–2003, 2016.