Appendix

Architecture and Hyperparameters

We considered multiple architectural variants for parameterizing an IQN. All of these build on the Q-network of a regular DQN (Mnih et al., 2015), which can be seen as the composition of a convolutional stack $\psi\colon \mathcal{X} \to \mathbb{R}^d$ and an MLP $f\colon \mathbb{R}^d \to \mathbb{R}^{|\mathcal{A}|}$, and extend it by an embedding of the sample point, $\phi\colon [0,1] \to \mathbb{R}^d$, and a merging function $m\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, resulting in the function

$$\mathrm{IQN}(x, \tau) = f(m(\psi(x), \phi(\tau))).$$

For the embedding ϕ, we considered a number of variants: a learned linear embedding, a learned MLP embedding with a single hidden layer of size n, and a learned linear function of n cosine basis functions of the form $\cos(\pi i \tau), i = 1, \ldots, n$. Each of those was followed by either a ReLU or sigmoid nonlinearity.

For the merging function m, the simplest choice would be a simple vector concatenation of $\psi(x)$ and $\phi(\tau)$. Note however, that the MLP f which takes in the output of m and outputs the action-value quantiles, only has a single hidden layer in the DQN network. Therefore, to force a sufficiently early interaction between the two representations, we also considered a multiplicative function $m(\psi, \phi) = \psi \odot \phi$, where $\odot$ denotes the element-wise (Hadamard) product of two vectors, as well as a 'residual' function $m(\psi, \phi) = \psi \odot (1 + \phi)$.

Early experiments showed that a simple linear embedding of τ was insufficient to achieve good performance, and the residual version of m didn't show any marked difference to the multiplicative variant, so we do not include results for these here. For the other configurations, Figure 5 shows pairwise comparisons between 1) a cosine basis function embedding and a completely learned MLP embedding, 2) an embedding size (hidden layer size or number of cosine basis elements) 32 and 64, 3) ReLU and sigmoid nonlinearity following the embedding, and 4) concatenation and a multiplicative interaction between $\psi(x)$ and $\phi(\tau)$.

Each comparison 'violin plot' can be understood as a marginalization over the other variants of the architecture, with the human-normalized performance at the end of training, averaged across six Atari 2600 games, on the y-axis. Each white dot corresponds to a configuration (each represented by two seeds), the black dots show the position of our preferred configuration. The width of the colored regions corresponds to a kernel density estimate of the number of configurations at each performance level.

Our final choice is a multiplicative interaction with a linear function of a cosine embedding, with $n = 64$ and a ReLU nonlinearity (see Equation 4), as this configuration yielded the highest performance consistently over multiple seeds. Also noteworthy is the overall robustness of the approach to these variations: most of the configurations consistently outperform the QR-DQN baseline shown as a grey horizontal line for comparison.

We give pseudo-code for the IQN loss in Algorithm 1. All other hyperparameters for this agent correspond to the ones used by Dabney et al. (2018). In particular, the Bellman target is computed using a target network. Notice that IQN will generally be more computationally expensive per-sample than QR-DQN. However, in practice IQN requires many fewer samples per update than QR-DQN so that the actual running times are comparable.

Algorithm 1 Implicit Quantile Network Loss

Require: N, N', K, κ and functions β, Z
input $x, a, r, x', \gamma \in [0, 1)$
  # Compute greedy next action
  $a^* \leftarrow \arg\max_{a'} \frac{1}{K} \sum_k^K Z_{\tilde{\tau}_k}(x', a'), \quad \tilde{\tau}_k \sim \beta(\cdot)$
  # Sample quantile thresholds
  $\tau_i, \tau'_j \sim U([0, 1]), \quad 1 \le i \le N, 1 \le j \le N'$
  # Compute distributional temporal differences
  $\delta_{ij} \leftarrow r + \gamma Z_{\tau'_j}(x', a^*) - Z_{\tau_i}(x, a), \quad \forall i, j$
  # Compute Huber quantile loss
output $\sum_{i=1}^N \mathbb{E}_{\tau'} \left[\rho^\kappa_{\tau_i}(\delta_{ij})\right]$

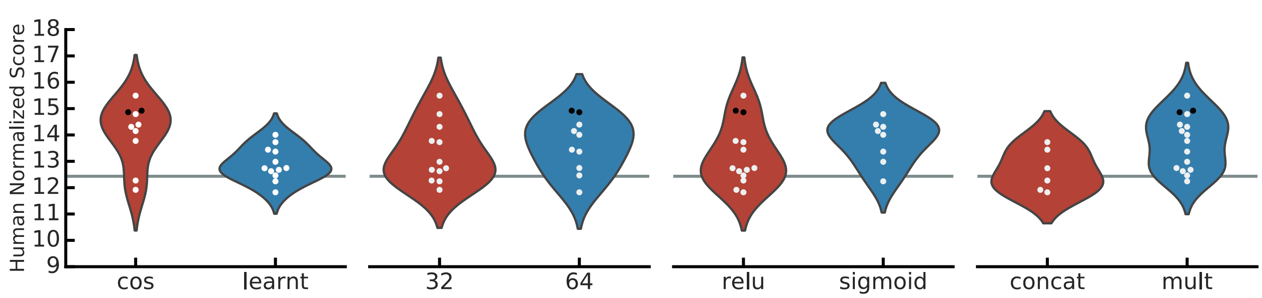

Figure 5. Comparison of architectural variants.

Evaluation

The human-normalized scores reported in this paper are given by the formula (van Hasselt et al., 2016; Dabney et al., 2018)

$$score = \frac{agent - random}{human - random},$$

where $agent$, $human$ and $random$ are the per-game raw scores (undiscounted returns) for the given agent, a reference human player, and random agent baseline (Mnih et al., 2015).

The 'human-gap' metric referred to at the end of Section 5 builds on the human-normalized score, but emphasizes the remaining improvement for the agent to reach super-human performance. It is given by $gap = \max(1 - score, 0)$, with a value of 1 corresponding to random play, and a value of 0 corresponding to super-human level of performance. To avoid degeneracies in the case of $human < random$, the quantity is being clipped above at 1.

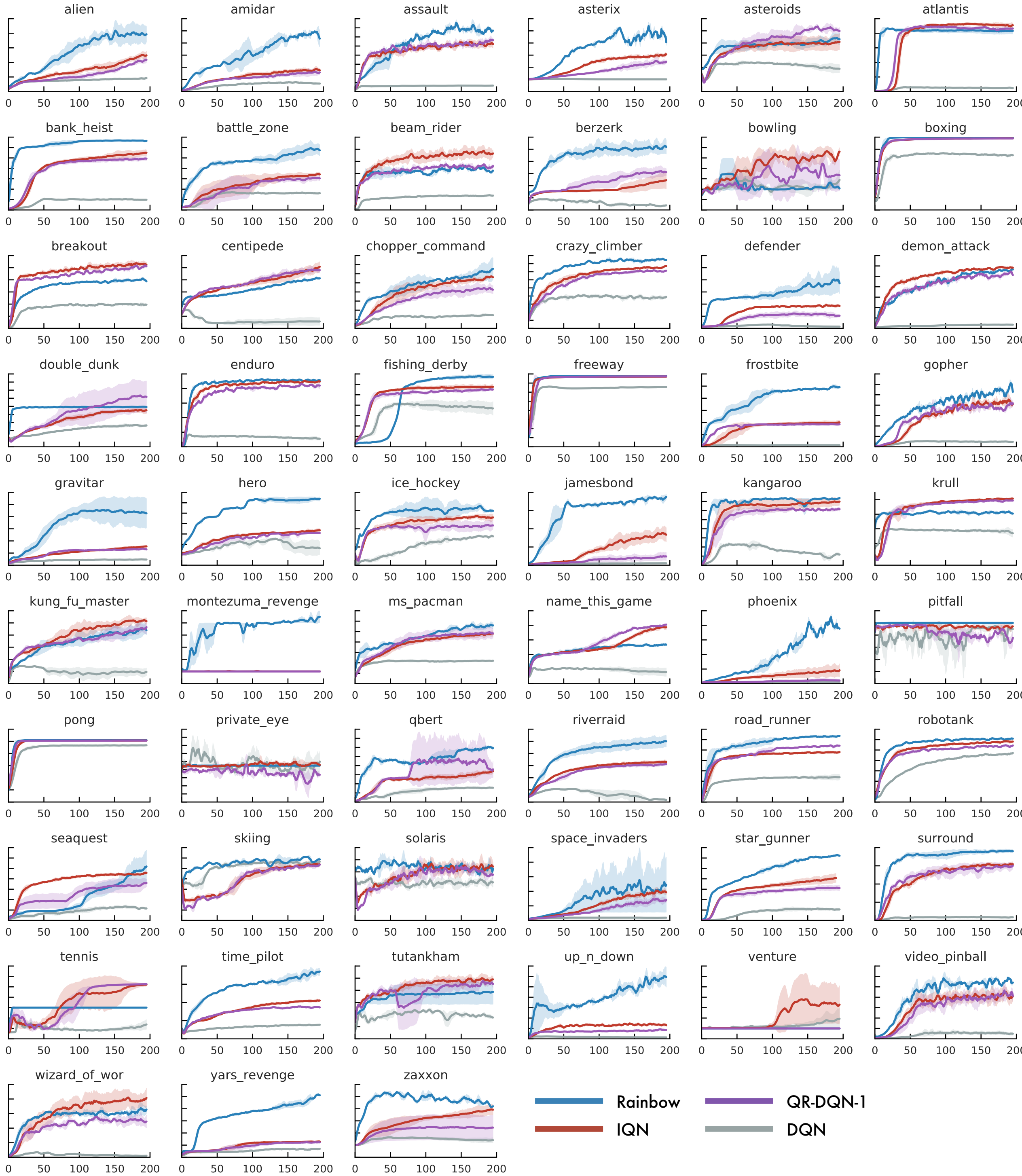

Figure 6. Complete Atari-57 training curves.

GAMES	RANDOM	HUMAN	DQN	PRIOR. DUEL.	QR-DQN	IQN
Alien	227.8	7,127.7	1,620.0	3,941.0	4,871	**7,022**
Amidar	5.8	1,719.5	978.0	2,296.8	1,641	**2,946**
Assault	222.4	742.0	4,280.4	11,477.0	22,012	**29,091**
Asterix	210.0	8,503.3	4,359.0	**375,080.0**	261,025	342,016
Asteroids	719.1	47,388.7	1,364.5	1,192.7	**4,226**	2,898
Atlantis	12,850.0	29,028.1	279,987.0	395,762.0	971,850	**978,200**
Bank Heist	14.2	753.1	455.0	**1,503.1**	1,249	1,416
Battle Zone	2,360.0	37,187.5	29,900.0	35,520.0	39,268	**42,244**
Beam Rider	363.9	16,926.5	8,627.5	30,276.5	34,821	**42,776**
Berzerk	123.7	2,630.4	585.6	**3,409.0**	3,117	1,053
Bowling	23.1	160.7	50.4	46.7	77.2	**86.5**
Boxing	0.1	12.1	88.0	98.9	**99.9**	99.8
Breakout	1.7	30.5	385.5	366.0	**742**	734
Centipede	2,090.9	12,017.0	4,657.7	7,687.5	**12,447**	11,561
Chopper Command	811.0	7,387.8	6,126.0	13,185.0	14,667	**16,836**
Crazy Climber	10,780.5	35,829.4	110,763.0	162,224.0	161,196	**179,082**
Defender	2,874.5	18,688.9	23,633.0	41,324.5	47,887	**53,537**
Demon Attack	152.1	1,971.0	12,149.4	72,878.6	121,551	**128,580**
Double Dunk	-18.6	-16.4	-6.6	-12.5	**21.9**	5.6
Enduro	0.0	860.5	729.0	2,306.4	2,355	**2,359**
Fishing Derby	-91.7	-38.7	-4.9	**41.3**	39.0	33.8
Freeway	0.0	29.6	30.8	33.0	**34.0**	**34.0**
Frostbite	65.2	4,334.7	797.4	**7,413.0**	4,384	4,324
Gopher	257.6	2,412.5	8,777.4	104,368.2	113,585	**118,365**
Gravitar	173.0	3,351.4	473.0	238.0	**995**	911
H.E.R.O.	1,027.0	30,826.4	20,437.8	21,036.5	21,395	**28,386**
Ice Hockey	-11.2	0.9	-1.9	-0.4	-1.7	**0.2**
James Bond	29.0	302.8	768.5	812.0	4,703	**35,108**
Kangaroo	52.0	3,035.0	7,259.0	1,792.0	15,356	**15,487**
Krull	1,598.0	2,665.5	8,422.3	10,374.4	**11,447**	10,707
Kung-Fu Master	258.5	22,736.3	26,059.0	48,375.0	**76,642**	73,512
Montezumas Revenge	0.0	4,753.3	**0.0**	**0.0**	**0.0**	**0.0**
Ms. Pac-Man	307.3	6,951.6	3,085.6	3,327.3	5,821	**6,349**
Name This Game	2,292.3	8,049.0	8,207.8	15,572.5	21,890	**22,682**
Phoenix	761.4	7,242.6	8,485.2	**70,324.3**	16,585	56,599
Pitfall!	-229.4	6,463.7	-286.1	**0.0**	**0.0**	**0.0**
Pong	-20.7	14.6	19.5	20.9	**21.0**	**21.0**
Private Eye	24.9	69,571.3	146.7	206.0	**350**	200
Q*Bert	163.9	13,455.0	13,117.3	18,760.3	**572,510**	25,750
River Raid	1,338.5	17,118.0	7,377.6	**20,607.6**	17,571	17,765
Road Runner	11.5	7,845.0	39,544.0	62,151.0	**64,262**	57,900
Robotank	2.2	11.9	**63.9**	27.5	59.4	62.5
Seaquest	68.4	42,054.7	5,860.6	931.6	8,268	**30,140**
Skiing	-17,098.1	-4,336.9	-13,062.3	-19,949.9	-9,324	**-9,289**
Solaris	1,236.3	12,326.7	3,482.8	133.4	6,740	**8,007**
Space Invaders	148.0	1,668.7	1,692.3	15,311.5	20,972	**28,888**
Star Gunner	664.0	10,250.0	54,282.0	**125,117.0**	77,495	74,677
Surround	-10.0	6.5	-5.6	1.2	8.2	**9.4**
Tennis	-23.8	-8.3	12.2	0.0	**23.6**	**23.6**
Time Pilot	3,568.0	5,229.2	4,870.0	7,553.0	10,345	**12,236**
Tutankham	11.4	167.6	68.1	245.9	**297**	293
Up and Down	533.4	11,693.2	9,989.9	33,879.1	71,260	**88,148**
Venture	0.0	1,187.5	163.0	48.0	43.9	**1,318**
Video Pinball	16,256.9	17,667.9	196,760.4	479,197.0	**705,662**	698,045
Wizard Of Wor	563.5	4,756.5	2,704.0	12,352.0	25,061	**31,190**
Yars Revenge	3,092.9	54,576.9	18,098.9	**69,618.1**	26,447	28,379
Zaxxon	32.5	9,173.3	5,363.0	13,886.0	13,112	**21,772**

Figure 7. Raw scores for a single seed across all games, starting with 30 no-op actions. Reference values from (Wang et al., 2016).