SUPPLEMENTARY MATERIAL OF "NON-DECREASING QUANTILE FUNCTION WITH EFFICIENT EXPLORATION FOR DISTRIBUTIONAL REINFORCEMENT LEARNING"

Anonymous authors
Paper under double-blind review

A HYPER-PARAMETERS SETTINGS FOR NDQFN WITH DPE

We consider multiple set of hyper-parameters, and the following one has the best performance in practice.

Parameter	Value
Min history to start learning	200K frames
Online network optimizer	Adam
Online Learning rate	0.0000625
Batch size	32
Number of $\boldsymbol{p}^*$	32
Number of sampled $\boldsymbol{\tau}$ and $\boldsymbol{\tau}'$	32
Discount factor	0.99
Step for n-step updates	3
Target network update frequency	40K frames
Predictor network learning rate	0.0000625
Predictor network optimizer	Adam

Table 1: Hyper-parameter Settings for NDQFN with DPE

B MAIN ALGORITHM OF NDQFN+DPE

Algorithm 1 NDQFN with DPE

Initialize replay memory D
Initialize online network Q_ω, target network $Q_{\omega'}$ and predictor network Q_{ω^*} randomly
For $episode = 1, M$ **do**
Initialize state x_1
  For $t = 1, T$ **do**
    Compute non-decreasing continuous quantile function by NDQFN
    Compute distributional prediction error $i(x_t, a_t)$
    Select $a_t = \arg\max_a [Q_{\omega'}(x_t, a) + c_t i(x_t, a)]$
    Execute action a_t in emulator and observe reward r_t and state x_{t+1}
    if $t > n$, **then**
      store transition $(x_{t-n}, a_{t-n}, r_{t-n+1}, \cdots, r_t, x_t)$ in D
    else
      store transition $(x_1, a_1, r_1, \cdots, r_t, x_t)$ in D
    end if
    Sample random minibatch of transitions $(x_j, a_j, r_j, \cdots, r_{j+n-1}, x_{j+n})$ from D
    Perform a gradient descent step on $L(x_j, a_j, r_j, \cdots, r_{j+n-1}, x_{j+n})$ with respect to the network parameters ω and ω^*
    Every C steps reset $\omega' = \omega$
  end for
end for

C COMPARISON OF DIFFERENT g INPUTS IN NDQFN

To investigate the empirical performance of the two different g function inputs mentioned in Section , we do some experiments on three Atari 2600 games (Breakout, Venture, Montezuma Revenge) for 100 million frames. The results are summarized in Figure 1.

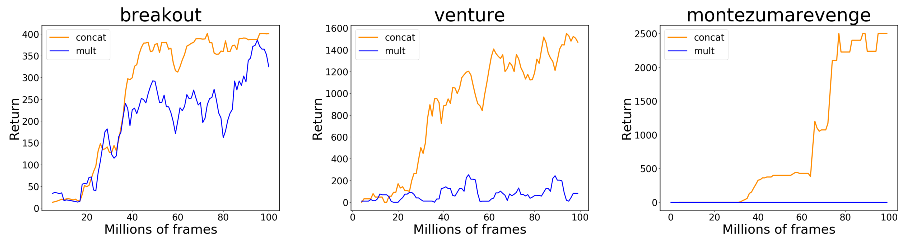

Figure 1: Performance comparison with different forms to combine $\psi(x)$, $\phi(p_i)$ and $\phi(p_{i-1})$, where "concat" denotes g takes $(\psi(x) \odot \phi(p_i)$ and $\phi(p_i) - \phi(p_{i-1})$ as input and "mult" denotes g takes $\psi(x) \odot \phi(p_{i-1}) \odot \phi(p_i)$ as input.

D DISTRIBUTION-BASED VS VALUE-BASED

To compare the empirical performance of distribution-based and value-based prediction error, three hard explored games are evaluated for 100 million frames. The exploration bonus of value-based prediction error is defined as follow:

$$i'(x_t, a_t) = |Q_{\omega'}(x_t, a_t) - Q_{\omega^*}(x_t, a_t)|,$$

where $Q_{\omega'}(\cdot)$ and $Q_{\omega^*}(\cdot)$ denote the state-action value function of the target and predictor network, respectively. Figure 2 illustrates that distribution-based prediction error outperforms value-based prediction error significantly on three hard-explored games under 100M training.

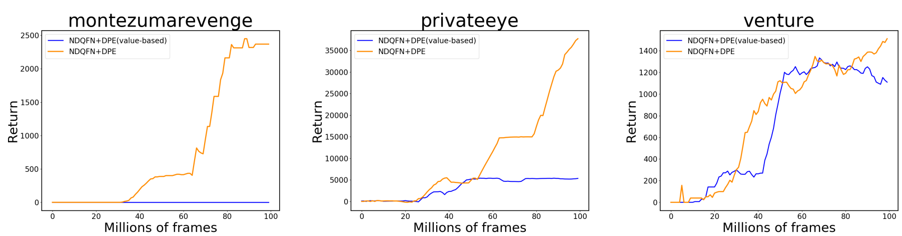

Figure 2: Comparison between distribution-based and value-based prediction error.

E Additional Results about NDQFN vs IQN

As a supplement of Section 6.1, Figure 3 visualizes the training curves of other Atari 2600 games for 100 million frames. The setting of experiments is aligned with Section 6.1. NDQFN significantly outperforms IQN with n-steps updates in most cases, especially in the early training stage with limited training samples.

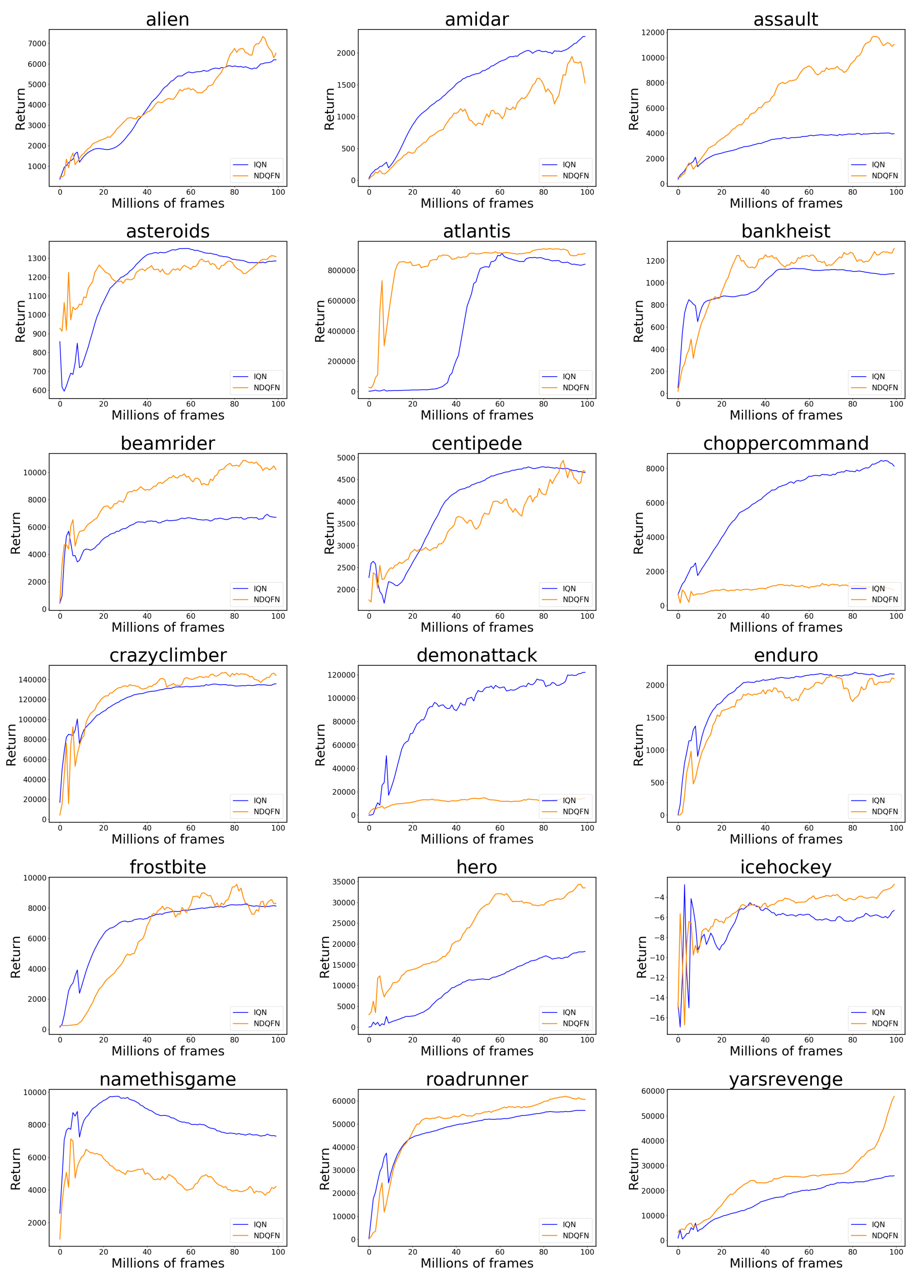

Figure 3: Training curve on Atari games for NDQFN and IQN.

F TRAINING CURVE

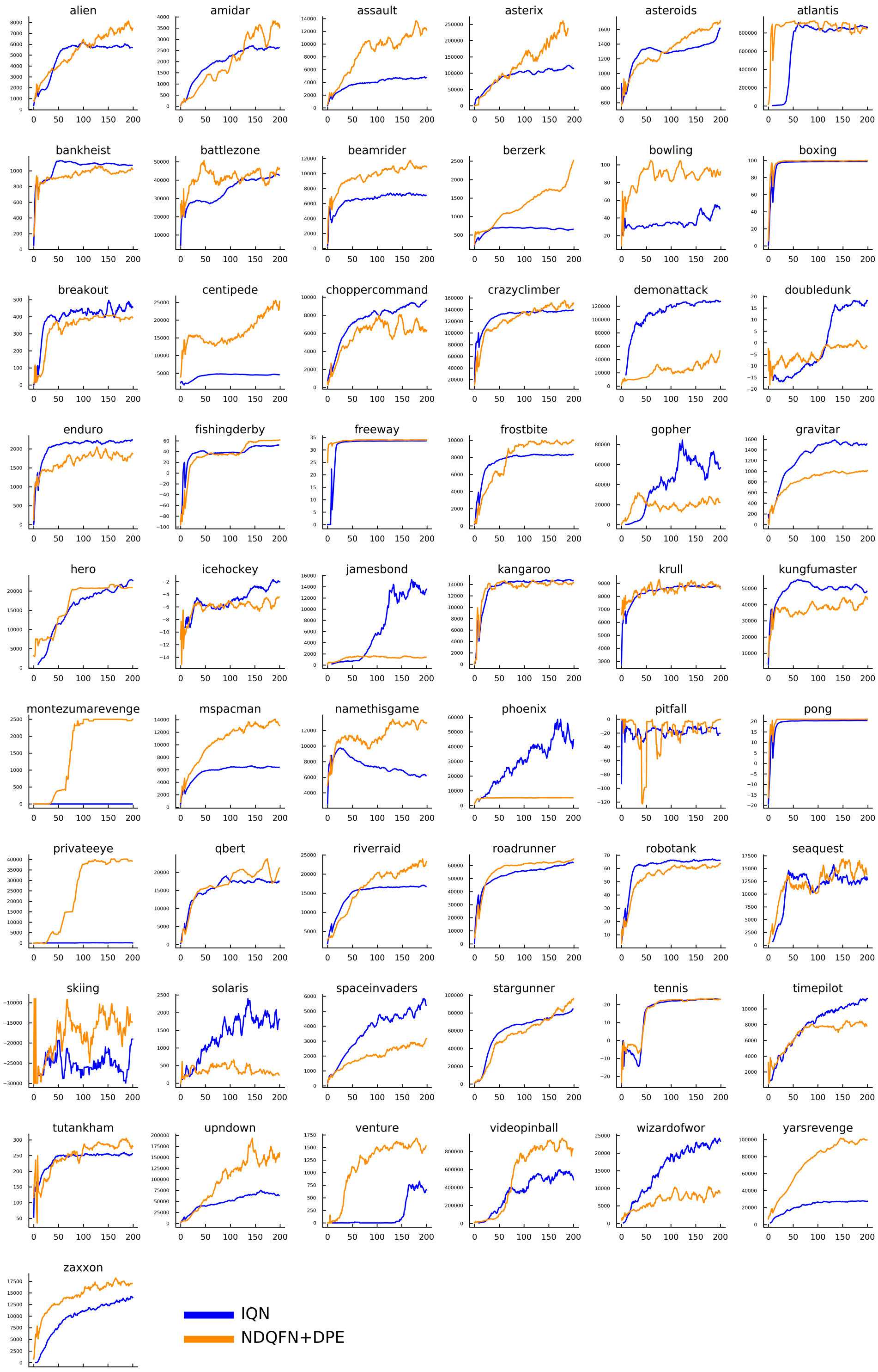

Figure 4: Complete Atari-55 training curves.

G RAW SCORE

GAMES	RANDOM	HUMAN	DQN	PRIOR.DUEL.	QRDQN	IQN	FQF	NDQFN+DPE
Alien	227.8	7127.7	1620.0	3941.0	4871.0	7022.0	**16754.6**	9398.9
Amidar	5.8	1719.5	978.0	2296.8	1641.0	2946.0	3165.3	**4503.0**
Assault	222.4	742.0	4280.4	11477.0	22012.0	**29091.0**	23020.1	14314.2
Asterix	210.0	8503.3	4359.0	375080.0	261025.0	342016.0	**578388.5**	308462.2
Asteroids	719.1	47388.7	1364.5	1192.7	4226.0	2898.0	**4553.0**	1848.7
Atlantis	12850.0	29028.1	279987.0	395762.0	971850.0	978200.0	957920.0	**1022680.0**
BankHeist	14.2	753.1	455.0	1503.1	1249.0	**1416.0**	1259.1	1203.7
BattleZone	2360.0	37187.5	29900.0	35520.0	39268.0	42244.0	**87928.6**	74326.1
BeamRider	363.9	16926.5	8627.5	30276.5	34821.0	**42776.0**	37106.6	11297.2
Berzerk	123.7	2630.4	585.6	3409.0	3117.0	1053.0	**12422.2**	2758.2
Bowling	23.1	160.7	50.4	46.7	77.2	86.5	102.3	**126.6**
Boxing	0.1	12.1	88.0	98.9	99.9	99.8	98.0	**100.0**
Breakout	1.7	30.5	385.5	366.0	742.0	734.0	**854.2**	416.1
Centipede	2090.9	12017.0	4657.7	7687.5	12447.0	11561.0	11526.0	**43555.0**
ChopperCommand	811.0	7387.8	6126.0	13185.0	14667.0	16836.0	**876460.0**	12066.7
CrazyClimber	10780.5	35829.4	110763.0	162224.0	161196.0	179082.0	**223470.6**	196887.5
DemonAttack	152.1	1971.0	12149.4	72878.6	121551.0	128580.0	**131697.0**	72048.3
DoubleDunk	-18.6	-16.4	-6.6	-12.5	21.9	5.6	**22.9**	0.8
Enduro	0.0	860.5	729.0	2306.4	2355.0	2359.0	2370.8	**2374.1**
FishingDerby	-91.7	-38.7	-4.9	41.3	39.0	33.8	52.7	**63.0**
Freeway	0.0	29.6	30.8	33.0	**34.0**	**34.0**	33.7	**34.0**
Frostbite	65.2	4334.7	797.4	7413.0	4384.0	4324.0	**16472.9**	10684.4
Gopher	257.6	2412.5	8777.4	104368.2	113585.0	118365.0	**121144.0**	51865.5
Gravitar	173.0	3351.4	473.0	238.0	995.0	911.0	**1406.0**	1098.1
Hero	1027.0	30826.4	20437.8	21036.5	21395.0	28386.0	**30926.2**	26074.3
IceHockey	-11.2	0.9	-1.9	-0.4	-1.7	0.2	**17.3**	-2.8
Jamesbond	29.0	302.8	768.5	812.0	4703.0	35108.0	87291.7	**2524**
Kangaroo	52.0	3035.0	7259.0	1792.0	15356.0	15487.0	15400.0	**15836.4**
Krull	1598.0	2665.5	8422.3	10374.0	11447.0	10707.0	10706.8	**10894.3**
KungFuMaster	258.5	22736.3	26059.0	48375.0	76642.0	73512.0	111138.5	53702.9
MontezumaRevenge	0.0	4753.3	0.0	0.0	0.0	0.0	0.0	**2500.0**
MsPacman	307.3	6951.6	3085.6	3327.3	5821.0	6349.0	7631.9	**16027.5**
NameThisGame	2292.3	8049.0	8207.8	15572.5	21890.0	**22682.0**	16989.4	15927.4
Phoenix	761.4	7242.6	8485.2	70324.3	16585.0	56599.0	**174077.5**	5500.0
Pitfall	-229.4	6463.7	-286.1	0.0	0.0	0.0	0.0	**0.5**
Pong	-20.7	14.6	19.5	20.9	**21.0**	**21.0**	**21.0**	**21.0**
PrivateEye	24.9	69571.3	146.7	206.0	350.0	200.0	140.1	**40500.0**
Qbert	163.9	13455.0	13117.3	18760.3	**572510.0**	25750.0	27524.4	24825.1
Riverraid	1338.5	17118.0	7377.6	20607.6	17571.0	17765.0	23560.7	**26982.1**
RoadRunner	11.5	7845.0	39544.0	62151.0	64262.0	57900.0	58072.7	**66475.5**
Robotank	2.2	11.9	63.9	27.5	59.4	62.5	**75.7**	67.8
Seaquest	68.4	42054.7	5860.6	931.6	8268.0	30140.0	29383.3	**30156.1**
Skiing	-17098.1	-4336.9	-13062.3	-19949.9	-9324.0	-9289.0	-9085.3	**-6450.0**
Solaris	1236.3	12326.7	3482.8	133.4	6740.0	**8007.0**	6906.7	1843.3
SpaceInvaders	148.0	1668.7	1692.3	15311.5	20972.0	28888.0	**46498.3**	3494.0
StarGunner	664.0	10250.0	54282.0	125117.0	77495.0	74677.0	**131981.2**	85136.0
Tennis	-23.8	-9.3	12.2	0.0	**23.6**	**23.6**	22.6	**23.6**
TimePilot	3568.0	5229.2	4870.0	7553.0	10345.0	12236.0	**14995.2**	9984.1
Tutankham	11.4	167.6	68.1	245.9	297.0	293.0	309.2	**332.2**
UpNDown	533.4	11693.2	9989.9	33879.1	71260.0	88148.0	75474.4	**333790.1**
Venture	0.0	1187.5	163.0	48.0	43.9	1318.0	1112	**1827.3**
VideoPinball	16256.9	17667.9	196760.4	479197.0	705662.0	698045.0	799155.6	**941443.2**
WizardOfWor	563.5	4756.5	2704.0	12352.0	25061.0	31190.0	**44782.6**	17744.1
YarsRevenge	3092.9	54576.9	18098.9	69618.1	26447.0	28379.0	27691.2	**105450.2**
Zaxxon	32.5	9173.3	5363.0	13886.0	13113.0	**21772.0**	15179.5	21144.3

Table 2: Raw scores across all games, averaged over 0.125M frames, from the agent snapshot that obtained the highest score during training.