A PROOFS

In this section, we provide proofs of theorems stated in Section 6. Recall from Section 3 that $\iota = \mathrm{polylog}(|\mathcal{S}|, (1-\gamma)^{-1}, N)$ is some constant. Our proofs rely on the following lemma, which bounds the estimation error due to using the empirical Bellman operator:

Lemma A.1. *For all state-action* $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ *such that* $n(\mathbf{s}, \mathbf{a}) \geq 1$*, function* Q*, and* $\delta \in (0, 1)$*, we have:*

$$\mathbb{P}\left(\left|\widehat{\mathcal{B}}^* Q(\mathbf{s}, \mathbf{a}) - \mathcal{B}^* Q(\mathbf{s}, \mathbf{a})\right| \leq \sqrt{\frac{\iota \log(1/\delta)}{n(\mathbf{s}, \mathbf{a})}}\right) \geq 1 - \delta\,.$$

The above lemma is a well-known result in reinforcement learning (Rashidinejad et al., 2021), whose derivation follows from Hoeffding's inequalities.

A.1 PROOF OF THEOREM 6.1

Without loss of generality, assume that $\delta_1, \delta_2 \leq \delta$ are the solution to the outer maximization of Equation 4 at convergence. Using Lemma A.1, we have that

$$\begin{aligned}
\widehat{Q}(\mathbf{s}, \mathbf{a}, \delta) &= \widehat{\mathcal{B}}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2) - \alpha \sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s}, \mathbf{a}) \wedge 1}} \\
&\leq \mathcal{B}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2) - \alpha \sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s}, \mathbf{a}) \wedge 1}} + \sqrt{\frac{\iota \log(1/\delta_1)}{n(\mathbf{s}, \mathbf{a})}} \leq \mathcal{B}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2) \quad \forall \mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}\,,
\end{aligned}$$

holds with probability at least $1 - \delta_1$ for any $\alpha \geq \iota^{1/2}$. Using Lemma 6.1, we have

$$\begin{aligned}
\widehat{Q}(\mathbf{s}, \mathbf{a}, \delta) \leq \mathcal{B}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta) &\implies \widehat{Q} \leq (I - \gamma P^*)^{-1} R \\
&\implies \widehat{Q}(\mathbf{s}, \mathbf{a}) \leq Q^*(\mathbf{s}, \mathbf{a}) \quad \forall \mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}\,,
\end{aligned}$$

holds with probability at least $1 - \delta_1 \geq 1 - \delta$, as desired.

A.2 PROOF OF THEOREM 6.2

Recall from Equation 5 that at convergence, we have,

$$\begin{aligned}
\widehat{Q}(\mathbf{s}, \mathbf{a}, \delta) = \arg\min_{Q} \max_{\delta_1, \delta_2} \max_{\pi} \,& \alpha \sqrt{\frac{\log(1/\delta_1)}{(n(\mathbf{s}) \wedge 1)}} \left(\mathbb{E}_{\mathbf{s} \sim D, \mathbf{a} \sim \pi(\mathbf{a}|s)}\left[Q(\mathbf{s}, \mathbf{a}, \delta)\right] - \mathbb{E}_{\mathbf{s}, \mathbf{a} \sim D}\left[Q(\mathbf{s}, \mathbf{a}, \delta)\right]\right) \\
&+ \frac{1}{2} \mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}' \sim D}\left[\left(Q(\mathbf{s}, \mathbf{a}, \delta) - \widehat{\mathcal{B}}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2)\right)^2\right] + \mathcal{R}(\pi) \\
\leq \max_{\delta_1, \delta_2} \max_{\pi} \arg\min_{Q} \,& \alpha \sqrt{\frac{\log(1/\delta_1)}{(n(\mathbf{s}) \wedge 1)}} \left(\mathbb{E}_{\mathbf{s} \sim D, \mathbf{a} \sim \pi(\mathbf{a}|s)}\left[Q(\mathbf{s}, \mathbf{a}, \delta)\right] - \mathbb{E}_{\mathbf{s}, \mathbf{a} \sim D}\left[Q(\mathbf{s}, \mathbf{a}, \delta)\right]\right) \\
&+ \frac{1}{2} \mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}' \sim D}\left[\left(Q(\mathbf{s}, \mathbf{a}, \delta) - \widehat{\mathcal{B}}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2)\right)^2\right] + \mathcal{R}(\pi)
\end{aligned}$$

For any $\delta_1, \delta_2 \leq \delta$ and π, we have that the solution to the inner-minimization over Q yields

$$\begin{aligned}
\tilde{Q}(\mathbf{s}, \mathbf{a}, \delta, \delta_1, \delta_2, \pi) = \arg\min_{Q} \,& \alpha \sqrt{\frac{\log(1/\delta_1)}{(n(\mathbf{s}) \wedge 1)}} \left(\mathbb{E}_{\mathbf{s} \sim D, \mathbf{a} \sim \pi(\mathbf{a}|s)}\left[Q(\mathbf{s}, \mathbf{a}, \delta)\right] - \mathbb{E}_{\mathbf{s}, \mathbf{a} \sim D}\left[Q(\mathbf{s}, \mathbf{a}, \delta)\right]\right) \\
&+ \frac{1}{2} \mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}' \sim D}\left[\left(Q(\mathbf{s}, \mathbf{a}, \delta) - \widehat{\mathcal{B}}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2)\right)^2\right] \\
\leq \,& \widehat{\mathcal{B}}^* \widehat{Q}(\mathbf{s}, \mathbf{a}, \delta_2) - \alpha \sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s})}} \left[\frac{\pi(\mathbf{a} \mid \mathbf{s})}{\pi_\beta(\mathbf{a} \mid \mathbf{s})} - 1\right].
\end{aligned}$$

This arises from taking the derivative of the minimization objective, and solving for Q that makes the derivative equal to 0. Note that we can simplify

$$\alpha\sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s})}}\left[\frac{\pi(\mathbf{a}\mid\mathbf{s})}{\pi_\beta(\mathbf{a}\mid\mathbf{s})}-1\right]=\alpha\sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s})}}\left[\frac{\pi(\mathbf{a}\mid\mathbf{s})-\pi_\beta(\mathbf{a}\mid\mathbf{s})}{\pi_\beta(\mathbf{a}\mid\mathbf{s})}\right]$$
$$=\alpha\sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s},\mathbf{a})}}\left[\frac{\pi(\mathbf{a}\mid\mathbf{s})-\pi_\beta(\mathbf{a}\mid\mathbf{s})}{\sqrt{\pi_\beta(\mathbf{a}\mid\mathbf{s})}}\right].$$

Without loss of generality, assume that $\delta_1, \delta_2 \le \delta$ and π are the solution to the outer-maximization. Substituting the previous result into the equation for $\widehat{Q}(\mathbf{s},\mathbf{a},\delta)$, and applying Lemma A.1 yields,

$$\widehat{Q}(\mathbf{s},\mathbf{a},\delta)\le\widehat{\mathcal{B}}^*\,\widehat{Q}(\mathbf{s},\mathbf{a},\delta_2)-\alpha\sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s},\mathbf{a})}}\left[\frac{\pi(\mathbf{a}\mid\mathbf{s})-\pi_\beta(\mathbf{a}\mid\mathbf{s})}{\sqrt{\pi_\beta(\mathbf{a}\mid\mathbf{s})}}\right]$$
$$\le\widehat{\mathcal{B}}^*\,\widehat{Q}(\mathbf{s},\mathbf{a},\delta_2)-\alpha\sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s},\mathbf{a})}}\left[\frac{\pi(\mathbf{a}\mid\mathbf{s})-\pi_\beta(\mathbf{a}\mid\mathbf{s})}{\sqrt{\pi_\beta(\mathbf{a}\mid\mathbf{s})}}\right]+\sqrt{\frac{\iota\log(1/\delta_1)}{n(\mathbf{s},\mathbf{a})}}.$$

Note that the middle term is not positive if $\pi(\mathbf{a}\mid\mathbf{s})<\pi_\beta(\mathbf{a}\mid\mathbf{s})$. However, we know that for $\mathbf{a}^*=\arg\max_{\mathbf{a}}\widehat{Q}(\mathbf{s},\mathbf{a},\delta)$ then $\pi(\mathbf{a}\mid\mathbf{s})\ge\pi_\beta(\mathbf{a}\mid\mathbf{s})$ by definition of π maximizing the learned Q-values. Therefore, we have

$$\widehat{V}(\mathbf{s},\delta)=\widehat{Q}(\mathbf{s},\mathbf{a}^*,\delta)\le\widehat{\mathcal{B}}^*\,\widehat{Q}(\mathbf{s},\mathbf{a}^*,\delta_2)-\alpha\sqrt{\frac{\log(1/\delta_1)}{n(\mathbf{s},\mathbf{a}^*)}}\left[\frac{\pi(\mathbf{a}^*\mid\mathbf{s})-\pi_\beta(\mathbf{a}^*\mid\mathbf{s})}{\sqrt{\pi_\beta(\mathbf{a}^*\mid\mathbf{s})}}\right]+\sqrt{\frac{\iota\log(1/\delta_1)}{n(\mathbf{s},\mathbf{a}^*)}}$$
$$\le\widehat{\mathcal{B}}^*\,\widehat{V}(\mathbf{s},\delta_2)\quad\forall\mathbf{s}\in\mathcal{S}$$

holds with probability at least $1-\delta_1$ for α satisfying

$$\alpha\ge\iota^{1/2}\max_{\mathbf{s},\mathbf{a}}\left[\frac{\pi(\mathbf{a}\mid\mathbf{s})-\pi_\beta(\mathbf{a}\mid\mathbf{s})}{\sqrt{\pi_\beta(\mathbf{a}\mid\mathbf{s})}}\right]^{-1}.$$

Then, using Lemma A.1, we have

$$\widehat{V}(\mathbf{s},\delta)\le\widehat{\mathcal{B}}^*\,\widehat{V}(\mathbf{s},\delta)\Longrightarrow\widehat{V}(\mathbf{s},\delta)\le V^*(\mathbf{s})\quad\forall\mathbf{s}\in\mathcal{S},$$

holds with probability at least $1-\delta_1\ge 1-\delta$, as desired.

B ATARI RESULTS

In this section, we provide per-game results across all Atari games that we evaluated on for the three considered dataset sizes. As mentioned in the main paper, we use the hyperparameter configuration detailed in Kumar et al. (2022) for our Atari experiments. For completion, we also reproduce the table in this section.

Table 3: Hyperparameters used by the offline RL Atari agents in our experiments. We follow the setup of Agarwal et al. (2020); Kumar et al. (2022).

Hyperparameter	Setting (for both variations)
Sticky actions	Yes
Sticky action probability	0.25
Grey-scaling	True
Observation down-sampling	(84, 84)
Frames stacked	4
Frame skip (Action repetitions)	4
Reward clipping	[-1, 1]
Terminal condition	Game Over
Max frames per episode	108K
Discount factor	0.99
Mini-batch size	32
Target network update period	every 2000 updates
Training environment steps per iteration	250K
Update period every	4 environment steps
Evaluation ϵ	0.001
Evaluation steps per iteration	125K
Q-network: channels	32, 64, 64
Q-network: filter size	8×8, 4×4, 3×3
Q-network: stride	4, 2, 1
Q-network: hidden units	512

Game	**REM**	**CQL**	**AEVL**	**Fixed-CCVL**	**CCVL**
Asterix	405.7 ± 46.5	821.4 ± 75.1	421.2 ± 67.8	874.0 ± 64.3	1032.1 ± 86.7
Breakout	14.3 ± 2.8	32.0 ± 3.2	7.4 ± 1.9	28.7 ± 2.8	31.2 ± 4.3
Pong	-7.7 ± 6.3	14.2 ± 3.3	-8.4 ± 6.8	14.7 ± 3.8	15.8 ± 4.4
Seaquest	293.3 ± 191.5	446.6 ± 26.9	320.6 ± 154.1	422.0 ± 21.9	551.2 ± 42.2
Qbert	436.3 ± 111.5	9162.7 ± 993.6	294.6 ± 100.3	9172.3 ± 907.6	9170.1 ± 1023.5
SpaceInvaders	206.6 ± 77.6	351.9 ± 77.1	224.2 ± 84.7	355.7 ± 80.2	355.4 ± 81.1
Zaxxon	2596.4 ± 1726.4	1757.4 ± 879.4	2467.8 ± 2023.4	1747.6 ± 894.3	2273.6 ± 1803.1
YarsRevenge	5480.2 ± 962.3	16011.3 ± 1409.0	4857.1 ± 1012.6	15890.7 ± 1218.2	20140.5 ± 2022.8
RoadRunner	3872.9 ± 1616.4	24928.7 ± 7484.5	5048.3 ± 2156.5	22590.3 ± 6860.2	26780.5 ± 10112.3
MsPacman	1275.1 ± 345.6	2245.7 ± 193.8	1164.7 ± 508.2	2542.3 ± 188.4	2673.2 ± 226.4
BeamRider	522.9 ± 42.2	617.9 ± 25.1	600.1 ± 57.3	645.3 ± 40.1	630.2 ± 37.8
Jamesbond	157.6 ± 65.0	460.5 ± 102.0	114.3 ± 56.7	462.1 ± 98.4	452.1 ± 153.9
Enduro	132.4 ± 16.1	253.5 ± 14.2	103.2 ± 10.1	244.8 ± 20.9	274.5 ± 23.8
WizardOfWor	1663.7 ± 417.8	904.6 ± 343.7	1640.7 ± 383.4	1488.1 ± 450.9	1513.8 ± 652.1
IceHockey	-9.1 ± 5.1	-7.8 ± 0.9	-10.4 ± 4.9	-7.6 ± 1.1	-7.1 ± 1.6
DoubleDunk	-17.6 ± 1.5	-14.0 ± 2.8	-16.8 ± 2.9	-14.1 ± 1.8	-13.4 ± 4.9
DemonAttack	162.0 ± 34.7	386.2 ± 75.3	183.2 ± 44.7	372.9 ± 81.7	570.3 ± 110.2

Table 4: Mean and standard deviation of returns per Atari game across 5 random seeds using 1% of replay dataset after 6.25M gradient steps. REM and CQL results are from Kumar et al. (2022).

Game	REM	CQL	AEVL	Fixed-CCVL	CCVL
Asterix	2317.0 ± 838.1	3318.5 ± 301.7	1958.9 ± 1050.2	3256.6 ± 395.1	5517.2 ± 1215.4
Breakout	33.4 ± 4.0	166.0 ± 23.1	16.7 ± 5.6	150.3 ± 17.8	172.5 ± 35.6
Pong	-0.7 ± 9.9	17.9 ± 1.1	-0.2 ± 4.7	17.6 ± 2.1	17.4 ± 2.8
Seaquest	2753.6 ± 1119.7	2030.7 ± 822.8	2853.0 ± 1089.2	2112.5 ± 856.4	2746.0 ± 1544.2
Qbert	7417.0 ± 2106.7	9605.6 ± 1593.5	5409.2 ± 3256.6	9750.7 ± 1366.8	10108.1 ± 2445.5
SpaceInvaders	443.5 ± 67.4	1214.6 ± 281.8	450.2 ± 101.3	1243.4 ± 269.8	1154.6 ± 302.1
Zaxxon	1609.7 ± 1814.1	4250.1 ± 626.2	1678.2 ± 1425.6	4060.3 ± 673.1	6470.2 ± 1512.2
YarsRevenge	16930.4 ± 2625.8	17124.7 ± 2125.6	17233.5 ± 2590.8	18040.5 ± 1545.9	19233.0 ± 1719.2
RoadRunner	46601.6 ± 2617.2	38432.6 ± 1539.7	45035.2 ± 3823.0	37945.7 ± 1338.9	42780.5 ± 4112.3
MsPacman	2303.1 ± 202.7	2790.6 ± 353.1	2148.8 ± 273.4	2501.5 ± 201.3	2680.4 ± 212.4
BeamRider	674.8 ± 21.4	785.8 ± 43.5	662.9 ± 50.7	782.3 ± 34.9	780.1 ± 40.8
Jamesbond	130.5 ± 45.7	96.8 ± 43.2	152.2 ± 58.2	112.3 ± 81.3	172.1 ± 153.9
Enduro	1583.9 ± 108.7	938.5 ± 63.9	1602.7 ± 135.5	913.2 ± 50.3	1376.2 ± 203.8
WizardOfWor	2661.6 ± 371.4	612.0 ± 343.3	1767.5 ± 462.1	707.4 ± 323.2	2723.1 ± 515.6
IceHockey	-6.5 ± 3.1	-15.0 ± 0.7	-9.1 ± 4.8	-17.6 ± 1.0	-10.2 ± 2.1
DoubleDunk	-17.6 ± 2.6	-16.2 ± 1.7	-19.4 ± 3.2	-15.2 ± 0.9	-9.8 ± 3.8
DemonAttack	5602.3 ± 1855.5	8517.4 ± 1065.9	2455.3 ± 1765.0	8238.7 ± 1091.2	9730.0 ± 1550.7

Table 5: Mean and standard deviation of returns per Atari game across 5 random seeds using 5% of replay dataset after 12.5M gradient steps. REM and CQL results are from Kumar et al. (2022).

Game	REM	CQL	AEVL	Fixed-CCVL	CCVL
Asterix	5122.9 ± 328.9	3906.2 ± 521.3	7494.7 ± 380.3	3582.1 ± 327.5	7576.0 ± 360.2
Breakout	96.8 ± 21.2	70.8 ± 5.5	97.1 ± 35.7	75.8 ± 6.1	121.4 ± 10.3
Pong	7.6 ± 11.1	5.5 ± 6.2	7.1 ± 12.9	5.2 ± 6.0	13.4 ± 6.1
Seaquest	981.3 ± 605.9	1313.0 ± 220.0	877.2 ± 750.1	1232.6 ± 379.3	1211.4 ± 437.2
Qbert	4126.2 ± 495.7	5395.3 ± 1003.64	4713.6 ± 617.0	5105.5 ± 986.4	5590.9 ± 2111.4
SpaceInvaders	799.0 ± 28.3	938.1 ± 80.3	692.7 ± 101.9	860.5 ± 77.3	1233.4 ± 103.1
Zaxxon	0.0 ± 0.0	836.8 ± 434.7	902.5 ± 895.2	904.1 ± 560.1	1212.2 ± 902.1
YarsRevenge	11924.8 ± 2413.8	12413.9 ± 2869.7	12508.5 ± 1540.2	11587.2 ± 2676.8	12502.6 ± 2349.2
RoadRunner	49129.4 ± 1887.9	45336.9 ± 1366.7	50152.9 ± 2208.9	44832.6 ± 1329.8	47972.1 ± 2991.3
MsPacman	2268.8 ± 455.0	2427.5 ± 191.3	2515.5 ± 548.0	2115.3 ± 108.9	2015.7 ± 352.8
BeamRider	4154.9 ± 357.2	3468.0 ± 238.0	4564.7 ± 578.4	3312.3 ± 247.3	3781.0 ± 401.8
Jamesbond	149.3 ± 304.5	89.7 ± 15.6	127.6 ± 414.8	91.9 ± 20.2	152.8 ± 42.8
Enduro	832.5 ± 65.5	1160.2 ± 81.5	959.2 ± 100.3	1204.6 ± 90.3	1585.0 ± 102.1
WizardOfWor	920.0 ± 497.0	764.7 ± 250.0	1184.3 ± 588.9	749.3 ± 231.8	1429.9 ± 751.4
IceHockey	-5.9 ± 5.1	-16.0 ± 1.3	-5.2 ± 7.3	-14.9 ± 2.5	-4.1 ± 5.9
DoubleDunk	-19.5 ± 2.5	-20.6 ± 1.0	-19.2 ± 2.2	-21.3 ± 1.7	-24.6 ± 6.2
DemonAttack	9674.7 ± 1600.6	7152.9 ± 723.2	10345.3 ± 1612.3	7416.8 ± 1598.7	12330.5 ± 1590.4

Table 6: Mean and standard deviation of returns per Atari game across 5 random seeds using initial 10% of replay dataset after 12.5M gradient steps. REM and CQL results are from Kumar et al. (2022).