Content:

1. Tabular Ablation Study, demonstrating stability of the bias-improvement of ADDQ with respect to Q, DQ, and WDQ across different parameters.

2. Deep Ablation Study, demonstrating stability of the scores of ADDQ across different parameters.

3. Comparison to more Algorithms (MaxMin, EBQL, and REDQ), demonstrating better behavior of ADDQ with respect to bias.

4. Relative variances vs. variances of game, demonstrating that the relative sample variances used in the paper proportionally relate to the real variance of the game.

5. Experimental confirmation of theoretical results for two-sided bandit MDP, demonstrating the connection of number of arms and variance to the overestimation error in QL as well as how ADDQ mitigates this in practice. Furthermore, it shows that there is not much difference between the more practical choice of ϵ-greedy exploration and the choice of uniform exploration rooted in theory.

1 Tabular Ablation Study

The choices of beta in the following Ablation study are named in the following way:

1. (Optional) First two letters: Left-tilted (lt), Right-tilted (rt)
2. First/Third letter: Neutral (n), Aggressive (a), Conservative (c)
3. Final digit: Refers to the number of intervals in the definition of Beta (3 or 5)

As in the paper, the intuition for aggressive, conservative, and neutral remains the same (no interpolation, just choosing which Algorithm's update to take vs. more interpolation, with neutral being in between the two choices.
Left- and Right-tilted refers to the shifted intervals for the relative Variance to fall into while choosing the interpolation coefficient. In the paper, only interval choices centered around 1 were considered, Left-tilted favors the Q update, Right-tilted the DQ update.
The concrete choices are:

$$\text{n3:} \quad \beta := \begin{cases} 0.75 & : S^2_{rel}(s,a) < 0.75 \\ 0.5 & : S^2_{rel}(s,a) \in [0.75, 1.25] \\ 0.25 & : S^2_{rel}(s,a) > 1.25 \end{cases}$$

$$\text{ltn3:} \quad \beta := \begin{cases} 0.75 & : S^2_{rel}(s,a) < 1.25 \\ 0.5 & : S^2_{rel}(s,a) \in [1.25, 1.75] \\ 0.25 & : S^2_{rel}(s,a) > 1.75 \end{cases}$$

$$\text{rtn3:} \quad \beta := \begin{cases} 0.75 & : S^2_{rel}(s,a) < 0.25 \\ 0.5 & : S^2_{rel}(s,a) \in [0.25, 0.75] \\ 0.25 & : S^2_{rel}(s,a) > 0.75 \end{cases}$$

$$\text{a3:} \quad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) < 0.99 \\ 0.5 & : S^2_{rel}(s,a) \in [0.99, 1.01] \\ 0 & : S^2_{rel}(s,a) > 1.01 \end{cases}$$

$$\text{lta3:} \quad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) < 1.49 \\ 0.5 & : S^2_{rel}(s,a) \in [1.49, 1.51] \\ 0 & : S^2_{rel}(s,a) > 1.51 \end{cases}$$

$$\text{rta3:} \quad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) < 0.49 \\ 0.5 & : S^2_{rel}(s,a) \in [0.49, 0.51] \\ 0 & : S^2_{rel}(s,a) > 0.51 \end{cases}$$

$$\text{c3:} \qquad \beta := \begin{cases} 0.6 & : S^2_{rel}(s,a) < 0.6 \\ 0.5 & : S^2_{rel}(s,a) \in [0.6, 1.4] \\ 0.4 & : S^2_{rel}(s,a) > 1.4 \end{cases}$$

$$\text{ltc3:} \qquad \beta := \begin{cases} 0.6 & : S^2_{rel}(s,a) < 1.1 \\ 0.5 & : S^2_{rel}(s,a) \in [1.1, 1.9] \\ 0.4 & : S^2_{rel}(s,a) > 1.9 \end{cases}$$

$$\text{rtc3:} \qquad \beta := \begin{cases} 0.6 & : S^2_{rel}(s,a) < 0.1 \\ 0.5 & : S^2_{rel}(s,a) \in [0.1, 0.9] \\ 0.4 & : S^2_{rel}(s,a) > 0.9 \end{cases}$$

$$\text{n5:} \qquad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) \leq 0.25 \\ 0.75 & : S^2_{rel}(s,a) \in (0.25, 0.75) \\ 0.5 & : S^2_{rel}(s,a) \in [0.75, 1.25] \\ 0.25 & : S^2_{rel}(s,a) \in (1.25, 1.75) \\ 0 & : S^2_{rel}(s,a) \geq 1.75 \end{cases}$$

$$\text{ltn5:} \qquad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) \leq 0.75 \\ 0.75 & : S^2_{rel}(s,a) \in (0.75, 1.25) \\ 0.5 & : S^2_{rel}(s,a) \in [1.25, 1.75] \\ 0.25 & : S^2_{rel}(s,a) \in (1.75, 2.25) \\ 0 & : S^2_{rel}(s,a) \geq 2.25 \end{cases}$$

$$\text{rtn5:} \qquad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) \leq -0.25 \\ 0.75 & : S^2_{rel}(s,a) \in (-0.25, 0.25) \\ 0.5 & : S^2_{rel}(s,a) \in [0.25, 0.75] \\ 0.25 & : S^2_{rel}(s,a) \in (0.75, 1.25) \\ 0 & : S^2_{rel}(s,a) \geq 1.25 \end{cases}$$

$$\text{a5:} \qquad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) \leq 0.99 \\ 0.75 & : S^2_{rel}(s,a) \in (0.99, 0.995) \\ 0.5 & : S^2_{rel}(s,a) \in [0.995, 1.005] \\ 0.25 & : S^2_{rel}(s,a) \in (1.005, 1.01) \\ 0 & : S^2_{rel}(s,a) \geq 1.01 \end{cases}$$

$$\text{lta5:} \qquad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) \leq 1.49 \\ 0.75 & : S^2_{rel}(s,a) \in (1.49, 1.495) \\ 0.5 & : S^2_{rel}(s,a) \in [1.495, 1.505] \\ 0.25 & : S^2_{rel}(s,a) \in (1.505, 1.51) \\ 0 & : S^2_{rel}(s,a) \geq 1.51 \end{cases}$$

$$\text{rta5:} \qquad \beta := \begin{cases} 1 & : S^2_{rel}(s,a) \leq 0.49 \\ 0.75 & : S^2_{rel}(s,a) \in (0.49, 0.495) \\ 0.5 & : S^2_{rel}(s,a) \in [0.495, 0.505] \\ 0.25 & : S^2_{rel}(s,a) \in (0.505, 0.51) \\ 0 & : S^2_{rel}(s,a) \geq 0.51 \end{cases}$$

$$\text{c5:} \qquad \beta := \begin{cases} 0.7 & : S^2_{rel}(s,a) \leq 0.1 \\ 0.6 & : S^2_{rel}(s,a) \in (0.1, 0.7) \\ 0.5 & : S^2_{rel}(s,a) \in [0.7, 1.3] \\ 0.4 & : S^2_{rel}(s,a) \in (1.3, 1.9) \\ 0.3 & : S^2_{rel}(s,a) \geq 1.9 \end{cases}$$

$$\text{ltc5:} \qquad \beta := \begin{cases} 0.7 & : S^2_{rel}(s,a) \leq 0.6 \\ 0.6 & : S^2_{rel}(s,a) \in (0.6, 1.2) \\ 0.5 & : S^2_{rel}(s,a) \in [1.2, 1.8] \\ 0.4 & : S^2_{rel}(s,a) \in (1.8, 2.4) \\ 0.3 & : S^2_{rel}(s,a) \geq 2.4 \end{cases}$$

$$\text{rtc5:} \qquad \beta := \begin{cases} 0.7 & : S^2_{rel}(s,a) \leq -0.4 \\ 0.6 & : S^2_{rel}(s,a) \in (-0.4, 0.2) \\ 0.5 & : S^2_{rel}(s,a) \in [0.2, 0.8] \\ 0.4 & : S^2_{rel}(s,a) \in (0.8, 1.4) \\ 0.3 & : S^2_{rel}(s,a) \geq 1.4 \end{cases}$$

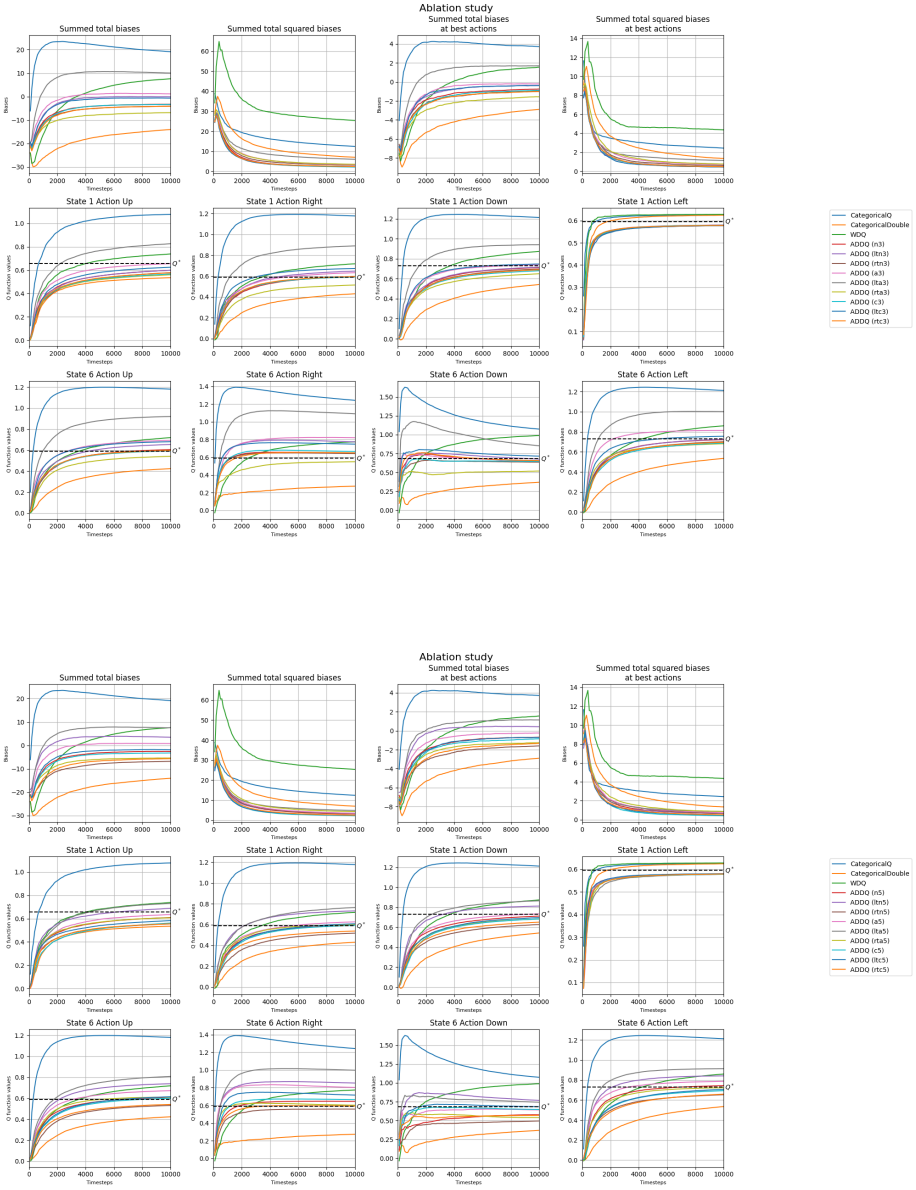

Figure 1: Tabular Ablation Study on GridWorld example from the Paper with parameter choices from above. The effect of hyperparameter choice of β is small with respect to the Bias improvement. Compared to Q, DQ, and WDQ the Bias is much lower. Conservative choices seem to work especially well. State 1 is adjacent to the Fake Goal, State 6 adjacent to the Stochastic Region.

2 Deep Ablation Study

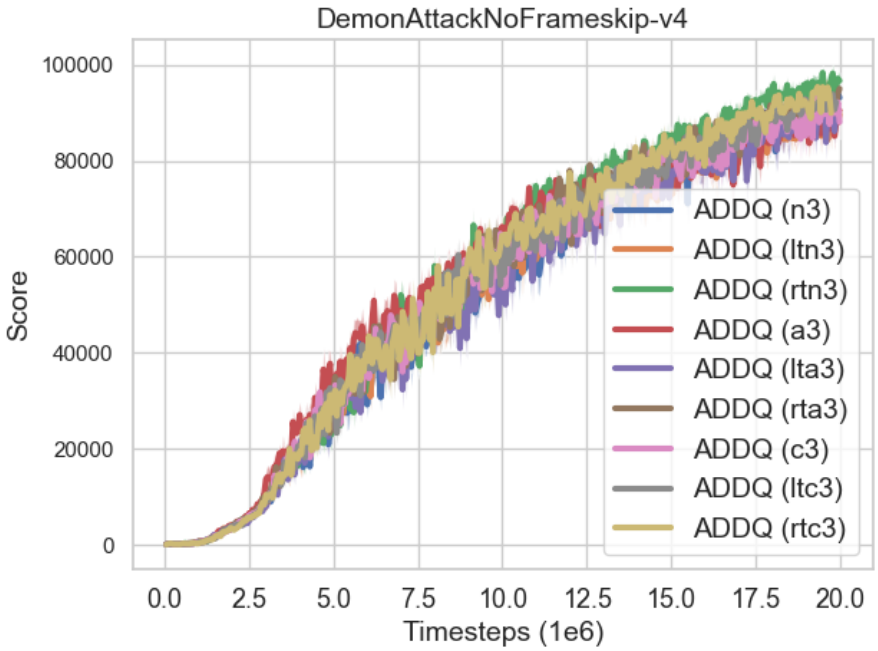

Figure 2: Deep Ablation Study on one Atari environment with 10 seeds and parameter choices as detailed in previous section. The effect of hyperparameter choice of β is marginal with respect to the score.

3 Comparison to more Algorithms

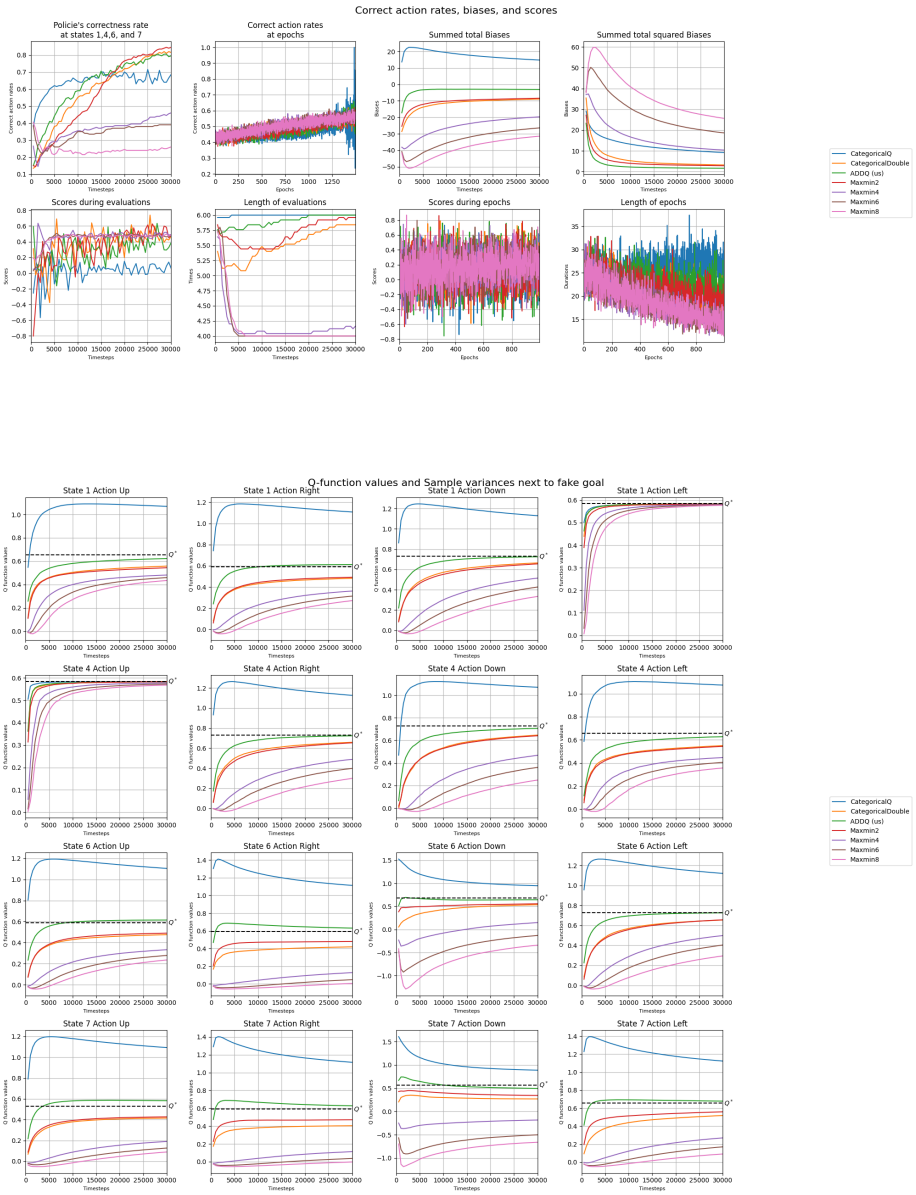

Figure 3: On the GridWorld example from the Paper, ADDQ decreases Biases compared to MaxMin Algorithm across different choices of Ensemble sizes. State 1 and 4 are adjacent to the Fake Goal, State 6 and 7 are adjacent to the Stochastic Region.

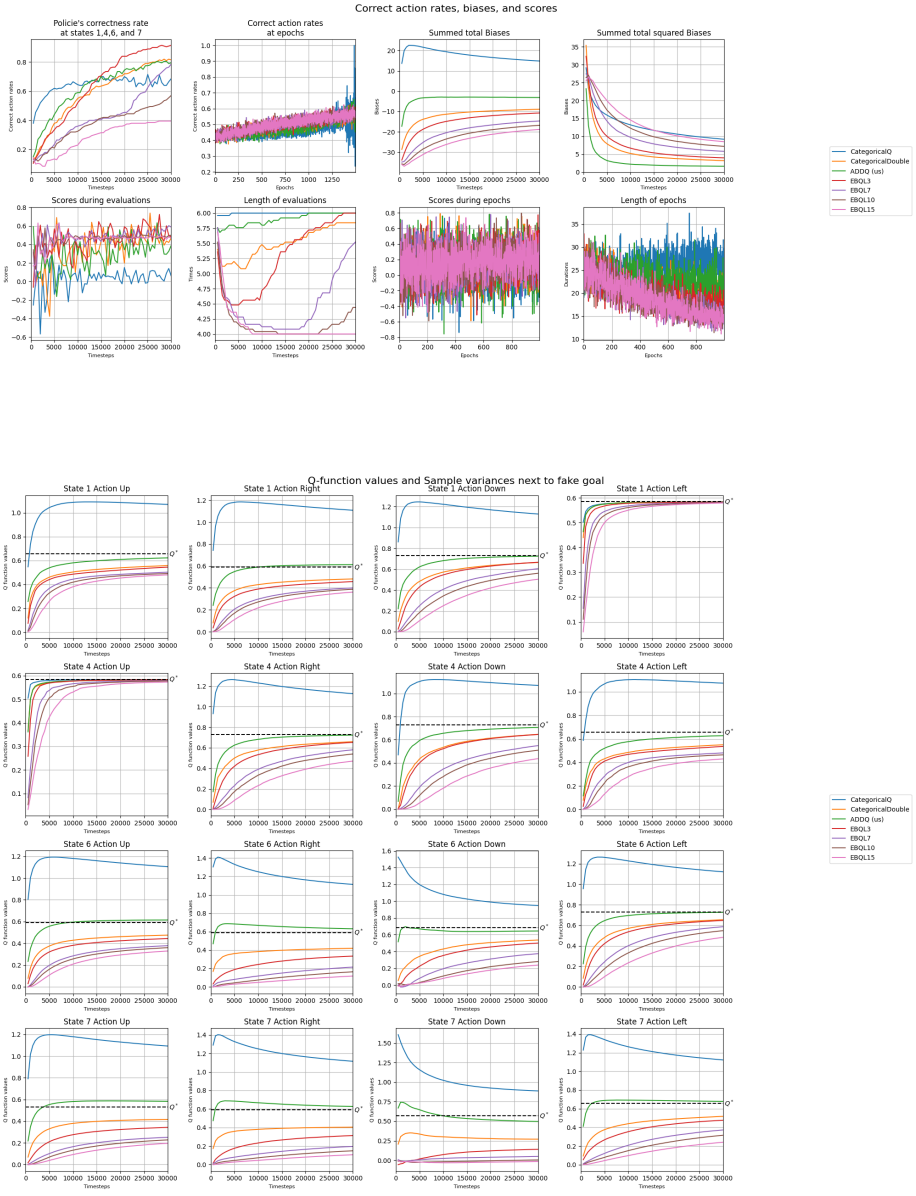

Figure 4: On the GridWorld example from the Paper, ADDQ decreases Biases compared to Ensemble Bootstrapped QL (EBQL) Algorithm across different choices of Ensemble sizes. State 1 and 4 are adjacent to the Fake Goal, State 6 and 7 are adjacent to the Stochastic Region.

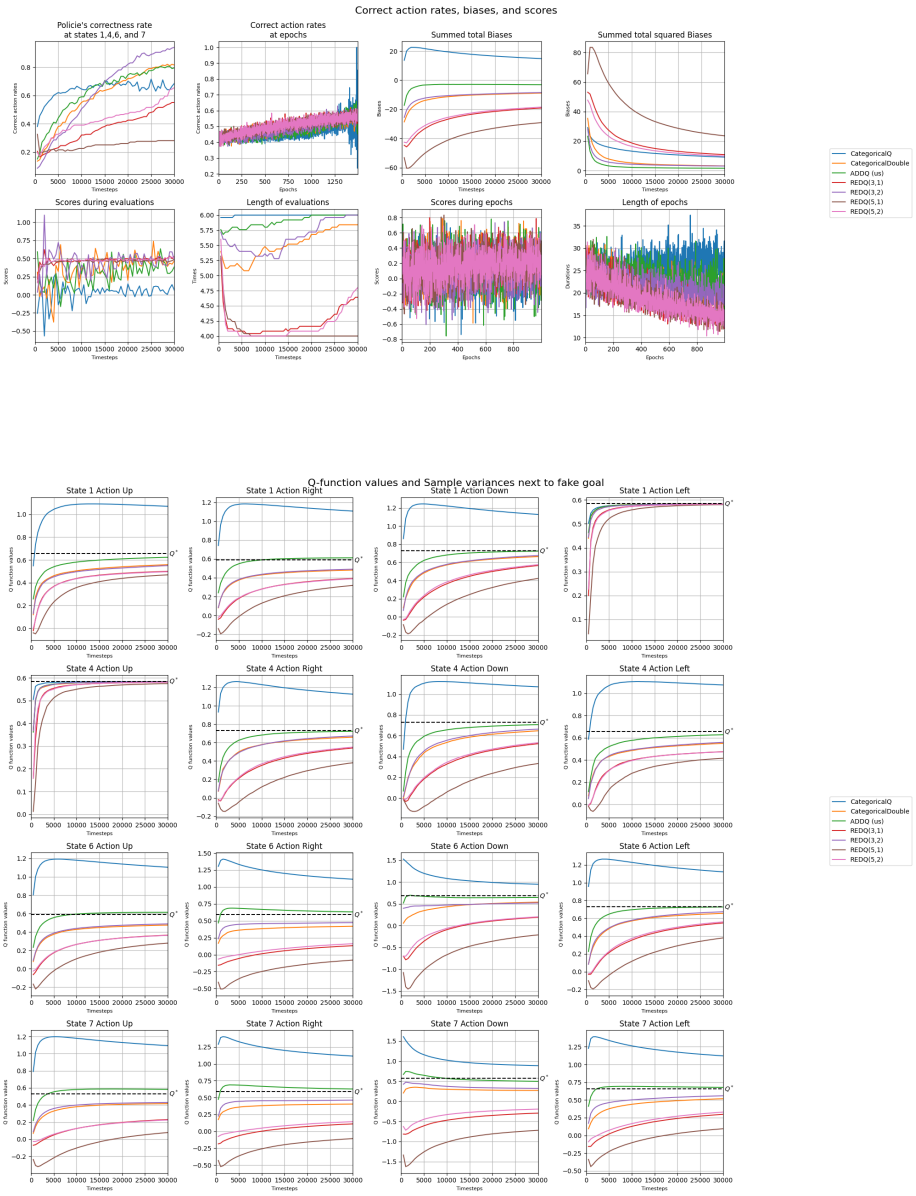

Figure 5: On the GridWorld example from the Paper, ADDQ decreases Biases compared to Randomized Ensemble DQL Algorithm across different choices of Ensemble sizes and random update subset sizes. The Bias is significantly lower. State 1 and 4 are adjacent to the Fake Goal, State 6 and 7 are adjacent to the Stochastic Region.

4 Relative variances vs. variances of game

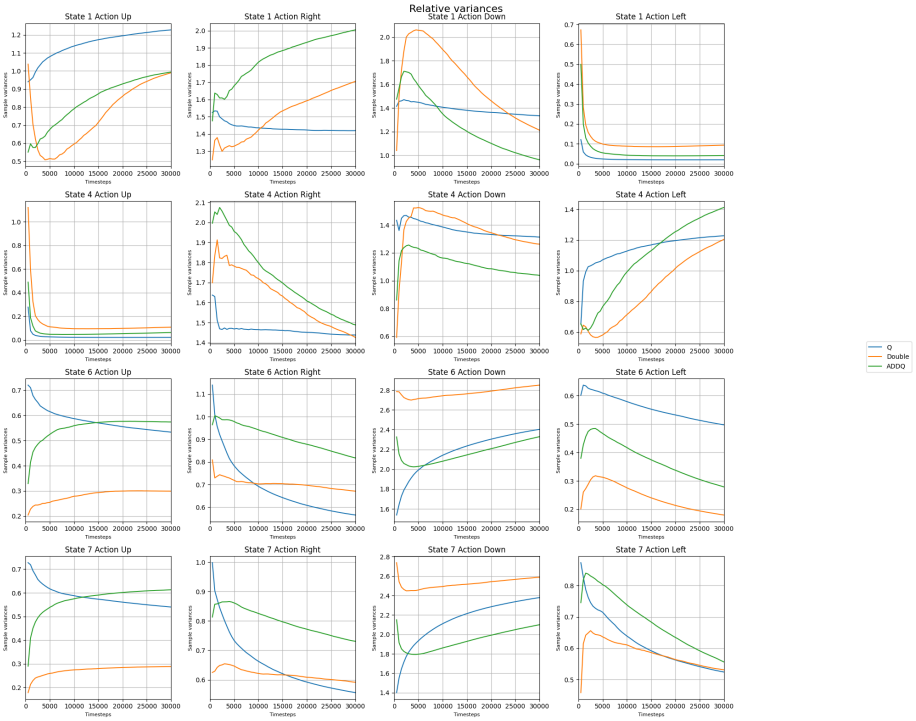

Figure 6: Relative variances of different state action pairs on the GridWorld example from the Paper. The relative sample variance is strongly determined by the relative variance of the next state. State 1 and 4 are to the right and the bottom of the Fake Goal respectively, coinciding with the Left/Up-action's relative sample variance being much smaller. Analogously, State 6 and 7 are above the Stochastic Region and the relative sample variance for going down is much higher. Interpolation coefficients will be chosen low (update leans towards DQL) whenever the relative sample variances (and thus the variances of the game) are high and vice-versa.

5 Experimental confirmation of theoretical results for two-sided bandit MDP

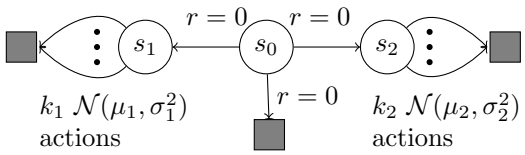

Figure 6: Two-sided bandit MDP, start in s_0, gray boxes terminal

Results are shown for $\mu_1 = -0.1$, $\mu_2 = 0.1$, $k_2 = 5$, $\sigma_2 = 1$, $\gamma = 0.9$, and learning rate $\frac{1}{n}$ in all cases; the correct decision is moving to the right in the Start State. The plot for correct action rates always refers to if the exploration was greedy. The following situations are shown:

1. QL and ADDQ with $\sigma_1 = 5$ iterating over $k_2 = 5, 10, 15, 20$, first with ϵ-greedy exploration (E) with ϵ linearly decreasing from 1 to 0.1 in 10000 steps and then with uniform exploration (U).

2. QL and ADDQ with $k_1 = 10$, iterating over $\sigma_1 = 2, 4, 6, 8$, first with ϵ-greedy exploration like above and then with uniform exploration.

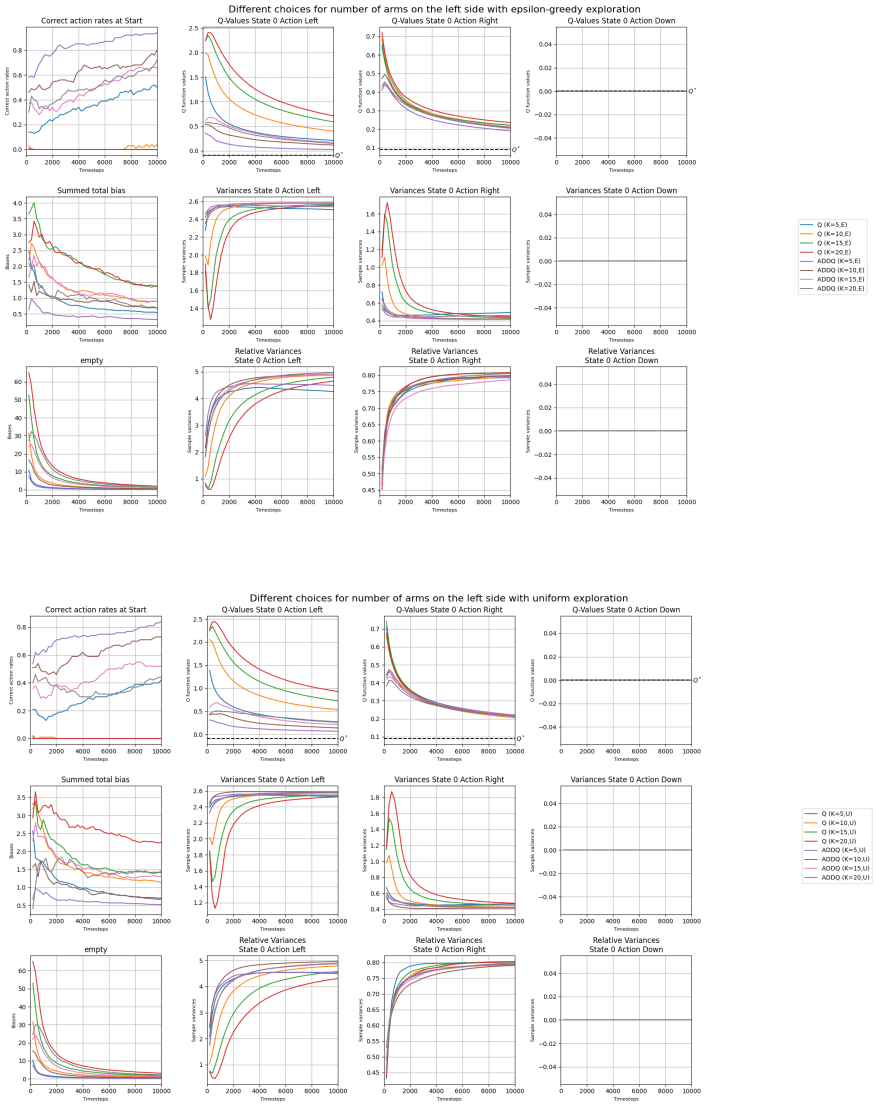

Figure 7: ADDQ and QL on two-sided bandit MDPs from the Paper with different number of arms on the left side, on the top with ϵ-greedy exploration (E) as described above, on the bottom with uniform exploration (U). The amount of overestimation of the Q-value is proportional to the number of arms to the left side of the MDP as suggested by the developed theory. ADDQ is much better in terms of bias, leveraging the local information given by the (relative) variances. There is almost no difference between ϵ-greedy exploration and uniform exploration, suggesting that the developed theory might generalize.

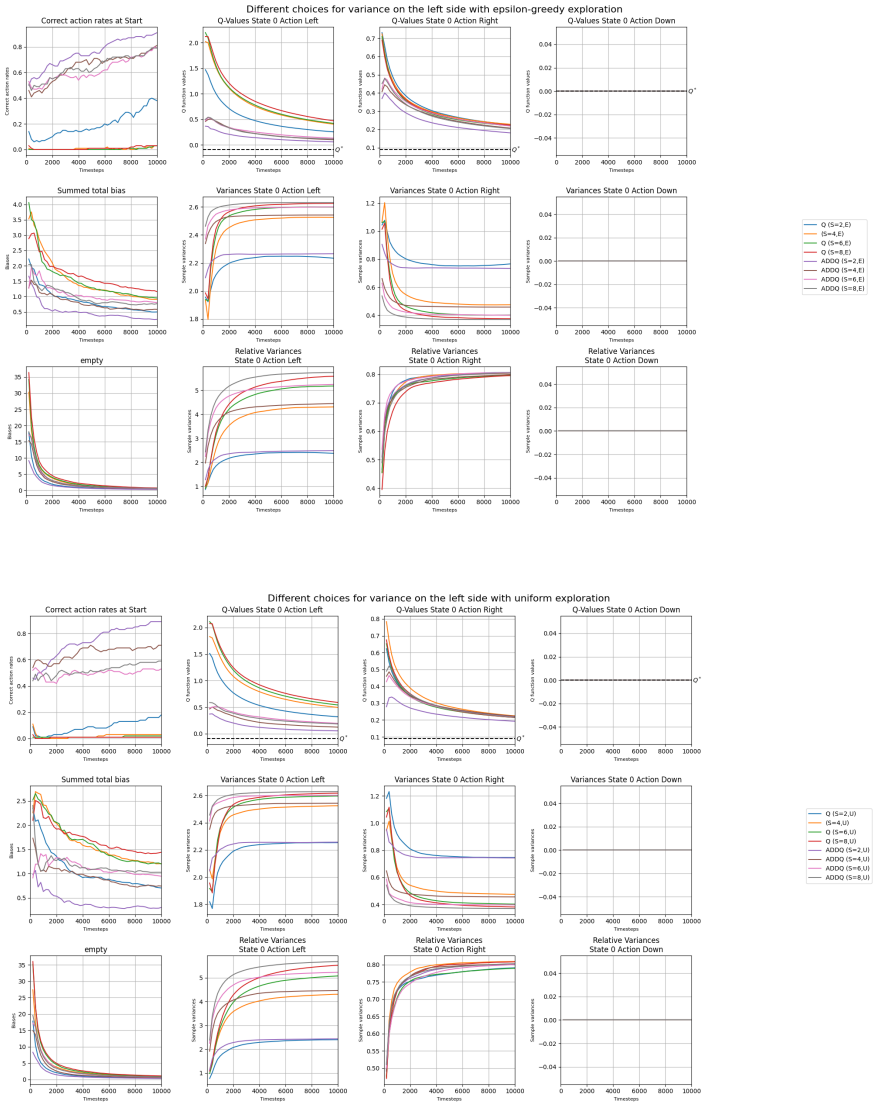

Figure 8: ADDQ and QL on two-sided bandit MDPs from the Paper with different variances on the left side, on the top with ϵ-greedy exploration (E) as described above, on the bottom with uniform exploration (U). The amount of overestimation of the Q-value is proportional to the number of arms to the left side of the MDP as suggested by the developed theory. ADDQ is much better in terms of bias, leveraging the local information given by the (relative) variances. There is almost no difference between ϵ-greedy exploration and uniform exploration, suggesting that the developed theory might generalize.