number of actions Figure 1: The orange bars show the bias in a single Q-

1.5

learning update when the action values are Q(s,a) = $V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically and in-

dependently. All bars are the average of 100 repetitions.

 $\max_{a} Q(s, a) - V_*(s)$ $Q'(s, \operatorname{argmax}_a Q(s, a)) - V_*(s)$