A. Proofs

Theorem 1. The proof of Theorem 1 follows directly from observing that $\hat{\mu}_*^{WE}$ is always smaller than $\hat{\mu}_*^{ME}$. In fact, the ME estimator can be seen as a weighted estimator that gives probability one to the variable associated to the largest sample mean $\hat{\mu}_i$, so that any other weighting cannot produce a larger value. □

Theorem 2. If we compare the expected value of DE reported in Equation (4) with the value of the estimator WE in Equation (3), we can notice strong similarities. The main difference is that in DE the sample mean of variable X_i and its probability of being the maximum are computed w.r.t. two independent set of samples, while in WE these two quantities are positively correlated. It follows that WE has a positive bias w.r.t. DE. □

Theorem 3. Starting from the definition of WE (3), we can derive the bound to the variance as follows

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_*^{WE}\right) &= \mathrm{Var}\left(\sum_{i=1}^{M}\hat{\mu}_i(S)w_i^S\right)\\ &\leq \mathrm{Var}\left(\sum_{i=1}^{M}\hat{\mu}_i(S)\right)\\ &= \sum_{i=1}^{M}\mathrm{Var}\left(\hat{\mu}_i(S)\right),\end{aligned}$$

where the inequality is a consequence of the maximization of each weight w_i^S with one and the last equality comes from the independence of the sample means. □

Theorem 4. Since the weights w_i computed by DWE are not random variables, it follows

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_*^{DWE}\right) &= \mathrm{Var}\left(\sum_{i=1}^{M}\hat{\mu}_i(S)w_i\right)\\ &= \sum_{i=1}^{M}w_i^2\mathrm{Var}\left(\hat{\mu}_i(S)\right)\\ &\leq \max_{i\in 1,\ldots,M}\frac{\sigma_i^2}{|S_i|},\end{aligned}$$

where the inequality is motivated by $w_i^2 \leq 1,\ \forall i$. □

B. Forex

The indicators chosen to define the states are: Moving Average Convergence/Divergence indicator, Relative Strength Index, Momentum, Channel Commodity Index, Stochastic Oscillator, Bollinger Bands, Moving Average Cross-Over The actions suggested by the indicators are computed by setting the parameters and the entry-exit conditions following the most used and common rules for these indicators. In particular all the signals are defined using implemented

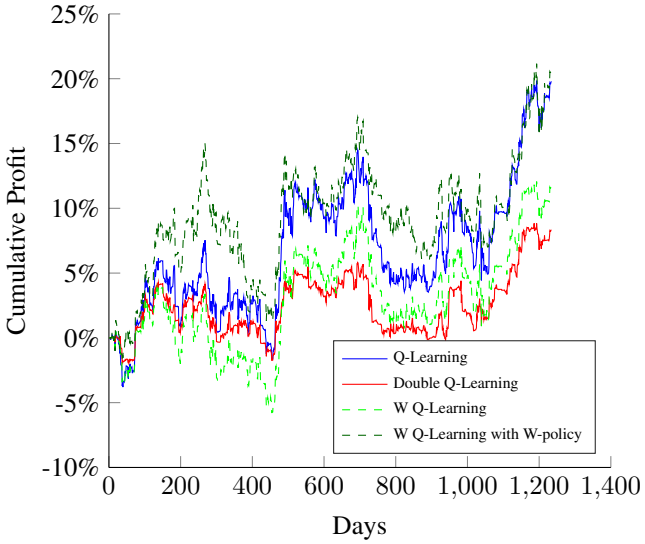

Figure 9. Cumulative profit in test set after 30 training episodes. Results are averaged over 100 experiments.

Matlab Financial Toolbox functions and setting parameters and conditions like below.

- Moving Average Covergence/Divergence indicator: the MACD is calculated by subtracting the 26-period exponential moving average from the 12-period moving average. A 9 period exponential moving average is used as signal line. When the MACD falls below the signal line a long position signal is produced. Otherwise if the MACD surpasses the signal line a short position signal is produced.

- Relative Strength Index: the period chosen for the RSI is 20 days. When the RSI is under the value of 30 an oversold market condition occurs, so a long position signal is produced. Similarly, when the RSI is over 70, a short position signal is generated. When the value is between these two bounds a close position signal is produced.

- Momentum: the momentum uses a period of 14 days. The strategy is to be always in the market. In particular if the value of the momentum is less than zero a long position signal is generated. Otherwise a short position signal is generated.

- Channel Commodity Index: the CCI is used with a 20 days period. The long position signal is produced when the CCI cross and surpasses the lower bound of -100. The short position is taken when the CCI falls behind the value of 100. Otherwise the close position signal is generated.

- Stochastic Oscillator: the Stochastic oscillator uses the high prices, low prices, and close prices with a 14 days period for the %K line and 3 for the %D line. If both lines are above the value of 80 and the %K line falls behind %D line, then a short position signal is generated. If both lines

are below 20 and the %K line surpasses the %D line, a long position signal is generated. In other conditions a close position signal is generated.

- Bollinger Bands: the Bollinger Bands signal is produced with a window size of 20. Long position signal is produced when the closing price surpasses the upper Bollinger Band. When the closing price falls down the lower band, then a short position signal is taken. Otherwise a close position signal is generated.

- Moving Average Cross-Over: we used two moving average of 20 and 200 periods. A long position signal is produced when the 20 period moving average falls behind the 200 periods one. Otherwise a short position signal is generated.

In Figure 9, we can compare the cumulative rewards of the algorithms on the test set after 30 training episodes. Observing the ranges of the values, we can see that the Double Q-learning agent gains and loses less than the other agents, so we can deduce that it entries in the market less often.