



Figure 2: Results from our directed 2-armed bandit test averaged across 10 runs with 1-std deviation (shaded). (a) plots the learning curves of the methods, (b) plots the avg. reward obtained by greedily running the policy computed 10 times (for clarity, the Oracle’s avg. reward is annotated with \times periodically). Vertical squiggly lines denote the step where a new task M_{i+1} and transition system δ_{i+1} were loaded ($M_i \neq M_{i+1}$ and $\delta_i \neq \delta_{i+1}$).

A Directed 2-armed Bandit Test

Fig. 2 shows the results obtained from our directed 2-armed bandit test. CLaP uses goodness of fit tests and thus is able to quickly identify that the distribution of the first lever has changed. Once the correct probabilities are learned for the first lever it computes a new policy that identifies that lever 2 is more lucrative.

B Notes on Source Code

We have included the source code as a part of the submission. However, as the file upload limit is only 5 MiB we had to remove essential libraries that are required for the code to work. We do however include the complete source code of our algorithm. Please refer to the Readme file that points to the files implementing CLaP and the baselines.

Similarly, our result CSV files are 7MB when compressed. We include data on only 5 of the 10 runs to fit within the file limit.