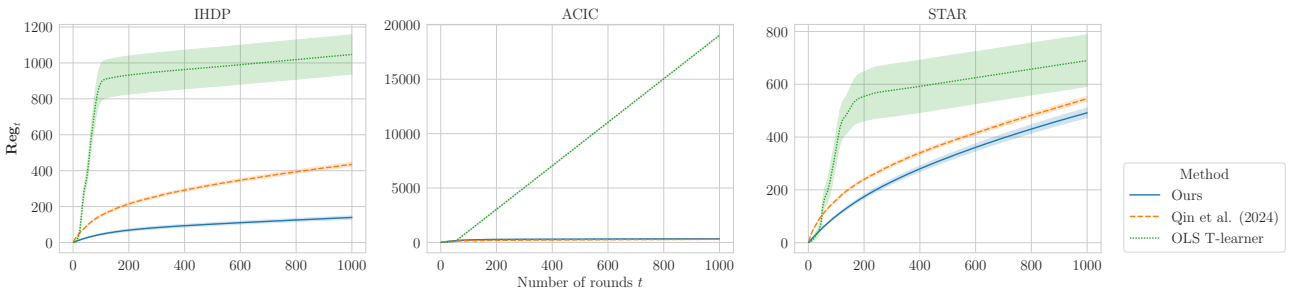

Figure 1. Regrets of our proposed method and two baselines. The OLS T-learner baseline is a static estimator retrained with all historical data at each round. The OLS baseline encounters numerical stability issues in some rounds and causes extreme values, so we clipped per-round regrets at 20 for better visualization in all figures. Environment parameters are set to $u = 0.5$, $v = 0.5$, and $P_T = 0$. Shaded bands indicate standard error.

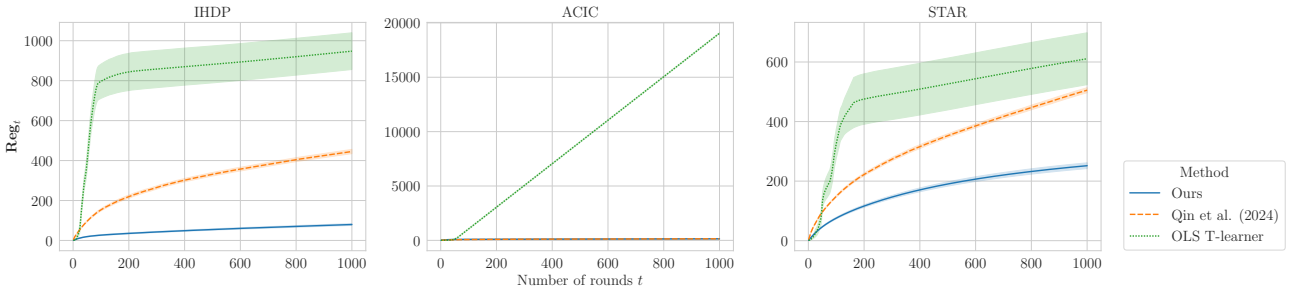

Figure 2. Regrets of our proposed method and two baselines. Environment parameters are set to $u = 0.5$, $v = 0.5$, and $P_T > 0$. Shaded bands indicate standard error.

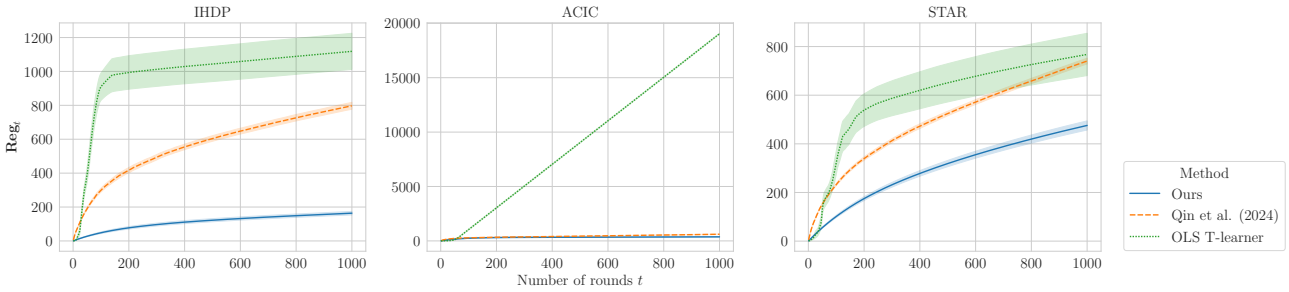

Figure 3. Regrets of our proposed method and two baselines in **a new environment**. The new environment changing mechanism replaces the $\delta(h(\boldsymbol{x}_t; \boldsymbol{\theta}_t))$ in the original mechanism with $\delta(MLP(\boldsymbol{x}_t; \boldsymbol{\theta}_t))$, where $MLP(\cdot)$ is a randomly initialized multi-layer perceptron with two hidden layers and ReLU activations meant to simulate more complex environments. Environment parameters are set to $u = 0.5$, $v = 0.5$, and $P_T = 0$. Shaded bands indicate standard error.

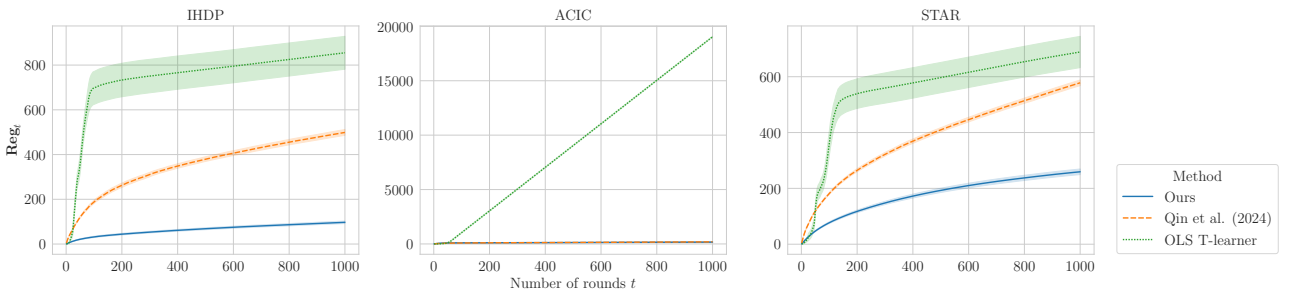

Figure 4. Regrets of our proposed method and two baselines in **a new environment** as in Fig. 3. Environment parameters are set to $u = 0.5$, $v = 0.5$, and $P_T > 0$. Shaded bands indicate standard error.