CS 332 Fall 2025

Project #2

Koshi Harashima, Ben Cole
Due Date: 10/22, 2025

1 Part 1 - A

Methods
In the Adversarial Fair Payoffs (AFP) setting, each round's payoff is given to the arm with the lowest total reward so far, creating an adversarial environment. We simulate $k = 3$, $n = 1000$ with learning rates $\epsilon = \{0.01, \sqrt{\log k/n}, 100\}$.
Results
FTL's regret grows linearly with rounds, while random and theoretical learning rates keep regret near zero and remain within the analytical bound. Total payoffs confirm the same pattern.
Takeaways
In adversarial, unpredictable environments, aggressive exploitation (FTL) performs poorly. Stable or random learning rates achieve low regret.

2 Part 1 - B

Methods
In the Bernoulli Payoffs (BP) setting, each arm's reward follows $v_j^i \sim B(p_j)$ with fixed probabilities p_j. We simulate $k = 3$, $n = 1000$ under full information.
Results
FTL rapidly identifies the best arm and achieves low regret as n increases. Random and theoretical learning rates fluctuate around higher regret levels.
Takeaways
When payoffs are stationary and one arm is truly optimal, FTL converges efficiently. Exploration offers no benefit in stable environments.

3 Part 2 - C

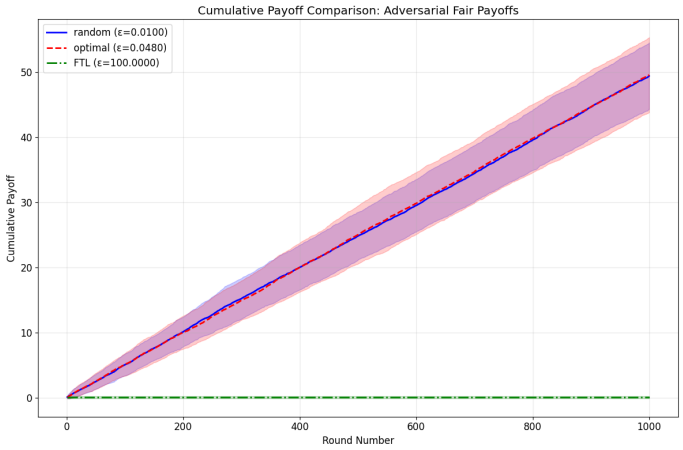
Methods
We apply the Exponential Weights (EW) algorithm to real-world Pachinko data, treating each store as an "arm." Daily ROI (Return on Investment) is normalized to $[0, 1]$, capturing store-level profitability under full information.
Model
Tokyo Pachinko stores are compared across days. Payoffs are nonstationary since stores adjust machine settings strategically to attract players.
Results
FTL quickly identifies and sticks to the best store, causing regret to converge to near zero. Random and theoretical learning rates fail to adapt to shifting ROI patterns and perform worse.
Takeaways
Even with noisy real-world data, simple FTL behavior is effective in our sample because the best store tends to persist long enough to be exploited, despite day-to-day noise

4 Part 2 - D

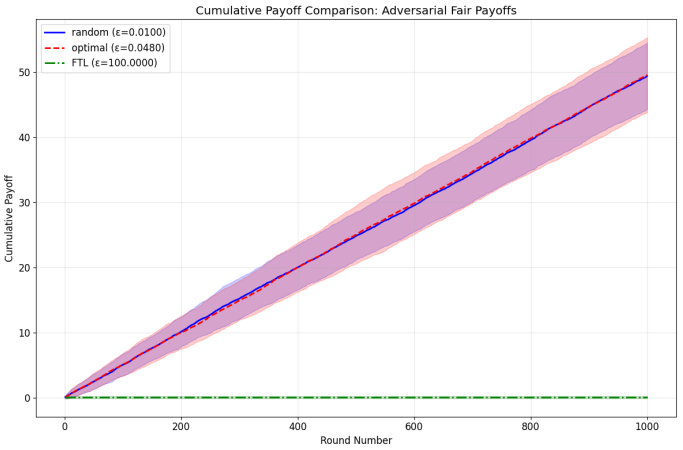
Methods
We build a randomized *Research Payoffs (RP)* model with clusters (correlated papers) and regime shifts. At round t, the agent allocates $w_t \in \Delta^N$ and receives $U_t = \alpha_t^\top w_t$.
Model
One cluster is high, one middle, others low; a hidden Markov state switches the dominant cluster with persistence p.
Algorithms
FTL (past leader), Uniform (even split), and EW/EG ($w_{t+1,i} \propto w_{t,i} e^{\varepsilon \alpha_{t,i}}$).
Results
FTL suffers spikes after switches; Uniform has steady underperformance; EW/EG (moderate ε) tracks the dominant cluster and achieves the lowest regret and highest average payoff.
Takeaways
Correlation + regime changes break both FTL and Uniform; multiplicative updates with a conservative but nonzero learning rate are robust.

Figures (Compact)

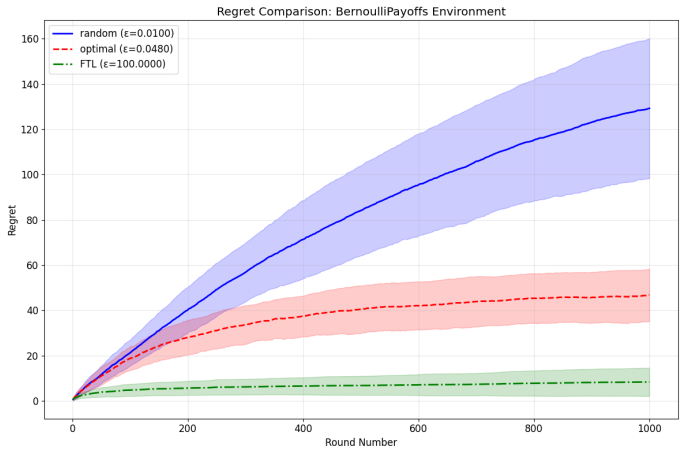
Note: AFP/BP use k=3, n=1000, h=1. Bound: $2h\sqrt{n \log k} \approx 110$.



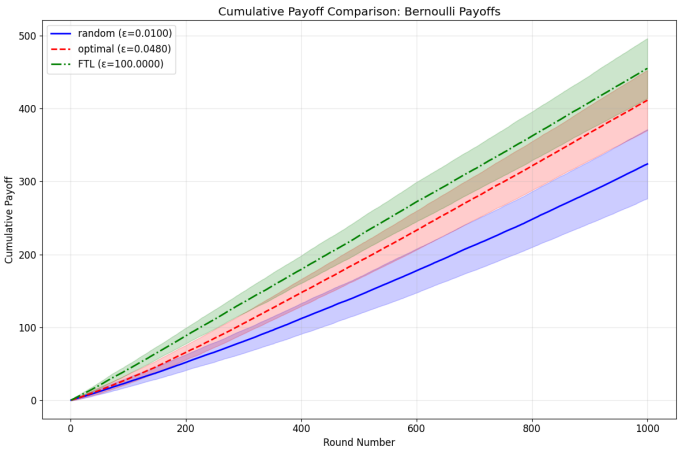
A: Regret vs. rounds



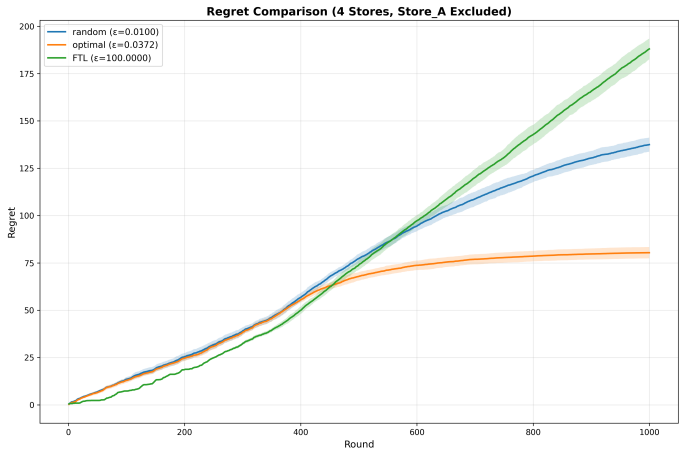
A: Total payoff



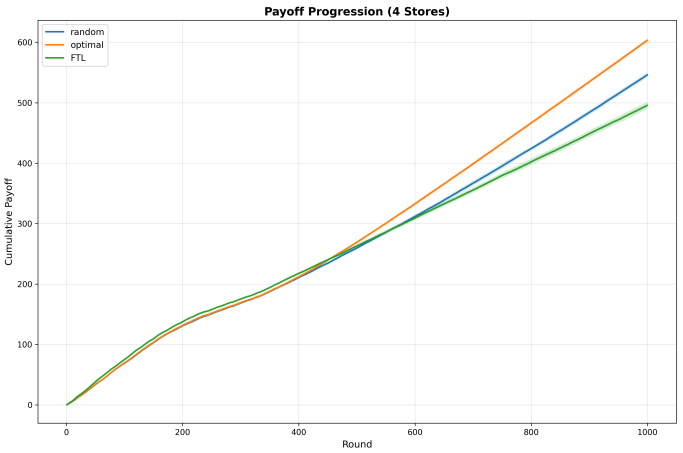
B: Regret vs. rounds



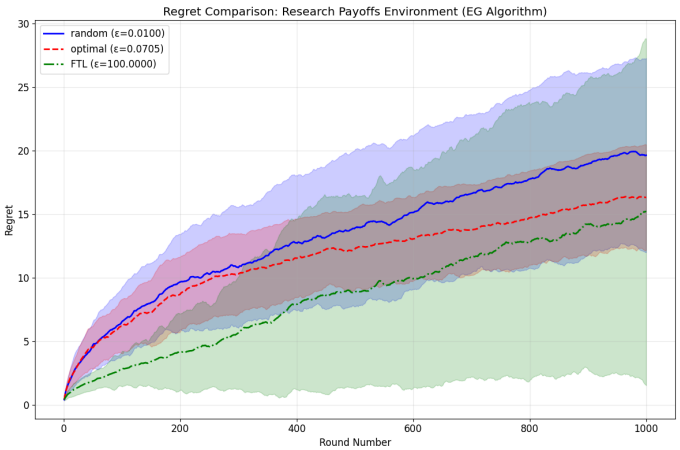
B: Total payoff



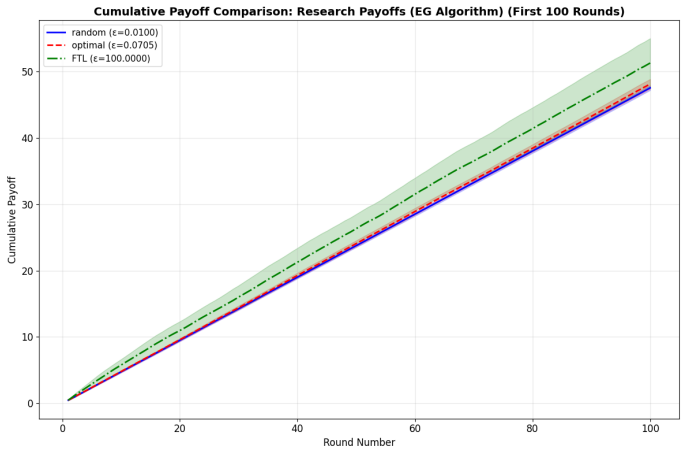
P: Regret vs. rounds



P: Total payoff



D: Regret vs. rounds



D: Total payoff