Project 2

CS 332, Fall 2025

Ben Cole Koshi Harashima

22 October, 2025

Basic Setting & Regret

Online learning setup

- k actions, n rounds; in round i we choose a_i, then observe full payoffs $v_1^i, \ldots, v_k^i \in [0, h]$ and obtain $v_{a_i}^i$.
- Algorithm payoff: $ALG = \sum_{i=1}^{n} v_{a_i}^i$; best-in-hindsight $OPT = \max_j \sum_{i=1}^{n} v_j^i$.
- Average regret: $\text{Regret}_n = \frac{1}{n}(OPT - ALG)$.

Exponential Weights (EW)

$$\pi_j^i = \frac{(1+\epsilon)^{V_j^{i-1}/h}}{\sum_{j'}(1+\epsilon)^{V_{j'}^{i-1}/h}}, \qquad V_j^i = \sum_{r=1}^{i} v_j^r$$

Bound: $\text{Regret}_n \le \epsilon n h + \frac{h \log k}{\epsilon}$; optimal $\epsilon = \sqrt{\frac{\log k}{n}}$ gives $2h\sqrt{n \log k}$.

EW: Rates & Monte Carlo Setup

Learning rates compared

- No learning: $\epsilon \approx 0$ (uniform random).
- Theoretical: $\epsilon = \sqrt{\ln k / n}$.
- FTL: $\epsilon \to \infty$ (or explicit Follow-The-Leader).

MC trials

- Fix $k = 5$, $n = 1000$; multiple runs; report mean regret with confidence intervals.
- Evaluate two models: Adversarial Fair Payoffs (AFP) and Bernoulli Payoffs (BP).

Part 1 - Summary

Methods
In AFP and BP, we first gain intuition from observation, then simulate them.

Results
In AFP, FTL works poorly, while other learning rates work well. In BP, FTL works better than the others.

Takeaways
FTL performs best when there is an optimal arm (stationary). Random and optimal learning rates perform best without a stable optimal arm.

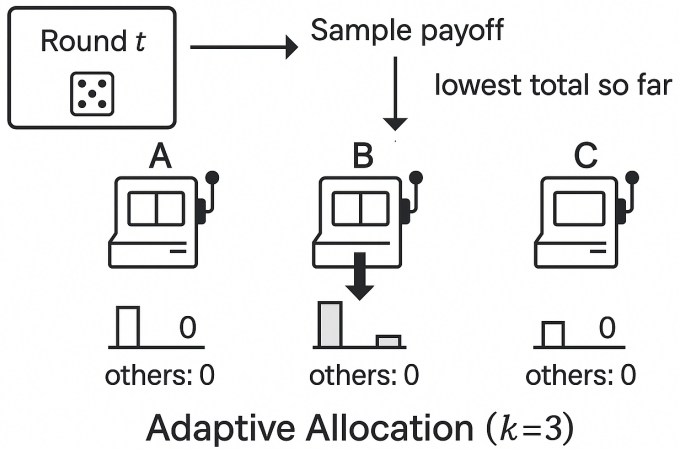
Part 1A (AFP): Setting & Intuition

Setting

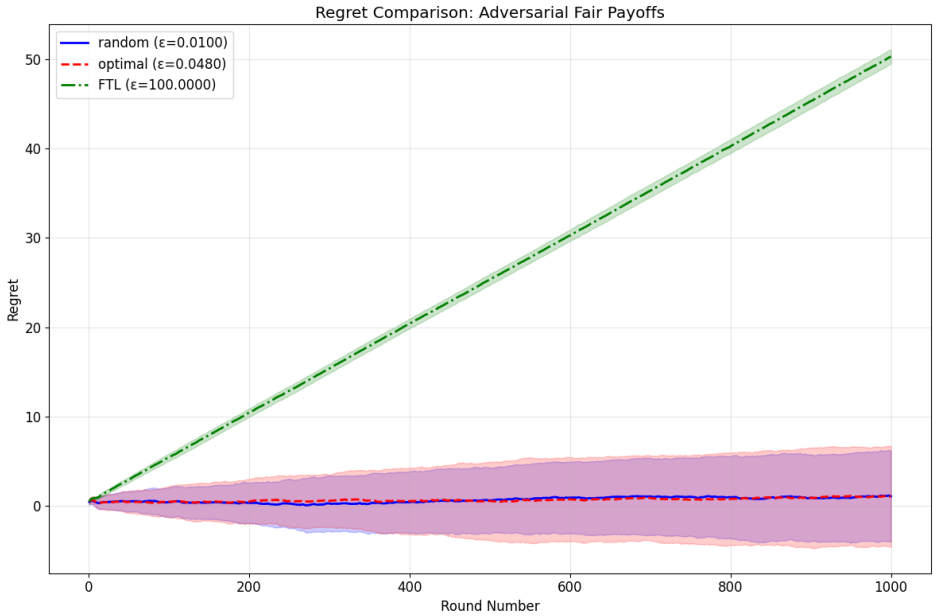
- Each round: draw $x \sim U[0,1]$; assign to arm $j^* = \arg\min_j V_j^{i-1}$ (others get 0).
- Fixed: $k = 5$, $n = 1000$.

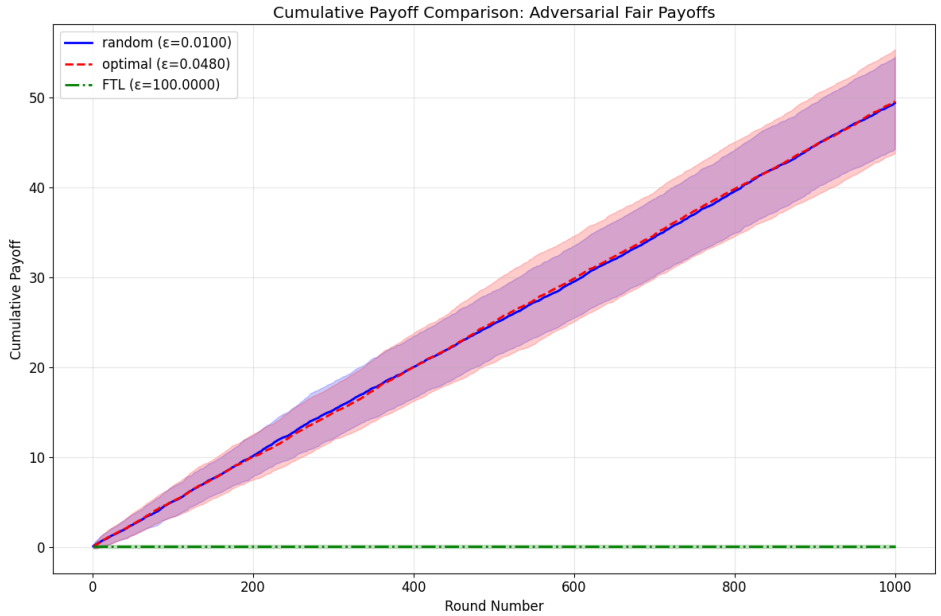
Intuition

- FTL chases the leader and is "punished" next round $\Rightarrow$ regret grows roughly linearly.
- Random or theoretical ϵ avoid overreaction $\Rightarrow$ near-zero regret.



Part 1A: Results





Note: With $\epsilon = \sqrt{\ln k / n}$, the bound is ≈ 110 for $h = 1, k = 5, n = 1000$; observed regret stays below this.

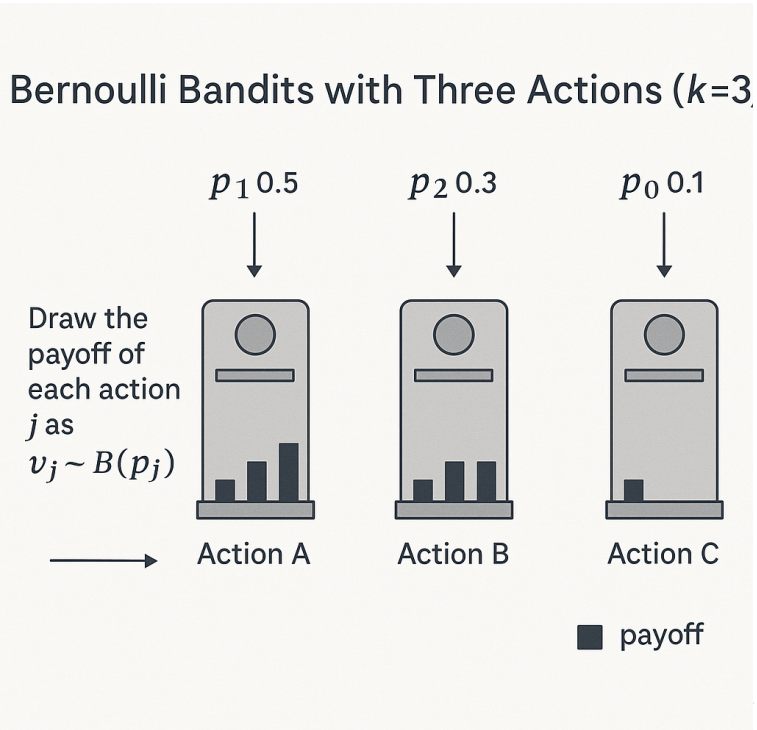
Part 1B (BP): Setting & Intuition

Setting

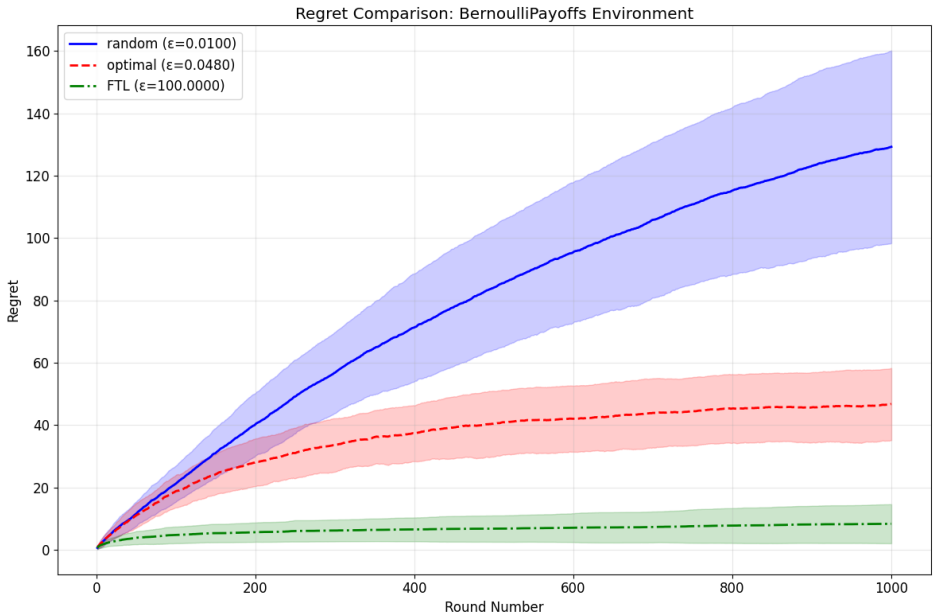
- Fix $p_j \in [0, 1/2]$; each round $v_j^i \sim \text{Bernoulli}(p_j)$; full information.
- $k = 5$, $n = 1000$.

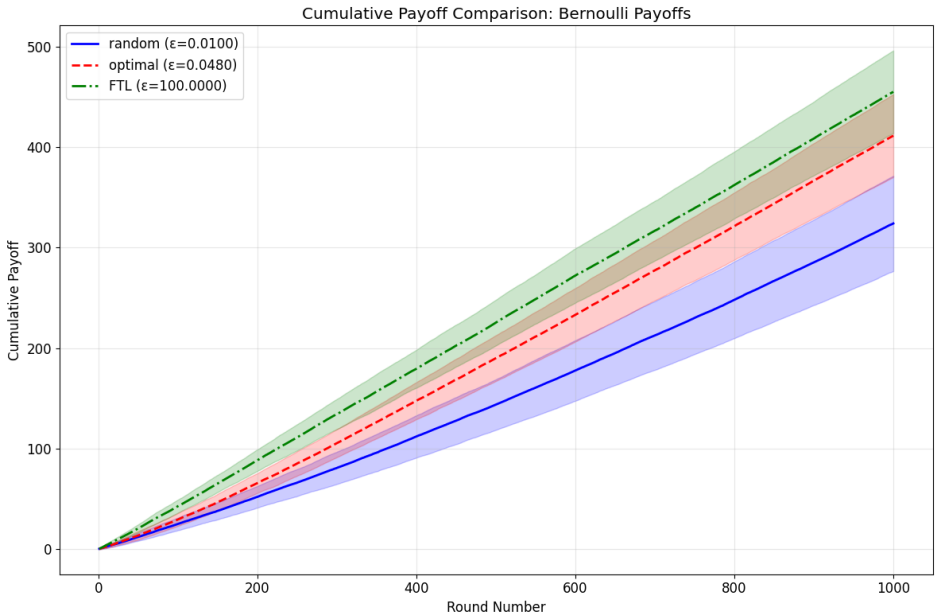
Intuition

- Stationary environment with a best arm $\Rightarrow$ FTL quickly locks in and achieves low regret.
- Random / theoretical ϵ explore more and lag.



Part 1B: Results





Note: Same theoretical ceiling applies; empirically FTL is best here.

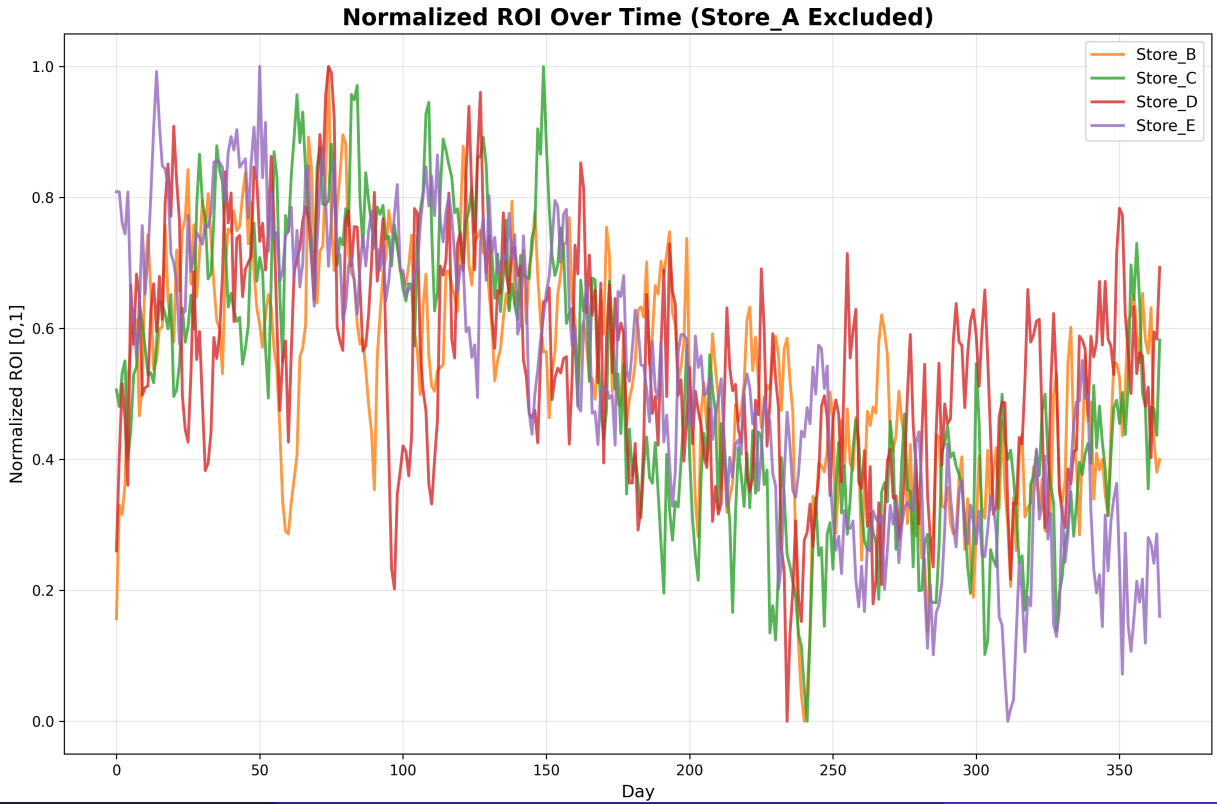
Part 2C (Pachinko): Methods, Data & Intuition

Methods

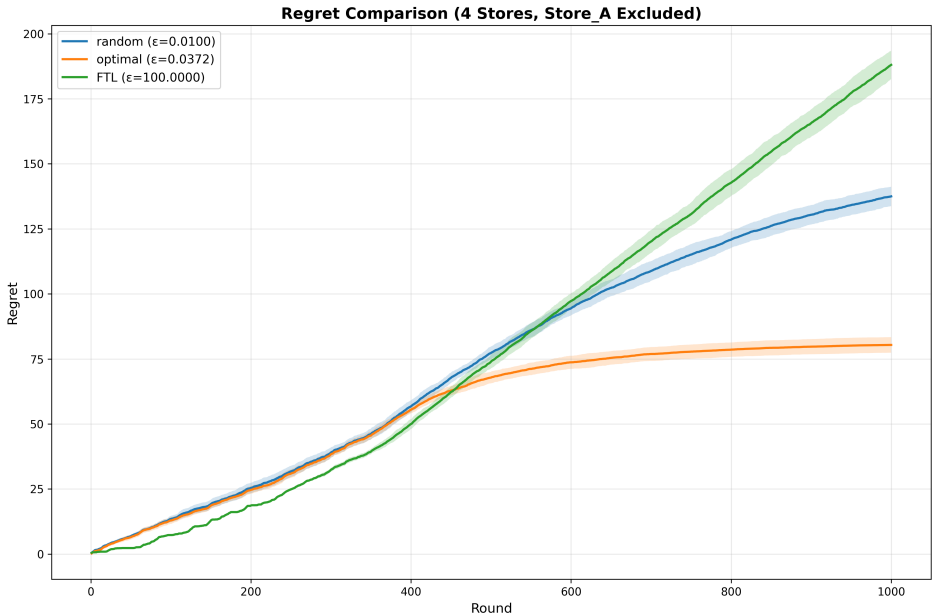
- Treat each store as an arm; daily ROI normalized to $[0, 1]$; full information.
- Four Tokyo stores; nonstationary ROI due to store setting changes.

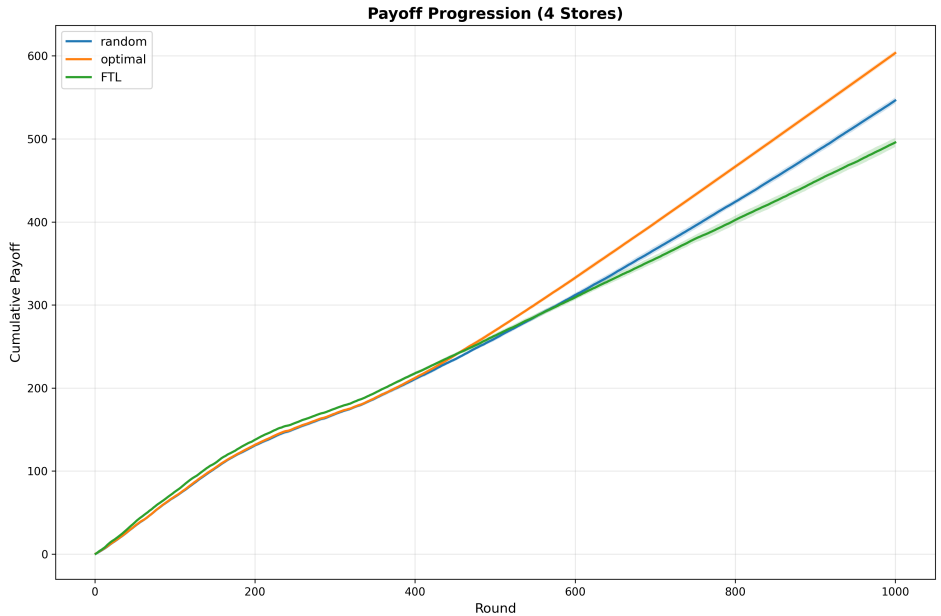
Intuition

- FTL (follow past best store) works if ROI drifts slowly.



Part 2C: Results





D - Setting

Here is our model in Part D. I model the environment in which researcher search the previous research papers for his work.
In each round i,

- There is hidden(unobserved) regime
- The researcher observes the clusters and the origin of the clusters of candidate papers
- The researcher chooses an **effort allocation vector**

$$w_t = (w_{t,1}, \ldots, w_{t,k}), \quad \text{such that } \sum_{j=1}^{k} w_{t,j} = 1,\ w_{t,j} \geq 0.$$

- The payoff is computed as a linear combination:

$$U_t = \alpha_t^{\mathsf{T}} w_t.$$

D - Techniques

In academic world, there are two often-said characteristics

- **High-quality papers are mass-produced by a cluster of researchers**
  - we can formulate this by introducing correlation of paper's value in a cluster
  - which disturbs uniform guessing to be optimal and to converge to the no-regret
- **Frequent innovations (regime changes) in research methods**
  - we can formulate this by introducing possibility of regime changes
  - which disturbs FTL to be optimal and to converge to the no-regret

Our technique leverages these well-known properties to formulate online learning in a way that suits the conditions of our problem.

D - Game structure and Intuition

Here's our intuition:

- FTL and uniform guessing will both work similarly to typical behaviors of researchers in the real world who choose the leading university (FTL) or collect from any/all universities (uniform random guessing).
- Between FTL and random guessing, maybe there is an optimal learning rate

D - Results (Regrets)

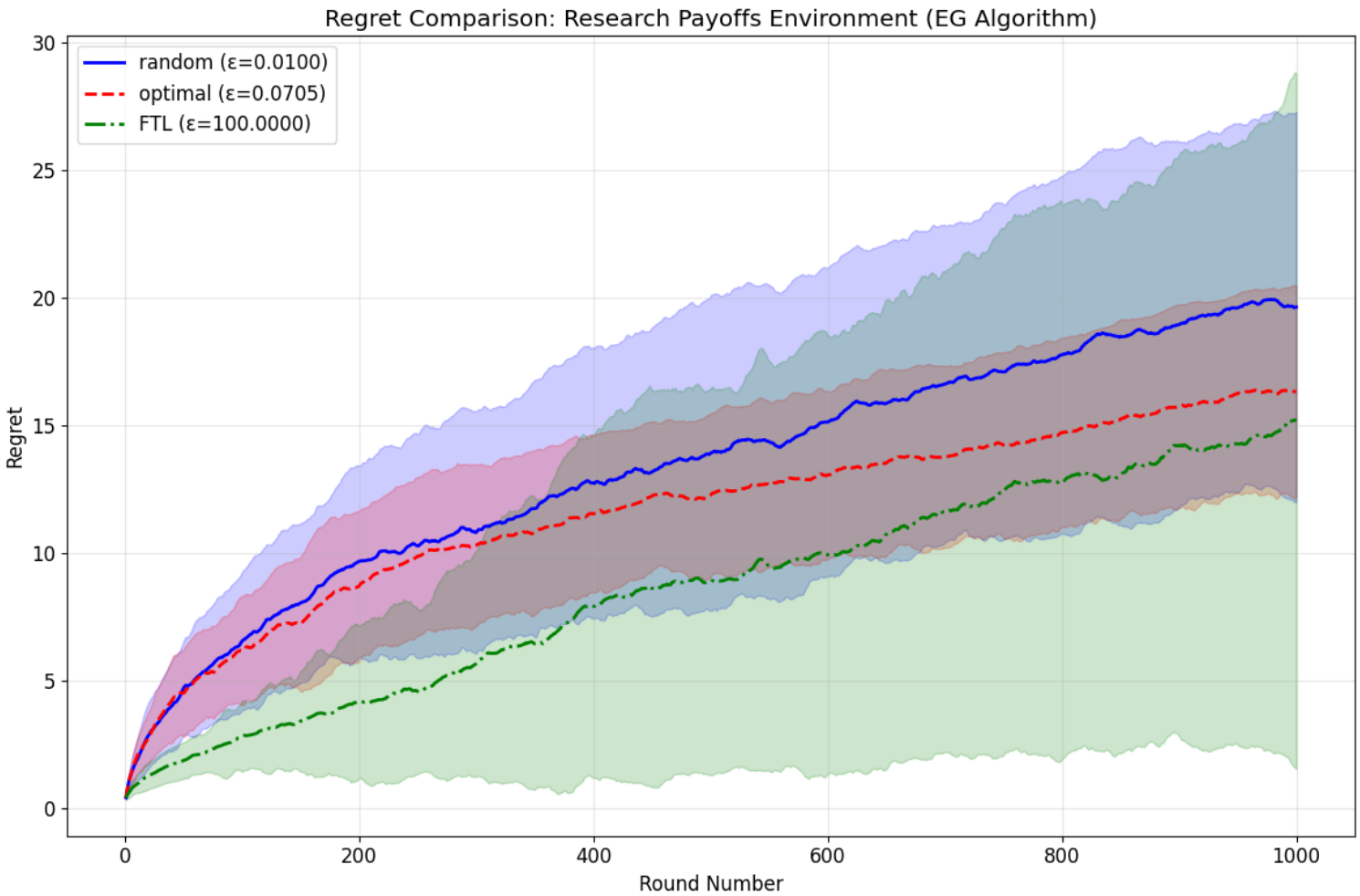


Figure: Regret

D - Results (Payoffs)

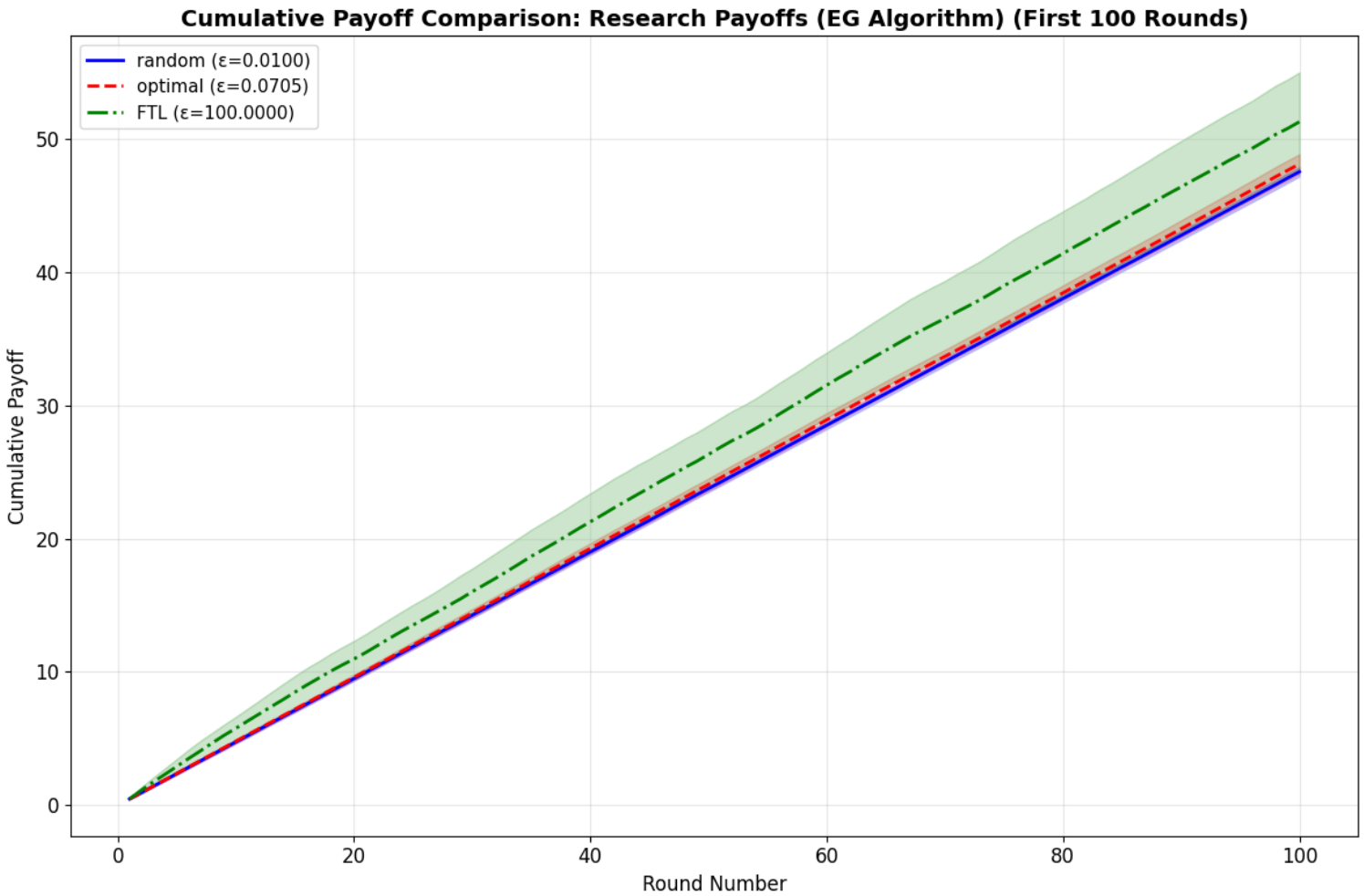


Figure: Payoff

Reference and Usage of AI

- **Pachinko store data:** Daily store-level "balls in/out" statistics obtained from *SloRepo* (`https://slorepo.com`). Data accessed on October 22, 2025 (America/Chicago).
- AI assistance was used for coding and figure generation; final verification, interpretation, and responsibility rest with the authors.

D - Appendix(Model)

- The research world is divided into several **clusters of researchers** (e.g., MIT, Stanford, Northwestern).
- Each cluster represents a correlated set of ideas or papers whose values move together (**intra-cluster correlation**).
- At every round, the researcher distributes total effort across candidate papers:

$$\sum_{i=1}^{N} w_{i,t} = 1, \quad w_{i,t} \geq 0.$$

- The payoff is the weighted sum of paper values:

$$U_t = \alpha_t^{\mathsf{T}} w_t.$$

D - AppendixModel

Regime Dynamics

- At any time, one cluster becomes **hegemony** (high valuation).
- The dominant cluster changes according to a **Markov transition** with probability p:
  for example, p = 0.7,

$$\Pr(z_t = z_{t-1}) = 0.7, \quad \Pr(z_t \neq z_{t-1}) = 0.3.$$

- In each regime:
  - One cluster has **High values** ($[0.7, 1.0]$)
  - Another has **Middle values** ($[0.5, 0.8]$)
  - Others remain **Low values** ($[0.0, 0.4]$)

D - Appendix (Example)

Example Setting

- Imagine three major research clusters: **MIT**, **Stanford**, and **Northwestern**.
- Each cluster represents a group of researchers whose papers are highly correlated in value (citations, attention, or impact).
- The researcher (our agent) allocates research effort across papers from these clusters every round.

Regime-Dependent Dominance

- At one period, the **MIT cluster** becomes dominant — its papers are cited frequently, gaining high valuation.
- Over time, attention shifts: Stanford's ideas rise in popularity, then Northwestern takes the lead.
- These shifts occur through a stochastic **Markov regime transition**, creating a dynamic environment.

Researcher's Behavior

- In each round, the researcher chooses how to distribute effort:

$$w_t = (w_{\text{MIT}}, w_{\text{Stanford}}, w_{\text{Northwestern}}), \quad \sum w_t = 1.$$

- The payoff is the weighted performance of papers:

$$U_t = \alpha_t^{\mathsf{T}} w_t.$$

- As the dominant cluster changes, the researcher must continuously reallocate effort to follow the new trend.

D- Appendix (Algorithm in the real world)

1. **Follow-The-Leader (FTL)**
   - **Intuitive meaning:**
     - "MIT is the most famous and successful group — just follow their papers."
     - Represents a researcher who always trusts the cluster that has performed best in the past.
   - **Problem:** When regimes shift (e.g., attention moves from MIT to Stanford), FTL reacts too slowly and suffers large regret.
2. **Uniform Guessing**
   - **Intuitive meaning:**
     - "Every cluster might have something interesting — I'll pick papers at random."
     - Represents an exploratory but non-learning researcher.
   - **Problem:** Ignores structure and fails to exploit high-performing clusters.

Summary about Algorithms

- Both FTL and Uniform Guessing resemble intuitive human search behaviors in academia.
- FTL overfits to the past (conservative imitation), while Uniform Guessing underfits (pure exploration).

D - Appendix(Compare EG and EW)

Extension

- The **Exponential Weights (EW)** algorithm updates a probability vector over discrete actions:

$$w_{t+1,i} = \frac{w_{t,i} \exp(\varepsilon \alpha_{t,i})}{\sum_{j=1}^{k} w_{t,j} \exp(\varepsilon \alpha_{t,j})}.$$

- EW assumes a **finite action set** and interprets weights as probabilities.

Exponentiated Gradient (EG)

- EG generalizes EW to an N-**dimensional continuous decision space**.
- The parameter vector $w_t \in \Delta^N$ represents a continuous allocation (effort, attention, or resource):

$$w_{t+1} = \frac{w_t \odot \exp(\varepsilon \nabla_t)}{\| w_t \odot \exp(\varepsilon \nabla_t) \|_1}, \quad \nabla_t = \alpha_t = \frac{\partial U(w_t)}{\partial w_t}.$$