Impact of Heavy-Tailed Rewards on Exploration Strategies in Insurance Underwriting

Johann Bartels, Karsten Bruns

1 Motivation

Insurance decision-making involves long-term consequences and uncertainty, making it ideal for RL. Real-world insurance rewards exhibit *heavy-tailed distributions* (rare, extreme losses) that may fundamentally alter optimal exploration strategies. This project examines how these domain-specific characteristics impact exploration efficiency in Q-learning agents, using a fixed delayed reward structure ($k = 10$) as context to isolate the heavy-tailed effect.

2 Refined Research Question

In a simulated insurance-underwriting task with heavy-tailed rewards and fixed delayed returns, does ez-greedy exploration yield higher discounted return and safer convergence than fixed ϵ-greedy?

3 Key Changes from Feedback

- Exclusive focus on heavy-tailed reward distributions with fixed delay ($k = 10$)
- Robust Pareto sampling: $x = \max(x_m, x_m/U^{1/\alpha})$
- Clear operationalization of exploration reduction threshold
- Optimized experimental scope with single-seed backup
- Defined analysis methods for insurance-specific insights

4 Experiments

Environment Design

Custom Gym environment with:

- **States**: 10 discrete customer profiles (3 age groups × 4 risk categories)
- **Actions**: Accept/reject decisions
- **Rewards**:
  - Fixed delay ($k = 10$ timesteps)
  - Heavy-tailed Pareto distribution: $p(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$ with $x_m = 1.0$
  - **Robust sampling**: $x = \max(1.0, 1.0/U^{1/\alpha})$

Compared Strategies

1. **Fixed ϵ-greedy** ($\epsilon = 0.3$)
2. **ez-greedy**:
   - **Reduces ϵ to 0.05 when**: $\frac{1}{|S|} \sum_s |Q_t(s) - Q_{t-100}(s)| < 0.01$
   - **Permanent reduction** (no cycling)

Experimental Settings

- **Core setting**: Extreme heavy-tailedness ($\alpha = 1.5$) with 5 seeds
- **Backup**: Moderate heavy-tailedness ($\alpha = 2.0$) with **1 seed (sensitivity check)**
- Fixed parameters: $\gamma = 0.99$, $\epsilon_0 = 0.3$, learning rate=0.1

Metrics & Analysis Methods

Primary metrics:

- Discounted return (last 100 episodes)
- **Episodes to threshold**: First episode where $\Delta Q < 0.01$ sustained for 100 episodes

Analysis methods:

- **Q-Q plots**: Visual verification of reward distributions
- **Boxplots**: Episode-to-threshold distribution across seeds
- **Welch's t-test**: Statistical comparison of final returns

Scope Limitations

- **Compute**: 10 core runs + 2 backup runs
- **Parallel execution**: Joblib with 4 cores ($\sim$5 hours worst-case)
- Tabular Q-learning with sensible defaults
- **Checkpointing**: Automatic experiment recovery

Hypothesis

We hypothesize that **ez-greedy** will:

- Achieve $\sim$18% higher discounted return in $\alpha = 1.5$ setting
- **Reduce exploration 30% earlier** while maintaining policy quality
- Show strongest advantage during extreme loss events

5 Optimized Timeline (11 Days)

- **Day 1-2**: Environment finalization (robust Pareto sampling, delay buffer)
- **Day 3**: Implement strategies with **ez-greedy algorithm**
- **Day 4**: **Smoke tests** (1 seed per strategy + extremeness check)
- **Day 5-6**: Core experiments (parallel execution)
- **Day 7-8**: Analysis (Q-Q plots, boxplots, statistical tests)
- **Day 9-10**: Report/poster with **domain focus**
- **Day 11**: Buffer and submission

Implementation Details

Report Structure Plan

Algorithm 1 ez-greedy Exploration Management

Initialize $\epsilon \leftarrow 0.3$, $Q_{prev} \leftarrow \emptyset$
for $episode = 1$ to 2000 **do**
  if $episode100 == 0$ **then**
    $\Delta \leftarrow \frac{1}{|S|} \sum_s |Q(s) - Q_{prev}(s)|$
    if $\Delta < 0.01$ **and** $\epsilon > 0.05$ **then**
      $\epsilon \leftarrow 0.05$ ▷ Permanent reduction
    end if
    $Q_{prev} \leftarrow Q$ ▷ Store current Q-values
  end if
  Execute standard Q-learning steps
end for

Table 1: Page budget allocation

Section	**Pages**
Introduction & Motivation	0.5
Related Work	0.5
Environment & Methods	1.5
Experiments & Results	1.5
Discussion & Outlook	0.5

Poster Focus

- **Central visualization**: Combined return curve + convergence boxplot
- **Insurance-specific insights**: Heavy-tail effects on exploration
- **Ez-greedy advantage**: Early exploration reduction mechanism