How Exploration Affects Long-Term Profitability on Insurance Contracts

Johann Bartels, Karsten Bruns

1 Motivation

In the insurance domain, decision-making often involves uncertainty and long-term consequences, ideal conditions for reinforcement learning (RL). A key challenge in RL is balancing exploration with exploitation. This project investigates whether adaptive exploration yields better results than fixed exploration in an insurance-like environment. This question is both practically relevant and scientifically testable.

2 Related Topics

This project directly relates to:

- Model-free control using ϵ-greedy exploration strategies
- Off-policy TD methods, especially Q-Learning
- Exploration-exploitation trade-offs in RL

It builds on RL concepts discussed in the course and applies them in a simplified real-world simulation context.

3 Idea

We propose to simulate an insurance environment where a reinforcement learning agent must decide whether to accept or reject customer contracts based on simplified customer features. The agent receives rewards based on long-term profitability outcomes (for example, accepting low-risk customers yields positive return). The research question is:

Does an ϵ-decay exploration strategy lead to better performance (measured by average episode return) compared to a fixed ϵ policy?

To answer this, we will implement two Q-learning agents. One with constant ϵ and one with a decaying ϵ and compare their results over multiple runs.

4 Experiments

Environments & Metrics

We will implement a custom Gym-style environment simulating an abstracted insurance decision process. States represent customer profiles; actions are accept/reject decisions. Rewards are based on simulated profit/loss per contract. Key metrics: average return, convergence rate, and policy quality over time.

Experimental Scope

We will run both fixed-ϵ and decaying-ϵ agents with:

- 5 seeds per configuration
- moderate grid search on learning rate and discount factor
- learning curves and reward distributions

Estimated Computational Load

The training episodes are lightweight and can be executed on a standard laptop. Each run is expected to take less than 10 minutes, resulting in a total of roughly 5–10 compute hours.

Expected Outcome

We hypothesize that the agent using a **decaying ϵ-greedy exploration strategy** will achieve a higher average reward per episode and more stable learning than an agent using a fixed ϵ.

The intuition behind this is that decaying ϵ enables sufficient exploration early in training, followed by stronger exploitation of learned policies later on. In contrast, fixed ϵ continues to introduce randomness even when the agent has learned a good policy.

Confidence in Hypothesis

We are moderately confident (roughly 70–80%) that the decaying-ϵ agent will outperform the fixed-ϵ baseline. However, since the effect size might depend on the exact parameter values (for example, decay rate, learning rate), some configurations might still show comparable performance.

5 Timeline

- **Week 1 (until July 12)**: Research and implementation of the environment
- **Week 2 (until July 19)**: Implementation of both exploration strategies
- **Week 3 (until July 26)**: Run experiments and visualize results
- **Week 4 (until August 02)**: Analysis and draft report/poster
- **Buffer (August)**: Adjustments, final report, and poster polish