Reinforcement Learning for Loan Pricing

Anuj Panwar

May 7, 2025

https://github.com/anujpanwarma2024/RLMidTermProject

Abstract

This report investigates RL for pricing loans using the Black-Scholes-Merton (BSM) model as a benchmark.

RL agents learn policies from borrower data to predict prices by minimizing pricing error.

Three algorithms are compared:

- REINFORCE
- REINFORCE with Baseline
- TD(0)

We evaluate them based on reward, convergence, and pricing accuracy using synthetic loan data.

Introduction

Traditional models like discounted cash flows are static and fail to reflect real-time borrower dynamics.

RL offers:

- Dynamic adaptation to borrower features
- Feedback-driven improvement over time

Our Setup:

- Environment: loan features
- Action: predicted price multiplier
- Reward: proximity to BSM benchmark

Problem Formulation

- **State:** $s \in \mathbb{R}^5$ (Loan Age, FICO, Term, Interest Rate, DTI)
- **Action:** $a \in [0, 1]$, drawn from $\pi(a|s) = \mathcal{N}(\theta^\top \phi(s), \sigma^2)$
- **Reward:** $r = -|a \cdot \text{LoanAmount} - \text{BSMPrice}|$

Goal: minimize expected deviation from BSM prices through policy learning.

Methods

REINFORCE:

$$\theta \leftarrow \theta + \alpha G_t \nabla_\theta \log \pi(a_t|s_t)$$

- Uses episode return
- High variance

REINFORCE with Baseline:

$$\theta \leftarrow \theta + \alpha(G_t - b_t)\nabla_\theta \log \pi(a_t|s_t), \quad b_{t+1} = (1-\beta)b_t + \beta G_t$$

- Reduces gradient variance
- Convergence can suffer if baseline is misestimated

TD(0):

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t), \quad \theta \leftarrow \theta + \alpha \cdot \delta_t \nabla_\theta \log \pi(a_t|s_t)$$

- Online updates

Experimentation

Environment:

- 200 episodes, 4 steps each
- Preprocessed synthetic loan data

Hyperparameters:

- $\alpha = 0.005$, $\gamma = 0.95$
- $\sigma = 0.2$, $\beta = 0.1$

Stabilization:

- Reward clipping to $[-10^5, 10^5]$
- Gradient clipping in $[-10, 10]$
- Actions clipped to $[0, 1]$

Results and Analysis

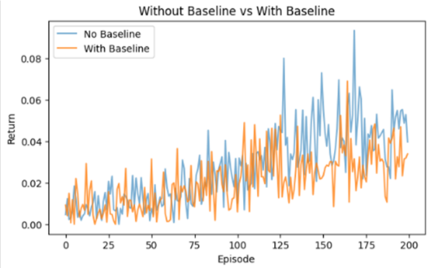
Mean Returns:

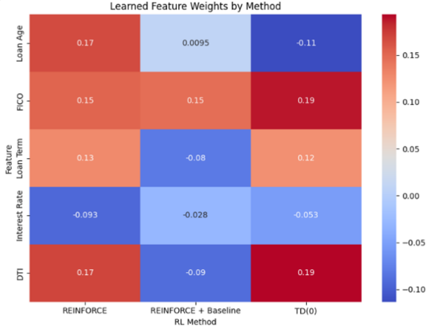
- REINFORCE: $-85{,}667$
- REINFORCE+Baseline: $-83{,}826$
- TD(0): $-\mathbf{75{,}579}$

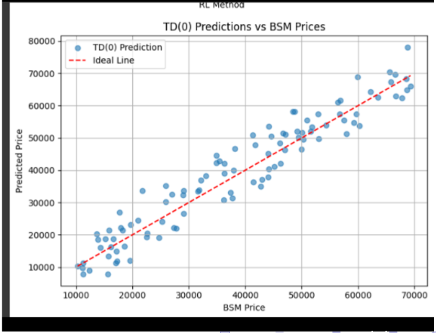
Observations:

- TD(0) showed stable convergence and better reward optimization
- Baseline method struggled with underfitting

Visual Insights







Conclusion and Limitations

Conclusion:

- TD(0) performed best due to lower variance and online learning
- RL successfully approximated BSM-based pricing

Limitations:

- Trained on synthetic BSM-based data
- Scalar output restricts pricing complexity

Future Work:

- Expand to multi-action price components
- Apply to real-world datasets
- Include risk-adjusted reward structures

References

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. `https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf`
- Black, F., & Scholes, M. (1973). *The Pricing of Options and Corporate Liabilities*. Journal of Political Economy. `https://www.jstor.org/stable/1831029`
- Williams, R. J. (1992). *Simple statistical gradient-following algorithms for connectionist reinforcement learning*. Machine Learning.
- Moody, J., Wu, L., Liao, Y., & Saffell, M. (1998). *Performance functions and reinforcement learning for trading systems and portfolios*. `https://www.researchgate.net/publication/221007119`
- RL for Finance Lecture Notes, Carnegie Mellon University. (2022)