RL Project Report: Reward Shaping Analysis

Submitted by: Alexandra Belkind 323419416 Daniel Chernov 208169805

0.Quick Start Guide

Clone the repository:

```
git clone https://github.com/DanielChernov99/RL-FrozenLake.git
cd RL-Frozen Lake
```

Install dependencies:

```
pip install -r requirements.txt
```

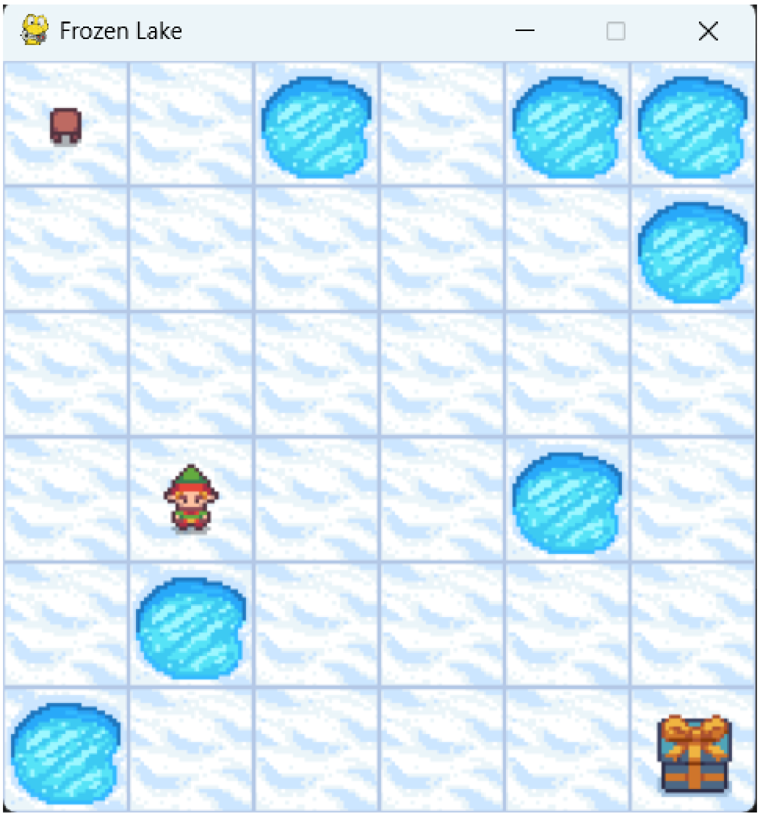
(Or py -m pip install -r requirements.txt if pip is not recognized)

Run the experiments:

```
py main.py
```

1.The Environment

- **Map Size:** Custom 6x6 grid (36 tiles).
- **Hole Density:** We placed 7 holes, resulting in a density of 19%.
- **Start/Goal:** Fixed Start (S) at (0,0) and Goal (G) at (5,5).
- **Dynamics:** The environment is slippery (is_slippery=True). However, we modified the transition probabilities to set a **Success Rate of 0.7** (probability of moving in the intended direction), with the remaining 0.3 split between orthogonal directions.



2.Reward Shaping Strategies

A. Baseline (No Shaping)

Formula: F(s,a,s') = 0

Intuition: Serves as the control group. The reward is sparse (plus 1 only at the goal), making early learning difficult as the agent relies solely on random exploration up until he finds the goal.

B. Step-Cost Shaping

Formula: F(s,a,s') = -0.001

Intuition: Adds a pressure to optimize efficiency. It penalizes loitering and encourages the agent to find the shortest path to the goal to minimize accumulated negative rewards.

C. Potential-Based Shaping

Formula: F(s,a,s') = $\gamma\Phi(s') - \Phi(s)$

Potential Function: $\Phi(s)$ = -ManhattanDistance(s, Goal)

Intuition: Acts as a compass. It provides dense feedback, rewarding the agent for reducing the distance to the goal.

D. Custom Strategy 1: "Safety First"

Formula: F(s,a,s') = -1.0 if s' is a hole, otherwise 0

Intuition: In a stochastic environment like FrozenLake, the primary risk is falling into holes. By assigning a distinct, heavy penalty for death, we aim to teach the agent risk aversion very early in the training process, effectively guiding exploration away from dangerous areas.

E. Custom Strategy 2: "Advanced Guidance"

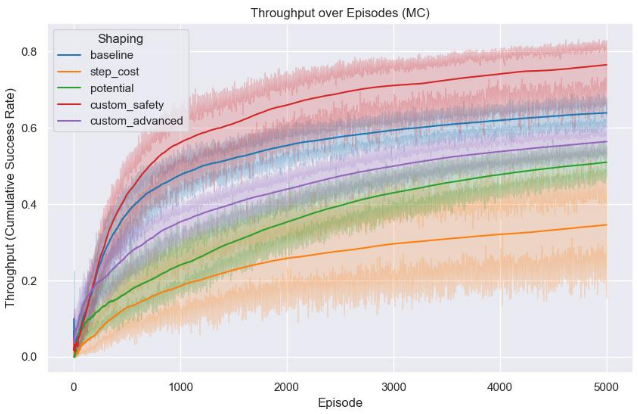
Formula: F = 1.0 * ($\gamma\Phi(s') - \Phi(s)$) + (-0.001) + HolePenalty(-0.5)

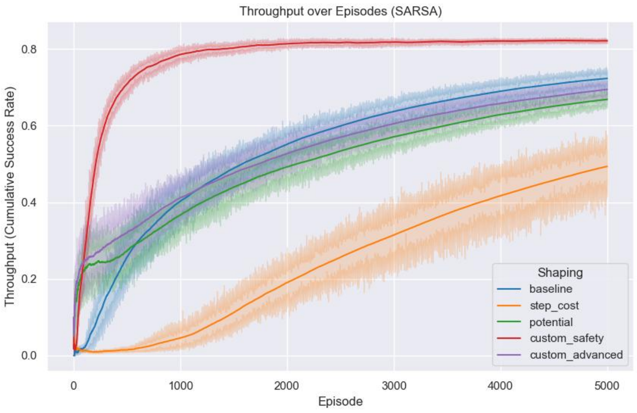
Intuition: This strategy aims to provide the richest signal possible by balancing three objectives: moving towards the goal, doing so quickly, and staying alive.

3. Results & Analysis

We compared the performance of SARSA and Monte Carlo (MC) agents across all shaping strategies over 5,000 episodes, averaged over 20 runs.

Throughput Analysis





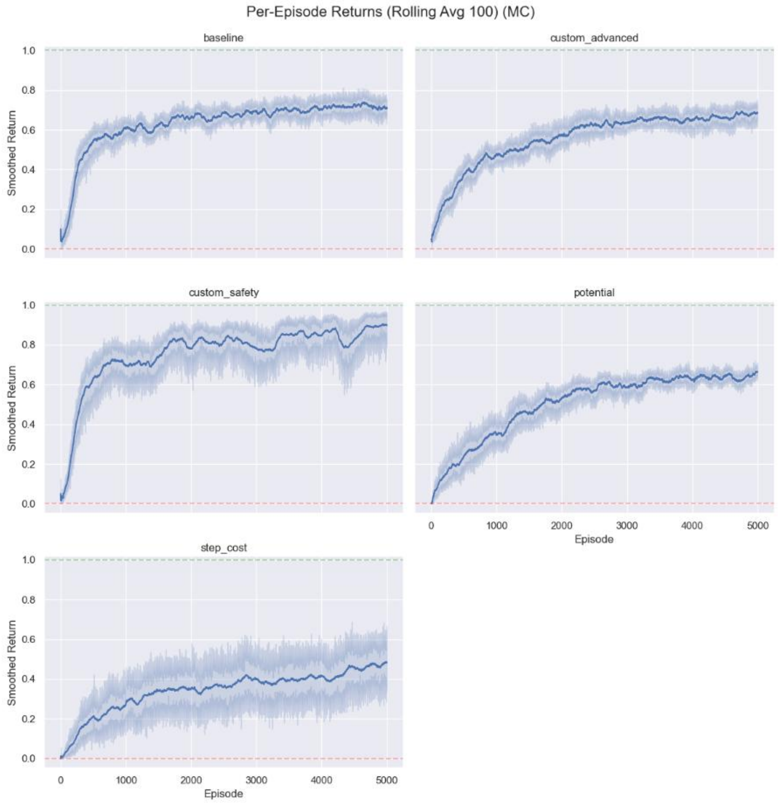
Early Learning: In both Monte Carlo and SARSA, the custom “Safety First” shaping strategy (red line) exhibits the fastest improvement during the early episodes. This effect is especially pronounced in SARSA, where the throughput rises sharply within the first few hundred episodes and quickly reaches a high success rate. The strong negative penalty for falling into holes provides a clear and immediate learning signal, encouraging safer trajectories early on. The “Advanced” strategy (purple line) also improves early learning compared to the Baseline, but its effect is more moderate and consistent rather than explosive, particularly in the MC setting.

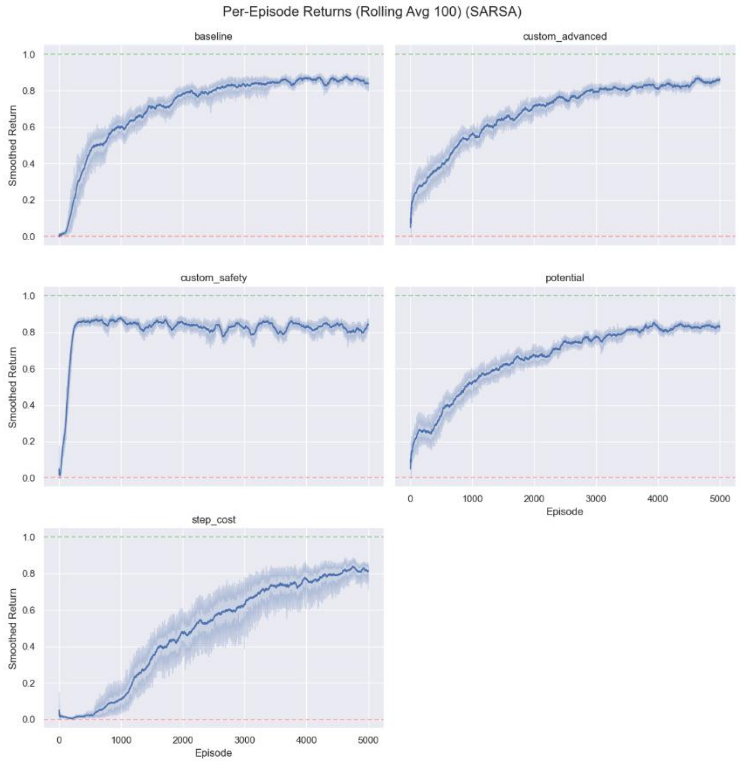
SARSA vs. MC: Across all shaping strategies, SARSA demonstrates faster and smoother learning than Monte Carlo. This is consistent with the theoretical difference between the algorithms: SARSA performs on-policy updates at every step using bootstrapping, allowing it to incorporate shaping rewards immediately and adjust its behaviour online. In contrast, Monte Carlo updates only at the end of each episode, which slows down the propagation of both rewards and penalties. As a result, shaping strategies especially those with strong immediate signals such as “Safety First” are significantly more effective under SARSA than under MC.

Per-Episode Returns

Convergence: Both Monte Carlo and SARSA converge to relatively high average returns, indicating that none of the shaping strategies prevent learning a near-optimal policy. However, SARSA converges more consistently and generally reaches higher final returns across most shaping methods. In Monte Carlo, convergence is slower and varies more between shaping strategies. In particular, step-cost shaping converges to a noticeably lower return compared to the other methods, while custom safety achieves the highest final performance.

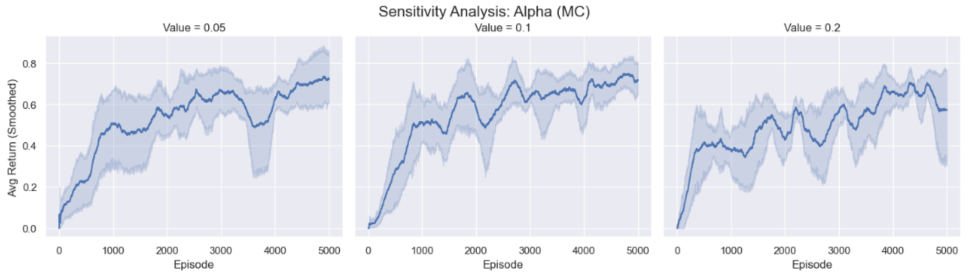
Stability: Shaping strategies generally reach a stable performance plateau earlier than the Baseline, though the effect differs between methods. The “Safety First” strategy stabilizes the fastest in both algorithms, with an especially sharp and early stabilization in SARSA. Potential-based and advanced shaping show smoother but slower stabilization. In Monte Carlo, stability is weaker overall, with higher variance persisting even after convergence. Overall, reward shaping improves stability and convergence speed more clearly in SARSA than in Monte Carlo.

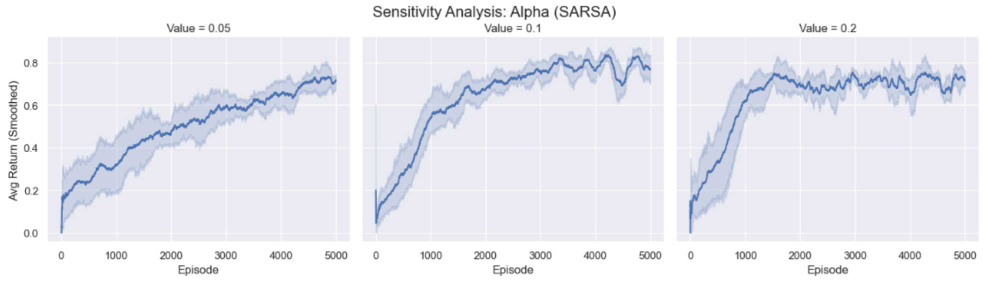




Sensitivity Analysis – Learning Rate (Alpha)

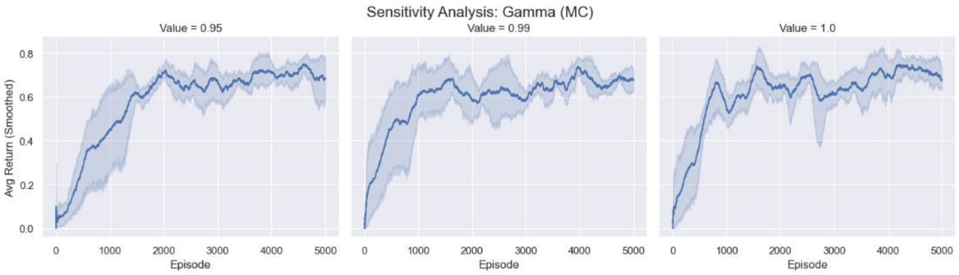
Learning Rate (Alpha) We examined the impact of the learning rate (alpha) on the agent's stability and convergence speed. As shown in the sensitivity plots, a low learning rate (alpha=0.05) resulted in a smooth but excessively slow learning curve, failing to reach optimal policy within the early episodes. Conversely, a higher learning rate (alpha=0.2) allowed for rapid initial learning but introduced significant variance (instability), causing the agent to overreact to the environment's stochastic transitions (slipping). The value alpha=0.1 was identified as the optimal trade-off, providing a stable yet efficient convergence trajectory.

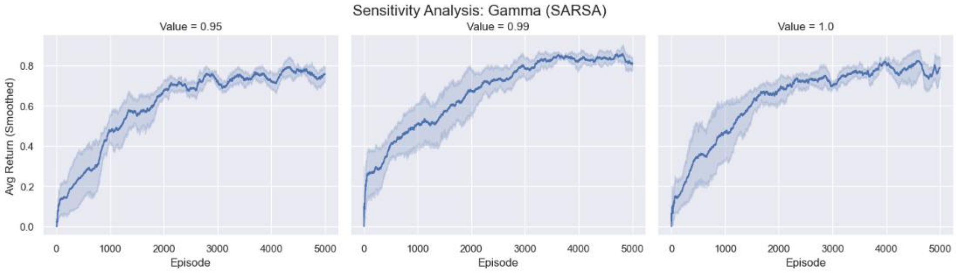




Sensitivity Analysis – Discount Factor (Gamma)

Discount Factor (Gamma) We further analysed the effect of the discount factor (gamma) on the agent's ability to solve the sparse reward problem. In the 6x6 Grid World, the path to the goal is relatively long, and the reward is only granted at the terminal state. Our results indicate that lower discount factors (gamma=0.95) caused the reward signal to decay too rapidly, preventing the agent from effectively valuing states at the beginning of the trajectory. A discount factor of gamma=1.0 (no discounting) yielded the best performance, ensuring that the positive signal from the goal was fully propagated back to the start state, which is crucial for long-horizon navigation tasks.





4. Discussion

Which shaping helped most?

The "Safety First" custom shaping provided the most significant boost in early learning. In grid worlds with holes and stochastic , learning what not to do proved more valuable initially than learning where to go using potential-based guidance.

Potential-Based Shaping Behaviour

While potential-based shaping theoretically preserves the optimal policy, it was often less effective than the explicit safety penalty in practice. In a slippery environment, reducing Manhattan distance is not always optimal if it involves stepping onto tiles with a high risk of falling into holes. The potential function does not account for this risk, whereas the safety shaping directly addresses it.

Exploration vs. Exploitation

A major challenge was balancing exploration with safety. With standard epsilon-greedy exploration, the agent frequently fell into holes, terminating episodes early and limiting its ability to reach the goal and observe positive rewards. The "Safety First" shaping mitigated this issue by providing a clear negative signal whenever the agent fell into a hole. This allowed the value function to learn which states were dangerous even before the goal was discovered, effectively guiding exploration toward safer paths.

Which shaping preserved final optimal behaviour?

Despite their different learning dynamics, all shaping strategies preserved a near-optimal final policy. Empirically, the “Safety First” shaping achieved the strongest final performance, reaching high returns without degrading the learned behaviour. This indicates that, in this environment, introducing an explicit safety penalty did not distort the optimal policy. Potential-based shaping, while not the best performer in practice, remains theoretically guaranteed to preserve optimal behaviour. The hybrid “Advanced Guidance” strategy converged to a similarly optimal policy but did not provide additional improvements in final performance, suggesting that its main benefit lies in guiding learning rather than enhancing the final solution.

Surprising results and failure cases

A surprising result was that the hybrid “Advanced Guidance” strategy, which combines directionality, efficiency, and safety, did not consistently outperform simpler shaping methods. Although it provides a richer learning signal, the interaction between multiple shaping terms may partially conflict. In particular, the efficiency penalty and safety penalty can compete with the potential-based guidance, leading to conservative behaviour that limits further improvement once a safe path is found.

Another notable failure case is the step-cost shaping. By applying a constant negative reward at every step, the agent is encouraged to terminate episodes as quickly as possible. In a stochastic environment with terminal states that yield zero reward, this effect is especially pronounced during the early stages of training, before the agent has discovered the goal. In this phase, falling into a hole can become preferable to prolonged movement, particularly under Monte Carlo learning. As a result, the agent may learn a form of “self-destructive” behaviour, minimizing cumulative penalty by ending the episode early rather than reaching the goal.

5. Summary

This project examined the impact of different reward shaping strategies on SARSA and Monte Carlo agents in a stochastic FrozenLake environment. The results show that reward shaping significantly affects early learning, with the custom “Safety First” strategy providing the strongest improvement by quickly guiding the agent away from holes. This effect was especially pronounced under SARSA.

All shaping strategies preserved near-optimal final behaviour. Potential-based shaping behaved as expected from theory, while the custom shaping methods achieved comparable or better empirical performance without distorting the learned policy. The hybrid “Advanced Guidance” strategy did not improve final performance beyond simpler approaches, indicating that its main benefit lies in accelerating learning rather than enhancing the final solution.

Finally, step-cost shaping proved to be a failure case during early training, particularly under Monte Carlo learning, where the constant negative reward encouraged premature episode termination before the goal was discovered. Overall, the findings highlight the importance of aligning reward shaping with environmental risks and learning dynamics.