Cliff Diving: Exploring Reward Surfaces in Reinforcement Learning Environments

Ryan Sullivan, J. K. Terry, Benjamin Black, John P. Dickerson

What is a “Reward Surface”?

Deep reinforcement learning methods indirectly attempt to optimize the expected cumulative discounted rewards achieved by policy.

$$J(\theta) = E_{\tau \sim \pi_\theta} R(\tau) \text{ where } R(\tau) = \sum_{t=0}^{n} \gamma^t r_t$$

This produces a “reward surface” in the high-dimensional parameter space of the policy network.

Why should we visualize Reward Surfaces?

- Visualizing surfaces has lead to fundamental insights for deep learning.
  - E.g. Li et al. 2018[1] visualized loss landscapes to show that that residual connections reduce the non-convexity of image classification tasks.
- Training RL agents can often be unstable with huge drops in performance.
  - We want to understand the cause of this issue.
  - Study failure modes of reinforcement learning.
- Policy gradient methods estimate the gradient of the reward surface.
  - Visualizing reward surfaces may lead to novel insights about policy gradients.

[1] Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In Neural Information Processing Systems (NeurIPS), 2018

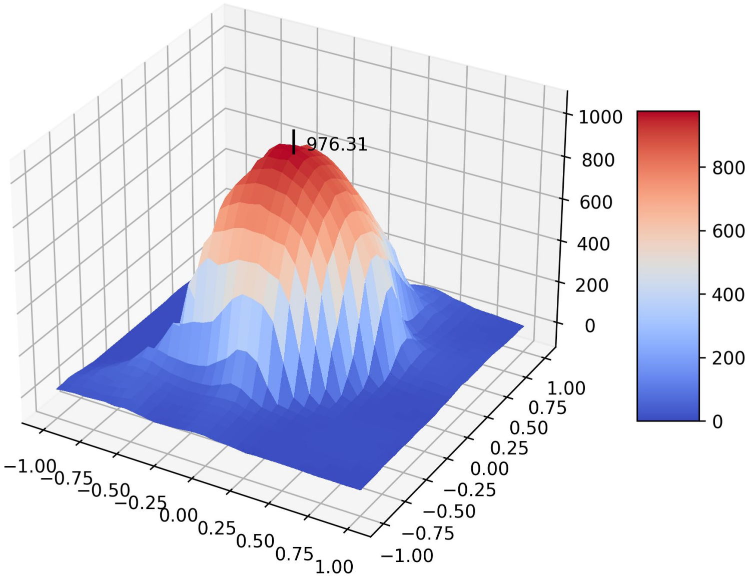
Overview

- Plot reward surfaces for 27 popular environments in OpenAI's Gym.
    - Demonstrate that the visualizations are consistent across multiple seeds.
    - Identified common characteristics across environments in the same set (Atari, Mujoco, etc).

- Discovered "cliffs" in reward surfaces using the gradient direction
    - Identified sudden, sharp decreases in reward in the policy gradient direction of almost every environment.
    - Demonstrated that the cliffs we visualize affect the performance of A2C.

Methodology: Reward Surfaces

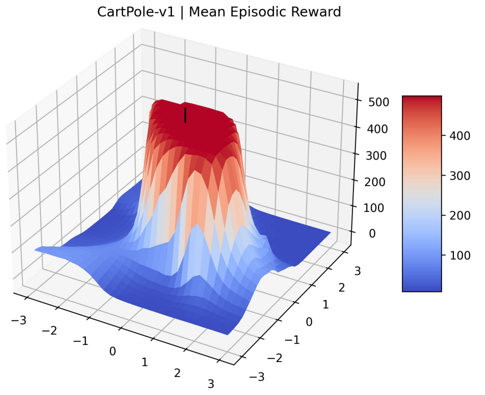
- Choose 2 filter-normalized directions[1] and plot empirical return of a policy network.
- Centered around a policy learned by PPO during training.
- The surface is specific to a particular environment and network architecture, and the center of the plot depends on the learning method and hyperparameters.
- Agents trained using tuned hyperparameters from RL Baselines3 Zoo to compare good regions of the parameter space.

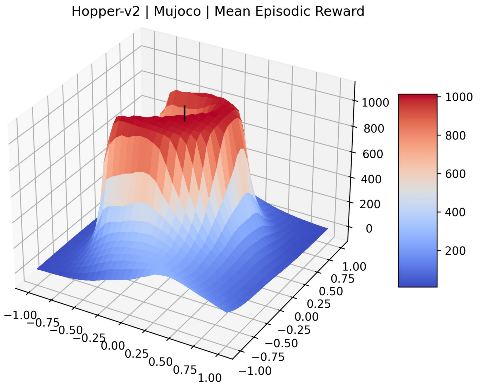


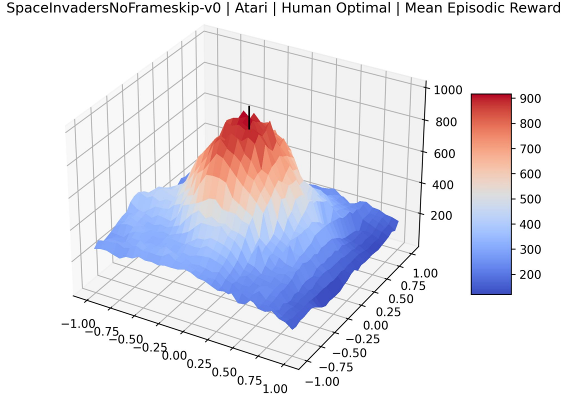
Atari Bank Heist

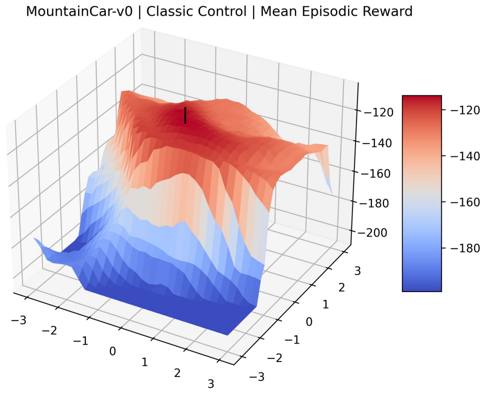
[1] Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In Neural Information Processing Systems (NeurIPS), 2018

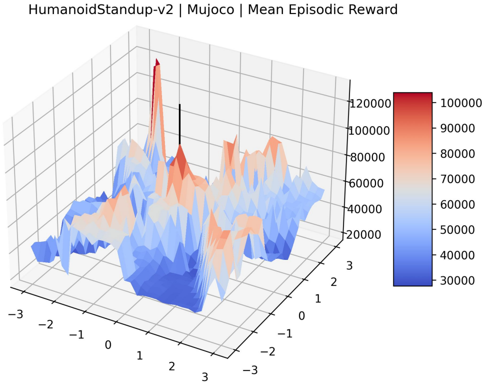
Reward Surface Results

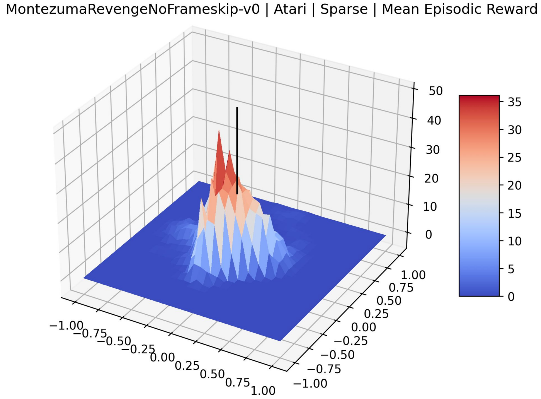




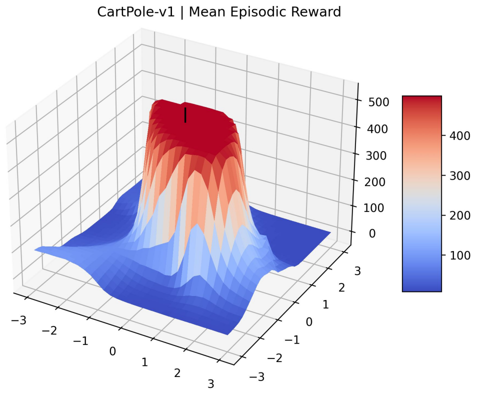


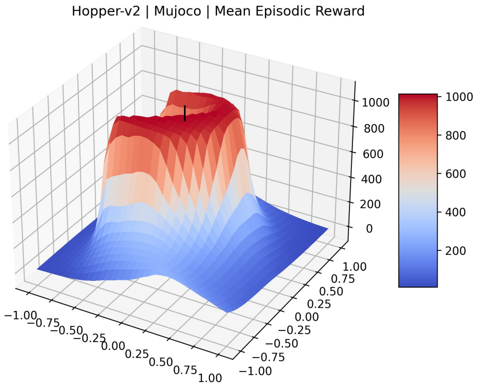


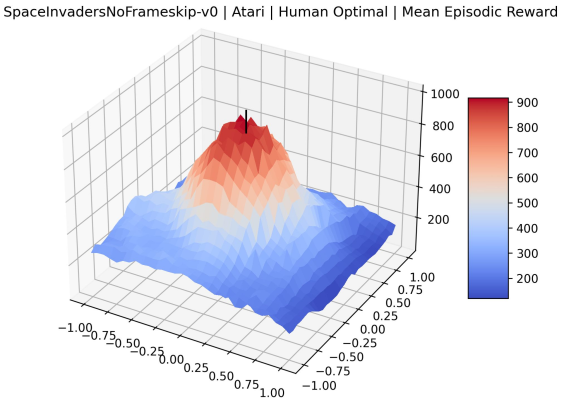


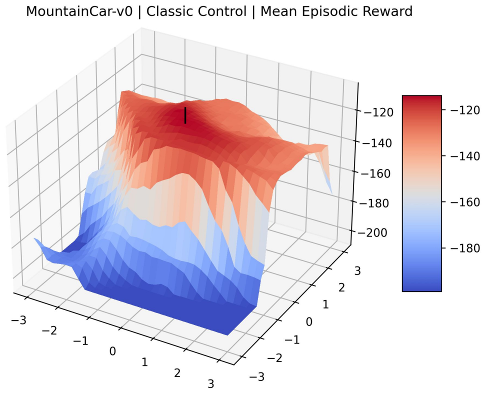


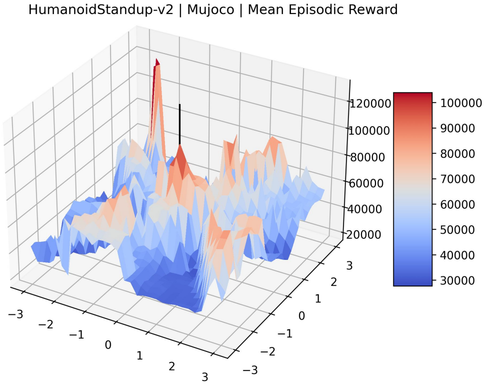
Reward Surface Results

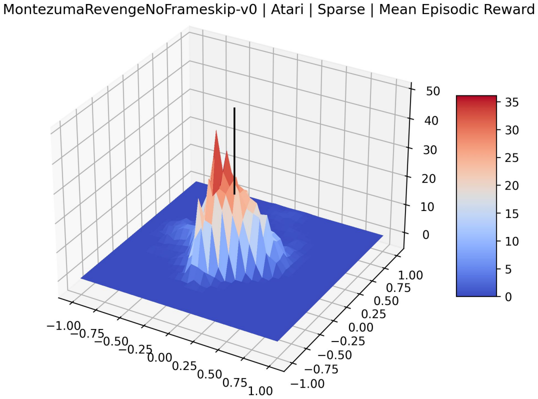






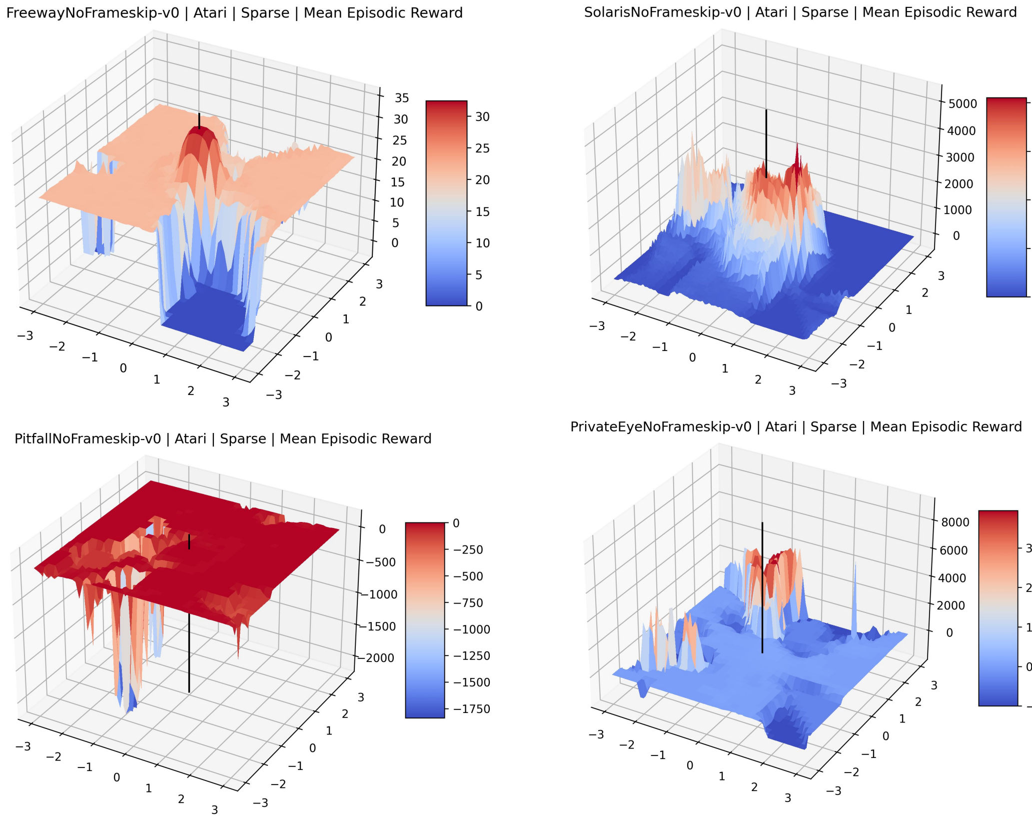






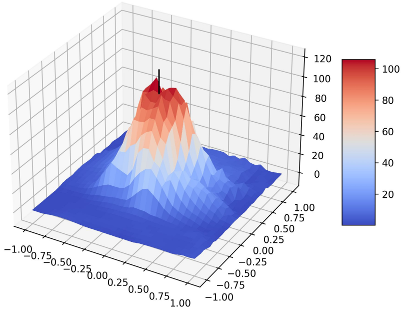
Sparse Rewards

- Sparse reward Atari environments have large flat regions.
- Large policy changes are required to see any variation in rewards.
- Maximizers are spiky even with extremely high sample size.

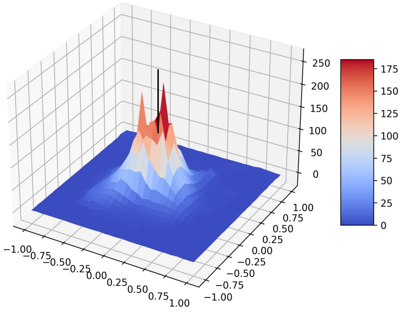


Reproducibility

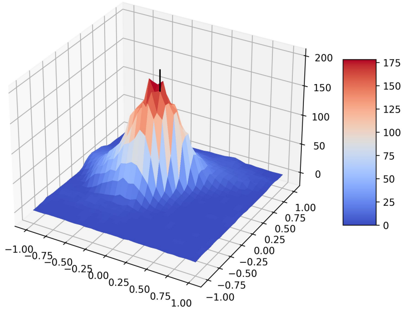
BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



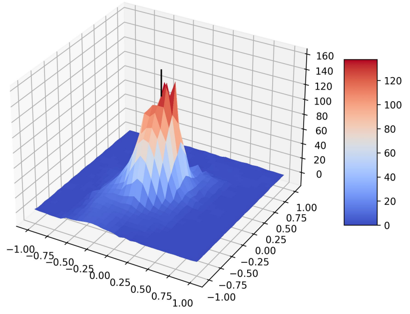
BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



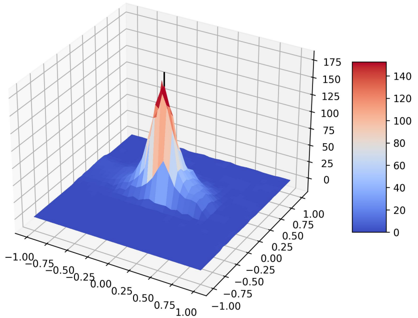
BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



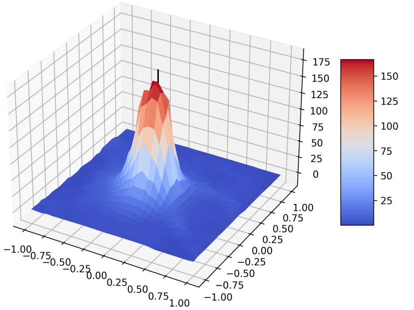
BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



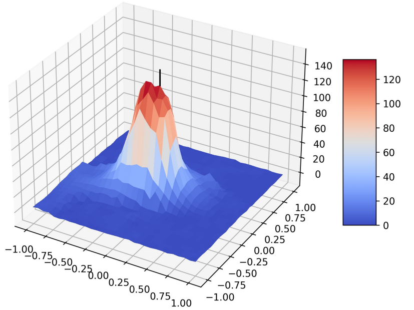
BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



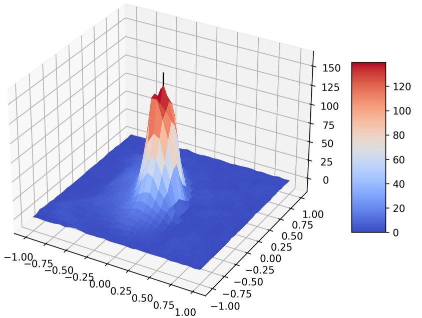
BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward



BreakoutNoFrameskip-v0 | Atari | Human Optimal | Mean Episodic Reward

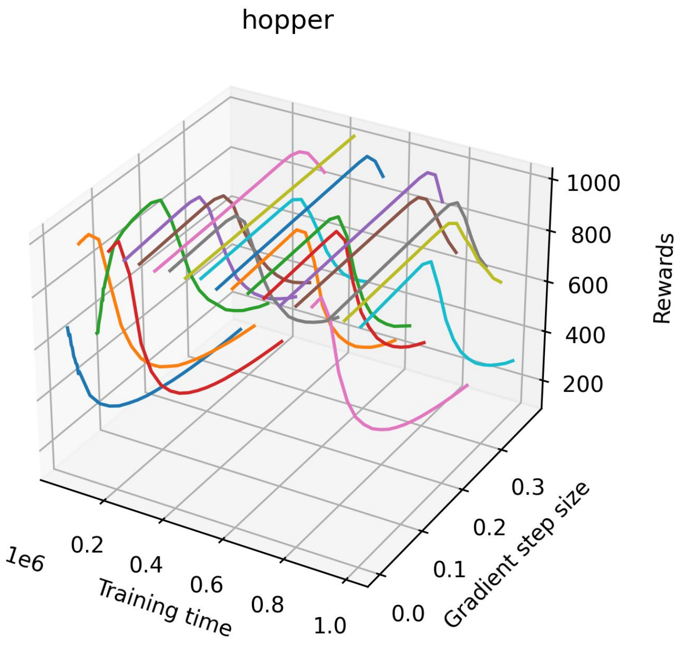


Reward Surfaces using the Gradient Direction

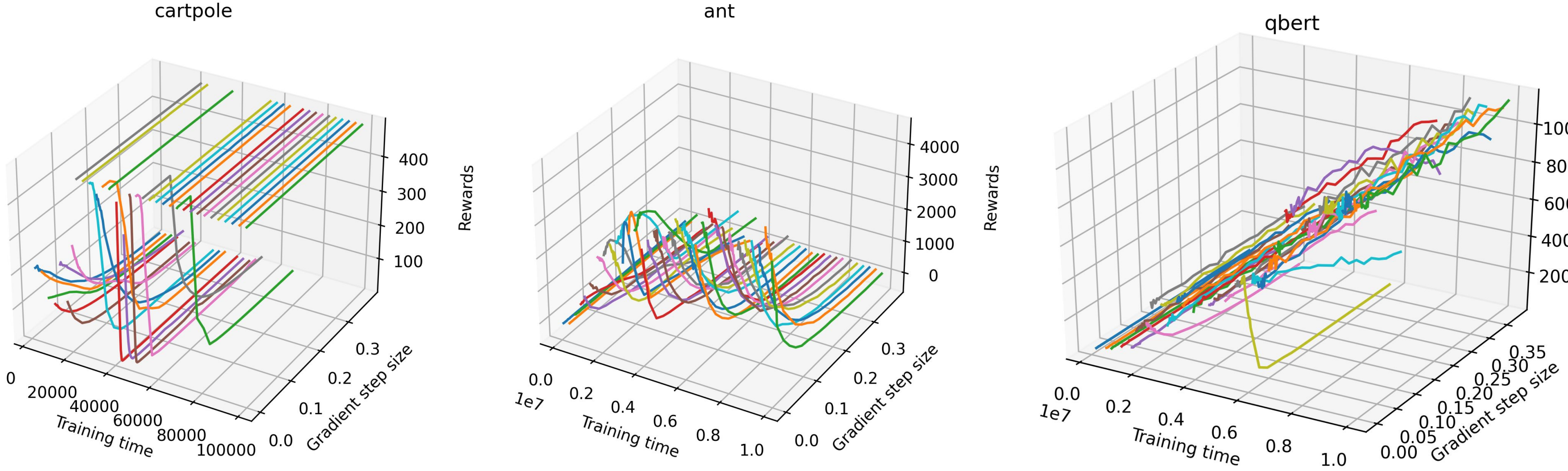
- Line plot of reward surface in a single dimension in the policy gradient direction.

- Individual line for many uniformly distributed checkpoints across training.

- Most environments have at least one if not many “cliffs” - sudden, sharp decreases in reward.



Cliffs in the Gradient Direction



A2C and PPO Cliff Performance

- Table shows percent change in reward after taking a few optimization steps at a checkpoint.

N-steps	Learning Rate	Method	Cliff	Non-Cliff
128	**0.000001**	**A2C**	**-0.3%**	**0.2%**
128	0.000001	PPO	0.03%	0.03%
128	**0.01**	**A2C**	**-0.3%**	**2.0%**
128	0.01	PPO	0.0%	0.04%
2048	**0.000001**	**A2C**	**-0.5%**	**0.2%**
2048	0.000001	PPO	-0.1%	-0.4%
2048	**0.01**	**A2C**	**-3.9%**	**2.9%**
2048	0.01	PPO	0.1%	0.1%

Library

- The library we used to produce these visualizations is available at:

  https://github.com/RyanNavillus/reward-surfaces

- Includes functions to plot 3D reward surfaces and line plots in filter normalized or gradient directions, as well as many other features:
    - Reward surfaces for value functions
    - GIFs of reward surfaces across training
    - Scripts for running experiments on multiple processors or slurm clusters