Exploring Reward Surfaces in Reinforcement Learning Environments

Ryan Sullivan[*12], **J. K. Terry**[*12], **Benjamin Black**[*12], **John P. Dickerson**[1]

[1] Department of Computer Science, University of Maryland
[2] Swarm Labs
rsulli@umd.edu, jkterry@umd.edu, bblack1@umd.edu, johnd@umd.edu

Abstract

Visualizing optimization landscapes has resulted in many fundamental insights in numeric optimization, specifically regarding failure modes of optimizations. However, visualizations of the surface of expected reward over policy parameters that reinforcement learning optimizes (the "reward surface") have only ever been generated to study specific questions with limited scope. This work presents reward surface and related visualizations of 17 of the most widely used reinforcement learning environments for the first time. Through this, we show a new result about reward surfaces – that the reward surfaces of environments with sparse rewards are flat and noisy – and provide confirmation for the first time that many popular reinforcement learning environments have "cliffs" in the reward surfaces, a suspected problem that drove the past research on trust-region policy gradient methods. These plots allow concrete visualizations of specific failure modes of reinforcement learning. We additionally introduce a highly extensible library that allows researchers to easily generate these visualizations in the future. [1]

Introduction

Reinforcement learning attempts to optimize the expected return over policy network parameters. Understanding this optimization surface, and how reinforcement learning algorithms behave on it is critical to understanding the successes and failures of deep reinforcement learning. Policy gradient methods attempt to optimize policies by approximating the gradient of this function. This means that known problems in non-convex optimization and known benefits of gradient descent in deep learning apply here.

A "reward surface" is the high dimensional surface of the reinforcement learning objective (the expected reward when following a given policy in an environment) over the policy network parameter space. These were first visualized by (Ilyas et al. 2018) to study the quality of policy gradient estimates. Li et al. (2017) pioneered filter-normalization, a method of effectively visualizing low dimensional surfaces of neural network loss functions.

Our work utilizes filter-normalization to visualize reward surfaces for a set of 17 reinforcement learning environments. To the best of our knowledge, this is the first time visualizations of these reward surfaces have been generated for large, diverse sets of environments, and our plots notably show that the reward surfaces of environments with sparse rewards are extremely flat and noisy. While this is an intuitive result, it has not been previously demonstrated and provides guidance on the scope of possible solutions to the problem of sparse rewards. We additionally identify several interesting aspects of these reward surfaces and what they reflect about the environments from the perspective of learning algorithms.

We additionally conduct a series of novel visualizations of the reward surface, finding evidence of steep "cliffs" in the reward surface plotted along the gradient direction of numerous environments. The assumption that these cliffs pose a challenge to reinforcement learning was the basis of the trust-region based family of policy gradient methods (notably TRPO (Schulman et al. 2015) and PPO (Schulman et al. 2017)), and our plots offer conclusive visual evidence that these cliff exist, as well as a method to study them further.

To encourage future research in this direction, we release a comprehensive, modular, and easy to use library for researchers to plot these reward surfaces and better understand their role during optimization.

Background and Related Work

Neural network visualization

A neural network objective is a high dimensional function $f(\theta)$ which takes in all of the parameters in the neural network and outputs a scalar. Unfortunately, doing standard functional analysis on this high dimensional space with very few theoretical guarantees is infeasible. However, we can explore specific directions in this high-dimensional space to gain some insight into the objective function's properties. In particular, we wish to explore the domain of this function in two dimensions so that we can construct a 3d visualization that is human interpretable.

Reward surfaces

The idea of a "reward surface" is the empirical expected return over a set of network parameters. The expected empiri-

[*] These authors contributed equally.

[1] A longer and more complete working version of this preliminary workshop paper is available upon request to rsulli@umd.edu. We expect to have a public draft shortly.

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

Reward surfaces were first visualized by Ilyas et al. (2018) to characterize problems with policy gradient estimates. The authors plotted a policy gradient estimate vs a random direction, showing via visually striking examples that low sample estimates of the policy gradient rarely guide the policy in a better direction than a random direction.

Later, Ota, Jha, and Kanezaki (2021) used the method from Li et al. (2017) directly to compare shallow neural network optimization surfaces to deep neural networks, showing that deep networks perform poorly because their loss surface has much more complex curvature. They used this visual insight to develop methods that can train deeper networks for reinforcement learning tasks. Bekci and Gümüş (2020) visualize the loss landscapes of actor-critic learning methods to see the effects of action smoothing and policy stochasticity for a specific inventory control task.

Sparse rewards

An environment with sparse rewards is one that rarely emits rewards, usually only when an agent transitions into a semantically defined "goal state" (e.g. "the door is opened" or "the opponent is dead"). Reward sparsity is given mathematical formalism by Riedmiller et al. (2018), and Atari environments are famously classified into a taxonomy (which we later use) of sparse or non-sparse rewards by Bellemare et al. (2016). This sparse reward structure is hard to learn because there are no shaping rewards leading up to the rewards from goal states, which makes policies that reach goal states difficult to discover. Even after a good trajectory is discovered, learning can still be difficult because sparsity exacerbates the credit assignment problem. This refers to the challenge of determining which specific actions were responsible for causing certain rewards in a given episode, and less frequent rewards make this attribution more challenging.

Visualizing a Diverse Set of Reward Surfaces

Methodology

A reward surface is a function from policy network parameters θ to mean episodic returns. Since the function domain is so large and complex, we focus our analysis around points in the policy space visited during training. Given training checkpoint θ, we are interested in understanding the local surface Returns$(\theta + d)$ for small perturbations d.

A key challenge in this work is to choose perturbations d that give an informative view of the actual local behavior of the neural network. For example, uniform random perturbations are known to be misleading in neural network analysis, because neural networks with ReLU activations have scale invariant weights (Li et al. 2017). To mitigate this problem, we use filter-normalized random directions (Li et al. 2017). As in that work, we view the policy neural network as a vector θ indexed by layer i and filter (not filter weight) j.[2] Then, we sample a random Gaussian direction d, and scale each filter to match the magnitude of the neural network parameters in that filter, by applying the following formula.

$$d_{i,j} = \frac{d_{i,j}}{\|d_{i,j}\|} \|\theta_{i,j}\|$$

To visualize this local space in 2 dimensions we plot reward against these filter normalized directions. The x and y axis are independently sampled filter normalized directions, and the surface is projected onto this plane. Note that since the dimension of the space is large, these directions are orthogonal with high probability. The plots are additionally scaled manually to highlight features of interest, so note the marks on the axes which indicate those manual scales.

A reward surface is specific to the chosen learning algorithm and hyperparameters, so for these experiments we chose to plot the reward surface of PPO agents using the tuned hyperparameters found in RL Zoo 3 (Raffin 2020). To understand what challenges RL algorithms face towards the end of training after sufficient exploration has occurred, we chose the best checkpoint during training, evaluated on 25 episodes, with a preference for later checkpoints when evaluations showed equal rewards. The best checkpoint was typically found during the last 10% of training.

Environment Selection

In exploring these reward surfaces, we sought to cover many widely used benchmark environments. As such we generated plots for all "classic control" environments in Gym (Brockman et al. 2016) and for popular Atari environments, but because of the large computational expense of generating reward surfaces we chose 12 environments instead of all of the Atari environments in Gym.

In the spirit of exploring very diverse reward schemes, we specifically picked six sparse reward environments (Montezuma's Revenge, Pitfall!, Solaris, Private Eye, Freeway, Venture), three dense reward environments (Bank Heist, Q*Bert, Ms. Pac-Man), and three popular easy exploration environments (Breakout, Pong and Space Invaders), per the standard taxonomy by Bellemare et al. (2016).

Plots

A sampling of the visualizations of the reward surfaces on the aforementioned environments can be seen in Figure 1, the classic control plots are shown in Figure 4, and the Atari plots can be found in Figure 5. We created plots for environments with extremely large rewards in log scale to make them easier to visually interpret.

Plot Variance and Repeatability

To demonstrate the consistency of these experiments across multiple random seeds, we repeated our reward surface plots 5 times for Breakout, Freeway, and Acrobot. For each trial, we trained and evaluated a new agent on a new seed. We can see from the plots in Figure 6 that the reward surfaces are extremely visually similar in each case, showing that training tends to converge to visually similar parts of the reward landscape, and that the characteristics of these plots are consistent across multiple seeds. Each point in these plots is an

[2]Note that this method also works for fully connected layers, which are equivalent to a convolutional layer with a 1x1 output feature map.

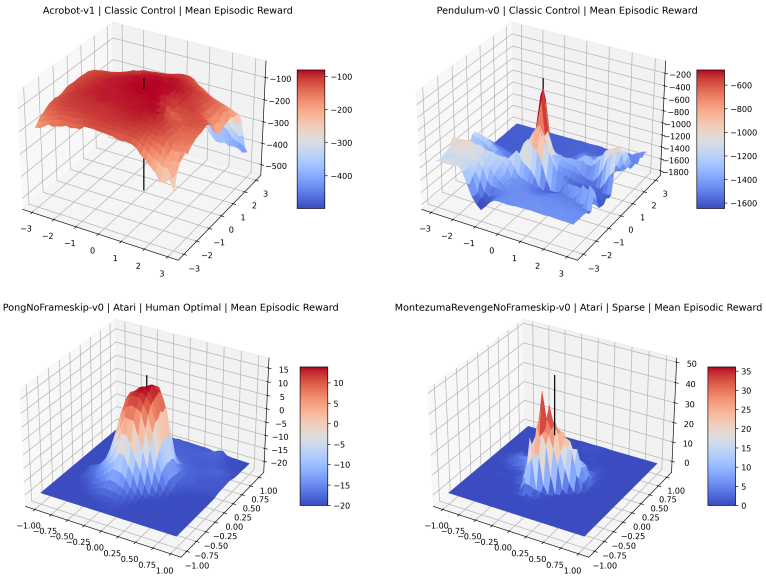

Figure 1: Reward surfaces of assorted Gym environments

estimate of the expected reward for the parameters at that point. We evaluated for at least 200,000 time steps at each point to ensure that the standard error for these estimates is small.

Findings in Plots

Flatness of Sparse Reward Environments The primary finding in the plots is the result that the reward surfaces of environments with sparse rewards are largely flat relative to the scale of rewards in the environment.

This result intuitively should be the case – for most actions in environments with sparse rewards, no reward is issued – however it is a previously undocumented visual phenomenon. This flatness suggests limitations of learning these environments, because a flat function cannot be iteratively optimized. These plots make it clear that sparse reward environments require better exploration methods than the simple heuristics built into PPO.

Other Observations One interesting observation is that the size of the maximizers present in the reward surfaces roughly correlates with the difficulty of the environment. Among the classic control environments, only Pendulum-v0 is considered unsolved on the OpenAI Gym Leaderboard, and it is the only environment of the five that does not have a local maximum spanning the entirety of the [-3, 3] range in each random direction. That being said, as a relatively easy Classic Control environment, we can see that its loss surface is still fairly smooth. We see a similar trend in the Atari environments where the dense reward and human optimal environments have comparatively large maximums while the sparse reward plots are spiky and have no clear, good maximizers. A metric based on this property could potentially be used to gauge the difficulty of an RL environment.

The sparse reward Atari environments are particularly interesting to examine. We see that Freeway's plot has a large canyon with a single smooth maximizer. Montezuma's Revenge, Private Eye, and Solaris all appear extremely noisy. The reward surface for Venture shows two neighboring maximizers, where the agent unfortunately converged to the lesser of the two peaks. And as its name suggests, the reward surface for Pitfall! is marred by several severe drops in reward. These plots seem to highlight different failure modes of sparse environments, either the surface is too flat, too noisy, or too non-convex to easily optimize.

Exploring the Gradient Direction

To better understand the optimization characteristics of these surfaces, we repeated these experiments using the gradient direction. In many of these plots, we find evidence of "cliffs" in the reward surface. These are directions in which rewards improve up to a point a very small distance away, and then sharply decreases past that point.

Gradient Directions

While filter normalized random directions provide a broad sense of the local optimization landscape, and are useful for analysis near the end of training, they are not necessarily very informative about the course or direction of training, as the directions sampled are likely orthogonal to the gradient direction used during training. To better understand the optimization trajectory, we evaluate and visualize the gradient direction against a filter normalized random direction.

One difficulty of plotting the gradient direction is that the gradient magnitudes vary drastically for different environments at different points in training. Additionally, any maximum in a reward surface can be made to look like a sharp cliff by using a large enough gradient scale, or like a large plateau by using a smaller gradient scale. To provide a com-

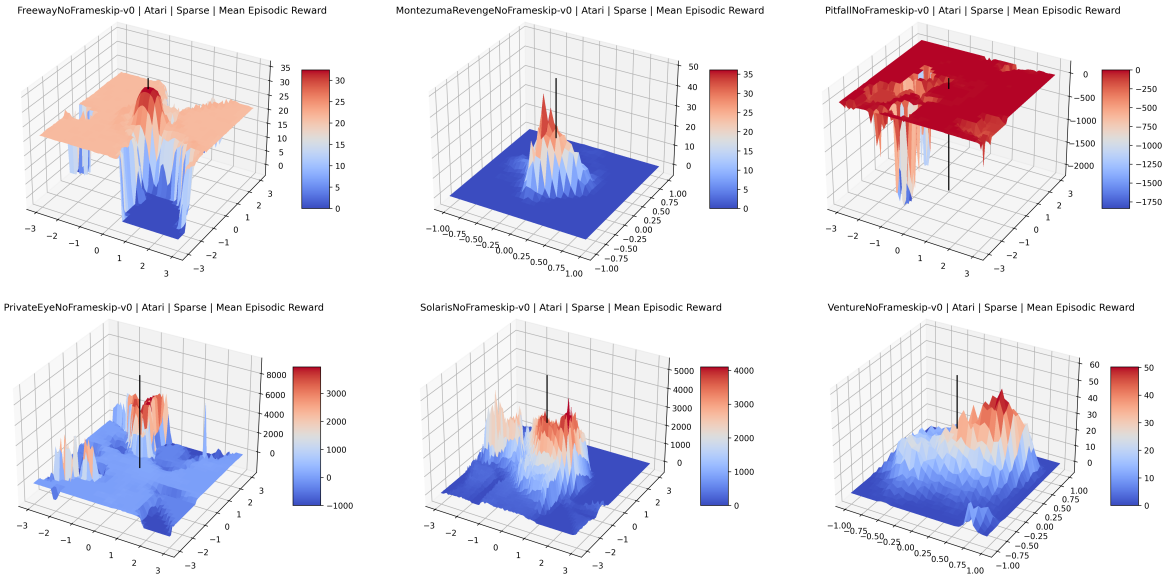

Figure 2: Reward surfaces of the Hard Exploration Atari environments

parable view of the gradient direction's sharpness, we normalize these directions instead of using manual scaling. This provides a less direct visualization of the individual learning dynamics of each environment, but allows for more fair comparisons between environments, and an unbiased visualization of sharpness in the plots.

Gradient Line Plots

We ran a second set of experiments to explore gradient directions across training, plotting the expected reward along the normalized gradient direction for evenly spaced checkpoints during a training run.

Methodology In order to understand the influence of the loss surface over the whole course of training, we plot a 1-dimensional projection of the rewards along the gradient direction vs. a series of checkpoints taken at uniform training step increments. Since the training checkpoints are relatively far apart from one another, the plot is somewhat discontinuous. However, since these checkpoints were not chosen by their performance, they should be representative of all points visited during training.

Observations We find that many of the same observations here as we do in the original reward surfaces. The gradient directions for dense reward and easy environments tend to point toward better rewards, and sparse reward environments have much noisier trajectories along the gradient directions. However, we also note some unique properties of the gradient direction. In some plots, for example in Pong, we see "cliffs" in the reward surface where the reward briefly increases, then sharply decreases. We find that these cliffs occur occasionally in almost every environment.

Gradient Heat Maps

To provide a more complete view of these cliffs, we produced heat map plots of the gradient direction against a random filter normalized direction.

Methodology Our method in this section is similar to Ilyas et al. (2018), except our plots have a much larger scale to visualize the long term dynamics of training, and we use filter normalization instead of uniform random directions. We first use a high-sample estimate of the gradient taken over 1 million timesteps, and plot steps in the gradient direction along the x axis. We then choose a filter normalized random direction to plot along the y axis. For environments where cliffs existed, we specifically chose to investigate those cliff-like checkpoints.

Observations We can see in these heat maps that the gradient direction has much sharper and more severe transitions in reward than the random normalized directions. This seems to show that to find better rewards, the agent must attempt to stay close to the edge of these cliffs without falling into a much worse parameter space.

Implications

The cliffs that we investigate in these plots may provide empirical evidence for why trust region and gradient clipping methods perform so well. The intuition that these cliffs exist was one of the motivations for developing more stable policy gradient methods like TRPO. We plan to perform further experiments to confirm our hypothesis that vanilla policy gradient methods fall off of these cliffs into poor regions of the parameter space, while TRPO and PPO avoid them.

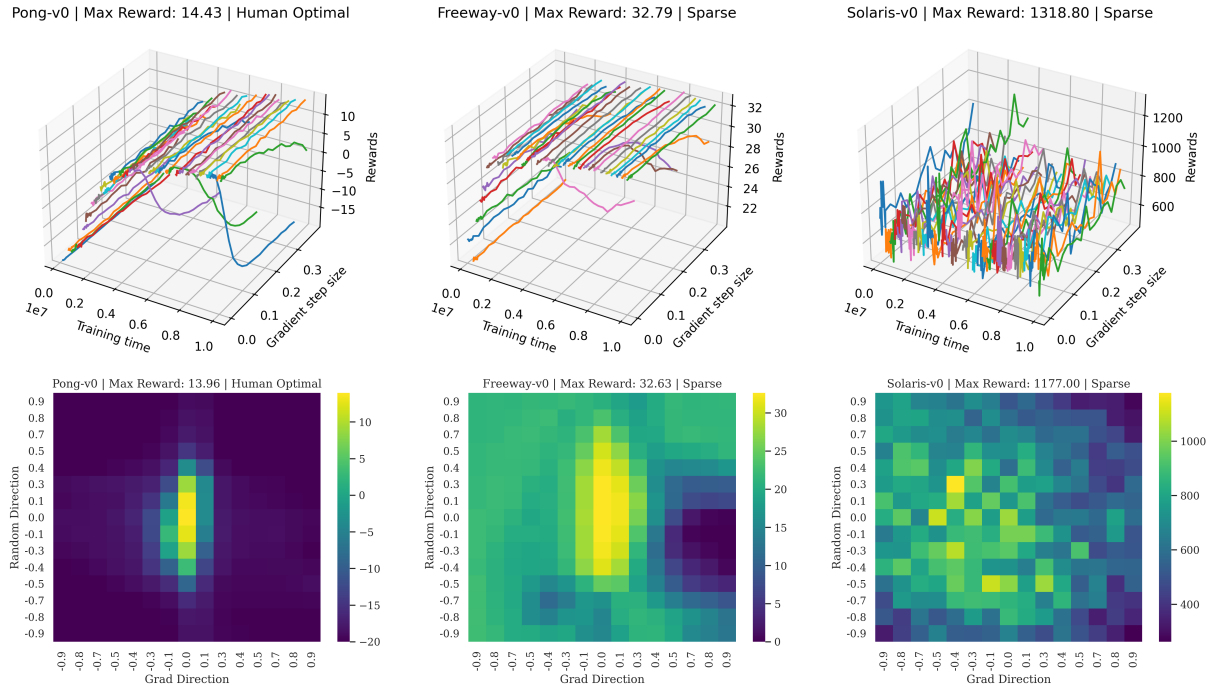

Figure 3: Line plots and heat maps of Pong, Freeway, and Solaris. Heat maps for Pong and Freeway were chosen specifically because they exhibit cliff-like behavior, while the checkpoint for Solaris is abritrary.

Library

To produce this work and encourage future research using these visualizations, we developed an extensive software library for plotting the reward surfaces of reinforcement learning agents. The library includes code for training agents using all of the options available in Stable Baselines 3 (Raffin et al. 2019) and hyperparameters from RL Zoo 3 (Raffin 2020). We provide algorithms for estimating the true policy gradient and the hessian of policy networks along with code for evaluating the rewards or discounted returns of trained agents. The entire code base supports the use of arbitrary directions for investigation, and specifically provides tools for using filter normalized and gradient directions. Finally, we include routines for plotting 3d reward surfaces, line plots, heat maps, and gifs of reward surfaces during training. The library is well organized and the experiments are documented to assist in future research. The library can be found at https://github.com/RyanNavillus/reward-surfaces.

Discussion

Our work primarily demonstrates the efficacy and reliability of using reward surfaces plotted against filter normalized directions to study the reward surfaces of RL agents. We hope that this work and our new library inspires future research into reward landscapes, so we propose a few interesting research directions. The paper that originally proposed filter normalization used it to study the effects of neural network architecture on the loss landscapes of image classification networks. Unlike most areas of deep learning, reinforcement learning has mostly failed to take advantage of deeper neural networks (Ota, Jha, and Kanezaki 2021). Our library and techniques could be used to visually study why larger networks might lead to instability during training. Similarly, our paper confirms failure modes of sparse reward environments for the first time. Our library could be used to study the effects of bonus-based exploration or intrinsic motivation on reward surfaces. Finally, this library provides interesting functionality such as hessian estimation, which allows us to study the curvature of reward surfaces, or reward surface gifs that can be used to investigate the nonstationarity of single or multi agent RL environments throughout training.

Conclusion

We generate numerous visualizations of the reward surfaces of most of the widely used reinforcement learning environments by researchers. In these plots, we clearly show "cliffs" in the reward surfaces for the first time, show various environment-specific behaviors of interest, show that the reward surfaces of sparse-reward environments are more flat than dense-reward environments, and introduce a library for easily generating more of these visualizations in the future. We hope that this steers researchers towards a better understanding of the challenges of reinforcement learning, that visual evidence of the cliffs that motivate trust region methods enables further advances in policy gradient methods, and that our library is used as an exploratory tool to study properties of individual environments.

References

Bekci, R. Y.; and Gümüş, M. 2020. Visualizing the Loss Landscape of Actor Critic Methods with Applications in Inventory Optimization. *arXiv preprint arXiv:2009.02391*.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29: 1471–1479.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Ilyas, A.; Engstrom, L.; Santurkar, S.; Tsipras, D.; Janoos, F.; Rudolph, L.; and Madry, A. 2018. A closer look at deep policy gradients. *arXiv preprint arXiv:1811.02553*.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2017. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*.

Ota, K.; Jha, D. K.; and Kanezaki, A. 2021. Training Larger Networks for Deep Reinforcement Learning. arXiv:2102.07920.

Raffin, A. 2020. RL Baselines3 Zoo. https://github.com/DLR-RM/rl-baselines3-zoo.

Raffin, A.; Hill, A.; Ernestus, M.; Gleave, A.; Kanervisto, A.; and Dormann, N. 2019. Stable Baselines3. https://github.com/DLR-RM/stable-baselines3.

Riedmiller, M.; Hafner, R.; Lampe, T.; Neunert, M.; Degrave, J.; Wiele, T.; Mnih, V.; Heess, N.; and Springenberg, J. T. 2018. Learning by playing solving sparse reward tasks from scratch. In *International Conference on Machine Learning*, 4344–4353. PMLR.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

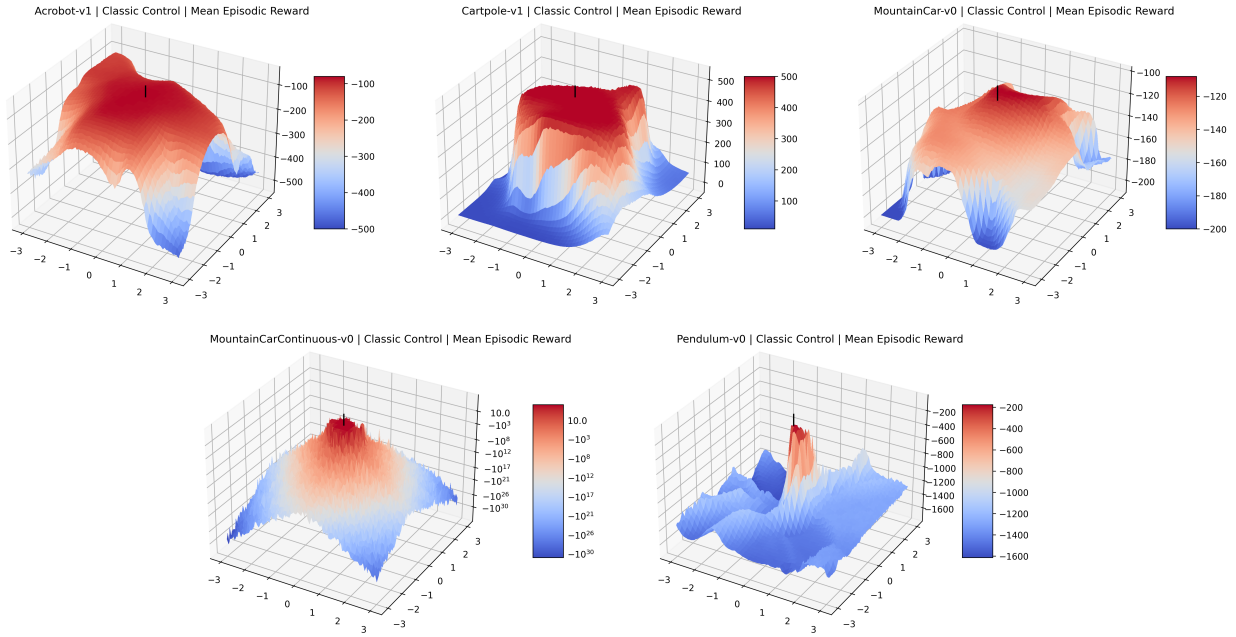

Figure 4: Reward surfaces for the 5 Classic Control environments in Gym.

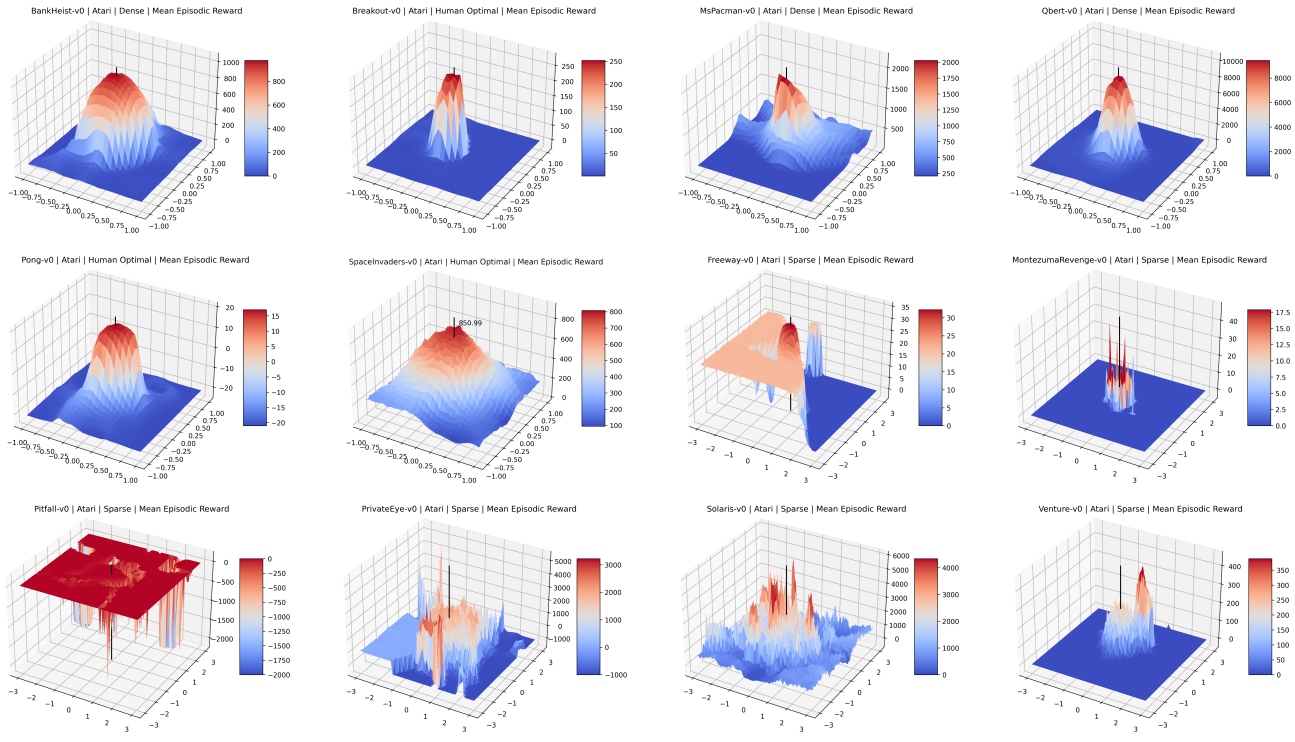

Figure 5: Reward surfaces for 12 Atari environments in Gym.

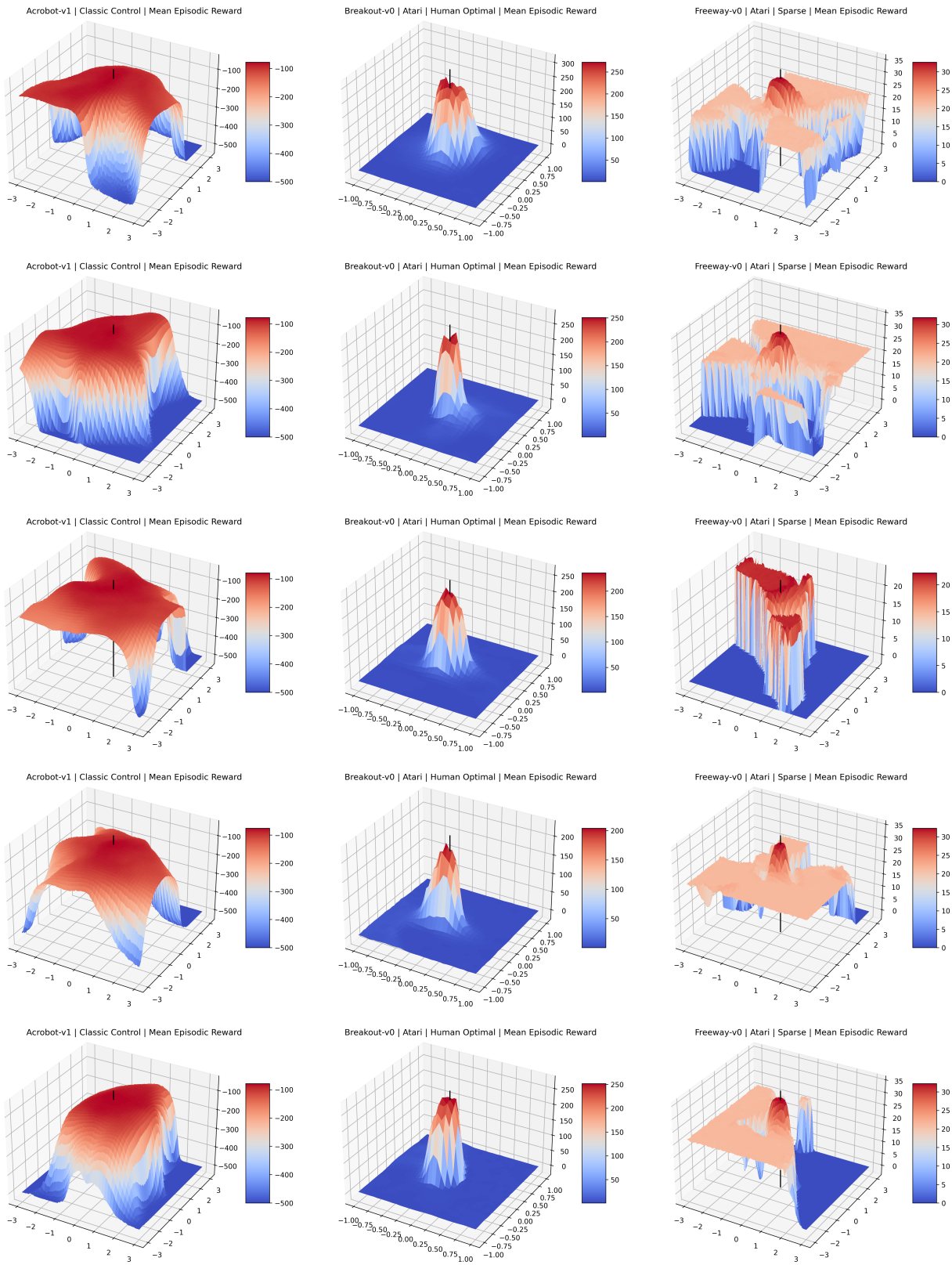

Figure 6: 5 runs of Acrobot, Breakout, and Freeway using different seeds. The plots of most runs are extremely visually similar.