Demystify the Gravity Well in the Optimization Landscape (Student Abstract)

Jason Xiaotian Dou [1] Runxue Bao [1] Susan Song [2] Shuran Yang [3] Yanfu Zhang [1] Paul Pu Liang [2] Haiyi Mao [1]

[1]University of Pittsburgh [2]Carnegie Mellon University [3]University of California, Berkeley

Abstract

We provide both empirical and theoretical insights to demystify the gravity well phenomenon in the optimization landscape. We start from describe the problem setup and theoretical results (an escape time lower bound) of the Softmax Gravity Well (SGW) in the literature. Then we move toward the understanding of a recent observation called ASR gravity well. We provide an explanation of why normal distribution with high variance can lead to suboptimal plateaus from an energy function point of view. We also contribute to the empirical insights of curriculum learning by comparison of policy initialization by different normal distributions. Furthermore, we provide the ASR escape time lower bound to understand the ASR gravity well theoretically. Future work includes more specific modeling of the reward as a function of time and quantitative evaluation of normal distribution's influence on policy initialization.

Introduction

A gravity well is the result of the pull of gravity caused by a body in space. In the context of reinforcement learning [6], the policy gradient optimizes a parameterized policy to maximize the long-term expected reward. Softmax transform is normally used to produce conditional action distributions, which results in the softmax policy gradient (SPG). Softmax transform is as the following. For $\theta \in \mathbb{R}^K$, $\pi_\theta = \mathrm{softmax}(\theta)$ is defined by

$$\pi_\theta(a) = \frac{\exp\{\theta(a)\}}{\sum_{a'} \exp\{\theta(a')\}} \tag{1}$$

for all $a \in \{1, \ldots, K\}$ [11].

The phenomenon "softmax gravity well (SGW)" is described as the following: gradient ascent trajectories are drawn toward suboptimal corners of the probability simplex and subsequently slow down in their progress toward the optimal vertex. The behavior of SPG is impacted by initialization [11]. It's similar to the concept of saddle point in the optimization landscape. An experiment on a Markov Decision Process (MDP) has been used to illustrate SGW in [11] very well. The reward r is defined as

$$r = (b + \Delta, b, \ldots, b)^\top \in [0, 1]^K \tag{2}$$

for some b, such that $\Delta > 0$ is the reward gap. And an escape time lower bound has been provided for the single state MDP setting:

Theorem

(SPG escape time lower bound [11]). In a single-state MDP, for any learning rate $\eta_t \in (0, 1]$, there exists an initialization of the policy π_{θ_1} and a positive constant C, such that SPG with full gradients cannot escape a suboptimal corner before time $t_0 := \frac{C}{\Delta \cdot \pi_{\theta_1}(a^*)}$:

$$\left(\pi^* - \pi_{\theta_t}\right)^\top r \geq 0.9 \cdot \Delta \tag{3}$$

A single state MDP with 6 actions is a simplified version of real-world reinforcement learning settings. In a three-action, multiple-state MDP setting [5], interestingly, a phenomenon similar to softmax gravity well is also observed, which is named ASR gravity well. The goal of this paper is to apply the theoretical and empirical insights of the softmax gravity well to demystify the ASR gravity well.

The ASR gravity well is illustrated in Figure 1. Details of the experimental setup are in dou2022sampling. When the policy is initialized with normal distribution with high variance, all the curves fall into a global minimum at around epoch 15. The "reverse peaks" are very significant within five (Recall@1, Recall@2, Recall@4, Recall@8, and NMI) out of six benchmarks. That means the phenomenon arises for both information retrieval and clustering applications [2, 10, 9, 3, 5, 8]. This empirical observation coincides elegantly with Theorem 1: a high standard deviation normal distribution can be *the policy initialization* of the policy π_{θ_1} which traps the algorithm into the plateaus which cannot escape within a time lower bound. To understand this more clearly, we first recall that the probability

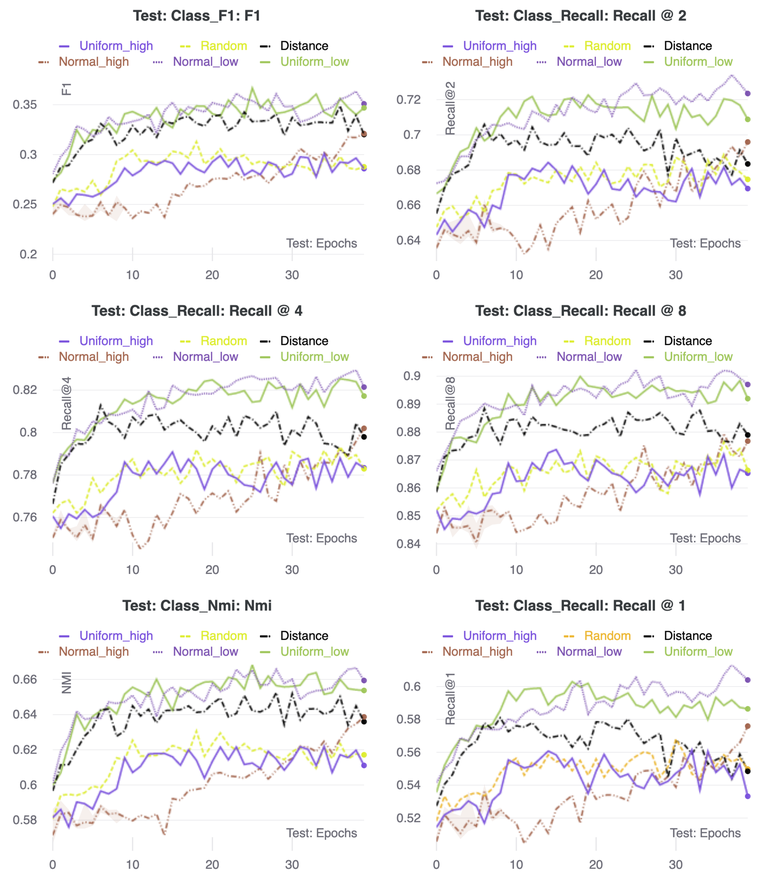

Figure 1. ASR Gravity Well [5]

density function of the normal distribution has μ as the mean and σ as the standard deviation.

From a statistical physics point of view, a high variance in the normal distribution often leads to high energy, making molecules trap into energy wells [13, 7].

Then we dig into the details of the normal distribution of the policy initialization models. In the Adaptive Sampling with Reward (ASR) setting [5], the policy initialization distribution guides a sampling distribution of $p(O_n \mid O_a)$, which builds on the distance between negative sample O_n and anchor sample O_a in triplet loss construction [2, 12, 4]. A high variance normal distribution policy initialization means that at the beginning of the sampling stage, the model will sample negative samples from a long range of distances from the anchor sample, which includes hard, semi-hard, and easy negative samples. However, in Figure 1, the gravity well only appears phenomenally in the normal high case, while for the normal high initialization, the curves are relatively flat.

Curriculum learning [1] offers two empirical insights: introducing gradually more difficult examples speeds up online training, and cleaner examples may yield better generation faster. Comparing the curves belonging to normal high and normal low distributions, we can see that from similar starting points, though the normal high curves fall into the gravity wells, they end up with higher benchmarks in all six benchmarks.

So we can add a piece new insight to curriculum learning, learning broadly (from hard negatives to easy negatives) may confuse the model in the short term and compromise downstream tasks' performance, but in the long run, learning broadly(normal high) will surpass learning within a narrow range (normal low).

we provide an ASR escape time lower bound which is adjusted from theorem 1. There are two key insights that distinguish the ASR escape time lower bound from the SPG escape time lower bound. First, $\Delta(s)$ becomes state-dependent. Second, the reward r becomes a bounded function of t, such that $r(t) = (b_1(t), b_2(t), \ldots, b_k(t))^\top$. As our reward is a weighted combination of recall and NMI evaluated on a validation dataset and it evolves over training. The proof of the theorem is in the supplementary material.

Theorem

(ASR escape time lower bound). In the multi-state MDP of the ASR setting, we use s to denote the current state. The reward $r(t)$ at time t is defined as $r(t) = (b_1(t) + \Delta(s), b_2(t), \ldots, b_k(t))^\top \in [0, 1]^K$ for $b_i(t), 1 \leq i \leq k$, such that $\Delta(s) > 0$ is the reward gap, were $b_i \in \mathcal{B} : \mathbb{R} \to \mathbb{R}$ and $\exists \beta$, such that $b_i(t) \leq \beta \leq b_1(t) + \Delta(s)$. For any learning rate $\eta_t \in (0, 1]$, there exists an initialization of the policy π_{θ_1} and a positive constant C, such that the ASR framework cannot escape a suboptimal plateau before time $t_0 := \frac{C}{\Delta(s) \cdot \pi_{\theta_1}(a^*)}$, i.e., it will hold that

$$\left(\pi^* - \pi_{\theta_t}\right)^\top r(t) \geq 0.9 \cdot \Delta(s) \tag{4}$$

Theorem 2 shows even ASR framework with relaxed conditions of reward, it still enjoys a similar lower bound.

Discussion

In the above section, we first demystify the ASR gravity well from the energy function's viewpoint. Then we complement curriculum learning's insights by comparison of different normal distribution initialization. We make a first step towards understanding the ASR gravity well theoretically by providing a new escape time lower bound. In future work, we plan to further enhance our understanding of the "wells" from two perspectives. First, we want to provide quantitative explanations of how normal distribution influences policy initialization. Second, we want to explore the approximation of the function $r(t)$ and $\Delta(s)$ to make the theory aligns more closely with empirical practice.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

[2] Shuo Chen, Lei Luo, Jian Yang, Chen Gong, Jun Li, and Heng Huang. Curvilinear distance metric learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Neurips*, volume 32. Curran Associates, Inc., 2019.

[3] Jason Xiaotian Dou, Minxue Jia, Nika Zaslavsky, Runxue Bao, Shiyi Zhang, Ke Ni, Paul Pu Liang, Haiyi Mao, and Zhihong Mao. Learning more effective cell representations efficiently. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022.

[4] Jason Xiaotian Dou, Lei Luo, and Raymond Mingrui Yang. An optimal transport approach to deep metric learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12935–12936, 2022.

[5] Jason Xiaotian Dou, Alvin Qingkai Pan, Runxue Bao, Haiyi Harry Mao, Lei Luo, and Zhihong Mao. Sampling through the lens of sequential decision making. *arXiv preprint arXiv:2208.08056*, 2022.

[6] Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*, 2021.

[7] Hailiang Liu and Xuping Tian. Sgem: stochastic gradient with energy and momentum. *arXiv preprint arXiv:2208.02208*, 2022.

[8] Jizhao Liu, Jing Lian, Julien Clinton Sprott, Qidong Liu, and Yide Ma. The butterfly effect in primary visual cortex. *IEEE Transactions on Computers*, 2022.

[9] Haiyi Mao, Minxue Jia, Jason Xiaotian Dou, Haotian Zhang, and Panayiotis V Benos. Coem: cross-modal embedding for metacell identification. *ICML Workshop on Computational Biology*, 2022.

[10] Haiyi Mao, Hongfu Liu, Jason Xiaotian Dou, and Panayiotis V. Benos. Towards cross-modal causal structure and representation learning. In *Machine Learning for Health*, 2022.

[11] Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Neurips*, 33, 2020.

[12] Jiexi Yan, Lei Luo, Chenghao Xu, Cheng Deng, and Heng Huang. Noise is also useful: Negative correlation-steered latent contrastive learning. In *CVPR*, pages 31–40, June 2022.

[13] Daniel M Zuckerman. *Statistical physics of biomolecules: an introduction*. CRC press, 2010.