Gradient–Momentum Correlation for Intrinsic Exploration in Deep Reinforcement Learning

Samuel Blad[1,2,*], Amy Loutfi[1]

[1]Fakultetsgatan 1, 701 82 Örebro, Örebro University
[2]Gustavsviksvägen 3, 702 44 Örebro, Nexer Group

Abstract

Curiosity-driven exploration in reinforcement learning often relies on prediction error or model loss, which can lead agents to focus on noisy or unlearnable data. We introduce a novel intrinsic reward based on the absolute dot product between a sample's gradient and the model's momentum. This measure identifies samples that contribute meaningfully to learning by aligning with the model's current update direction. Experiments on modified MNIST and CIFAR-10 tasks demonstrate that our method improves learning efficiency, remains robust to label noise, and induces emergent curriculum learning. Compared to uniform and curiosity-based sampling, our approach offers a simple and effective alternative.

Keywords

Reinforcement Learning, Exploration, Curriculum learning, Curiosity

1. Introduction and Related Work

Exploration in reinforcement learning (RL), enables agents to discover new knowledge and improve their behavior over time. One widely adopted strategy, especially in curiosity-driven learning, rewards the agent for encountering unexpected situations using prediction errors or model loss as proxies for novelty. However, these methods frequently struggle to distinguish between genuine learning opportunities and randomness. For example, unpredictable noise can generate large losses, misleading the agent into prioritizing uninformative experiences [1].

In this work, we propose an intrinsic reward based on the correlation between an individual sample's gradient and the momentum of the model's parameters. The intuition is that samples contributing to meaningful learning will have gradients that align (or counter-align) with the overall direction of parameter updates. In contrast, noisy or overly difficult samples will appear misaligned or inconsistent, reducing their influence.

We show that this gradient–momentum correlation can act as a signal for identifying learnable experiences, guiding the agent's exploration in a way that is both robust to noise and conducive to emergent curriculum learning. Through experiments on modified MNIST and CIFAR-10 tasks, we demonstrate how this method improves sampling efficiency and overall performance compared to curiosity- and uniform-based strategies.

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.
*Corresponding author.
samuel.blad@oru.se (S. Blad); amy.loutfi@oru.se (A. Loutfi)
https://mpi.aass.oru.se/samuel-blad/ (S. Blad); https://mpi.aass.oru.se/amy-loutfi/ (A. Loutfi)
0000-0003-1913-882X (S. Blad); 0000-0002-3122-693X (A. Loutfi)

The motivation behind this work is to improve exploration in reinforcement learning by distinguishing between meaningful learning signals and noise. We address this by proposing a simple intrinsic reward that better highlights learnable experiences, leading to more efficient and robust training.

We theorize that there can be progress in learning whilst the loss is not decreasing, therefore challenging a core assumption in many reinforcement learning (RL) exploration methods—namely, that progress in learning is always reflected by a decrease in loss. Rather, the paper justifies using a more nuanced signal—like gradient–momentum correlation—to detect where real learning occurs supporting the development of more robust exploration strategies that avoid being misled by noise or temporarily stagnant loss.

In previous work, one exploration approach which has sparked interest is to analyze the loss of a surrogate model, for instance a model of the dynamics of the environment. Curiosity driven approaches [2] [3] directly correlate exploration to the value of the loss. However, these approaches assume that high losses correlate with high interest in visitation. This causes problems in environments with epistemic uncertainty [1], which will be referred to in this paper as "noise".

2. Method

Momentum, a standard component in optimization algorithms such as Adam [4], helps smooth and accelerate learning by tracking recent gradients. It represents the accumulated direction of learning. By comparing a current sample's gradient of the loss with this momentum, we can assess whether the sample contributes meaningfully to the ongoing learning direction. In this work, we measure how well the gradient of each sample correlates with the momentum to quantify its contribution to learning.

To avoid bias, our approach treats both positive and negative correlations with the learning direction equally. This ensures that the model gives equal importance to data that either reinforces or counterbalances its current trajectory, helping it find a more accurate balance along that learning axis. As training progresses, this axis—representing the direction of ongoing change in parameter space—shifts over time, reflecting the model's evolving understanding.

We formalize this gradient–momentum correlation as an exploration bonus reward $r_{exp} = |\sum_i \frac{g_i \cdot m_i}{v_i}|$ where g is the gradient of a particular sample, m is the momentum, and $v > 0$ is its second moment, as defined in the Adam optimizer, used here for normalization. This value can be computed during backpropagation with no added computational cost. We incorporate r_{exp} as an intrinsic reward for each transition in the reinforcement learning environment.

3. Experiments

3.1. Experiment Setup

We evaluate our approach using modified versions of the MNIST and CIFAR-10 datasets. In our setup, an agent selects which class to sample from for training a classifier. This process is

analogous to a reinforcement learning (RL) environment without extrinsic rewards, where each episode spans three steps and corresponds to a single image-label pair.

At step one, the agent (actor) chooses a class. At step two, the environment returns an image from the chosen class as an observation. At step three, the environment returns the correct label of the previous image as an observation, terminating the episode. The classifier serves as a dynamics model, attempting to predict the next observation - the label of the image. The agent is trained to maximize an intrinsic reward r_{exp} using policy gradient methods.

Both the classifier and the agent are implemented as small multilayer perceptrons (MLPs). The intrinsic reward r_{exp} is derived from the classifier during training and provides the learning signal for the agent.

To introduce controlled variation in task difficulty and label noise, we group the 10 dataset classes into four clusters: $\{0\}, \{1, 2\}, \{3, 4, 5\}$, and $\{6, 7, 8, 9\}$. Two distinct experimental conditions are designed:

Experiment 1 – Curriculum: At the beginning of training, each sample's label is randomly reassigned to another label within its group. This creates a moderately perturbed label space while preserving group-level semantic coherence.

Experiment 2 – Noise: Labels are randomized within their group each time a sample is drawn, simulating stochastic and potentially noisy transitions. Here the agent chooses one of the four class groups for sampling instead of selecting individual classes.

We compare our method against two baselines: (1) Uniform sampling (no agent), and (2) Curiosity-based sampling, where the classifier's loss replaces r_{exp} as the agent's intrinsic reward.

3.2. Experiment Results

Figures 1 and 2 present the outcomes for MNIST under the Curriculum setting and CIFAR-10 under the Noise setting, respectively[1].

In the Curriculum setting, our method exhibits a clear progression through the groups, prioritizing them in the same order as their final losses, indicating an emergent curriculum. Unlike curiosity-based exploration, which emphasizes high-loss samples, our method selects samples based on where the learning gradient is steepest. This allows easier groups to improve quickly early in training while harder groups are addressed later. As a result, overall convergence is accelerated compared to uniform sampling.

Notably, our method stops focusing on the easiest group before it reaches its minimum loss, suggesting that it prioritizes regions of steep learning over low absolute loss. In later training stages, sampling becomes more proportional to loss magnitudes, resembling curiosity-based behavior, but still differs by not entirely neglecting lower-loss groups.

In the Noise setting, our method maintains stable performance across groups. On CIFAR-10, it prioritizes groups roughly in accordance with their noise levels, with less attention given to

[1]For brevity we omitted results for CIFAR-10 on experiment 1, and MNIST on experiment 2 as these results were very similar to their counterpart on the same experiment.

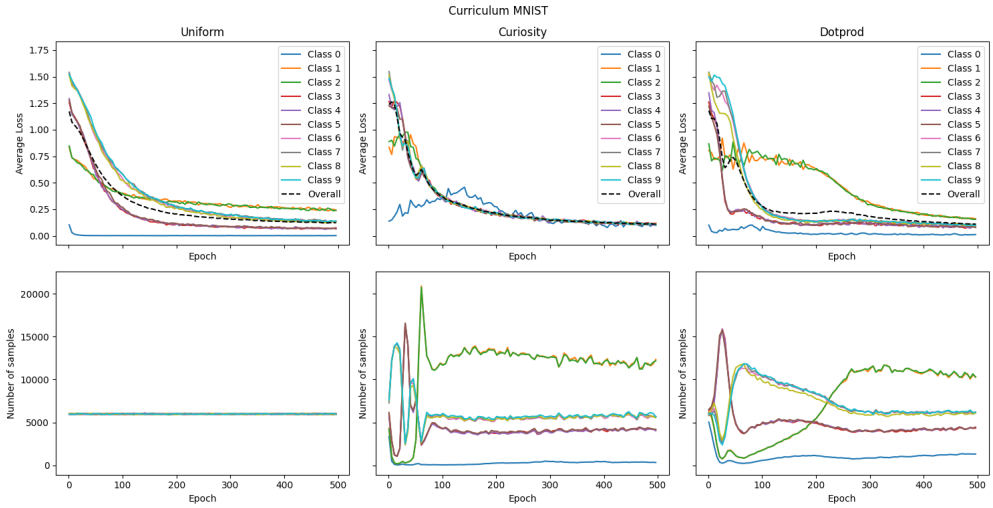

Figure 1: Comparison of losses and sampling between all methods on MNIST variant 1 ”Curriculum”.

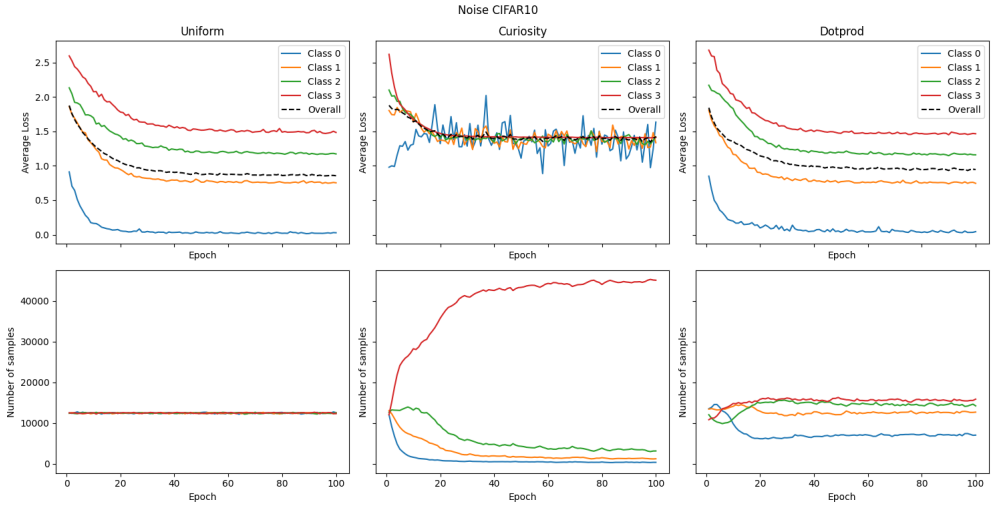

Figure 2: Comparison of losses and sampling between all methods on CIFAR variant 2 ”Noise”.

the no-noise group—a sign of resilience in differentiating learnable patterns from randomness. On MNIST, it samples more evenly among the noisy groups while still avoiding overfitting to the easiest class.

Curiosity-driven sampling, in contrast, tends to focus solely on high-loss groups. This results in degraded overall performance, particularly in the Noise condition, as it fails to distinguish between informative difficulty and pure noise.

In Figure 2 for uniform, we can notice that the slope of groups 2 and 4 is slightly steeper than that of group 3. Our method takes notice of this, and gives group 3 a slightly lower priority at the start, similar to the curriculum behavior we saw in variant 1.

4. Discussion

Our experiments show that gradient–momentum correlation may be a reliable signal for guiding exploration. Unlike curiosity-based methods that treat all high-loss samples equally, our approach successfully downweights uninformative or noisy data. Near convergence, our method slightly favored samples with noise over those with zero loss (Figure 2). While this might seem suboptimal, it may preserve the chance to extract overlooked structure, rather than discarding samples that appear noisy but may still be learnable.

We also observed emergent curriculum learning, where the agent progressed through sample groups in a manner that mirrored their difficulty. This property may be valuable in RL domains where complex tasks require mastering foundational ones first [5, 6].

5. Future work

Future work could explore alternative formulations or transformations of the signal other than the dot product with the gradient momentum. Additionally, tuning the momentum parameters or adapting them during training may influence the quality of the intrinsic reward and warrants further investigation. Nevertheless, this paper presents a novel but simple approach, leaving room for broader exploration of its underlying concept. Our next steps include applying this method in larger and more complex RL environments, with a focus on domains that require advanced exploration strategies such as Crafter [7] or Procgen [8].

References

[1] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, A. A. Efros, Large-scale study of curiosity-driven learning, arXiv preprint arXiv:1808.04355 (2018).

[2] D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, Curiosity-driven exploration by self-supervised prediction, in: International conference on machine learning, PMLR, 2017, pp. 2778–2787.

[3] Y. Burda, H. Edwards, A. Storkey, O. Klimov, Exploration by random network distillation, arXiv preprint arXiv:1810.12894 (2018).

[4] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[5] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[6] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, P. Stone, Curriculum learning for reinforcement learning domains: A framework and survey, Journal of Machine Learning Research 21 (2020) 1–50.

[7] D. Hafner, Benchmarking the spectrum of agent capabilities, arXiv preprint arXiv:2109.06780 (2021).

[8] K. Cobbe, C. Hesse, J. Hilton, J. Schulman, Leveraging procedural generation to benchmark reinforcement learning, in: International conference on machine learning, PMLR, 2020, pp. 2048–2056.