FACEBOOK AI

Return Dispersion as an Estimator of Learning Potential for Prioritized Level Replay

Iryna Korshunova, Minqi Jiang, Jack Parker-Holder, Tim Rocktäschel, Edward Grefenstette

Facebook AI Research, University College London

UCL

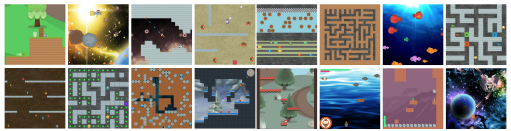
In procedurally-generated environments with algorithmically created levels, sampling training levels in proportion to their learning potential can improve generalization and sample efficiency of RL agents.

An intuitive way of estimating the learning potential of a level is to measure the dispersion of agent's returns across episodes.

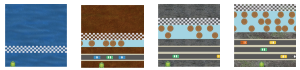
We explore the behaviour of this dispersion-based score and connect it to the existing value-based level scoring function previously used for prioritized level replay.

OpenAI Procgen Benchmark [1]

16 simple-to-use procedurally-generated environments



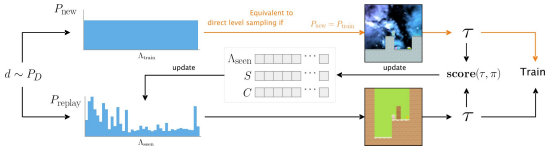
Screenshots from each Procgen environment



Examples of levels from the Leaper environment

Prioritized Level Replay [2]

automatic curriculum discovery for procedurally-generated environments



Overview of Prioritized Level Replay. The next level is either sampled from a distribution with support over unseen levels (top), or from the replay distribution, which prioritizes levels based on their learning potential (bottom). In both cases, we collect a trajectory τ and update the level's score and the replay distribution. This update depends on the lists of previously seen levels, their latest scores S, and last sampled timestamps C.

PLR computes level scores as an average magnitude of the generalized advantage estimate (GAE) [3]:

$$\text{score}(\tau, \pi) = \frac{1}{T}\sum_{t=0}^{T}\left|\hat{A}_t^{\text{GAE}(\gamma,\lambda)}\right| = \frac{1}{T}\sum_{t=0}^{T}\left|\sum_{k=t}^{T}(\gamma\lambda)^{k-t}\delta_k\right|$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

In PPO [4], this score corresponds to the average L1 value loss.

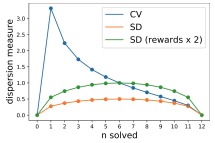
Settings where L1 value loss might lead to inappropriate scores:

- stochastic rewards
- partially-observable environments
- observational feature aliasing in the value network

Returns dispersion as a PLR score

an alternative heuristic to the L1 value loss

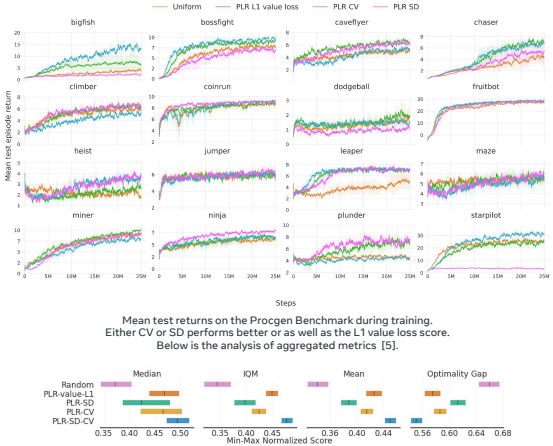
Intuition: levels with high learning potential are those that the agent cannot solve consistently. On such levels, the dispersion (variability) of returns across episodes will be high.



Measures of dispersion: standard deviation (SD, σ) and coefficient of variation (CV, σ/μ) of returns across 12 episodes as a function of the number of solved episodes, where max return is +1.
SD, unlike CV, is not scale-invariant, so multiplying rewards by a factor of two leads to larger SD values.

Experiments

comparing PLR variants with different scoring functions



Mean test returns on the Procgen Benchmark during training.
Either CV or SD performs better or as well as the L1 value loss score.
Below is the analysis of aggregated metrics [5].

Conclusions

what did we learn?

- We can theoretically connect value losses to the variance of returns
- In practice, we see inconsistent gains from using dispersion-based scores, however we can explain most of the failure cases
- Potential improvements can be gained from better variance estimators [6]
- The problem of dealing with the aleatoric uncertainty remains

References

[1] K. Cobbe, C. Hesse, J. Hilton, J. Schulman. "Leveraging Procedural Generation to Benchmark Reinforcement Learning". ICML, 2020

[2] M. Jiang, E. Grefenstette, T. Rocktäschel. "Prioritized Level Replay". ICML, 2021

[3] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel. "High-dimensional continuous control using generalized advantage estimation". ICLR, 2016

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov. "Proximal policy optimization algorithms". ArXiv preprint, 2017

[5] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, M. G. Bellemare. "Deep reinforcement learning at the edge of the statistical precipice". NeurIPS, 2021

[6] A. Tamar, D. Di Castro, S. Mannor. " Learning the variance of the reward-to-go". JMLR, 2016