The 23rd Annual Signal and Image Sciences Workshop at Lawrence Livermore National Laboratory (LLNL), Center for Advanced Signal Image Sciences (CASIS)

Towards a Representative Metric of Behavior Style in Imitation and Reinforcement Learning

Igor Borovikov†, Jesse Harder†, Michael Sadovsky*, Ahmad Beirami†

†Electronic Arts, EADP Data and AI, Redwood Shores, California 94065
{iborovikov, jharder, abeirami}@ea.com

*Institute of Computational Modelling, Krasnoyarsk, Russia 660036
msad@icm.krasn.ru .

Motivation

Problem: For automated playtesting and game balancing in video games, we need to train an agent that behaves like a human player.

Challenges:

- Agents are hard to train using Reinforcement Learning (RL), and they don't reproduce organic play style.
- Imitation Learning (IL) reproduces the style but doesn't deliver target performance.

Hypothesis: Shaping a training reward in RL with a style term can address both challenges.

Question: How to compute the style term in a training reward?

Approach

- Collect n-grams of actions in an episode E as in NLP-like statistical model of language style,
- Shape episode Q reward $R'(Q) = R(Q) + \omega D(Q, P)$, where:
  - $R(Q)$ - regular reward from the episode,
  - weighted with ω style reward $D = D_{\lambda,N}(P, Q)$ parameterized by $0 < \lambda < 1$ (larger λ emphasises longer n-grams),

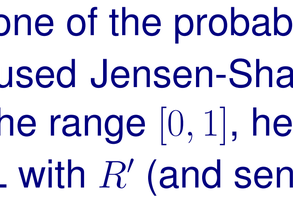
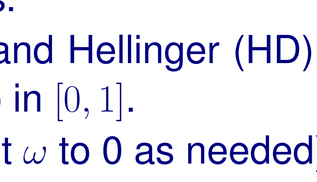
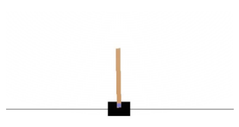
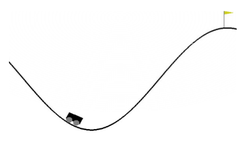
$$D(Q, P) = \frac{\lambda}{1-\lambda} \sum_{n=0}^{N} \lambda^n d(q_n, p_n) + \frac{\lambda^{N+1}}{1-\lambda} d(q_N, p_N)$$

  - p_n (q_n) is n-gram distribution of P (Q),
  - d is one of the probability distances.
  - We used Jensen-Shannon (JSD) and Hellinger (HD); both are in the range $[0, 1]$, hence D is also in $[0, 1]$.
- Run RL with R' (and send style weight ω to 0 as needed).

Training Algorithms

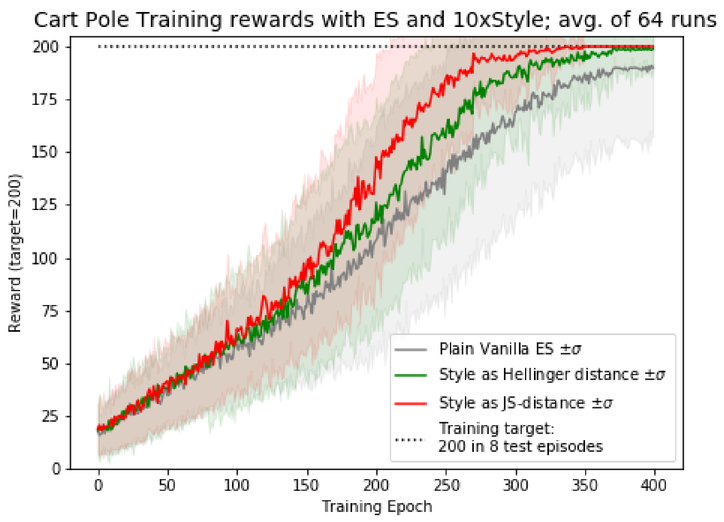
- **DQN** is hard to inject demonstrations into replay buffer, slow convergence, reward shaping is difficult (*discarded*).
- **Evolution Strategies** (ES), a black box stochastic optimization over model parameters, robust convergence and easy reward shaping.

Experiments in Two Environment Types

"Reactive"	"Strategic"
Combat in First Person Shooter	Puzzle solving, Exploration
	
Strong dense rewards signal	Infrequent noisy rewards
OpenAI gym: Cart Pole	OpenAI gym: Mountain Car
	

"Reactive" Environment: OpenAI Gym Cart Pole

- **Demonstrations** P**:** "Limited repetitions" sub-optimal policy implementing simple heuristic avoiding over-regulation.
- **Reward:** $R'(Q) = R + k(R_0 - R)D(Q, P)$, where R_0 is the target reward equal number of frames until the pole falls down.

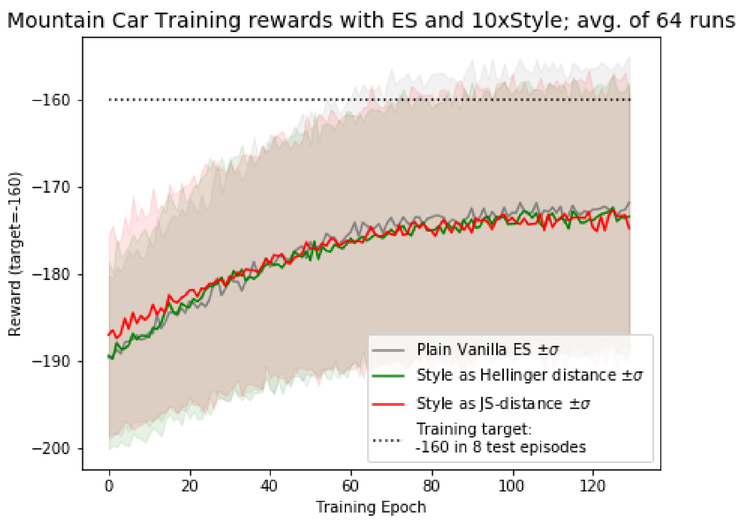


Training from scratch, $k = 10$, shows that both JSD and HD behave similarly and give minor speedup of training when approaching the target reward. HSD allowed to to complete training with perfect results at epoch ∼350 while other methods didn't reach target reward 200 at the end of all training runs. ES used population 5 (a very small one), yet successfuly converged.

"Strategic" Environment: OpenAI Gym Mountain Car

- **Demonstrations** P**:** Near-optimal policy adding mechanical energy to the system.
- **Reward:** $R'(Q) = R + kD(Q, P)$, k - tunable parameter.

No convergence for ES from scratch with or without style! Training quickly gets into local maximum, with style component making such a behavior worse. Reason: the "good" parameters are only a small subset of the entire space, which is hard to find by random ES exploration.



Training with ES from scratch for Mountain Car fails due to sparsity of the reward signal. Starting from partially pretrained with IL model, shown on the figure, and $k = 10$ we observe that style contribution is neutral in the cases of both HD and JSD. The results suggest that rewards with naive style component do not contribute to RL training of an agent as hypothesised in the "strategic" type of environments.

Conclusion

- Training agents with style component in the reward may speed up training in certain type of environments ("reactive").
- More often than not, the difference from style-less training is marginal.
- In sparse reward environments ("strategic"), style may prematurely lead to local extremum.

Future Work

- Explore additional style metrics (e.g., based on automata and HMM).
- Can adding noise to style-influenced training help to avoid local extrema?