Analysis of Curriculum Learning methods in Reinforcement Learning

Orel Danyil
2nd year Master's student
Computer Science
National University of Kyiv-Mohyla Academy
Mentor: Glybovets Mykola

Current directions

X. Wang at al. mentions **benchmark evaluation** within Curriculum Learning (CL) methods as promising scientific direction (2021)

For now, there are no certain rules, but only preferences in making decision, what CL method to use. These preferences were derived for several CL methods in general usage not paying attention to the specific domains and datasets properties

X. Wang et al. asks still a few open questions (2021)

- How to choose **a proper CL method** for a certain task having certain dataset properties?
- What are pros and cons for different **dataset configurations**? (in terms of feature space, sparsity of the dataset, etc.)

Aim: provide a comprehensive comparison of CL methods in RL domain across various environments

Object: set of CL methods including Pre-defined, Self-paced, Transfer, Teacher, and Anti-CL methods

Objectives:

- Review existing industry knowledge about benchmarking CL methods in RL
- Conduct experiments and measure performance metrics for CL methods
- Develop generalized and intuitive strategies for selecting most suitable CL method under specific environmental properties of RL task

Solution #1

Isolated environments

Comparison should encounter different **environmental configurations**:

- State space complexity
- Action space diversity
- Sparsity of the rewards
- Temporal dynamics of game play

For such purpose, this research includes: CartPole, MountainCar, and Atari games Boxing environments

Solution #2

Metric design

Comparison of CL methods considering various **RL-specific metrics**:

During **training** phase:

- Learning Stability

During **evaluation** phase:

- Mean of Rewards
- Standard Deviation of Rewards
- Average Adjusted Returns (AAR)
- Safe Exploration Score (SES)
- Convergence Speed

Solution #3

Agent architectures

Comparison of CL methods considering various **RL-specific architectures.** Selection of RL method per environment is based on dimensionality of state:

- Small: Q-Learning, DQN, PPO
- Large: DQN, PPO

For making precise measurements, architectures **best practices** and **hyperparameters** included in the research:

- SOTA optimization algorithm: Adam
- Learning rate scheduling: Exponential
- Gradient clipping

Experimental configurations

Software and Hardware:

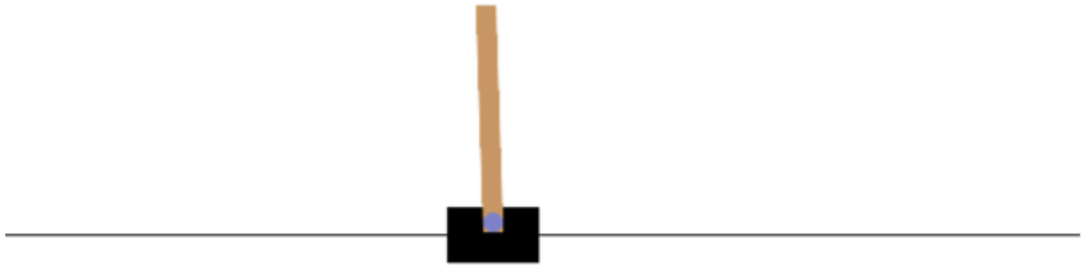
- **Framework**: PyTorch
- **Hardware**: Single GPU core based of NVIDIA RTX 3090 machine
- **Dataset**: OpenAI Gym
- **Other**: Jupyter Notebook

Theory:

- Based on Henderson at al. (2017), benchmarking should not be representable on $N < 5$ trials due to variability of RL training
- This research held experiments via $N = 5$ repetitions for each agent and corresponding CL method

CartPole

- Space complexity: position, velocity, pole angle, angular velocity
- Action complexity: move left / move right
- Rewards: not sparse
- Terminal state: mean reward 500.0 during evaluation
- Curriculum parameter: pole length



CartPole: results

Based on Q-Learning and DQN agents' performance:

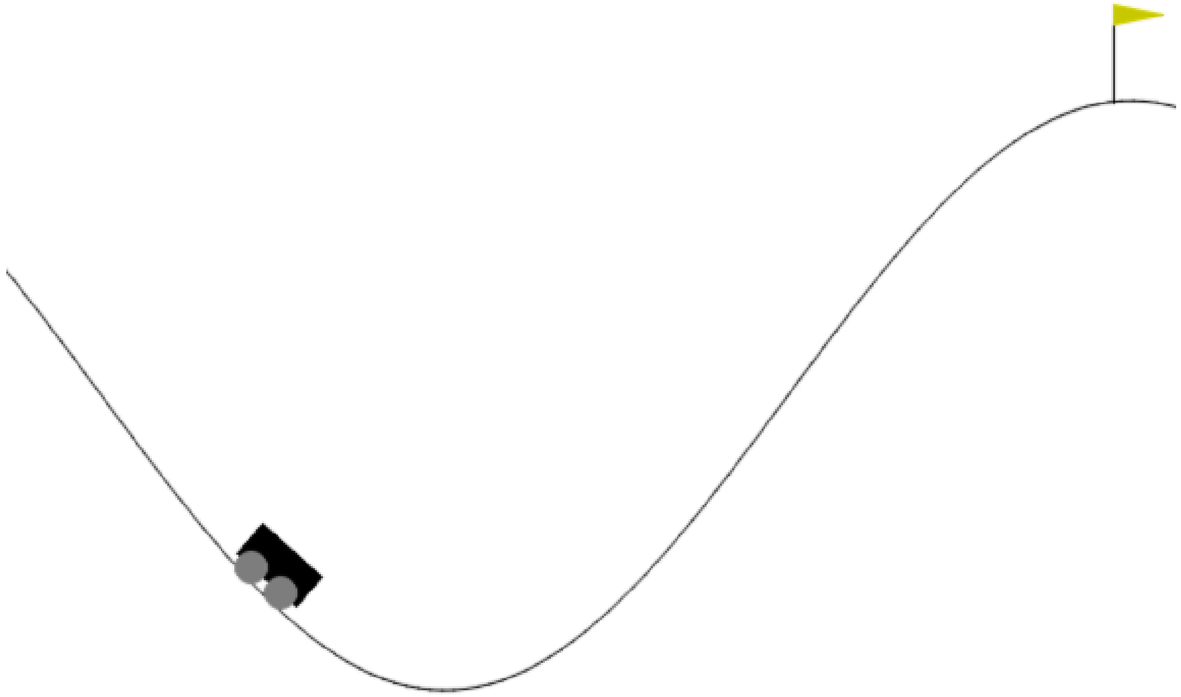
- Teacher learning the only converged CL method with the highest performance metrics
- Root-p shows most consistent and stable learning
- Anti-CL has the highest SES

DQN	Metrics						
	AAR	SES	Learning Stability	Mean Reward	Std Reward	Max Reward	Convergence Speed
Anti-curriculum	1.39	0.98	40.75	123.25	16.00	376.40	N/A
Baseline	1.00	**0.99**	53.62	60.57	16.31	119.14	N/A
Hard	1.34	0.98	29.15	95.22	12.71	159.42	N/A
Linear	1.39	0.98	53.20	78.34	18.03	155.68	N/A
Logarithmic	1.12	0.98	55.65	128.79	14.31	255.58	N/A
Logistic	1.43	**0.99**	61.53	178.47	18.01	355.94	N/A
Mixture	1.51	**0.99**	52.65	73.23	17.03	146.46	N/A
One-pass	1.38	0.98	54.41	32.48	15.34	64.96	N/A
Polynomial	**1.57**	0.98	42.31	232.81	15.46	465.62	N/A
Root-p	1.33	**0.99**	41.38	42.83	**10.51**	85.66	N/A
Teacher learning	1.44	0.98	51.61	**250.00**	12.82	**500.00**	**3**
Transfer learning	1.19	0.98	**23.44**	222.44	12.10	444.88	N/A

Table 2: Comparative analysis of CL methods based on DQN agent trained in CartPole environment. The table displays the AAR, SES, Learning Stability, Mean Reward, Standard Deviation of Reward (Std Reward), Maximum Reward, and Convergence Speed. Highlighted values indicate the highest performance in each metric, illustrating the efficacy and particular strengths of each configuration in specific aspects of learning and exploration.

MountainCar

- Space complexity: position, velocity
- Action complexity: move left / move right, do nothing
- Rewards: extremely sparse, given after achieving a flag
- Terminal state: mean reward -110.0 during evaluation
- Curriculum parameter: gravity



MountainCar: results

Based on Q-learning, DQN, and PPO agents' performance:

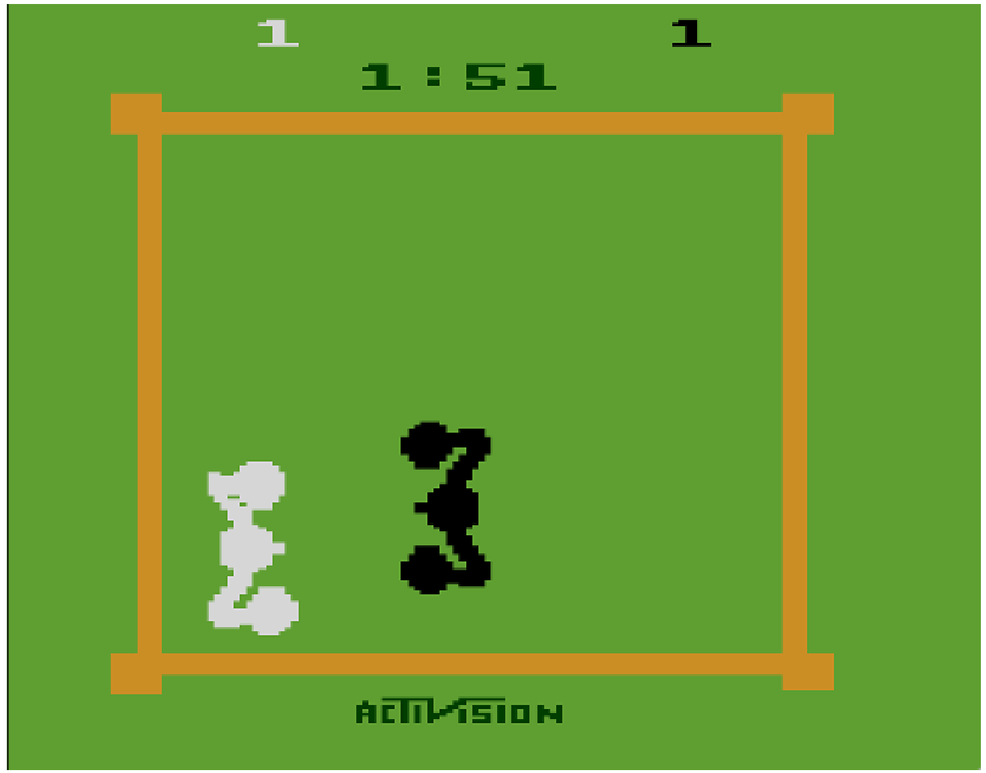
- Non-CL Baseline and CL-based Polynomial turned out to be the most adaptive methods
- Pre-defined (Root-p and One-pass) shows the most consistent and stable learning confirming Naverkar et al. thoughts (2020)

DQN	Metrics		
	AAR	SES	Learning Stability
Anti-curriculum	-32.81	**1.00**	116.99
Baseline	**-16.34**	**1.00**	302.12
Hard	-66.00	0.99	234.70
Linear	-36.86	**1.00**	131.89
Logarithmic	-20.13	**1.00**	153.42
Logistic	-48.94	**1.00**	134.76
Mixture	-69.28	0.99	225.66
One-pass	-36.82	**1.00**	142.64
Polynomial	-58.08	**1.00**	122.45
Root-p	-34.00	**1.00**	**103.50**
Teacher learning	-957.09	**1.00**	120.05
Transfer learning	-31.11	**1.00**	221.41

Table 4: Comparative Analysis of CL methods performance metrics based on DQN agent performance in MountainCar environment. The table displays the AAR, SES, and Learning Stability for various CL strategies. Highlighted values indicate the highest performance in each metric, illustrating the strengths of each configuration in adaptability and exploration.

Boxing

- Space complexity: image 210x160 in RGB format of game screen
- Action complexity: move left, move right, punch, and block
- Rewards: not sparse, but dynamic
- Terminal state: mean reward 0.0 during evaluation
- Curriculum parameter: skip frame



Boxing: results

Based on DQN and PPO agents' performance:

- Transfer learning prevails other CL and non-CL methods in terms of performance and adaptability which extending Taylor et al. (2009) ideas
- One-pass shows most consistent and stable learning
- SPL-based methods and Anti-CL prevail in speed of convergence

DQN	Metrics						
	AAR	SES	Learning Stability	Mean Reward	Std Reward	Max Reward	Convergence Speed
Anti-curriculum	-0.00094	0.99922	6.62	-6.90	5.97	1.2	**0**
Baseline	-0.00087	0.99967	7.83	-8.70	8.20	0.8	8
Hard	-0.00086	0.99761	6.20	-6.82	7.27	-3.0	N/A
Linear	-0.00555	0.99950	16.37	-28.90	14.85	1.0	**0**
Logarithmic	-0.00421	0.99915	15.09	-27.31	11.19	-7.2	N/A
Logistic	-0.00280	**0.99999**	9.29	-20.82	8.15	-11.6	N/A
Mixture	-0.00116	0.99955	**5.56**	-9.70	4.81	-2.8	N/A
One-pass	-0.00134	0.99968	6.49	-5.50	**4.58**	-2.2	N/A
Polynomial	-0.00231	0.99932	10.99	-14.80	10.49	-7.6	N/A
Root-p	-0.00120	0.99870	7.64	-5.75	6.33	3.4	**0**
Teacher learning	-0.00135	0.99840	9.78	-9.21	7.33	-0.6	N/A
Transfer learning	**-0.00015**	0.99888	6.03	**-1.88**	6.11	**8.8**	2

Table 6: Comparative analysis of CL methods based on DQN agent performance for Atari games Boxing environment. The table displays the AAR, SES, Learning Stability, Mean Reward, Standard Deviation of Reward (Std Reward), Maximum Reward, and Convergence Speed for distinct CL methods measured during DQN agent learning. Highlighted values indicate the highest performance in each metric, illustrating the efficacy and particular strengths of each configuration in specific aspects of learning and exploration.

Conclusions

1. Complexity of dynamics of environment:
   - Simple (e.g., CartPole): SPL-based methods are prefered due to overall high performance and rapid adaptation
   - Environments with sparse rewards (e.g., MountainCar): non-CL or SPL-based methods should be preferred as methods excel in balancing exploration and stability
   - Complex (e.g., Boxing): Transfer learning or Anti-CL methods tuned for efficient decision making and reward maximization within analyzed set of CL methods

2. **Rapid convergence:** Teacher learning and SPL-based Linear methods demonstrated quick convergence to termination conditions within set of environments

3. **Adaptation with negative outcomes**: SPL-based methods prevail in accomodation to punishing rewarding mechanisms

Thank you for listening!