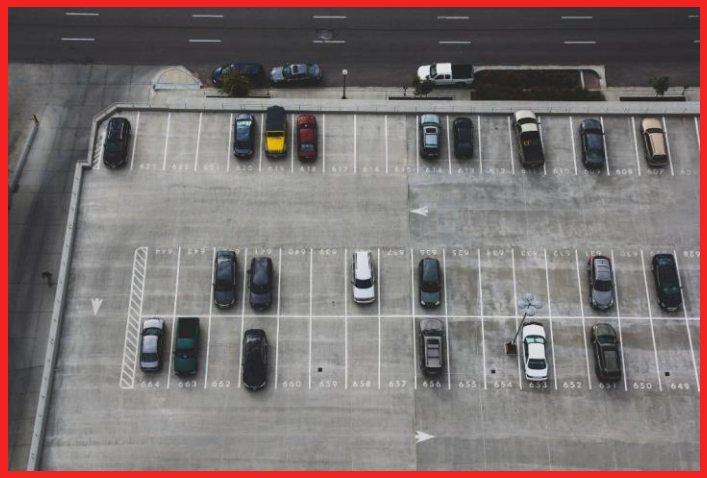
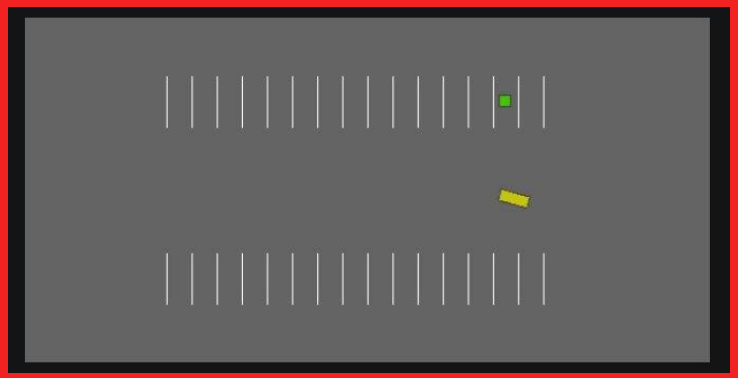
Deep Reinforcement Learning Implementation on a Parking Environment

- Team Members:
  - Aditya Aspat
  - Atharva Jamsandekar



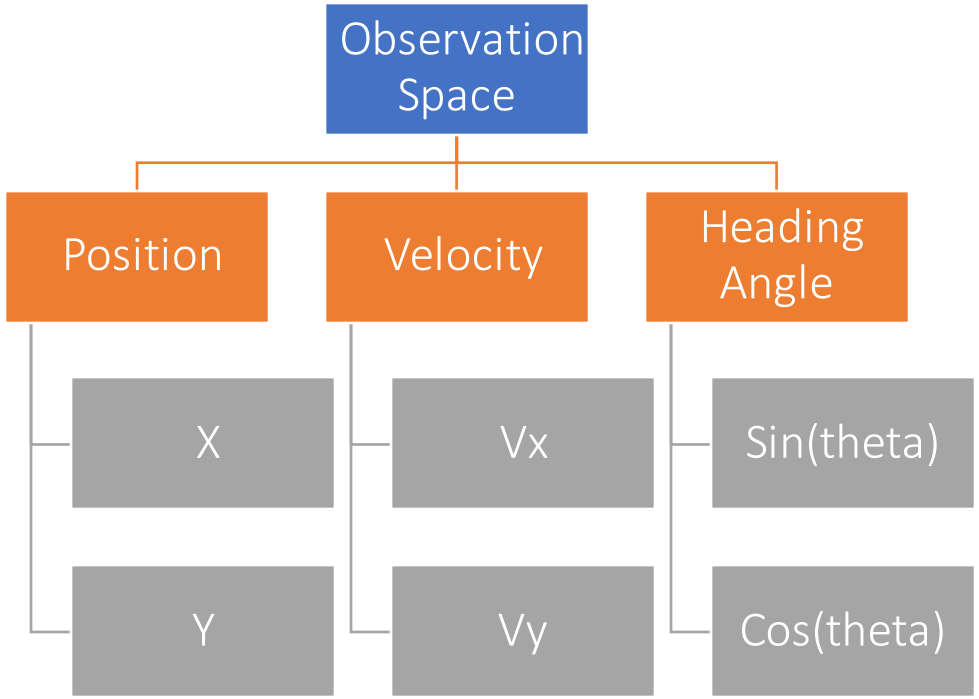


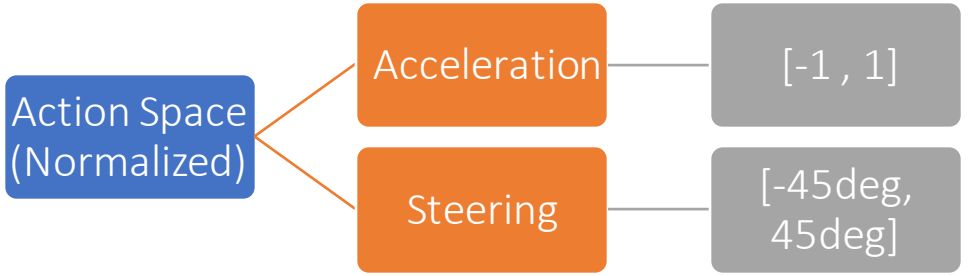
Northeastern University

Problem Statement & Environment

- **Objective:** To park the Agent in the highlighted parking spot in the parallel parking orientation.
  - Continuous action spaces with multiple actions like steering and acceleration pose challenges due to high dimensionality, requiring precise control, complex policy representation, and effective exploration strategies in reinforcement learning tasks.
- **Motivation:**
  - Action Space is identical to the control space of real-world Vehicles.
  - The domain is a stepping stone to understand the research in the continuous action space domain.
- **Rewards:**
  - The Lp norm is calculated between the achieved state and the desired state after multiplying with a reward weights vector.
  - Collision Reward: -5
  - Success Reward: 0.12

$$\ell_p = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}, \text{for } p \geq 1$$





Deep Deterministic Policy Gradient (DDPG)

- **Features:**
  - Actor-Critic with Target Network
  - Continuous Action & Observation Space
  - Maximizes the expected cumulative long-term reward
  - Off Policy Algorithm
  - Replay Buffer
  - Soft Update of Target Networks

$$J_Q = \frac{1}{N}\sum_{i=1}^{N}(r_i + \gamma(1-d)Q_{targ}(s_i', \mu_{targ}(s_i')) - Q(s_i, \mu(s_i))^2$$

$$J_\mu = \frac{1}{N}\sum_{i=1}^{N} Q(s_i, \mu(s_i))$$

$$\theta^{\mu}{}_{targ} \leftarrow \tau\theta^{\mu}{}_{targ} + (1-\tau)\theta^{\mu}$$

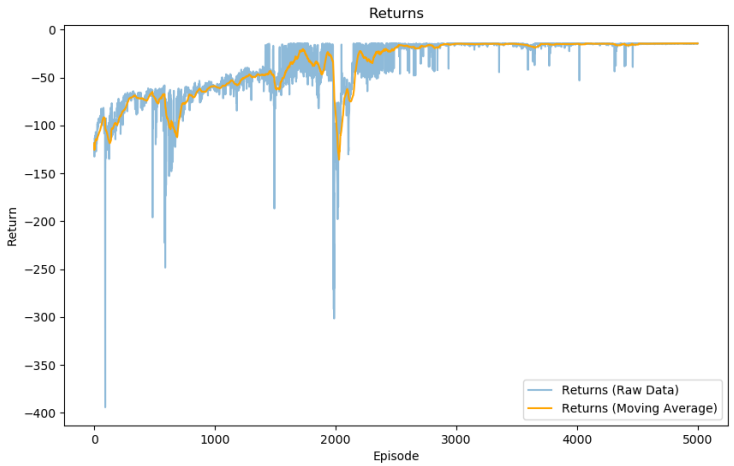
$$\theta^{Q}{}_{targ} \leftarrow \tau\theta^{Q}{}_{targ} + (1-\tau)\theta^{Q}$$

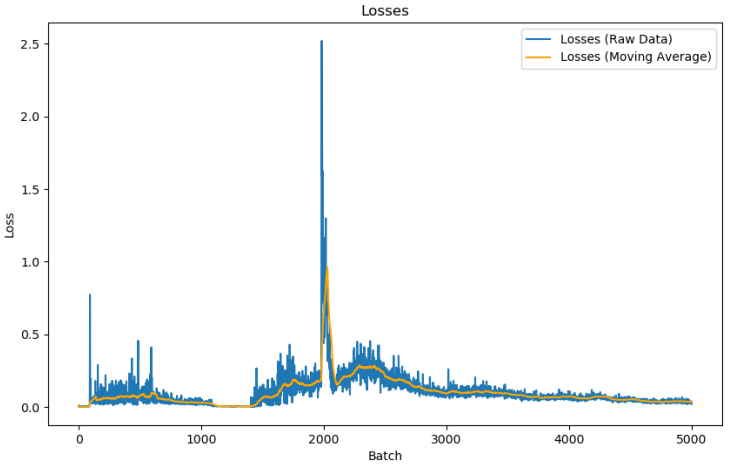
- **Hyperparameters:**
  - OU Noise (mu=0, theta=0.15, sigma=0.2)

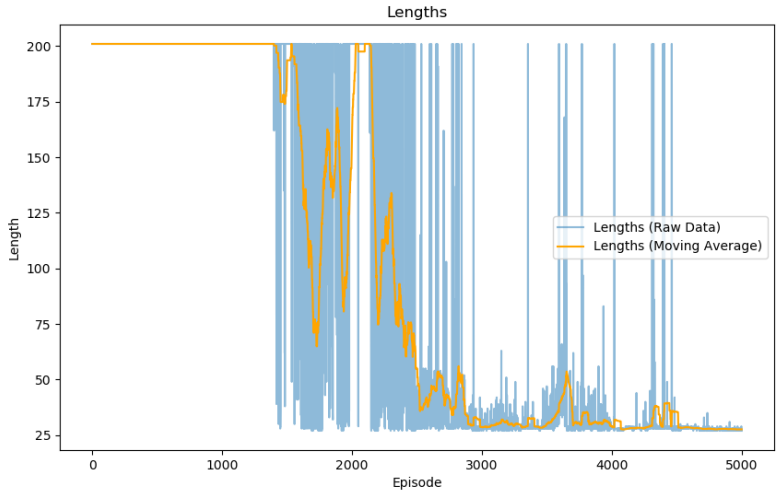
$$dx_t = \theta(\mu - x_t)dt + \sigma dW_t$$

  - Gamma=0.99
  - Tau=0.005

- **Results:**







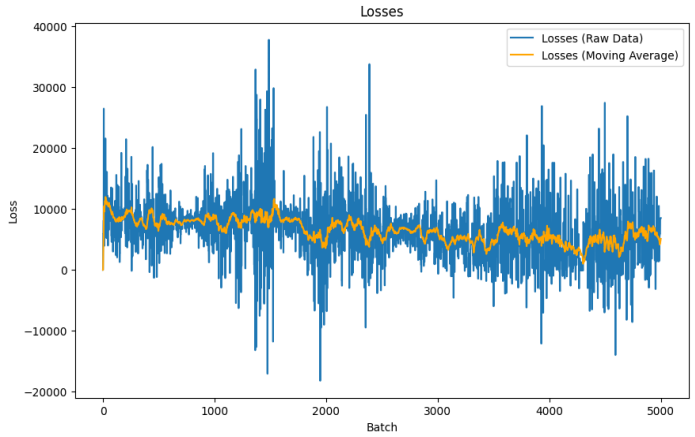
Proximal Policy Optimization (PPO)

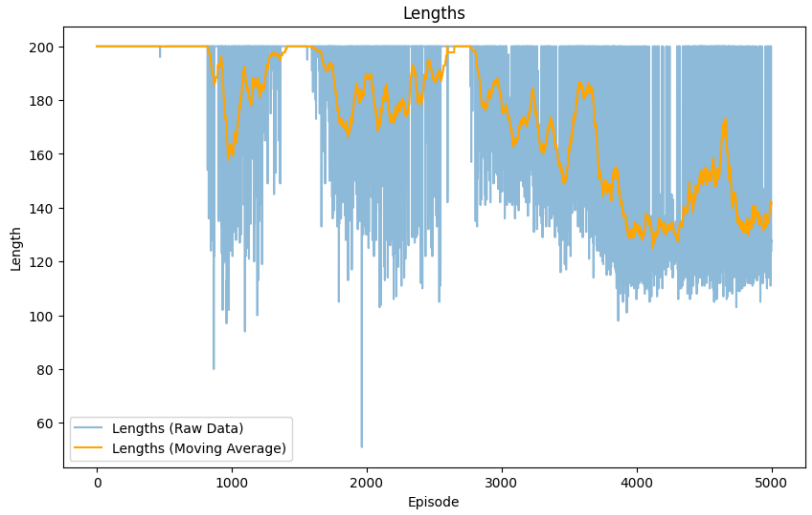
- **Features:**
  - Actor-Critic Structure
  - Continuous Action & Observation Space
  - On Policy Algorithm
  - Replay Buffer
  - Clipped Objective function.

$$\mathcal{L}_{\theta_k}^{CLIP}(\theta) = \mathop{\mathrm{E}}_{\tau \sim \pi_k}\left[\sum_{t=0}^{T}\left[\min(r_t(\theta)\hat{A}_t^{\pi_k}, \mathrm{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t^{\pi_k})\right]\right]$$

- **Hyperparameters:**
  - K_epochs = 40 (update policy for K epochs)
  - eps_clip = 0.2 (clip parameter for PPO)
  - gamma = 0.99 (discount factor)
  - lr_actor = 0.0003 (learning rate for actor network)
  - lr_critic = 0.001 (learning rate for critic network)

- **Results:**





Soft Actor Critic (SAC) [object File]

- **Features:**
  - 1 Actor 2 Critic Network Structure
  - Continuous Action & Observation Space
  - Maximize the entropy of the policy
  - Off Policy Algorithm
  - Replay Buffer
  - Soft Update of Target Networks

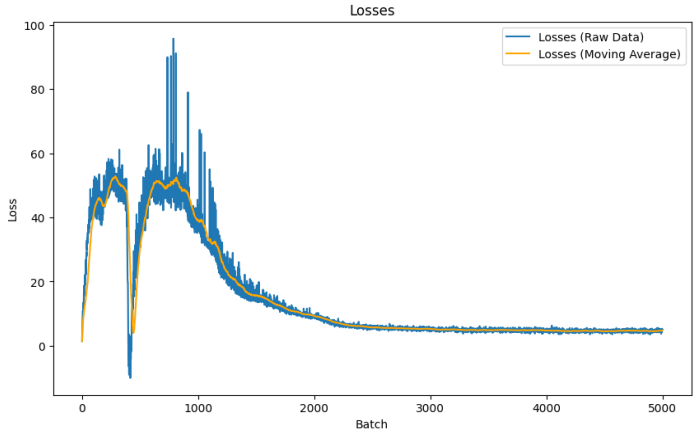
$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\mathbf{s}_t)) \right].$$

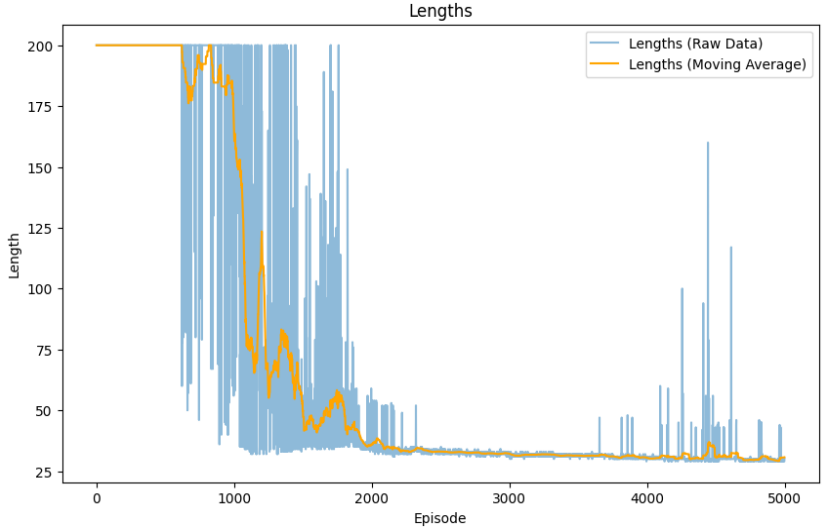
$$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(\mathbf{s}_t) \left(V_\psi(\mathbf{s}_t) - Q_\theta(\mathbf{s}_t, \mathbf{a}_t) + \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) \right)$$

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right]$$

- **Hyperparameters:**
  - alpha = 0.2
  - gamma = 0.99
  - tau = 0.005
  - Learning rate= 0.0003
  - batch_size = 64

- **Results:**





Conclusion

[object File]

- DDPG performs the best amongst the implemented algorithms.
- PPO has decent performance that gives near-optimal results.
- Feature Association was implemented but it gave sub-optimal results.
- SAC converges slower than others but has good performance.
- Future Work:
  - Try to improve the results in Feature Association Model.
  - Apply the implementations to similar domains in our fields of research interests.