R02/Sarthak Mittal/200050129

January 11, 2023

The paper demonstrates the possibilities and challenges of using **deep reinforcement learning (DRL)** to control **complex dynamical systems** in domains where agents must respect imprecisely specified **human norms**. They show the capabilities of their agent using Gran Turismo game **(automobile racing)**, having real-time non-linear control challenges and complex multi-agent interactions. They chose this domain because highly realistic **simulations** exist.

They claim that racing at the absolute limits of friction leaves little margin for **errors**, and other drivers put even greater demands on **accuracy**. The past approaches have achieved success in solo driving or simple passing scenarios, but not racing at the highest levels. They classify success of racers into **4 skills - control** (understanding of dynamics of vehicle and idiosyncrasies of track), **tactics** (pass and defend against opponents and execute precise manoeuvres), **etiquette** (imprecisely specified sportsmanship rules) and **strategy** (modelling opponents to attempt a pass). Although lacking strategist skill, their model-free off-policy DRL agent, **GT Sophy**, achieved notable advances in first 3 skills.

They made GT Sophy control up to **20** cars, and used a GPU machine to asynchronously update neural networks of the **10-20** PlayStations. The core actions were mapped to two continuous-valued dimensions: **changing velocity** (accelerating or braking) and **steering** (left or right). They encoded the approaching course segment as **60** equally spaced **3D** points along each edge and centre line. They fetched state information of the racer and all opponents through an **API** and maintained **proximity** lists of opponents. They trained GT Sophy using a new DRL algorithm they call **Quantile Regression Soft Actor-Critic (QR-SAC)**. This approach learns a **policy (actor)** on the basis of the agent's observations and has a **value function (critic)** that estimates the future rewards. QR-SAC handles **N-step returns** and uses a representation of the **probability distributions** of future rewards.

The reward function was a hand-tuned **linear combination** of reward components computed on the transition between the previous state s and current state s'. The reward components were: course progress (Rcp), off-course penalty (RsocorRloc), wallpenalty (Rw), tyre-slip penalty (Rts), passing bonus (Rps), any-collision penalty (Rc), rear-end penalty (Rr) and unsporting-collision penalty (Ruc). Not surprisingly, the hardest chicanes and corners are the places where GT Sophy has the **biggest performance gains**.

The progress reward and penalties helped GT Sophy learn to get around the track in only a few hours and be faster than **95%** of the humans in their reference dataset. As in previous work, adding rewards specifically for passing helped the agent learn to **overtake** other cars. Although, there is a requirement of **human judges**, whose judgements incorporate a lot of **context**. They experimented with several approaches to encode etiquette as **instantaneous penalties** on the basis of situational analysis of the collisions. They opted for a conservative approach that penalized the agent for **any collision** in which it was involved in, with some extra penalties if the collision was likely considered **unacceptable**.

The straightforward application of **self-play** was inadequate in this setting. By racing against only copies of itself, the agent was ill-prepared for the **imprecision** it would see with human opponents. This feature of racing, that one player's sub-optimal choice causes the other player to be penalized, is not a feature of zero-sum games such as Go and Chess. Hence they used a **mixed-population** of opponents, including agents curated from previous experiments and the game's (relatively slower) built-in AI. In addition, there was the **exposure problem**: certain states are inaccessible without opponent cooperation. So they developed a process that they called **mixed-scenario** training, where they configured scenarios presenting noisy variations of critical situations. They used a form of **stratified sampling** to ensure that situational diversity was present throughout training.

To evaluate GT Sophy, they raced the agent in two events against top GT drivers. The first event was on **2 July 2021** and involved both time-trial and head-to-head races. Although the human drivers were allowed to see a 'ghost' of GT Sophy as they drove around the track, GT Sophy won all three time-trial matches, while the human drivers won the team event on 2 July 2021 by a score of **86-70**. They examined GT Sophy's performance, improved the training regime, increased the network size, made small modifications to some features and rewards and improved the population of opponents. Following that, GT Sophy handily won the rematch held on **21 October 2021** by an overall team score of **104-52**.

One of the advantages of using deep RL to develop a racing agent is that it eliminates the need to program how and when to execute the skills needed - as long as it is exposed to the right conditions, the agent learns to do the right thing by trial and error. Although GT Sophy demonstrated enough tactical skill to beat expert humans in head-to-head racing, there are many areas for improvement, particularly in **strategic decision-making**.

The success of GT Sophy in this environment shows, for the **first time**, that it is possible to train AI agents that are better than the top human **racers** across a range of car and track types. This result can be seen as another important step in the continuing progression of competitive tasks that computers can beat the **very best people** at, such as Chess, Go, Jeopardy, Poker and StarCraft. In the context of previous landmarks of this kind, GT Sophy is the **first** that deals with head-to-head, competitive, high-speed racing, which requires **advanced tactics** and subtle **sportsmanship** considerations.