# Reinforcement Learning with Hard Constraints for Autonomous Driving: CS234 project milestone

**Vaishali Kulkarni** [*1] **Philippe Weingertner** [*1]
**Ramtin Keramati (Project Mentor)** [1]

## 1. Introduction

Reinforcement Learning (RL) has demonstrated its capability to learn efficient strategies on many different and complex tasks. In particular, in games like chess and go, the best human players have lost against RL algorithms (Silver et al. [9]). There is a growing traction for applying such RL algorithms to complex robotics tasks like Autonomous Driving. Nevertheless with Autonomous Driving we are dealing with additional challenges. We are in a partially observable environment where enforcing safety is of paramount importance. As a consequence, considering safety via a reward and the optimization of a statistical criteria is not sufficient. Hard Constraints have to be enforced all the time. We propose to study how the RL optimization criteria can be modified to deal with hard constraints; how algorithms like DQN could be modified to cope with such hard constraints and more generally how an RL agent could be integrated in a Decision Making module for Autonomous Driving to provide efficient and scalable strategies while still providing safety guarantees. So we propose to address the following problem formulation:

$$\max_{\theta} \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi_\theta(s_t))\right]$$

$$\text{s.t. lower\_bound}(C_i(s_t, a_t)) \geq \text{Margin}_i \; \forall i \in [\![1, K]\!]$$

where the expectation corresponds to the statistical RL objective subject to a set of safety constraints.

We tackle the problem of safe control in physical systems where certain quantities have to be kept constrained. For an autonomous vehicle we must always maintain its distance from obstacles above some margin. But in fact the real state of the world is only partially observable and the Driving Models of surrounding cars are not known exactly: so we are dealing with uncertainty and our constraints are actually a set of random variables $C_i$ which we want to lower bound. Note that in most of the references the constraints are only considered in expectation via a constraint of type $J^{\pi}_{C_i} = E_{\pi}\left[C_i(s,a)\right]$ whereas here we are interested in enforcing stronger constraints.

**Main Project Milestone Updates:**

- We will define a robust MDP objective taking into account model uncertainty (related to the uncertainty of the prediction of vehicles trajectories) to derive a safer RL algorithm. The robust MDP setting has been so far mainly considered in planning with RL model based approaches. One notable exception is Tamar, Mannor, and Xu [10].

- We will use a single hard constraint which is related to the min Time To Collision that is computed between the ego vehicle and the predicted trajectories of the surrounding vehicles. This min TTC shall be greater or equal than a threshold margin.

## 2. Background/Related Work

In Mirchevska et al. [8] a DQN network is used for tactical decision making in an autonomous driving pipeline but the DQN algorithm itself is not modified to handle hard constraints and the safety is guaranteed by checking the output of the RL algorithm. Our objective here, in contrast, would be to have an RL algorithm that is directly dealing with hard constraints to avoid frequent and sub-optimal actions masking. A review of the different safe RL techniques has been done in García and Fernández [6]. Some techniques mainly deal with soft constraints by either reshaping the reward or trying to minimize the variance related to the risk of making unsafe decisions, while other try to handle hard constraints. Garcia et al. have analyzed and categorized safe RL techniques in two families of approaches: one consists in modifying the exploration process while the other consists in modifying the optimality criterion. In Leurent

[*]Equal contribution [1]Department of Computer Science, Stanford University, California, USA. Correspondence to: Vaishali Kulkarni <vaishali@stanford.edu>, Philippe Weingertner <pweinger@stanford.edu>, Ramtin Keramati <keramati@stanford.edu>.

*Proceedings of the $35^{th}$ International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. 

et al. [7] the RL objective is replaced by a surrogate objective which captures hard constraints and handles model uncertainty by defining a lower bound of the expectation objective. In Achiam et al. [1] constrained policy optimization is solved with a modified trust-region policy gradient. The algorithm's update rule projects the policy to a safe feasibility set in each iteration. But the policy is kept within constraints only in expectation. In Dalal et al. [5] they directly add to the policy a safety layer that analytically solves an action correction formulation per each state. This safety layer is learned beforehand but is approximated by a first order linear approximator. In Tessler, Mankowitz, and Mannor [11] and in Bohez et al. [2] the proposed approaches are completely in line with our objective here: modifying the RL objective such that it deals directly with hard constraints. But there is no closed form solution for such a problem and a Lagrangian relaxation technique is used for solving the constrained optimization problem. Given a Constrained Markov Decision Process (CMDP), the unconstrained problem is transformed to $\min_{\lambda \geq 0} \max_{\theta} L(\lambda, \theta) = \min_{\lambda \geq 0} \max_{\theta} [J_R^{\pi_\theta} - \lambda(J_C^{\pi_\theta} - \alpha)]$ where $L$ is the Lagrangian and $\lambda$ the Lagrange multiplier (a penalty coefficient). We propose to study how such techniques could be applied to the Decision Making process of an Autonomous Driving pipeline and we will benchmark different RL algorithms, modified to cope with hard constraints, in an Anti Collision Tests setting.

## 3. Approach

A Policy Gradient network will be used as a baseline. We will consider 3 type of modifications. From the most simple, conceptually, to the most complex, we will:

1. Propose a post RL safety check. We want to override the actions that are considered unsafe based on some pre-defined rules. This type of approach is used in a paper from BMW Group by Mirchevska et al. [8].

2. Modify the exploration process so that only safe actions are explored. Similar to Bouton et al. [3] the idea is to derive an exploration strategy that constrains the agent to choose among actions that satisfy safety criteria. Hence the search space of policies is restricted to a "safe" or safer subspace of policies.

3. Replace the RL objective $\max_\theta \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi_\theta(s_t))\right]$ by an objective taking into account hard constraints

$$\max_{\theta} \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi_\theta(s_t))\right]$$

$$\text{s.t. lower\_bound}\left(C_i(s_t, a_t)\right) \geq \text{Margin}_i \; \forall i \in [\![1, K]\!]$$

and study how RL algorithms should be modified to account for this new objective. In a recent paper from DeepMind from Bohez et al. [2], this type of approach is applied to a realistic, energy-optimized robotic locomotion task, using the Minitaur quadruped developed by Ghost Robotics. But the constraints were considered only in expectation.

**Main Project Milestone Updates:**

- We will focus on step 3. We decided to use Policy Gradient algorithms as they can handle the continuous action problems applicable to autonomous driving. We will use the actor critic to optimize the objective function in step 3 such that a robust objective taking into account model uncertainty plus a constraint bound (to avoid collisions) are satisfied. To enforce the hard constraint, we will explore algorithms along the lines of (either or both, to be refined) Lagrangian methods for CMDP (Constrained MDP) or TRPO (or PPO) where constraints threshold are checked at every parameter $\theta$ update. We will study the asymptotic policy convergence in the presence of the hard constraints using Policy Gradient algorithm.

- The surrounding vehicles will move according to an IDM driving model. This driving model depends on 5 parameters. The ego vehicle does not know (precisely) the parameters used by the surrounding vehicles. This is the main source of uncertainty.

- In a similar way to Leurent et al. [7], section 3, we will define a surrogate lower bound objective by bounding the collected rewards by their minimum over the interval hull of $S(t)$ which is the smallest convex set containing the union of reachability sets. Concretely we want to capture the minimum Time To Collision and account for interval uncertainty when predicting trajectories and hence Time To Collision with other vehicles. In reality we are not dealing with a MDP problem, but with a POMDP (Partially Observable MDP) problem and we want to define a more robust MDP objective taking into account the level of uncertainties related to future trajectories.

- The hard constraint will be min TTC $\geq 10$ where the Time To Collision is computed based on:
  - $\text{acceleration}_{\text{ego-vehicle}} = \pi_\theta(s_t)$ which depends on the parameters of the Neural Network
  - Predicted trajectories of surrounding vehicles which depend on interval estimates of 5 IDM parameters per surrounding vehicle.
  - So the hard constraint will involve the parameters $\theta$ of $\pi_\theta(s_t)$ such that min Time To Collision is above some threshold.

- We may leverage on the following code tensorflow_constrained_optimization from Google research.

## 4. Experimental Results

**Main Project Milestone Updates:**

- The legacy simulator code has been ported from Julia to Python (for use in CS234 Project)
- The simulator code (cf github CS234_Project) has been upgraded for IDM driving model. In previous CS238 project a simple Constant Velocity model was used.

### 4.1. Simulator

We have upgraded the Anti Collision Tests environment, ACT, developed for a previous CS238 Stanford project. A vehicle has to drive from a point A to a point B as fast as possible, while avoiding other vehicles that are crossing its path and trying to minimize hard braking decisions. It is a multi-objectives task where efficiency, comfort and safety have to be optimized. We use Intelligent Driver Models, IDM, depending on 5 parameters unknown to the ego vehicle, and different per vehicle. So it is a model-free setting: we do not know the model of the environment, the driving models of others, and we would like to learn to drive efficiently and safely in this context.

### 4.2. Evaluation Metrics

In order to measure success and benchmark different versions of the algorithms, we will use 3 metrics: a safety metric (percentage of collisions), an efficiency metric (time to goal), and a comfort metric (number of hard braking decisions or jerk). We want to enforce safety while not compromising efficiency or comfort.

## 5. Milestones and Next Steps

- Python code for ACT test framework with IDM driving models (February 25: DONE)
- Establish a baseline with RL Policy Gradient algorithm trained on ACT simulator framework (March 5 ?)
- Experiment with Policy Gradient and a Hard Constraint on Mujoco Cheetah environment; upgrading Assignment 3 (First results March 5 ?)
- Meeting to be planned $5^{th}$ or $6^{th}$ of March with Vaishali, Ramtin, Philippe. Philippe who is located in France will be in California these days.
- Complete code for robust MDP objective with Policy Gradient on ACT test framework (March 12 ?)
- Complete code to handle Hard Constraint. Penalty methods with PG and either Lagrangian or CPO with TRPO or PPO (March 12 ?)
- Combine Policy Gradient algorithm with Robust MDP objective and min TTC Hard Constraint in ACT test framework (March 16)
- Results analysis and writeup (March 20)

## References


[1] Joshua Achiam et al. "Constrained Policy Optimization". In: (2017). URL: http://arxiv.org/abs/1705.10528.

[2] Steven Bohez et al. *Success at any cost: value constrained model-free continuous control*. 2019. URL: https://openreview.net/forum?id=rJlJ-2CqtX.

[3] Maxime Bouton et al. "Reinforcement learning with probabilistic guarantees for autonomous driving". In: *Conference on Uncertainty in Artifical Intelligence (UAI)*. 2018. URL: https://drive.google.com/open?id=1d2tl4f6GQgH1SERveTmPAR42bMVwHZAZ.

[4] Yinlam Chow et al. "A Lyapunov-based Approach to Safe Reinforcement Learning". In: *CoRR* abs/1805.07708 (2018).

[5] Gal Dalal et al. "Safe Exploration in Continuous Action Spaces". In: (2018). URL: http://arxiv.org/abs/1801.08757.

[6] Javier García and Fernando Fernández. "A Comprehensive Survey on Safe Reinforcement Learning". In: *Journal of Machine Learning Research* (2015). URL: http://jmlr.org/papers/v16/garcia15a.html.

[7] Edouard Leurent et al. "Approximate Robust Control of Uncertain Dynamical Systems". In: 2018.

[8] Branka Mirchevska et al. "High-level Decision Making for Safe and Reasonable Autonomous Lane Changing using Reinforcement Learning". In: 2018. DOI: 10.1109/ITSC.2018.8569448.

[9] David Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* (2017). URL: http://dx.doi.org/10.1038/nature24270.

[10] Aviv Tamar, Shie Mannor, and Huan Xu. "Scaling Up Robust MDPs Using Function Approximation". In: *Conference on Machine Learning*. ICML'14. Beijing, China, 2014. URL: http://dl.acm.org/citation.cfm?id=3044805.3044913.

[11] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. "Reward Constrained Policy Optimization". In: abs/1805.11074 (2018). URL: http://arxiv.org/abs/1805.11074.