Safe Reinforcement Learning Framework Under a Linear Programming Formulation

Reinforcement learning (RL) has achieved remarkable progress in recent years, driven largely by advances in deep learning. However, as RL agents are increasingly deployed in safety-critical domains such as robotics and healthcare, the need to ensure safety has received limited attention. Standard RL formulations focus on maximising expected cumulative reward, but real-world environments rarely reduce to a single objective. Instead, agents must balance reward maximisation and constraint satisfaction, such as avoiding unsafe states or limiting exposure to risk in hazardous environments. A lot of recent safe RL methods modify state-of-the-art RL algorithms for the unconstrained setting to a Lagrangian version. This entails iteratively updating the Lagrangian multiplier as a weighting factor for the constraints and then optimising a policy over a weighted combination of reward maximisation and constraint reduction. However, their drawbacks are not sustainable for safety-demanding applications: they lack stability in performance and require extensive hyperparamater fine-tuning. One cause of these problems stem from the non-convex optimisation over policy parameters, often resulting in instability, hyperparameter sensitivity, and limited convergence guarantees.

Instead of optimising the RL objective over the policy space, an alternative perspective rephrases the RL problem as a linear program (LP), where we now optimise over the occupancy measure $\mu(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_\pi(s = s_t, a = a_t)$, for a given state-action pair (s,a) and γ being the discount factor. Intuitively, the occupancy measure can be seen as a long-term visitation measure of a policy π for a given state-action pair. In this convex optimisation framework, the objective is to maximise the (discounted) expected return with respect to the occupancy measure, subject to linear constraints that preserve the flow of the occupancy measure between state-action pairs.

Most of the recent work relating to LP-based RL focuses on the theoretical properties of such algorithms in the unconstrained setting. Our contribution lies in the investigation of the empirical benefits of LP-based RL for the constraint setting, two aspects which have been underexplored in existing literature. In a very minimal and modular way, we can add safety constraints that are also linear with respect to the occupancy measure. This yields a constrained optimisation problem that remains convex and linear, significantly simplifying the optimisation process.

To test our approach, we propose a simple stochastic primal-dual algorithm that alternates between updates of the primal and dual variables based on samples from the environment. The Lagrangian multiplier is learned as part of the dual variables and remains an interpretable signal of constraint importance, revealing the trade-off between performance and safety. Preliminary experiments validate this formulation on various tasks representing reward-based and state constraints but also less explored directions such as constraints on risk-preferences.

As seen in Figure 1, our method consistently learns behaviour that respects safety constraints, which in this case equates to reaching a constraint value of 0, while achieving near-optimal returns. In comparison, popular baselines converge to a constraint-violating policy and exhibit unstable learning. Moreover, we find that our LP-based approach is significantly more robustness to hyperparameter initialisation. To show this, we train our method and PPOLag, a popular baseline, on different hyper-parameter initialisations. We then compare the distribution of their return and constraint values across initialisations. For the constraint values, the interquartile range for PPOLag spans from 2 to 3. In contrast, the interquartile range from our approach ranges from 0.000 to 0.031. A similar dispersion pattern can be observed for the return. This indicates that our approach addresses the limitation of hyperparameter sensitivity of primal-dual methods in constrained settings.

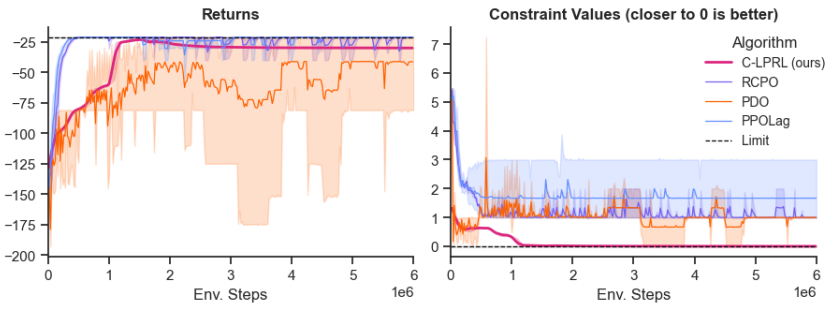

Figure 1: Performance on an obstacle navigation task with 3 obstacles. For clarity, only the best performing algorithms are visualised.

These results yield a promising proof of concept that we want to build on. Our study highlights that safety constraints can be incorporated into RL in a natural and modular way under the LP framework, preserving interpretability and theoretical soundness. We believe that this perspective opens the door to a broader class of safe RL algorithms which encapsulate an even wider range of constraint types which are both practical and reliable, and we currently extend this line of work to more complex domains and settings.