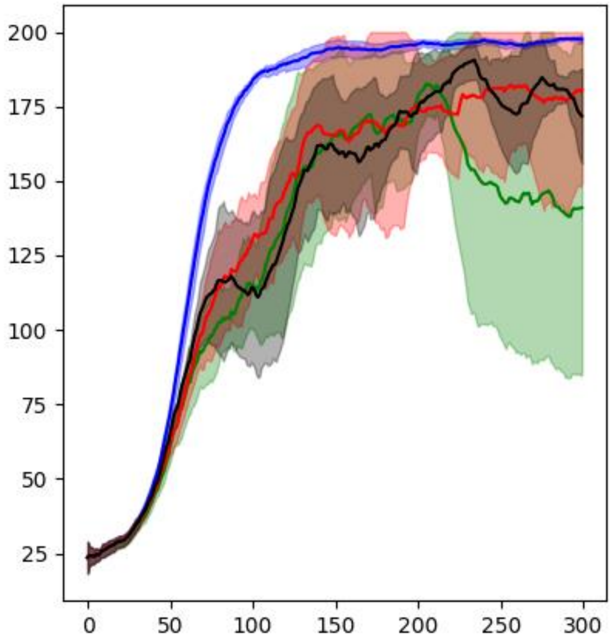

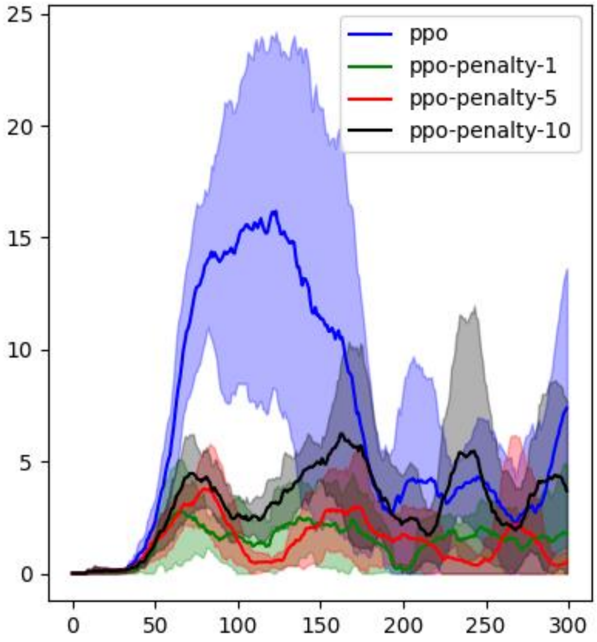

The left picture is the episodic reward graph to compare PPO against various PPO penalties (Beta = 1, 5, 10). The right picture is the episodic constraint graph.

In left picture, we can see that for ppo-penalties, averaged rewards are lower compared to ppo. This is because for ppo-penalties, constraints under cost are taken into account.

From the episodic constraint graph (graph on the right side), we see that the various ppo-penalty curves hover between average cumulative constraint value of 0 to 5, whereas PPO initially goes way over the constraint value for the initial episodes. As ppo does not take into account the constraints, ppo costs goes way over the hovering thresholds maintained by ppo-penalty graphs.

We can see with lower Beta values, the average episodic rewards are lower compared to higher beta values and ppo graph. To explain this, let’s revisit the constraint RL goal:

Goal: $\pi^* = argmax_\pi E_\pi\left[\sum_{t=0}^{h} \gamma^t r_t\right]$

such that $\mathrm{E}_\pi\left[\sum_{t=0}^{h} \gamma^t c_t^{(i)}\right] \leq \beta_i \forall i$

Our goal now is to select a policy π that maximizes the average rewards, while also keeping the average cost under specified Beta values. When beta is low, this means we have to maintain a lower average costs. This affects the average rewards more severely compared to higher beta, because now we have a tighter bound on average costs.

This is the reason why in the right picture, ppo-penalty with beta value 1 is below higher beta values for average costs. Because Beta value is low, the average cost has to be lowered compared to ppo and ppo-penalties of 5 and 10. As beta values are increased (5 and 10), we see improvement in rewards as we are relaxing the constraints by allowing it to be costlier. And with ppo, as we are not imposing any penalty on costs – in can obtain the best average rewards at the expense of very high cost.