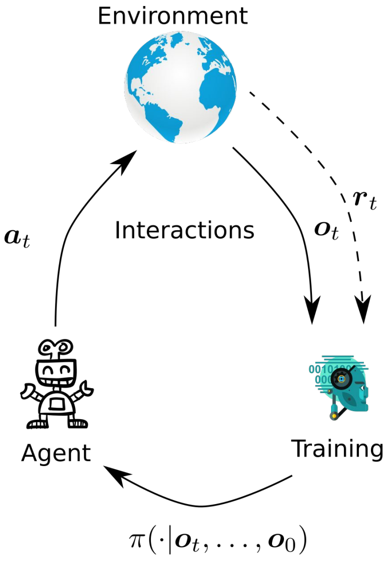
Effects of Safety State Augmentation on Safe Exploration

Aivar Sootla, Alexander Cowen-Rivers, Jun Wang and Haitham Bou-Ammar

Safe Reinforcement learning



$$\max_{\pi} \mathbb{E}_s^{\pi} J_{\text{task}}, J_{\text{task}} \triangleq \sum_{t=0}^{\infty} \gamma_r^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$$

$$\text{s. t.: } g(J_{\text{safety}}) \geq 0, J_{\text{safety}} \triangleq d - \sum_{t=0}^{\infty} \gamma_l^t \boldsymbol{l}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$$

We can model different constraints:

1. Fuel constraints,
2. Physical constraints for damage prevention
3. Comfort constraints

Unresolved problems:

1. Safe deployment
2. Safety guarantees after training
3. **Safety during training**

Two problems:

$$\max_{\pi} \ \mathbb{E}_{\boldsymbol{s}}^{\pi} J_{\text{task}}, \ J_{\text{task}} \triangleq \sum_{t=0}^{\infty} \gamma_r^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$$

$$\text{s. t.: } g(J_{\text{safety}}) \geq 0, \ J_{\text{safety}} \triangleq d - \sum_{t=0}^{\infty} \gamma_l^t \boldsymbol{l}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$$

$$g_{\text{po}}(z) = \mathbb{P}(z \geq 0) - 1$$

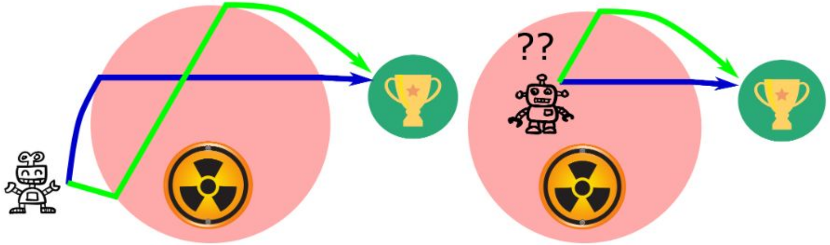
Safety with probability one

$$g_{\text{av}}(z) = \mathbb{E} z$$

Safety on average

Key observations

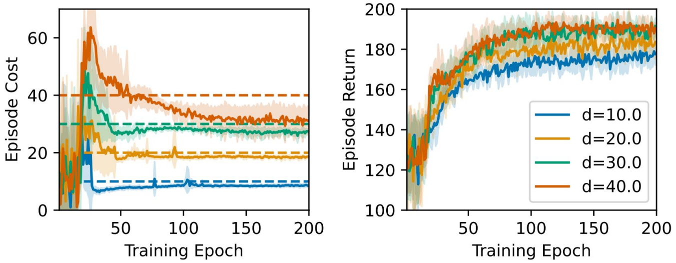
Safety constraint tracking is required:



Solution: Safety state augmentation

$$z_t = \gamma^{-t}(d - \sum_{k=0}^{t-1} \gamma_l^k l(\boldsymbol{s}_k, \boldsymbol{a}_k, \boldsymbol{s}_{k+1}))$$

Initial safety budget leads to different safety violations, e.g. in safe pendulum:



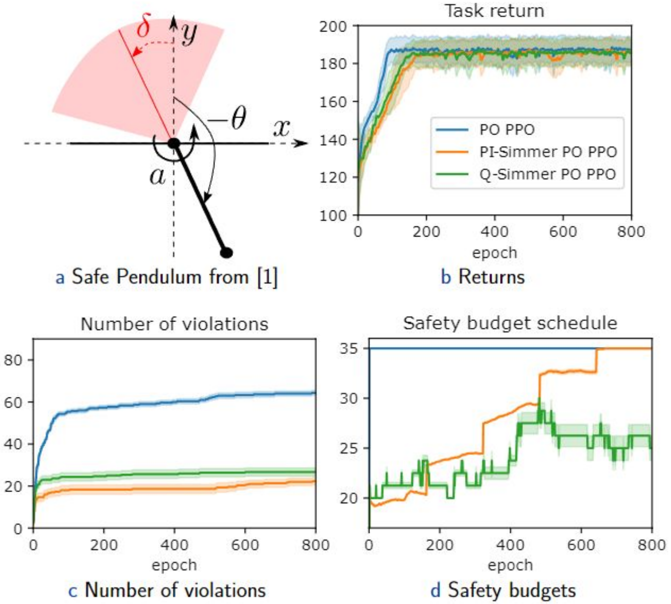
Solution: Schedule initial safety budget

Safety with probability one

$$\max_{\boldsymbol{u}_k\in[-\delta \boldsymbol{d},\delta \boldsymbol{d}]} \; -\sum_k \mathrm{relu}\left(-\hat{g}_{\mathcal{D}_k}(\boldsymbol{z}_T^k(\boldsymbol{d}^{\text{target}}))\right),$$
$$\boldsymbol{d}_{k+1} = \mathrm{clip}\left(\boldsymbol{d}_k + \boldsymbol{u}_k, \boldsymbol{d}^{\text{start}}, \boldsymbol{d}^{\text{target}}\right), \boldsymbol{d}_0 = \boldsymbol{d}^{\text{start}}$$

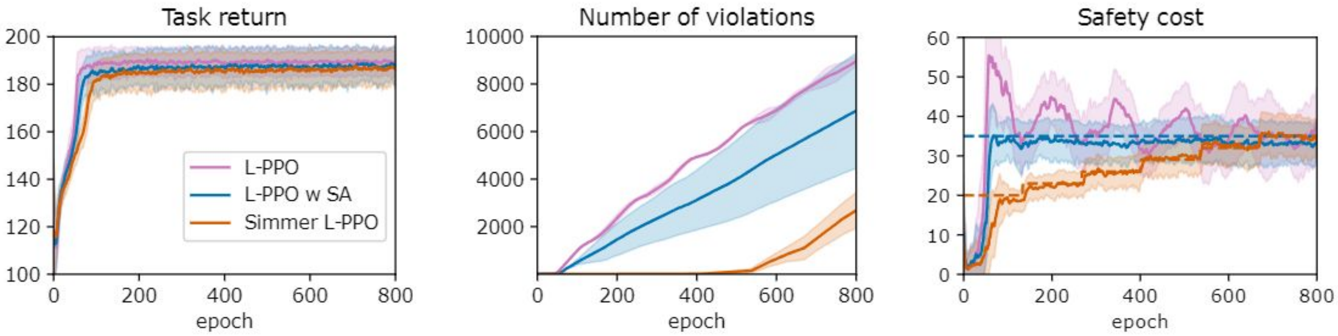
This is a hard to solve bilevel RL problem and we derive two heuristics:

a) PI controller based
b) Q learning based



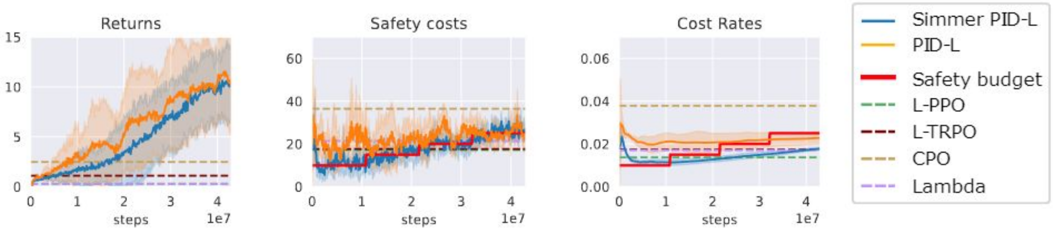
a Safe Pendulum from [1]
b Returns
c Number of violations
d Safety budgets

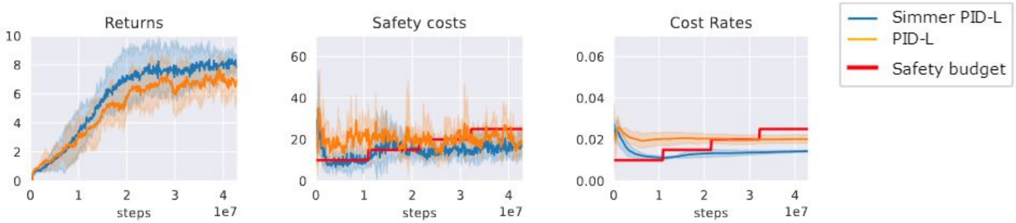
Safety on average for safe pendulum



Naive scheduling works!

Safety on average for point and car push





Conclusion

1. We argue that the optimal policy must depend on the safety state to improve safety;

2. Safety state augmentation and simmering show superior performance on pendulum swing-up and safety gym tasks for average constrained problems

3. Simmering RL algorithms with probability one constraints can significantly reduce safety violations during training in an online fashion

4. Please see our paper for unabridged quantitative results and further experiments of safety gym environments. Parts of the illustrations are downloaded from http://www.freepik.com