Do Androids Dream of Electric Fences?

Safe Reinforcement Learning with Imagination-Based Agents

Peter He, Borja Gonzalez Leon, Francesco Belardinelli
Peter He
Imperial College London
University College London
peter.he.21@ucl.ac.uk

Borja Gonzalez Leon
Imperial College London

Francesco Belardinelli
Imperial College London
Coming Up In Today’s Presentation…

SOTA (Unsafe) RL Agent

Our Method

Reinforcement Learning

Environment

Agent

Observation, Reward

Action

Definitions  Background  Approach  Findings  Future
Reinforcement Learning

\[ \mathcal{T} : S \times A \times S \rightarrow [0, 1] \]
\[ \mathcal{R} : S \times A \times S \rightarrow \mathbb{R} \]

Agent

Environment

Observation, Reward

Action

Definitions Background Approach Findings Future
Reinforcement Learning

\[ \mathcal{T} : S \times A \times S \to [0, 1] \]
\[ \mathcal{R} : S \times A \times S \to \mathbb{R} \]

Observation, Reward

Environment

Action

\[ \pi : S \times A \to [0, 1] \]
Reinforcement Learning

\[ T : S \times A \times S \rightarrow [0, 1] \]
\[ R : S \times A \times S \rightarrow \mathbb{R} \]

Agent

Environment

Observation, Reward

Action

\[ \pi : S \times A \rightarrow [0, 1] \]

Maximise

\[ \mathbb{E} \left[ \sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \right] \]
Safety

“Bad things shouldn’t happen”
Safety

“Bad things shouldn’t happen”
Safety

“Bad things shouldn’t happen”

\[ \phi, \phi' ::= true \mid p \mid \neg \phi \mid \phi \land \phi' \mid \Box \phi \mid \phi \cup \phi' \]
Safety

\[ \phi, \phi' ::= \text{true} \mid p \mid \neg \phi \mid \phi \land \phi' \mid \bigcirc \phi \mid \phi \cup \phi' \]

\[ \neg \text{agent\_in\_red\_square} \cup \text{episode\_ended} \]
Safety

\[ \neg \text{agent.in.red.square} \cup \text{episode-ended} \]

```
0
\[ \neg \text{agent.in.red.square} \]
\[ \neg \text{episode-ended} \]

\[ \text{agent.in.red.square} \]

1
\[ \text{agent.in.red.square} \]
\[ \text{episode-ended} \]
\[ \neg \text{agent.in.red.square} \]
\[ \neg \text{episode-ended} \]

2
\[ \text{agent.in.red.square} \]
\[ \text{episode-ended} \]
\[ \neg \text{agent.in.red.square} \]
\[ \neg \text{episode-ended} \]
```
Shielding

Environment <-> Agent <-> Shield

Observation, Reward <-> Safe Action <-> Action

M. Alshiekh et al. Safe Reinforcement Learning via Shielding. 2018
Shielding

M. Alshiekh et al. *Safe Reinforcement Learning via Shielding*, 2018
Shielding

Definitions

Background

Approach

Findings

Future

Given some finite trace \( \rho = s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} \ldots \xrightarrow{a_{n-1}} s_n \).

And the set of all finite traces of length \( H \) from state \( s \), \( \varrho_H(s) \).

A policy \( \pi \) is \( H \)-bounded safe iff.

\[
\forall s \in S. (\exists \rho = (s, a) \in \varrho_H(s) \land S(\rho, \phi) \land \pi(s_0) = a_0) \quad \lor \quad \forall \rho \in \varrho_H(s). \neg S(\rho, \phi)
\]

If there is a safe trace of length \( H \), we take it.
Our Approach

- **Agent**
- **Environment**
- **Action**
- **Observation, Reward**
- **Shield**
- **Safe Action**
- **Learned Model of the Environment**

Definitions, Background, Approach, Findings, Future
Model-Based RL to the Rescue!

1. Learn a model of the environment
2. Learn a policy inside the model of the environment
3. Collect data in the real environment using the learned policy
4. Repeat until convergence
Model-Based RL to the Rescue!

1. Learn a model of the environment
2. Learn the policy using the model of the environment
3. Collect data in the real environment using the learned policy
4. Also use the model of the environment to keep the agent safe
5. Repeat until convergence
Recurrent State Space Model

Definitions

Background

Approach

Findings

Future

$h_t = f(h_{t-1}, z_{t-1}, a_{t-1})$
$z_t \sim q(z_t|h_t, o_t)$
$\hat{z}_t \sim p(\hat{z}_t|h_t)$
$\hat{o}_t \sim p(\hat{o}_t|h_t, z_t)$
$\hat{r}_t \sim p(\hat{r}_t|h_t, z_t)$

Safety
Recurrent
State
Space
Model

Definitions
Background
Approach
Findings
Future
SRSSM with Neural Networks

\[
\begin{align*}
\hat{z}_0 &= D + h_0 \\
\hat{z}_1 &= D + h_1 \\
\hat{z}_2 &= D + h_2 \\
\end{align*}
\]

\[
\begin{align*}
\hat{\lambda}_0 &= \hat{\lambda}_1 \\
\hat{\lambda}_2 &= \hat{\lambda}_1 + \hat{\lambda}_2 \\
\end{align*}
\]

\[
\begin{align*}
\min_{\theta} \mathcal{L}_{\text{model}} &= \mathcal{L}_{\text{observation}} + \mathcal{L}_{\text{reward}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{violation}}
\end{align*}
\]
Latent Shielding

- Environment
- Agent
- Shield
- Observation, Reward
- Safe Action
- Action

Multiple possible futures

Imagination (SRSSM)
Approximate Bounded Prescience

\[ \hat{a}_n = \pi(\hat{s}_{h-1}^n) + \eta \]

Definitions
Background
Approach
Findings
Future
ABP Shielding for Latent Trajectories

\[ \hat{a}_h^n = \pi(\hat{s}_{h-1}^n) + \eta \]

Definitions, Background, Approach, Findings, Future

Noise Term
ABP Shielding for Latent Trajectories

\[ \pi' (s_t) = \begin{cases} \pi(s_t) & \text{if } P(s_t) + 1 \text{ is not a safety violation} \\ \zeta(s_t) & \text{otherwise} \end{cases} \]

Fewer than 14,000,605 * \( \varepsilon \)

Background
Safe RL
Approach
Findings
Future
ABP Shielding for Latent Trajectories
Training an Agent with Latent Shielding

1. Learn a **SRSSM** model of the environment
2. Learn the policy using the model of the environment, assigning a punishment to violation states
3. Collect data in the real environment using the learned policy **with the shield**
4. Repeat until convergence
But It’s Not All Fun and Games…

An inaccurate internal model of the environment can lead to the latent shield hindering exploration!
But It’s Not All Fun and Games…

An inaccurate internal model of the environment can lead to the latent shield hindering exploration
But It’s Not All Fun and Games…

In fact, even bounded prescience shielding can hinder exploration
Shield Introduction Schedules

Potential implementations:
+ A gradually decaying probability of disabling the shield with respect to time
+ Enabling the shield once the change in dynamics model loss falls to below some threshold
+ Simply enabling shielding after a certain number of training episodes have been completed
Performance Evaluation

Seed 1

Seed 2

BPS (Giacobbe et al.)

Latent Shield (ours)

Baseline (Hafner et al.)

Performance Evaluation


Static Gridworld

(see paper for MORE graphs)
## Performance Evaluation

<table>
<thead>
<tr>
<th>Flavour</th>
<th>Metric</th>
<th>Latent</th>
<th>Unshielded</th>
<th>BPS</th>
<th>CPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>Testing Reward</td>
<td>15067 (434)</td>
<td>13148 (249)</td>
<td>12468 (620)</td>
<td>-2925 (1065)</td>
</tr>
<tr>
<td></td>
<td>Training Violations</td>
<td>0.30 (0.76)</td>
<td>2.25 (1.60)</td>
<td>0 (0)</td>
<td>13.43 (19.25)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1262 (172)</td>
<td>2306 (833)</td>
<td>0 (0)</td>
<td>16455 (1435)</td>
</tr>
<tr>
<td>Procedural</td>
<td>Testing Reward</td>
<td>8084 (2221)</td>
<td>6825 (1427)</td>
<td>1938 (3552)</td>
<td>-1588 (2051)</td>
</tr>
<tr>
<td></td>
<td>Training Violations</td>
<td>4.50 (3.59)</td>
<td>33.7 (16.28)</td>
<td>0 (0)</td>
<td>19.60 (13.83)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14018 (1852)</td>
<td>15309 (4686)</td>
<td>0 (0)</td>
<td>18705 (3756)</td>
</tr>
<tr>
<td>Cliff</td>
<td>Testing Reward</td>
<td>8.57 (2.96)</td>
<td>10.76 (3.29)</td>
<td>10.50 (3.28)</td>
<td>7.56 (2.86)</td>
</tr>
<tr>
<td>$p_{stick} = 0.1$</td>
<td>Training Violations</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>3.40 (1.91)</td>
</tr>
<tr>
<td></td>
<td>Training Violations</td>
<td>58.2 (9.60)</td>
<td>90.0 (9.10)</td>
<td>24.0 (13.02)</td>
<td>973.0 (357.7)</td>
</tr>
<tr>
<td>Driver</td>
<td>Testing Reward</td>
<td>8.10 (4.99)</td>
<td>6.63 (8.07)</td>
<td>7.10 (9.52)</td>
<td>6.44 (3.00)</td>
</tr>
<tr>
<td>$p_{stick} = 0.5$</td>
<td>Training Violations</td>
<td>0.18 (0.84)</td>
<td>0.54 (1.53)</td>
<td>0.22 (1.18)</td>
<td>0.48 (1.24)</td>
</tr>
<tr>
<td></td>
<td>Training Violations</td>
<td>91.8 (16.85)</td>
<td>157.6 (18.4)</td>
<td>80.4 (17.43)</td>
<td>3126 (2823)</td>
</tr>
</tbody>
</table>

Examining Latent Dynamics

BPS

Dreamer

Ours
Open Questions

+ What’s the best Shield Introduction Schedule?
+ How might we leverage uncertainty?
+ How might we leverage offline pre-training?
Takeaways

+ **Latent shielding** lets you shield agents in high-dimensional environments without knowledge of the dynamics *a priori*.

+ It does this by learning the environment model rather than having it be handcrafted.

+ Shielding can harm model-based DRL algorithms - introduce the shield gently with a **Shield Introduction Schedule**.