Theory and Applications of Approximate Model-Based Shielding for Safe Reinforcement Learning

Alex Goodall Francesco Belardinelli

Department of Computing
Imperial College London

a.goodall22@imperial.ac.uk

SAFE & TRUSTED AI
UKRI CENTRE FOR DOCTORAL TRAINING

UK Research and Innovation

King's College London

Imperial College London

Overview

1. Overview of shielding for RL
2. Bounded Safety
3. Approximate Model-based Shielding (AMBS)
4. Experiments and Applications

Section 1

Preface

Shielding for Reinforcement Learning

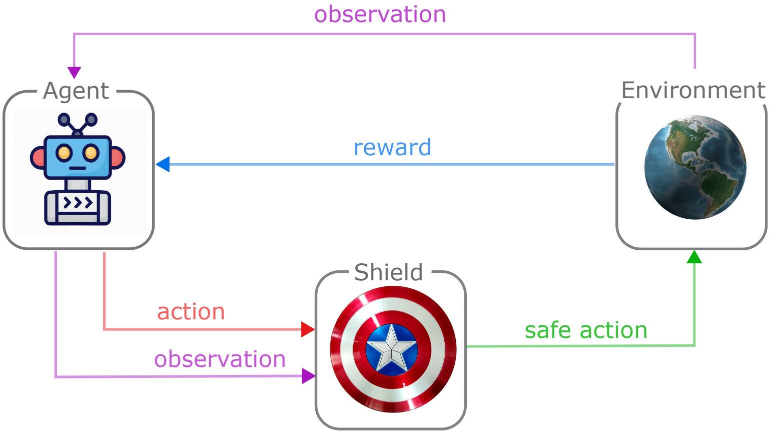


Figure: Shielding for reinforcement learning framework [Alshiekh et al., 2018]

Shielding for Reinforcement Learning (Continued)

1. The shield is represented as a **finite state reactive system** $\mathcal{S}$.
2. To construct the shield we require a **safety automaton** φ^s, a safety game is solved such that $\mathcal{S}$ realises the safety specification φ encoded by φ^s (i.e. $\mathcal{S} \models \varphi$).

Shielding for Reinforcement Learning (Continued)

1. The shield is represented as a **finite state reactive system** $\mathcal{S}$.
2. To construct the shield we require a **safety automaton** φ^s, a safety game is solved such that $\mathcal{S}$ realises the safety specification φ encoded by φ^s (i.e. $\mathcal{S} \models \varphi$).

Limitations

1. Knowledge of the safety relevant dynamics of the environment are required **a priori** to construct the safety automaton.
2. Solving the safety game can be computationally expensive without additional assumptions.

Bounded Prescience Shielding

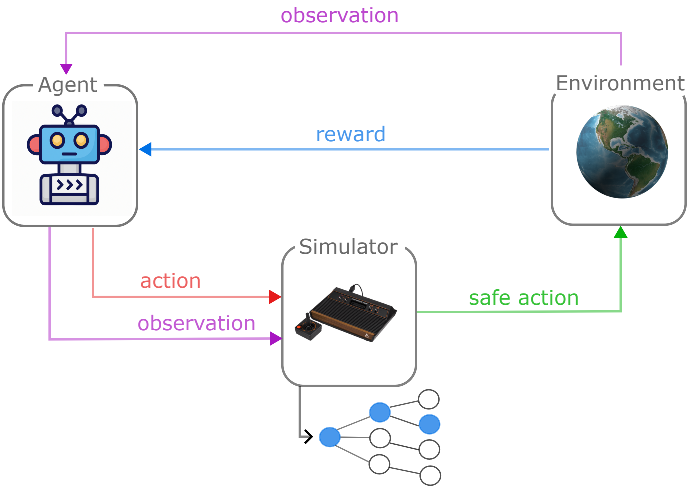


Figure: Bounded Prescience Shielding (BPS) [Giacobbe et al., 2021]

Bounded Prescience Shielding (Continued)

1. A high-fidelity **simulator** is used to verify the safety and shield policies up to some bounded **look-ahead** horizon.

Bounded Prescience Shielding (Continued)

1. A high-fidelity **simulator** is used to verify the safety and shield policies up to some bounded **look-ahead** horizon.

Limitations

1. Rolling out a simulator can be computationally expensive.
2. In most cases BPS **cannot** be used during training, and only during deployment for short horizons ($H = 5$).

Latent Shielding

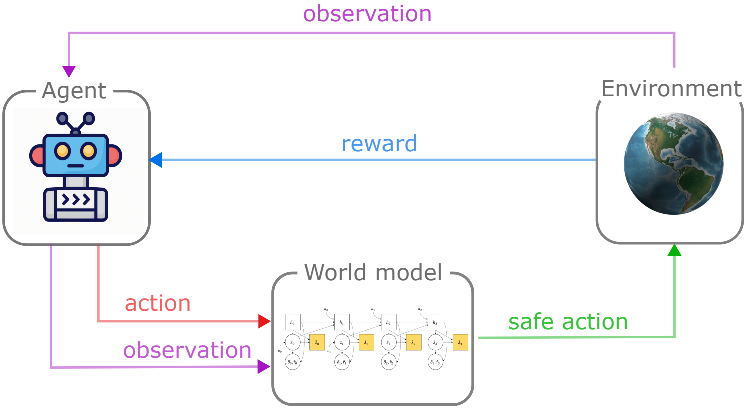


Figure: Latent shielding [He et al., 2021]

Latent Shielding (Continued)

1. During training the learned **world model** is used to estimate the probability of a violation in the near future.

Latent Shielding (Continued)

1. During training the learned **world model** is used to estimate the probability of a violation in the near future.

Limitations

1. Injecting noise into the action **overestimates** the probability - leading to overly conservative behaviour.
2. **Intrinsic punishment** and **shield introduction schedules** are required to resolve this, but they can be tricky to tune.
3. Limited by how far the world model can be rolled out (i.e. **short horizons** $H = 15$).

Limitations

1. Classical shielding for RL operates with quite **restrictive assumptions**: access to the safety-relevant dynamics of the MDP.

Limitations

1. Classical shielding for RL operates with quite **restrictive assumptions**: access to the safety-relevant dynamics of the MDP.
2. Methods such as BPS still **assume** access to a black-box simulator of the environment.

Limitations

1. Classical shielding for RL operates with quite **restrictive assumptions**: access to the safety-relevant dynamics of the MDP.
2. Methods such as BPS still **assume** access to a black-box simulator of the environment.
3. Other methods (e.g. **latent shielding**) have a relatively short look-ahead horizon (H=15).

Limitations

1. Classical shielding for RL operates with quite **restrictive assumptions**: access to the safety-relevant dynamics of the MDP.
2. Methods such as BPS still **assume** access to a black-box simulator of the environment.
3. Other methods (e.g. **latent shielding**) have a relatively short look-ahead horizon (H=15).
4. Most shielding methods only evaluate on quite **simple grid-world** environments.

Contributions

1. We will introduce a **safe exploration** problem based on the satisfaction of **bounded safety** defined in **probabilistic computation tree logic** (PCTL).
2. We propose **approximate model-based shielding** (AMBS), a safe exploration strategy and model-based RL algorithm that leverages **world models**, and improves **latent shielding** by using safety critics, a cost predictor and a learned backup policy.

Contributions

1. We will introduce a **safe exploration** problem based on the satisfaction of **bounded safety** defined in **probabilistic computation tree logic** (PCTL).
2. We propose **approximate model-based shielding** (AMBS), a safe exploration strategy and model-based RL algorithm that leverages **world models**, and improves **latent shielding** by using safety critics, a cost predictor and a learned backup policy.
3. We provide PAC-style probabilistic bounds on the probability of accurately detecting a safety violation and develop a strong theoretical justification for the use of **world models**.

Contributions

1. We will introduce a **safe exploration** problem based on the satisfaction of **bounded safety** defined in **probabilistic computation tree logic** (PCTL).
2. We propose **approximate model-based shielding** (AMBS), a safe exploration strategy and model-based RL algorithm that leverages **world models**, and improves **latent shielding** by using safety critics, a cost predictor and a learned backup policy.
3. We provide PAC-style probabilistic bounds on the probability of accurately detecting a safety violation and develop a strong theoretical justification for the use of **world models**.
4. We apply AMBS to a variety of **visual input** settings, such as classic **Atari games** and continuous control problems from the **Safety Gym** suite.

Section 2

Preliminaries

POMDP with Labels

Visual RL setting are typically modelled as partially observable MDP (POMDP). For our purposes we also extend the POMDP tuple to include state-dependent labels.

Definition (POMDP with labels)

A POMDP with labels is a 9-tuple $\mathcal{M} = (S, A, p, \iota_{init}, R, \Omega, O, AP, L)$ where, S is the set of states, A is the set of actions, $p : S \times A \times S \to [0, 1]$ is the **transition function**, $\iota_{init} : S \to [0, 1]$ is the **initial state distribution** such that $\int_{s \in S} \iota_{init}(s) = 1$, $R : S \times A \to \mathbb{R}$ is the **reward function**, Ω is a set of observations, $O : S \times A \times \Omega \to [0, 1]$ is the **observation function**, which defines the probability of an observation conditional on the previous state-action pair, AP is a set of **atomic propositions** which maps to the set of states by the 'expert' **labelling function** $L : S \to 2^{AP}$.

At each timestep t the agent receives an observation $o_t \in \Omega$, a reward $r_t \in \mathbb{R}$ and a set of labels $L(s_t) \in 2^{AP}$.

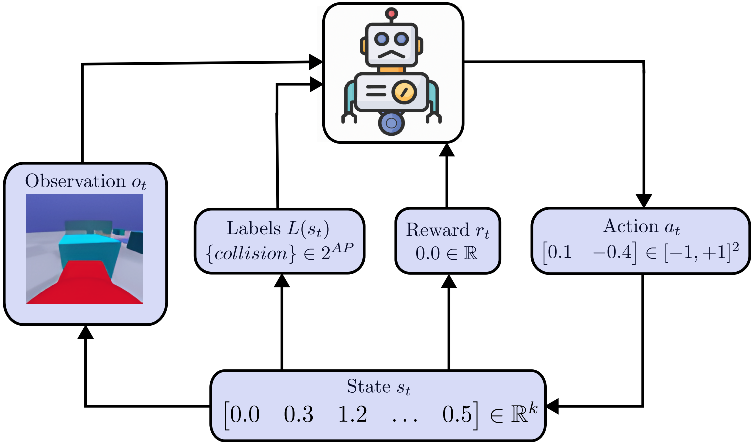


Figure: Visual representation of a POMDP with labels

Safe Exploration

In addition we are given a **propositional safety-formula** Ψ, e.g.

$$\Psi = \neg\textbf{collision} \wedge (\textbf{red-light} \Rightarrow \textbf{stop})$$

for $AP = \{$**collision, red-light, stop**$\}$. A state s is called **safe** if it satisfies the safety-formula Ψ, denoted $s \models \Psi$, which is determined by applying the **satisfaction relation** (from propositional logic),

$$\begin{aligned} s \models a &\text{ iff } a \in L(s) \\ s \models \neg\Psi &\text{ iff } s \not\models \Psi \\ s \models \Psi_1 \wedge \Psi_2 &\text{ iff } s \models \Psi_1 \text{ and } s \models \Psi_2 \end{aligned}$$

Safe Exploration

In addition we are given a **propositional safety-formula** Ψ, e.g.

$$\Psi = \neg\textbf{collision} \land (\textbf{red-light} \Rightarrow \textbf{stop})$$

for $AP = \{$**collision, red-light, stop**$\}$. A state s is called **safe** if it satisfies the safety-formula Ψ, denoted $s \models \Psi$, which is determined by applying the **satisfaction relation** (from propositional logic),

$$s \models a \text{ iff } a \in L(s)$$
$$s \models \neg\Psi \text{ iff } s \not\models \Psi$$
$$s \models \Psi_1 \land \Psi_2 \text{ iff } s \models \Psi_1 \text{ and } s \models \Psi_2$$

Goal

Find a policy π that maximises reward, that is $\pi^* = \arg\max_\pi \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \cdot r_t]$, while minimising the cumulative number of violations of the safety-formula Ψ during training and deployment.

Bounded Safety

Consider some fixed (stochastic) policy π and POMDP $\mathcal{M}$. Together π and $\mathcal{M}$ define a **transition system** $\mathcal{T} : S \times S \to [0, 1]$, where $\int_{s' \in S} \mathcal{T}(s, s') = 1$.

Definition (Bounded Safety)

A finite trace with length n of the transition system $\mathcal{T}$, is a sequence of states $s_0 \to s_1 \to \ldots \to s_n$ denoted τ, the i^{th} state of τ is given by $\tau[i]$. A trace τ satisfies bounded safety if and only if all of its states satisfy the state formula Ψ that encodes our safety constraints.

$$\tau \models \Box^{\leq n}\Psi \quad \text{iff} \quad \text{for all } 0 \leq i \leq n, \tau[i] \models \Psi$$

where $\Box$ is the common temporal operator 'always' (or 'globally') [Baier and Katoen, 2008] and n is some look-ahead horizon.

Δ-Bounded Safety

We can formalise Δ-**Bounded Safety** in PCTL.

Definition (Δ-Bounded Safety)

A state $s \in S$ satisfies Δ-bounded safety as follows,

$$s \models \mathbb{P}_{\geq 1-\Delta}(\Box^{\leq n}\Psi) \text{ iff } \mu_s(\{\tau \mid \tau[0] = s, \text{ for all } 0 \leq i \leq n, \tau[i] \models \Psi\}) \in [1-\Delta, 1] \quad (1)$$

where μ_s is a well-defined probability measure induced by the transition probabilities $\mathcal{T}$, over the set of traces staring from s and with finite length n.

Motivating Example

Example: avoiding irrecoverable states [Thomas et al., 2021].

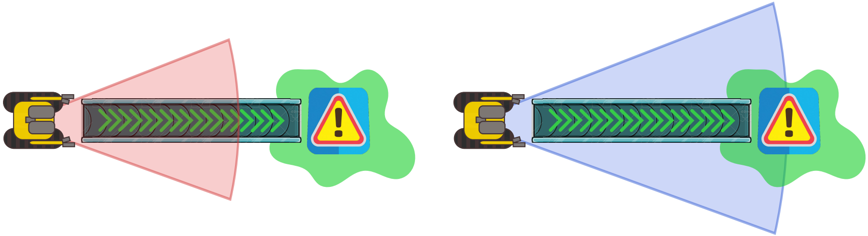


Figure: To detect the **unsafe state** (pool of acid) at the end of the conveyor belt the agent needs a sufficient look-ahead horizon. During exploration the first agent (**left**) may fail to detect the pool of acid at the end of the conveyor belt and unknowingly venture down an irrecoverable. The second agent (**right**) has a sufficient look-ahead horizon and can avoid the pool of acid during exploration.

Section 3

Approximate Model Based Shielding (AMBS)

Key Ideas

- By using [Hafner et al., 2023] we can learn an **approximate** transition system $\widehat{\mathcal{T}} \approx \mathcal{T}$ that captures the underlying dynamics of the POMDP.

Key Ideas

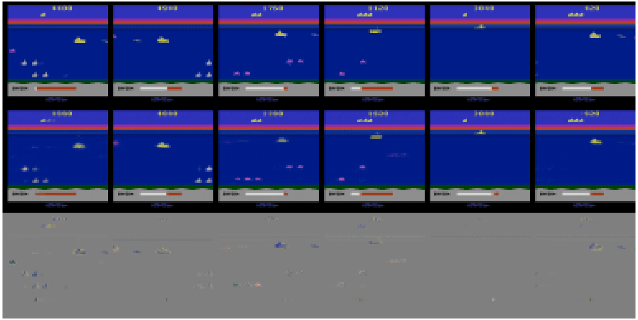
- By using [Hafner et al., 2023] we can learn an **approximate** transition system $\widehat{\mathcal{T}} \approx \mathcal{T}$ that captures the underlying dynamics of the POMDP.
- By sampling traces $\tau \in \mathcal{T}$ we can check **Δ-bounded safety**, i.e, $s \models \mathbb{P}_{\geq 1-\Delta}(\Box^{\leq H}\Psi)$; the Δ parameter meaningfully trades-off safety and exploration.

Key Ideas

- By using [Hafner et al., 2023] we can learn an **approximate** transition system $\widehat{\mathcal{T}} \approx \mathcal{T}$ that captures the underlying dynamics of the POMDP.
- By sampling traces $\tau \in \mathcal{T}$ we can check **Δ-bounded safety**, i.e, $s \models \mathbb{P}_{\geq 1-\Delta}(\square^{\leq H}\Psi)$; the Δ parameter meaningfully trades-off safety and exploration.
- During training and deployment we can **shield** the agent by overriding 'unsafe' actions when necessary.

World Models

"DreamerV3 learns a world model from experiences and uses it to train an actor critic policy from imagined trajectories. The world model encodes sensory inputs into categorical representations and predicts future representations and rewards given actions." [Hafner et al., 2023]



Recurrent State Space Model (RSSM)

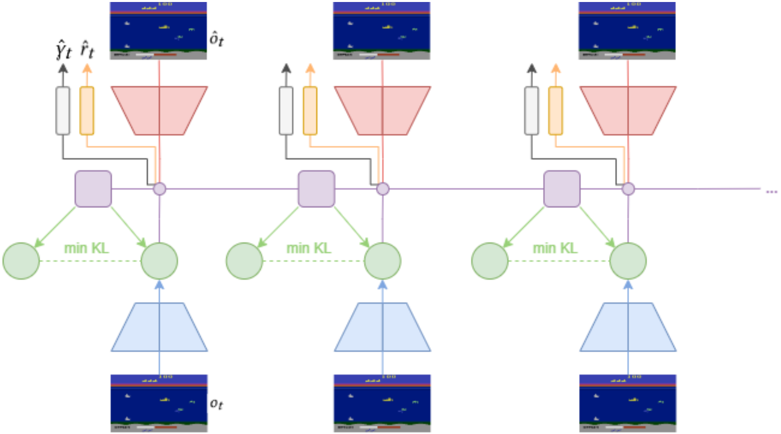


Figure: Recurrent State Space Model (RSSM) [Hafner et al., 2023]

Additional Components (Cost Function)

- Cost predictor $\hat{c}_t \sim p_\theta(\cdot \mid h_t, \hat{z}_t)$ implemented as an MLP mapping the learnt latent space $\hat{s}_t = (h_t, \hat{z}_t)$ to estimated costs $\hat{c}_t$.
- Targets for the cost predictor,

$$c_t = \begin{cases} 0, & \text{if } s_t \models \Psi \\ C, & \text{otherwise} \end{cases} \tag{2}$$

  where $C > 0$ is a hyperparameter.
- Trained with the usual log likelihood gradients.

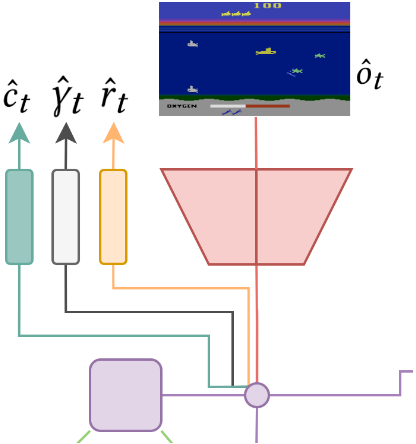


Figure: RSSM with costs

Additional Components (Safety Critics & Backup Policy)

Safety Critics:

- Trained with a TD3-style algorithm [Fujimoto et al., 2018] to estimate the cost value function,

$$V^C(s) = \mathbb{E}_{\pi^{\text{task}}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot c_t \;\middle|\; s_0 = s\right] \tag{3}$$

- Used to check Δ-bounded safety with a longer horizon.

Additional Components (Safety Critics & Backup Policy)

Safety Critics:

- Trained with a TD3-style algorithm [Fujimoto et al., 2018] to estimate the cost value function,

$$V^C(s) = \mathbb{E}_{\pi^{\text{task}}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot c_t \;\middle|\; s_0 = s\right] \tag{3}$$

- Used to check Δ-bounded safety with a longer horizon.

Backup Policy:

- The task policy π^{task} is trained to maximise reward, the backup policy π^{safe} is used as a **default safe policy**, it can be constructed in advance, however in most cases it must be trained with RL (to minimise costs),

$$\min \mathbb{E}_{\pi^{\text{safe}}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot c_t\right] \tag{4}$$

The Shielding Procedure

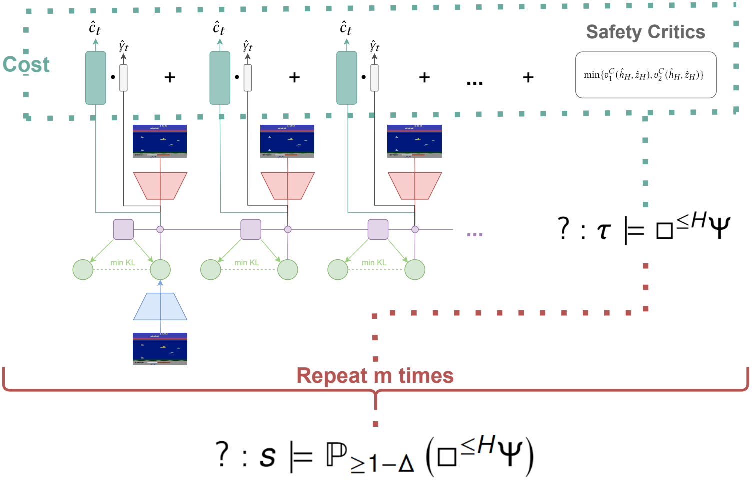


Figure: Checking Δ-bounded safety

Full Algorithmic Overview

1. Initialise replay buffer $\mathcal{D}$ with M random episodes.
2. Repeat until convergence:
    3. Sample a batch $B \sim \mathcal{D}$ and update the RSSM with representation learning.
    4. Sample latent trajectories with π^{task} and train π^{task} with RL to maximise reward.
    5. Using the same trajectories, train the safety critics with maximum likelihood to predict the expected discounted cost.
    6. Sample latent trajectories with π^{safe} and train π^{safe} with RL to minimise cost.
    7. For K environment interactions:
        8. Samples m trajectories in the world model with π_{task} to check if $s \models \mathbb{P}_{\geq 1-\Delta}(\Box^{\leq H}\Psi)$
        9. If $\Pr_{\pi^{\text{task}}}\left[\Box^{\leq H}\Psi\right] < 1 - \Delta$, then sample an action $a \sim \pi^{\text{safe}}$ play with the backup policy, else sample an action the task policy $a \sim \pi^{\text{task}}$.
        10. Play a in the environment and observe o', r, $L(s)$, append the experience to $\mathcal{D}$.

Probabilistic Guarantees

Let $\mu_{s \models \phi}$ denote the probability that $s \models \phi$, where $\phi = \Box^{\leq H}\Psi$, note that $s \models \mathbb{P}_{\geq 1-\Delta}(\Box^{\leq H}\Psi)$ (Δ-bounded-safety) if and only if $\mu_{s \models \phi} > 1 - \Delta$.

Probabilistic Guarantees

Let $\mu_{s\models\phi}$ denote the probability that $s \models \phi$, where $\phi = \square^{\leq H}\Psi$, note that $s \models \mathbb{P}_{\geq 1-\Delta}(\square^{\leq H}\Psi)$ (Δ-bounded-safety) if and only if $\mu_{s\models\phi} > 1 - \Delta$.

Theorem 1 (Fully observable case)

Let $\epsilon > 0$, $\delta > 0$, $s \in S$ be given. With access to the **true** transition system $\mathcal{T}$, with probability $1 - \delta$ we can obtain an **ϵ-approximate estimate** of the measure $\mu_{s\models\phi}$, by sampling m traces $\tau \sim \mathcal{T}$, provided that,

$$m \geq \frac{1}{2\epsilon^2} \log\left(\frac{2}{\delta}\right) \tag{5}$$

Probabilistic Guarantees

Let $\mu_{s\models\phi}$ denote the probability that $s \models \phi$, where $\phi = \square^{\leq H}\Psi$, note that $s \models \mathbb{P}_{\geq 1-\Delta}(\square^{\leq H}\Psi)$ (Δ-bounded-safety) if and only if $\mu_{s\models\phi} > 1 - \Delta$.

Theorem 1 (Fully observable case)

Let $\epsilon > 0$, $\delta > 0$, $s \in S$ be given. With access to the **true** transition system $\mathcal{T}$, with probability $1 - \delta$ we can obtain an **ϵ-approximate estimate** of the measure $\mu_{s\models\phi}$, by sampling m traces $\tau \sim \mathcal{T}$, provided that,

$$m \geq \frac{1}{2\epsilon^2} \log\left(\frac{2}{\delta}\right) \tag{5}$$

This result gives us a **sample complexity bound**, that dictates how many traces we need to sample (from $\mathcal{T}$) to check Δ-bounded safety with **high probability** (i.e $1 - \delta$).

Probabilistic Guarantees (Continued)

Suppose we only have access to an **approximate transition system** $\widehat{\mathcal{T}}$. We provide the following sample complexity bound.

Probabilistic Guarantees (Continued)

Suppose we only have access to an **approximate transition system** $\widehat{\mathcal{T}}$. We provide the following sample complexity bound.

Theorem 2

Let $\epsilon > 0$, $\delta > 0$ be given. Suppose that for all $s \in S$, the total variation (TV) distance between $\mathcal{T}(s' \mid s)$ and $\widehat{\mathcal{T}}(s' \mid s)$ is **upper bounded** by some $\alpha \leq \epsilon/n$. That is,

$$D_{\text{TV}}\left(\mathcal{T}(s' \mid s), \widehat{\mathcal{T}}(s' \mid s)\right) \leq \alpha \; \forall s \in S \tag{6}$$

Then with probability $1 - \delta$ we can obtain an ϵ**-approximate estimate** of the measure $\mu_{s \models \phi}$, by sampling m traces $\tau \sim \widehat{\mathcal{T}}$, provided that,

$$m \geq \frac{2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \tag{7}$$

Probabilistic Guarantees (Tabular Case)

When does $D_{\text{TV}}\left(\mathcal{T}(s' \mid s), \widehat{\mathcal{T}}(s' \mid s)\right) \leq \alpha$?

Probabilistic Guarantees (Tabular Case)

When does $D_{\text{TV}}\left(\mathcal{T}(s' \mid s), \widehat{\mathcal{T}}(s' \mid s)\right) \leq \alpha$?

Theorem 3

Let $\alpha > 0$, $\delta > 0$, $s \in S$ be given. With probability $1 - \delta$ the total variation (TV) distance between $\mathcal{T}(s' \mid s)$ and $\widehat{\mathcal{T}}(s' \mid s)$ is **upper bounded** by α, provided that all actions $a \in A$ with non-negligable probability $\eta \geq \alpha/(|A||S|)$ (under π) have been picked from s at least m times, where

$$m \geq \frac{|S|^2}{\alpha^2} \log\left(\frac{2|A||S|}{\delta}\right) \tag{8}$$

Upper Bound (Partially Observable Case)

Theorem 4

Let b_t be a latent representation (belief state) such that $p(s_t \mid o_{\leq t}, a_{\leq t}) = p(s_t \mid b_t)$. Let the fixed policy $\pi(\cdot \mid b_t)$ be a general probability distribution conditional on belief states b_t. Let f be a generic f-divergence measure (TV or similar). Then the following holds:

$$D_f(\mathcal{T}(s' \mid b), \widehat{\mathcal{T}}(s' \mid b)) \leq D_f(\mathcal{T}(b' \mid b), \widehat{\mathcal{T}}(b' \mid b))$$

where $\mathcal{T}$ and $\widehat{\mathcal{T}}$ are the 'true' and approximate transition system respectively, defined now over both states s and belief states b.

Upper Bound (Partially Observable Case)

Theorem 4

Let b_t be a latent representation (belief state) such that $p(s_t \mid o_{\leq t}, a_{\leq t}) = p(s_t \mid b_t)$. Let the fixed policy $\pi(\cdot \mid b_t)$ be a general probability distribution conditional on belief states b_t. Let f be a generic f-divergence measure (TV or similar). Then the following holds:

$$D_f(\mathcal{T}(s' \mid b), \widehat{\mathcal{T}}(s' \mid b)) \leq D_f(\mathcal{T}(b' \mid b), \widehat{\mathcal{T}}(b' \mid b))$$

where $\mathcal{T}$ and $\widehat{\mathcal{T}}$ are the 'true' and approximate transition system respectively, defined now over both states s and belief states b.

This result motivates the use of **world models** since the RHS appears in the RSSM loss function.

Section 4

Experiments

Atari Games

Environment	Safety formula Ψ
Assault	$\neg\textbf{hit} \land \neg\textbf{overheat}$
DoubleDunk	$\neg\textbf{out-of-bounds} \land \neg\textbf{shoot-bf-clear}$
Enduro	$\neg\textbf{crash-car}$
KungFuMaster	$\neg\textbf{loose-life} \land \neg\textbf{energy-loss}$
Seaquest	$(\textbf{surface} \Rightarrow \textbf{diver})\neg\textbf{hit} \land \neg\textbf{out-of-oxygen}$

Figure: Assault

Figure: Double Dunk


Figure: Enduro

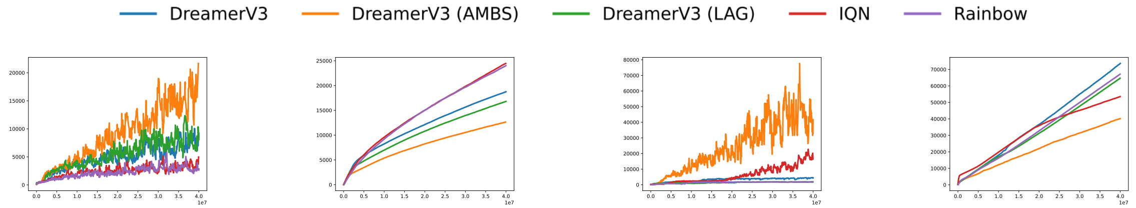
Figure: Kung Fu Master

Figure: Seaquest

Results [Goodall and Belardinelli, 2023]

Table: Episode return and cumulative violations at the end of training.

		DreamerV3	DreamerV3 (AMBS)	DreamerV3 (LAG)	IQN	Rainbow
Assault	Best Score ↑	14738	**44467**	19832	9959	9632
	# Violations ↓	18745	**12638**	16802	24462	24019
DoubleDunk	Best Score ↑	**24**	**24**	**24**	**24**	-
	# Violations ↓	877499	**66248**	359018	188363	-
Enduro	Best Score ↑	2369	2367	2365	2375	**2383**
	# Violations ↓	167933	*132147*	174217	129012	**108000**
KungFuMaster	Best Score ↑	97000	**117200**	97200	51600	59500
	# Violations ↓	427476	**10936**	567559	284909	612762
Seaquest	Best Score ↑	4860	**145550**	1940	34150	1900
	# Violations ↓	73641	**40147**	64679	53516	67101



(a) Reward (Assault)
(b) Violations (Assault)
(c) Reward (Seaquest)
(d) Violations (Seaquest)

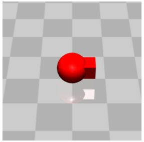
Qualitative results

Shielded:

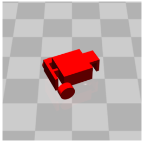
Unshielded:

Safety Gym

Safety Gym vehicles:

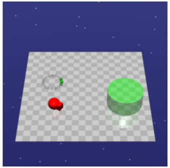


(a) Point

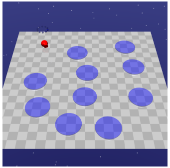


(b) Car

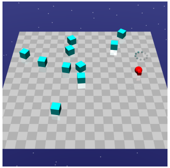
Safety Gym tasks and constraints:



(a) Goal positions



(b) Hazardous areas

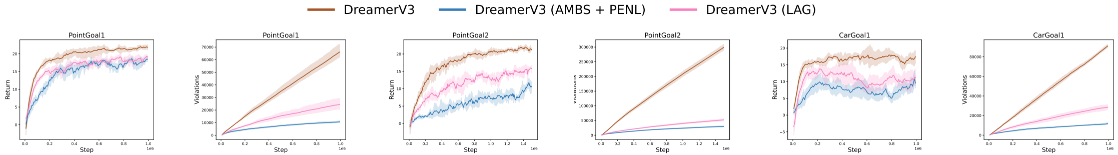


(c) Vases

Results [Goodall and Belardinelli, 2024]

Table: Episode return and cumulative violations at the end of training.

		AMBS + PENL	DreamerV3 + LAG	DreamerV3
PointGoal1	Episode Return ↑	17.32 ± 3.29	19.15 ± 0.92	**21.85 ± 1.26**
(1M)	# Violations ↓	**9354 ± 3734**	24996 ± 6627	66146 ± 8873
PointGoal2	Episode Return ↑	10.64 ± 2.61	15.78 ± 1.84	**21.25 ± 0.65**
(1.5M)	# Violations ↓	**29720 ± 3850**	52157 ± 6151	292606 ± 16062
CarGoal1	Episode Return ↑	8.87 ± 2.95	11.23 ± 4.10	**17.42 ± 2.96**
(1M)	# Violations ↓	**11423 ± 1479**	28639 ± 4644	87917 ± 2750



(a) PointGoal1 (b) PointGoal1 (c) PointGoal2 (d) PointGoal2 (e) CarGoal1 (f) CarGoal1

Conclusions

1. AMBS is a general purpose framework for shielding RL policies by simulating and verifying possible futures with a learned dynamics model or **world model** [Hafner et al., 2023].

Conclusions

1. AMBS is a general purpose framework for shielding RL policies by simulating and verifying possible futures with a learned dynamics model or **world model** [Hafner et al., 2023].
2. In contrast to **latent shielding** [He et al., 2021] our algorithm requires minimal hyperparameter tuning and **no schedules** and obtains further look-ahead capabilities with **safety critics**.

Conclusions

1. AMBS is a general purpose framework for shielding RL policies by simulating and verifying possible futures with a learned dynamics model or **world model** [Hafner et al., 2023].
2. In contrast to **latent shielding** [He et al., 2021] our algorithm requires minimal hyperparameter tuning and **no schedules** and obtains further look-ahead capabilities with **safety critics**.
3. We also develop a rigorous set of theoretical results that underpin AMBS.

Conclusions

1. AMBS is a general purpose framework for shielding RL policies by simulating and verifying possible futures with a learned dynamics model or **world model** [Hafner et al., 2023].
2. In contrast to **latent shielding** [He et al., 2021] our algorithm requires minimal hyperparameter tuning and **no schedules** and obtains further look-ahead capabilities with **safety critics**.
3. We also develop a rigorous set of theoretical results that underpin AMBS.
4. Our empirical results demonstrate that agents can benefit from shielding (AMBS) in both **discrete** (Atari) and **continuous** (Safety Gym) safety-critical domains.

Future Work

1. What are the challenges associated with more sophisticated safety properties, e.g. regular safety properties, LTL safety properties? (currently working on this)

Future Work

1. What are the challenges associated with more sophisticated safety properties, e.g. regular safety properties, LTL safety properties? (currently working on this)
2. Investigate different shielding procedures – how can we best leverage the backup policy, maybe integrate it into the policy gradient of the task policy? Can we use **model predictive control** (MPC) or planning as the backup policy?

Future Work

1. What are the challenges associated with more sophisticated safety properties, e.g. regular safety properties, LTL safety properties? (currently working on this)
2. Investigate different shielding procedures – how can we best leverage the backup policy, maybe integrate it into the policy gradient of the task policy? Can we use **model predictive control** (MPC) or planning as the backup policy?
3. Can we incorporate uncertainty estimation, **Bayesian world models** [As et al., 2022], to improve the agent learning and develop an 'uncertainty aware' shielding approach?

References

Alshiekh, Mohammed and Bloem, Roderick and Ehlers, Rüdiger and Könighofer, Bettina and Niekum, Scott and Topcu, Ufuk (2018)
Safe reinforcement learning via shielding
Proceedings of the AAAI Conference on Artificial Intelligence 32, 1 (2018)

Giacobbe, Mirco and Hasanbeig, Mohammadhosein and Kroening, Daniel and Wijk, Hjalmar (2021)
Shielding atari games with bounded prescience
arXiv preprint arXiv:2101.08153 (2021)

He, Chloe and León, Borja G and Belardinelli, Francesco (2021)
Do Androids Dream of Electric Fences? Safety-Aware Reinforcement Learning with Latent Shielding
arXiv preprint arXiv:2112.11490 (2021)

Baier, Christel and Katoen, Joost-Pieter (2008)
Principles of model checking
MIT press (2008)

Garrett Thomas, Yuping Luo, and Tengyu Ma. (2021)
Safe reinforcement learning by imagining the near future.
Advances in Neural Information Processing Systems 34 (2021), 13859–13869.

Hafner, Danijar and Pasukonis, Jurgis and Ba, Jimmy and Lillicrap, Timothy (2023)
Mastering Diverse Domains through World Models
arXiv preprint arXiv:2301.04104 (2023)

Fujimoto, Scott and Hoof, Herke and Meger, David (2018)
Addressing function approximation error in actor-critic methods
International conference on machine learning 1587–1596 (2018)

References

Goodall, Alexander W., and Francesco Belardinelli (2023)
Approximate Model-Based Shielding for Safe Reinforcement Learning
arXiv preprint arXiv:2308.00707 (2023)

Goodall, Alexander W., and Francesco Belardinelli (2023)
Leveraging Approximate Model-based Shielding for Probabilistic Safety Guarantees in Continuous Environments
arXiv preprint arXiv:2402.00816 (2024)

Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause (2022)
Constrained policy optimization via bayesian world models arXiv preprint arXiv:2201.09802 (2022)

The End

Thank you for listening!