A Algorithms

A.1 DreamerV3 with AMBS

Algorithm 2 DreamerV3 [22] with AMBS

Initialise: replay buffer $\mathcal{D}$ with S random episodes and RSSM parameters θ randomly.

while not converged **do**
- *// World model learning*
- Sample B sequences $\{\langle o_t, a_t, r_t, c_t, \gamma_t^{\text{safe}}, o_{t+1}\rangle_{t=k}^{k+H}\} \sim \mathcal{D}$.
- Update RSSM parameters θ with representation learning [22].
- *// Task policy optimisation*
- 'Imagine' trajectory σ from every o_t in every seq. with π^{task}.
- Train π^{task} with TD-λ actor-critic to optim. Eq. 2.
- *// Safety critic optimisation*
- Train v_1^C and v_2^C with maximum likelihood to estim. Eq. 11.
- *// Safe policy optimisation*
- 'Imagine' trajectory σ from every o_t in every seq. with π^{safe}.
- Train π^{safe} with TD-λ actor-critic to optim. Eq. 10.
- *// Environment interaction*
- **for** $k = 1, ..., K$ **do**
  - Observe o_t from environment and compute $\hat{s}_t = (z_t, h_t)$.
  - Sample action $a \sim \pi^{\text{task}}$ with the task policy.
  - Shield the proposed action $a' = \text{shield}(a)$ and play a'.
  - Observe r_t, o_{t+1} and $L(s_t)$ and construct c_t and γ_t^{safe}.
  - Append $\langle o_t, a_t, r_t, c_t, \gamma_t^{\text{safe}}, o_{t+1}\rangle$ to $\mathcal{D}$.
- **end for**

end while

B Proofs

We provide proofs for the theorems introduced in Section 3.

B.1 Proof of Theorem 1

Theorem 1 Restated. *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,*

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

Proof. We estimate $\mu_{s\models\phi}$ by sampling m traces $\langle\tau_j\rangle_{j=1}^m$ from $\mathcal{T}$. Let $X_1, ..., X_m$ be indicator r.v.s such that,

$$X_j = \begin{cases} 1 & \text{if } \tau_j \models \square^{\leq n}\Psi, \\ 0 & \text{otherwise} \end{cases}$$

Let,

$$\tilde{\mu}_{s\models\phi} = \frac{1}{m}\sum_{j=1}^{m} X_j, \text{ where } \mathbb{E}_{\mathcal{T}}[\tilde{\mu}_{s\models\phi}] = \mu_{s\models\phi}$$

Then by Hoeffding's inequality,

$$\mathbb{P}\left[|\tilde{\mu}_{s\models\phi} - \mu_{s\models\phi}| \geq \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right)$$

Bounding the RHS from above with δ and rearranging completes the proof. □

The only caveat is arguing that $\tau_j \models \square^{\leq n}\Psi$ is easily checkable. Indeed this is the case (for polynmial n) because in the fully observable setting we have access to the state and so we can check that $\forall i\ \tau[i] \models \Psi$.

B.2 Proof of Theorem 2

We split the proof into two parts for Theorem 2, first we present the *error amplification* lemma [36], followed by the full proof of Theorem 2.

Lemma 7 (error amplification)**.** *Let $\mathcal{T}(s' \mid s)$ and $\widehat{\mathcal{T}}(s' \mid s)$ be two transition systems with the same initial state distribution ι_{init}. Let $\mathcal{T}^t(s)$ and $\widehat{\mathcal{T}}^t(s)$ be the marginal state distribution at time t for the transitions systems $\mathcal{T}$ and $\widehat{\mathcal{T}}$ respectively. That is,*

$$\mathcal{T}^t(s) = \mathbb{P}_{\tau\sim\mathcal{T}}[\tau[t] = s]$$
$$\widehat{\mathcal{T}}^t(s) = \mathbb{P}_{\tau\sim\widehat{\mathcal{T}}}[\tau[t] = s]$$

Suppose that,

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

then the marginal distributions are bounded as follows,

$$D_{TV}\left(\mathcal{T}^t, \widehat{\mathcal{T}}^t\right) \leq \alpha t\ \forall t$$

Proof. First let us fix some $s \in S$. Then,

$$\begin{aligned}
\left|\mathcal{T}^t(s) - \widehat{\mathcal{T}}^t(s)\right| &= \left|\sum_{\bar{s}\in S}\mathcal{T}(s \mid \bar{s})\mathcal{T}^{t-1}(\bar{s}) - \sum_{\bar{s}\in S}\widehat{\mathcal{T}}(s \mid \bar{s})\widehat{\mathcal{T}}^{t-1}(\bar{s})\right| \\
&\leq \sum_{\bar{s}\in S}\left|\mathcal{T}(s \mid \bar{s})\mathcal{T}^{t-1}(\bar{s}) - \widehat{\mathcal{T}}(s \mid \bar{s})\widehat{\mathcal{T}}^{t-1}(\bar{s})\right| \\
&\leq \sum_{\bar{s}\in S}\left|\mathcal{T}^{t-1}(\bar{s})\left(\mathcal{T}(s \mid \bar{s}) - \widehat{\mathcal{T}}(s \mid \bar{s})\right)\right| \\
&\quad + \left|\widehat{\mathcal{T}}(s \mid \bar{s})\left(\mathcal{T}^{t-1}(\bar{s}) - \widehat{\mathcal{T}}^{t-1}(\bar{s})\right)\right|
\end{aligned}$$

Using the above inequality we get the following,

$$\begin{aligned}
2D_{\text{TV}}\left(\mathcal{T}^t, \widehat{\mathcal{T}}^t\right) &= \sum_{s\in S}\left|\mathcal{T}^t(s) - \widehat{\mathcal{T}}^t(s)\right| \\
&\leq \sum_{\bar{s}\in S}\mathcal{T}^{t-1}(\bar{s})\sum_{s\in S}\left|\mathcal{T}(s \mid \bar{s}) - \widehat{\mathcal{T}}(s \mid \bar{s})\right| \\
&\quad + \sum_{\bar{s}\in S}\left|\mathcal{T}^{t-1}(\bar{s}) - \widehat{\mathcal{T}}^{t-1}(\bar{s})\right| \\
&\leq 2\alpha + 2D_{\text{TV}}\left(\mathcal{T}^{t-1}, \widehat{\mathcal{T}}^{t-1}\right) \\
&\leq 2\alpha t
\end{aligned}$$

The final inequality holds by applying the the recursion obtained on t until $t = 0$ where $\mathcal{T}$ and $\widehat{\mathcal{T}}$ start from the same initial state distribution ι_{init}. □

Theorem 2 Restated. *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 bounded by some $\alpha \leq \epsilon/n$. That is,*

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

Now fix an $s \in S$, 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,

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

Proof. Recall that,

$$\mu_{s\models\phi} = \mu_s(\{\tau \mid \tau[0] = s, \text{ for all } 0 \le i \le n, \tau[i] \models \Psi\})$$

where $\tau \sim \mathcal{T}$. Equivalently we can write,

$$\mu_{s\models\phi} = \mathbb{P}_{\tau\sim\mathcal{T}}[\tau \models \square^{\le n}\Psi]$$

Similarly, let $\hat{\mu}_{s\models\phi}$ be defined as the *true* probability under $\widehat{\mathcal{T}}$,

$$\hat{\mu}_{s\models\phi} = \mathbb{P}_{\tau\sim\widehat{\mathcal{T}}}[\tau \models \square^{\le n}\Psi]$$

Let the following denote the average state distribution for $\mathcal{T}$ and $\widehat{\mathcal{T}}$ respectively,

$$\rho_{\mathcal{T}}(s) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\tau\sim\mathcal{T}}(\tau[i] = s)$$

$$\rho_{\widehat{\mathcal{T}}}(s) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\tau\sim\widehat{\mathcal{T}}}(\tau[i] = s)$$

By following the simulation lemma [28], we get the following,

$$\begin{aligned}
\left|\mu_{s\models\phi} - \hat{\mu}_{s\models\phi}\right| &= \left|\mathbb{P}_{\tau\sim\mathcal{T}}[\tau \models \square^{\le n}\Psi] - \mathbb{P}_{\tau\sim\widehat{\mathcal{T}}}[\tau \models \square^{\le n}\Psi]\right| \\
&\le 1 \cdot D_{TV}(\rho_{\mathcal{T}}, \rho_{\widehat{\mathcal{T}}}) \\
&= \frac{1}{2}\sum_{s\in S}\left|\rho_{\mathcal{T}}(s) - \rho_{\widehat{\mathcal{T}}}(s)\right| \\
&= \frac{1}{2n}\sum_{s\in S}\Big|\sum_{i=1}^{n}\mathbb{P}_{\tau\sim\mathcal{T}}(\tau[i] = s) - \mathbb{P}_{\tau\sim\widehat{\mathcal{T}}}(\tau[i] = s)\Big| \\
&\le \frac{1}{2n}\sum_{s\in S}\sum_{i=1}^{n}\left|\mathbb{P}_{\tau\sim\mathcal{T}}(\tau[i] = s) - \mathbb{P}_{\tau\sim\widehat{\mathcal{T}}}(\tau[i] = s)\right| \\
&\le \frac{1}{2n}\sum_{i=1}^{n}\alpha n \qquad \text{(Using Lemma 7)} \\
&= \frac{\alpha n}{2}
\end{aligned}$$

Now we have that $\left|\mu_{s\models\phi} - \hat{\mu}_{s\models\phi}\right| \le (\alpha n)/2 \le \epsilon/2$. It remains to obtain an $\epsilon/2$-approximation of $\hat{\mu}_{\models\phi}$. Using the exact same reasoning as in the proof of Theorem 1, we estimate $\hat{\mu}_{\models\phi}$ by sampling m traces $\langle\tau_j\rangle_{j=1}^{m}$ from $\widehat{\mathcal{T}}$. Then provided,

$$m \ge \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$$

with probability $1-\delta$ we obtain an $\epsilon/2$-approximation of $\hat{\mu}_{\models\phi}$ and by extension an ϵ-approximation of $\mu_{\models\phi}$. □

B.3 Proof of Theorem 3

Theorem 3 Restated. *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 $\eta \ge \alpha/(|A||S|)$ (under π) have been picked from s at least m times, where*

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

Proof. Recall that $p(s' \mid s, a)$ denotes the probability of transitioning to s' from s when action a is played. Lets first fix s' and just consider approximating one of these probabilities. Let,

- $p = p(s' \mid s, a)$ and
- m be the number of times a is played from s.
- Let $X_1, ..., X_m$ be the indicator r.v.s such that,

$$X_i = \begin{cases} 1 & \text{if } s' \text{ given } (s,a) \\ 0 & \text{otherwise} \end{cases}$$

- $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$, where $\mathbb{E}[\hat{p}] = p$.

By Hoeffding's Inequality we have,

$$\mathbb{P}\left[|\hat{p} - p| \ge \frac{\alpha}{|S|}\right] \le 2 \cdot \exp\left(-2m\frac{\alpha^2}{|S|^2}\right)$$

Bounding the RHS from above by $\delta/(|S||A|)$ gives us,

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

And so with probability $1 - \delta/(|S||A|)$ we have an $\alpha/|S|$-approximation for p provided m satisfies the above bound. Taking a union bound over all $s' \in S$ and all $a \in A$ we have with probability at least $1-\delta$ an $\alpha/|S|$-approximation $p(s' \mid s, a)$ for all $s' \in S$ and all actions $a \in A$ with non-negligible probability $\eta \ge \frac{\alpha}{|A|}$ (under π). It remains to show that the TV distance between $\mathcal{T}(s' \mid s)$ and $\widehat{\mathcal{T}}(s' \mid s)$ is upper bounded by α,

$$\begin{aligned}
&2D_{TV}\left(\mathcal{T}(s' \mid s), \widehat{\mathcal{T}}(s' \mid s)\right) \\
&= \sum_{s'\in S}\left|\mathcal{T}(s' \mid s) - \widehat{\mathcal{T}}(s' \mid s)\right| \\
&= \sum_{s'\in S}\sum_{a\in A}\left|p(s' \mid s, a)\pi(a \mid s) - \hat{p}(s' \mid s, a)\pi(a \mid s)\right| \\
&= \sum_{s'\in S}\Bigg(\sum_{a\in A:\pi(a|s)\ge\eta}\left|p(s' \mid s, a)\pi(a \mid s) - \hat{p}(s' \mid s, a)\pi(a \mid s)\right| \\
&\qquad + \sum_{a\in A:\pi(a|s)<\eta}\left|p(s' \mid s, a)\pi(a \mid s) - \hat{p}(s' \mid s, a)\pi(a \mid s)\right|\Bigg) \\
&= \sum_{s'\in S}\Bigg(\sum_{a\in A:\pi(a|s)\ge\eta}\pi(a \mid s)\left|p(s' \mid s, a) - \hat{p}(s' \mid s, a)\right| \\
&\qquad + \sum_{a\in A:\pi(a|s)<\eta}\pi(a \mid s)\left|p(s' \mid s, a) - \hat{p}(s' \mid s, a)\right|\Bigg) \\
&\le \sum_{s'\in S}\left(\sum_{a\in A:\pi(a|s)\ge\eta}\pi(a \mid s)\frac{\alpha}{|S|} + \sum_{a\in A:\pi(a|s)<\eta}\pi(a \mid s)\right) \\
&\le \sum_{s'\in S}\left(\frac{\alpha}{|S|} + |A| \cdot \eta\right) \\
&\le \sum_{s'\in S}\left(\frac{\alpha}{|S|} + \frac{\alpha}{|S|}\right) \\
&\le 2\alpha
\end{aligned}$$

□

B.4 Proof of Theorem 4

Theorem 4 Restated. *Let b_t be a latent representation (belief state) such that $p(s_t \mid o_{t\le t}, a_{\le t}) = p(s_t \mid b_t)$. 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.

Proof. First for clarity, we define the transitions systems $\mathcal{T}$ and $\widehat{\mathcal{T}}$ over states s and belief states b. First we have as before

$$\mathcal{T}(s' \mid b) = \mathbb{P}_{\pi,p}[s_t = s' \mid b_{t-1} = b]$$
$$\widehat{\mathcal{T}}(s' \mid b) = \mathbb{P}_{\pi,\hat{p}}[s_t = s' \mid b_{t-1} = b]$$

Additionally, let,

$$\mathcal{T}(b' \mid b) = \mathbb{P}_{\pi,p}[b_t = b' \mid b_{t-1} = b]$$
$$\widehat{\mathcal{T}}(b' \mid b) = \mathbb{P}_{\pi,\hat{p}}[b_t = b' \mid b_{t-1} = b]$$

From these definitions, note that we can immediately define conditional and joint probabilities (i.e. $\mathcal{T}(s', b' \mid b)$, $\mathcal{T}(s' \mid b)$, $\mathcal{T}(b' \mid s', b)$ and similarly for $\widehat{\mathcal{T}}$) using the standard laws of probability.

Now we are ready to apply the data-processing inequality [2] for f-divergences as follows,

$$\begin{aligned}
&D_f(\mathcal{T}(b' \mid b), \widehat{\mathcal{T}}(b' \mid b)) \\
&= \mathbb{E}_{b' \sim \widehat{\mathcal{T}}}\left[f\left(\frac{\mathcal{T}(b' \mid b)}{\widehat{\mathcal{T}}(b' \mid b)}\right)\right] \\
&= \mathbb{E}_{s', b' \sim \widehat{\mathcal{T}}}\left[f\left(\frac{\mathcal{T}(s', b' \mid b)}{\widehat{\mathcal{T}}(s', b' \mid b)}\right)\right] \\
&= \mathbb{E}_{s' \sim \widehat{\mathcal{T}}}\left[\mathbb{E}_{b' \sim \widehat{\mathcal{T}}} f\left(\frac{\mathcal{T}(s', b' \mid b)}{\widehat{\mathcal{T}}(s', b' \mid b)}\right)\right] \\
&\geq \mathbb{E}_{s' \sim \widehat{\mathcal{T}}}\left[f\left(\mathbb{E}_{b' \sim \widehat{\mathcal{T}}} \frac{\mathcal{T}(s', b' \mid b)}{\widehat{\mathcal{T}}(s', b' \mid b)}\right)\right] \qquad \text{(Jensen's)} \\
&= \mathbb{E}_{s' \sim \widehat{\mathcal{T}}}\left[f\left(\mathbb{E}_{b' \sim \widehat{\mathcal{T}}} \frac{\mathcal{T}(s', b' \mid b)\widehat{\mathcal{T}}(b' \mid s', b)}{\widehat{\mathcal{T}}(s', b' \mid b)\mathcal{T}(b' \mid s', b)}\right)\right] \\
&= \mathbb{E}_{s' \sim \widehat{\mathcal{T}}}\left[f\left(\mathbb{E}_{b' \sim \widehat{\mathcal{T}}} \frac{\mathcal{T}(s' \mid b)}{\widehat{\mathcal{T}}(s' \mid b)}\right)\right] \\
&= \mathbb{E}_{s' \sim \widehat{\mathcal{T}}}\left[f\left(\frac{\mathcal{T}(s' \mid b)}{\widehat{\mathcal{T}}(s' \mid b)}\right)\right] \\
&= D_f(\mathcal{T}(s' \mid b), \widehat{\mathcal{T}}(s' \mid b))
\end{aligned}$$

□

C Code and Hyperparameters

Here we specify the most important hyperparameters for each of the agents in our experiments. For more precise implementation details we refer the reader to https://github.com/sacktock/AMBS.

World Model. As discussed we leverage DreamerV3 [22]. For all architectural details and hyperparameter choices please refer to [22].

Predictor Heads The cost function and safety discount predictor heads are implemented as neural network architectures similar to those used for the reward predictor and termination predictor of DreamerV3 [22]. Specifically the cost function predictor head is implemented in exactly the same was as the reward predictor head except it is used to predict cost signals instead of reward signals. Similarly, the safety discount predictor is a binary classifier implemented in the exact same was as the termination predictor, except one predicts safety-violations and the other predicts episode termination.

Safe Policy As discussed, the safe policy is trained with the same TD-λ style actor-critic algorithm that is used for the standard reward maximising (task) policy of DreamerV3 [22]. The actor and critic are implemented as neural networks with the same architecture that is used for the task policy and the hyperparameters are consistent between the safe policy and the task policy. Other details are outlined in Table 3.

Safety Critics The safety critics are trained with a TD3-style algorithm [13]. The two critics themselves (and their target networks) are implemented as neural networks with the same architectures used for the critics that help train the task policy and safe policy. The hyperparameters also are mostly the same and any changes are outlined in Table 4.

C.1 DreamerV3 Hyperparameters

Table 3. DreamerV3 hyperparameters [22]. Other methods built on DreamerV3 such as AMBS and LAG use this set of hyperparameters as well, unless otherwise specified.

Name	Symbol	Value
General		
Replay capacity	-	10^6
Batch size	B	16
Batch Length	-	64
Num. Envs	-	8
Train ratio	-	64
MLP Layers	-	5
MLP Units	-	512
Activation	-	LayerNorm + SiLU
World Model		
Num. latents	-	32
Classes per latent	-	32
Num. Layers	-	5
Num. Units	-	1024
Recon. loss scale	β_{pred}	1.0
Dynamics loss scale	β_{dyn}	0.5
Represen. loss scale	β_{rep}	0.1
Learning rate	-	10^{-4}
Adam epsilon	ϵ_{adam}	10^{-8}
Gradient clipping	-	1000
Actor Critic		
Imagination horizon	H	15
Discount factor	γ	0.997
TD lambda	λ	0.95
Critic EMA decay	-	0.98
Critic EMA regulariser	-	1
Return norm. scale	S_{reward}	$\text{Per}(R, 95) - \text{Per}(R, 5)$
Return norm. limit	L_{reward}	1
Return norm. decay	-	0.99
Actor entropy scale	η_{actor}	$3 \cdot 10^{-4}$
Learning rate	-	$3 \cdot 10^{-5}$
Adam epsilon	ϵ_{adam}	10^{-5}
Gradient clipping	-	100

C.2 AMBS Hyperparameters

Table 4. AMBS hyperparameters. We note that $m > 512$ is sufficient for $\Delta = 0.1$, $\epsilon = 0.09$, $\delta = 0.01$ using a bound similar to Eq. 5 that gives a bound on overestimating $\mu_{s \models \phi}$.

Name	Symbol	Value
Shielding		
Safety level	Δ	0.1
Approx. error	ϵ	0.09
Num. samples	m	512
Failure probability	δ	0.01
Look-ahead horizon	T	30
Cost Value	C	10
Safe Policy		
See 'Actor Critic' table 3		
...		
Safety Critic		
Type	-	TD3-style [13]
Slow update freq.	-	1
Slow update fraction	-	0.02
EMA decay	-	0.98
EMA regulariser	-	1
Cost norm. scale	S_{cost}	$\text{Per}(R, 95) - \text{Per}(R, 5)$
Cost norm. limit	L_{cost}	1
Cost norm. decay	-	0.99
Learning rate	-	$3 \cdot 10^{-5}$
Adam epsilon	ϵ_{adam}	10^{-5}
Gradient clipping	-	100

C.3 LAG Hyperparameters

Table 5. LAG hyperparameters [37]. We use the default hyperparameters provided in [6] for the safety gym benchmark [37].

Name	Symbol	Value
Penalty Multiplier	μ_k	$5 \cdot 10^{-9}$
Lagrange Multiplier	λ^k	10^{-6}
Penalty Power	σ	10^{-5}
Safety Horizon	T	30
Cost Value	C	10
Cost Threshold	$d = C \cdot \gamma^T$	≈ 9

C.4 IQN Hyperparameters

Table 6. IQN hyperparameters [12] adapted for a fairer comparison as in [10].

Name	Symbol	Value
Replay capacity	-	10^6
Batch size	B	32
IQN kappa	κ	1.0
Num. τ samples	-	64
Num. τ' samples	-	64
Num. quantile samples	-	32
Discount factor	γ	0.99
Update horizon	n_{hor}	3
Update freq.	-	4
Target update freq.	-	8000
Epsilon greedy min	ϵ_{min}	0.01
Epsilon decay period	-	250000
Optimiser	-	Adam
Learning rate	-	$5 \cdot 10^{-5}$
Adam epsilon	ϵ_{adam}	$3.125 \cdot 10^{-4}$

C.5 Rainbow Hyperparameters

Table 7. Rainbow hyperparameters as recommended in [25].

Name	Symbol	Value
Replay capacity	-	10^6
Batch size	B	32
Discount factor	γ	0.99
Update horizon	n_{hor}	3
Update freq.	-	4
Target update freq.	-	8000
Epsilon greedy min	ϵ_{min}	0.01
Epsilon decay period	-	250000
Optimiser	-	Adam
Learning rate	-	$6.25 \cdot 10^{-5}$
Adam epsilon	ϵ_{adam}	$1.5 \cdot 10^{-4}$

D Learning Curves

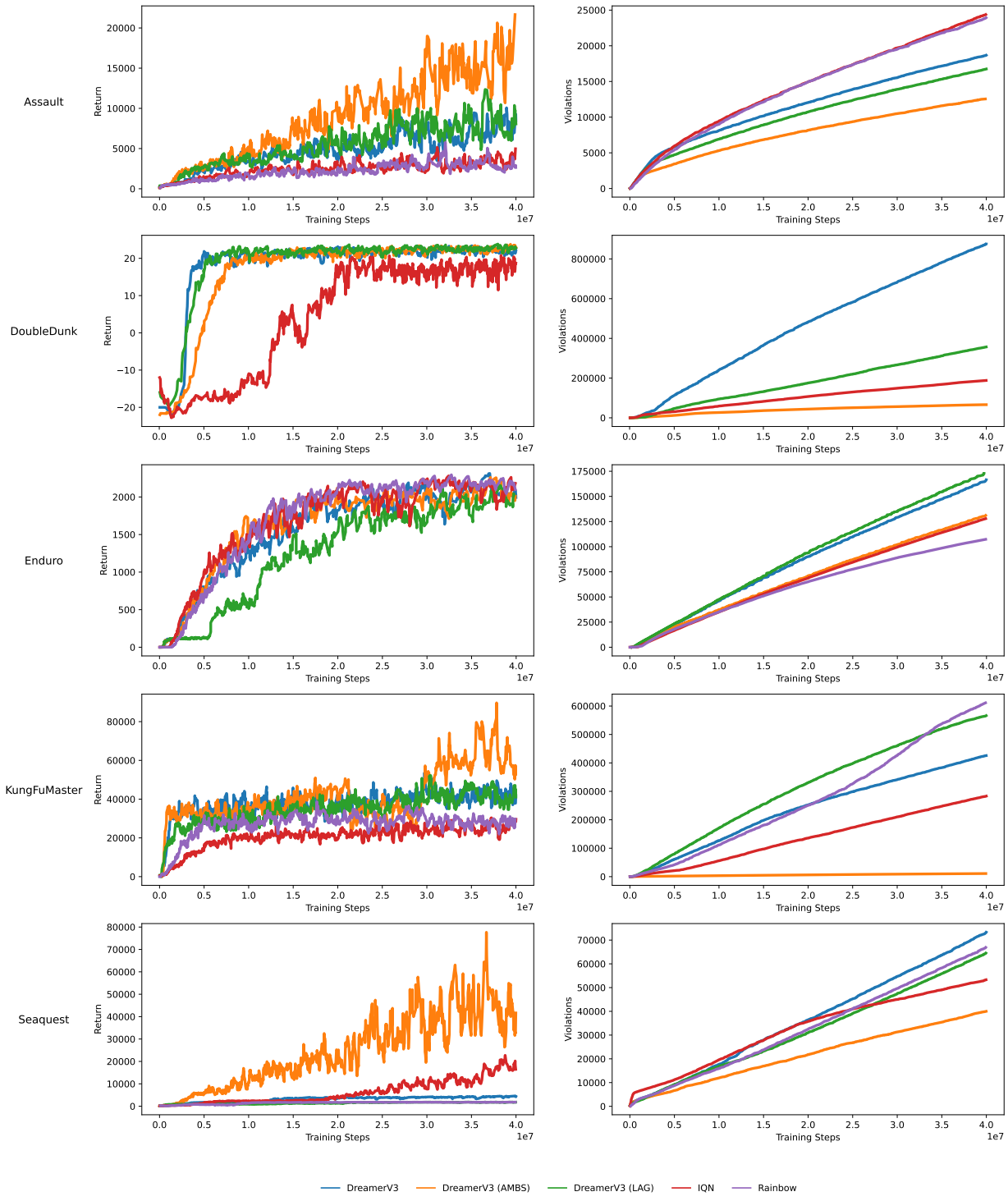

Figure 2. Learning curves for each of the five agents on a small set of Atari games. Each line represents one run over 10M environment interactions (40M frames). The left plots represent the episode return and the right plots represent the cumulative safety-violations during training.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel, 'Constrained policy optimization', in *International conference on machine learning*, pp. 22–31. PMLR, (2017).

[2] Syed Mumtaz Ali and Samuel D Silvey, 'A general class of coefficients of divergence of one distribution from another', *Journal of the Royal Statistical Society: Series B (Methodological)*, **28**(1), 131–142, (1966).

[3] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu, 'Safe reinforcement learning via shielding', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, (2018).

[4] Eitan Altman, *Constrained Markov decision processes: stochastic modeling*, Routledge, 1999.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, 'Concrete problems in ai safety', *arXiv preprint arXiv:1606.06565*, (2016).

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

[7] Christel Baier and Joost-Pieter Katoen, *Principles of model checking*, MIT press, 2008.

[8] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, 'The arcade learning environment: An evaluation platform for general agents', *Journal of Artificial Intelligence Research*, **47**, 253–279, (jun 2013).

[9] Roderick Bloem, Bettina Könighofer, Robert Könighofer, and Chao Wang, 'Shield synthesis: Runtime enforcement for reactive systems', in *International conference on tools and algorithms for the construction and analysis of systems*, pp. 533–548. Springer, (2015).

[10] Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare, 'Dopamine: A Research Framework for Deep Reinforcement Learning', (2018).

[11] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh, 'Lyapunov-based safe policy optimization for continuous control', *arXiv preprint arXiv:1901.10031*, (2019).

[12] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos, 'Implicit quantile networks for distributional reinforcement learning', in *International conference on machine learning*, pp. 1096–1105. PMLR, (2018).

[13] Scott Fujimoto, Herke Hoof, and David Meger, 'Addressing function approximation error in actor-critic methods', in *International conference on machine learning*, pp. 1587–1596. PMLR, (2018).

[14] Tanmay Gangwani, Joel Lehman, Qiang Liu, and Jian Peng, 'Learning belief representations for imitation learning in pomdps', in *Uncertainty in Artificial Intelligence*, pp. 1061–1071. PMLR, (2020).

[15] M Giacobbe, Mohammadhosein Hasanbeig, Daniel Kroening, and Hjalmar Wijk, 'Shielding atari games with bounded prescience', in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, (2021).

[16] Alexander W Goodall and Francesco Belardinelli, 'Approximate shielding of atari agents for safe exploration', *arXiv preprint arXiv:2304.11104*, (2023).

[17] William H Guss, Cayden Codel, Katja Hofmann, Brandon Houghton, Noboru Kuno, Stephanie Milani, Sharada Mohanty, Diego Perez Liebana, Ruslan Salakhutdinov, Nicholay Topin, et al., 'Neurips 2019 competition: the minerl competition on sample efficient reinforcement learning using human priors', *arXiv preprint arXiv:1904.10079*, (2019).

[18] David Ha and Jürgen Schmidhuber, 'Recurrent world models facilitate policy evolution', in *Advances in Neural Information Processing Systems*, eds., S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, volume 31. Curran Associates, Inc., (2018).

[19] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi, 'Dream to control: Learning behaviors by latent imagination', in *International Conference on Learning Representations*, (2020).

[20] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson, 'Learning latent dynamics for planning from pixels', in *International conference on machine learning*, pp. 2555–2565. PMLR, (2019).

[21] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba, 'Mastering atari with discrete world models', in *International Conference on Learning Representations*, (2021).

[22] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap, 'Mastering diverse domains through world models', *arXiv preprint arXiv:2301.04104*, (2023).

[23] Alexander Hans, Daniel Schneegaß, Anton Maximilian Schäfer, and Steffen Udluft, 'Safe exploration for reinforcement learning.', in *ESANN*, pp. 143–148. Citeseer, (2008).

[24] P He, B Gonzalez Leon, and F Belardinelli, 'Do androids dream of electric fences? safety-aware reinforcement learning with latent shielding'. CEUR Workshop Proceedings.

[25] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver, 'Rainbow: Combining improvements in deep reinforcement learning', in *Proceedings of the AAAI conference on artificial intelligence*, volume 32, (2018).

[26] Wassily Hoeffding, 'Probability inequalities for sums of bounded random variables', *The collected works of Wassily Hoeffding*, 409–426, (1994).

[27] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine, 'When to trust your model: Model-based policy optimization', *Advances in neural information processing systems*, **32**, (2019).

[28] Michael Kearns and Satinder Singh, 'Near-optimal reinforcement learning in polynomial time', *Machine learning*, **49**, 209–232, (2002).

[29] Jonathan N Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang, 'Learning in pomdps is sample-efficient with hindsight observability', *arXiv preprint arXiv:2301.13857*, (2023).

[30] Yongshuai Liu, Jiaxin Ding, and Xin Liu, 'Ipo: Interior-point policy optimization under constraints', in *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4940–4947, (2020).

[31] Zuxin Liu, Hongyi Zhou, Baiming Chen, Sicheng Zhong, Martial Hebert, and Ding Zhao, 'Constrained model-based reinforcement learning with robust cross-entropy method', *arXiv preprint arXiv:2010.07968*, (2020).

[32] Yuping Luo and Tengyu Ma, 'Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations', *Advances in Neural Information Processing Systems*, **34**, 25621–25632, (2021).

[33] Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling, 'Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents', *Journal of Artificial Intelligence Research*, **61**, 523–562, (2018).

[34] Haritz Odriozola-Olalde, Maider Zamalloa, and Nestor Arana-Arexolaleiba, 'Shielded reinforcement learning: A review of reactive methods for safe learning', in *2023 IEEE/SICE International Symposium on System Integration (SII)*, pp. 1–8. IEEE, (2023).

[35] Martin L Puterman, 'Markov decision processes', *Handbooks in operations research and management science*, **2**, 331–434, (1990).

[36] Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar, 'A game theoretic framework for model based reinforcement learning', in *International conference on machine learning*, pp. 7953–7963. PMLR, (2020).

[37] Alex Ray, Joshua Achiam, and Dario Amodei, 'Benchmarking safe exploration in deep reinforcement learning', *arXiv preprint arXiv:1910.01708*, **7**(1), 2, (2019).

[38] Richard S Sutton, 'Dyna, an integrated architecture for learning, planning, and reacting', *ACM Sigart Bulletin*, **2**(4), 160–163, (1991).

[39] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[40] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al., 'Deepmind control suite', *arXiv preprint arXiv:1801.00690*, (2018).

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

[42] Christopher KI Williams and Carl Edward Rasmussen, *Gaussian processes for machine learning*, volume 2, MIT press Cambridge, MA, 2006.

[43] Jorge Nocedal Stephen J Wright. Numerical optimization, 2006.

[44] Wenli Xiao, Yiwei Lyu, and John Dolan, 'Model-based dynamic shielding for safe and efficient multi-agent reinforcement learning', *arXiv preprint arXiv:2304.06281*, (2023).

[45] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge, 'Projection-based constrained policy optimization', in *International Conference on Learning Representations*.