Supplementary Material

Appendix

Proofs of the Theorems

Lemma 1 (Stochastic lower bound $F_{L,\gamma}$) *Consider the experiment of randomly sampling two times m points of the initial ball $\mathcal{B}_0$: $(a_1, \dots, a_m)$ and $(b_1, \dots, b_m)$. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function and $X = \max_{i=1}^{m} |g(a_i) - g(b_i)|/\|a_i - b_i\|$ be a random variable with the unknown cumulative distribution function F. Let $(x_1, \dots, x_n) \sim X$ be independent, identically distributed samples with the empirical distribution function $\hat{F}_n(x) = \sum_{i=1}^{n} 1_{x_i \leq x}$. Let G_n be a generalized extreme value distribution fitted to the empirical distribution function $\hat{F}_n$ and let D_n^- describe the goodness of fit, being the one-sided Kolmogorov–Smirnov statistic:*

$$D_n^- = \sup_x (G_n(x) - \hat{F}_n(x)) \tag{S1}$$

Given the confidence level γ and $\alpha = \min(\gamma, 0.5)$, then let us define $\epsilon_{n,\gamma}$ and $F_{L,\gamma}$ as follows:

$$\epsilon_{n,\gamma} = \sqrt{\frac{\ln \frac{1}{\alpha}}{2n}} \tag{S2}$$

$$F_{L,\gamma}(x) = G_n(x) - \epsilon_{n,\gamma} - D_n^- \tag{S3}$$

Then it holds that:

$$\Pr(\sup_x (F_{L,\gamma}(x) - F(x)) \leq 0) \geq 1 - \gamma, \tag{S4}$$

which intuitively means that $F_{L,\gamma}$ is a lower bound of F with confidence γ.

Proof. The Fisher-Tippett-Gnedenko theorem states that the distribution of a normalized maximum converges to the generalized extreme value distribution, if the distribution of the normalized maximum does converge. So intuitively that theorem is similar to the central limit theorem for the averages, but for the normalized maxima. Consequently, we start by fitting the empirical distribution function $\hat{F}_n$ by a generalized extreme value distribution G_n and compute Eq. (S1).

The Dvoretzky-Kiefer-Wolfowitz inequality (**?**) with a tight constant determined by (**?**), states that for all $\epsilon \geq \sqrt{\frac{1}{2n} \ln 2}$, it holds that:

$$\Pr(\sup_x (\hat{F}_n(x) - F(x)) > \epsilon) \leq e^{-2n\epsilon^2} \tag{S5}$$

Solving $\gamma = e^{-2n\epsilon^2}$ for ϵ and considering Massarts lower bound for ϵ, yields:

$$\Pr(\sup_x (\hat{F}_n(x) - F(x)) > \epsilon_{n,\gamma}) \leq \gamma, \tag{S6}$$

with $\epsilon_{n,\gamma}$ as defined in Eq. (S2). We use the triangular inequality for supremum and the monotony of the probability measure as follows:

$$\Pr\left(\sup_x \left(G_n(x) - F(x)\right) > \epsilon_{n,\gamma} + D_n^-\right) = \tag{S7}$$

$$= \Pr\Big(\sup_x \big(G_n(x) - \hat{F}_n(x) + \tag{S8}$$

$$+ \hat{F}_n(x) - F(x)\big) > \epsilon_{n,\gamma} + D_n^-\Big) \tag{S9}$$

$$\leq \Pr\Big(\sup_x \big(G_n(x) - \hat{F}_n(x)\big) + \tag{S10}$$

$$+ \sup_x \big(\hat{F}_n(x) - F(x)\big) > \epsilon_{n,\gamma} + D_n^-\Big) \tag{S11}$$

$$\overset{\text{(S1)}}{=} \Pr\left(\sup_x \left(\hat{F}_n(x) - F(x)\right) > \epsilon_{n,\gamma}\right) \tag{S12}$$

$$\overset{\text{(S6)}}{\leq} \gamma, \tag{S13}$$

from which it follows directly, that Eq. (S4) hold.

Theorem 1 (Radius of Stochastic Lipschitz Caps) *Given a continuous-depth model f from Eq. (1) in the main paper ($\partial_t x = f(x)$ with $x(t_0) \in B(x_0, \delta_0)$), $\gamma \in (0, 1)$, $\mu > 1$, target time t_j, the set of all sampled points $\mathcal{V}$, the number of sampled points $N = |\mathcal{V}|$, the sample maximum $\bar{m}_{j,\mathcal{V}} = \max_{x \in \mathcal{V}} d_j(x)$, the IVP solutions $\chi(t_j, x)$, and the corresponding stretching factors $\lambda_x = \|\partial_x \chi(t_j, x)\|$ for all $x \in \mathcal{V}$. Let us define $\hat{\gamma} = 1 - \sqrt{1 - \gamma}$. Let $\Delta\lambda_{\mathcal{V}}$ be the $\sqrt{1-\gamma}$-quantile of a stochastic lower bound $F_{L,\hat{\gamma}}$ as defined in Eq.* (S3) *of Lemma 1:*

$$\Delta\lambda_{\mathcal{V}}(\gamma) = F_{L,\hat{\gamma}}^{-1}(\sqrt{1-\gamma}), \tag{S14}$$

Let r_x be defined as:

$$r_x = \frac{\left(-\lambda_x + \sqrt{\lambda_x^2 + 4 \cdot \Delta\lambda_{\mathcal{V}} \cdot (\mu \cdot \bar{m}_{j,\mathcal{V}} - d_j(x))}\right)}{2 \cdot \Delta\lambda_{\mathcal{V}}}, \tag{S15}$$

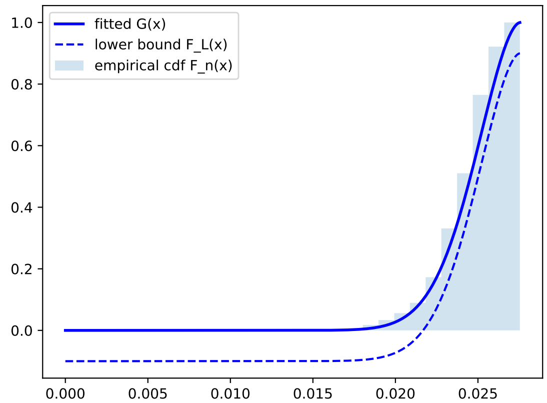

Figure S1: Visualisation of the stochastic lower bound $F_{L,\gamma}$ of Lemma 1.

with $\mathcal{C}_{x,r_x} = B(x,r_x)^S$*, it holds that:*

$$\Pr\left(\max_{i=1}^{m} d_j(y_i) \leq \mu \cdot \bar{m}_{j,\mathcal{V}}\right) \geq 1-\gamma \quad \forall y_i \in \mathcal{C}_{x,r_x}, \tag{S16}$$

and thus that $B(x,r_x)^S$ *is a* γ, t_j*-Lipschitz cap.*

Proof. Let $\{\lambda_1, \ldots, \lambda_n\}$ be n independent experiments by sampling from $X = \max_{i=1}^{m} |\lambda_{a_i} - \lambda_{b_i}|/\|a_i - b_i\|$ as defined in Lemma 1, where each variable is the maximum of m executions. From Eq. (S4) it follows that:

$$\Pr(F_{L,\hat{\gamma}(\lambda)} \leq F(\lambda)) \geq 1-\hat{\gamma} = \sqrt{1-\gamma} \tag{S17}$$

Let us now derive the probability of X being less or equal to $\Delta\lambda_{\mathcal{V}}$ defined by Eq. (S14). For any sets A, B it holds that $\Pr(A) \geq \Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B)$, thus:

$$\Pr(X \leq \Delta\lambda_{\mathcal{V}}) \geq \tag{S18}$$

$$\geq \Pr\left(X \leq \Delta\lambda_{\mathcal{V}} | F_{L,\hat{\gamma}(\lambda)} \leq F(\lambda)\right) \cdot \tag{S19}$$

$$\cdot \Pr\left(F_{L,\hat{\gamma}(\lambda)} \leq F(\lambda)\right) \tag{S20}$$

Let us have a look on Eq. (S19): As $\Pr(X \leq \Delta\lambda) = F(\Delta\lambda)$ and we are looking for the conditional probability depending on $F_{L,\hat{\gamma}(\lambda)} \leq F(\lambda)$, we can use $F_{L,\hat{\gamma}}(\Delta\lambda)$ as a lower bound of Eq. (S19) and thus, using Eq. (S17):

$$\Pr(X \leq \Delta\lambda_{\mathcal{V}}) \geq F_{L,\hat{\gamma}}(\Delta\lambda_{\mathcal{V}}) \cdot \sqrt{1-\gamma} \tag{S21}$$

As Eq. (S14) defines $\Delta\lambda_{\mathcal{V}}$ as the $\sqrt{1-\gamma}$-quantile of $F_{L,\gamma}$, we can further state that

$$\Pr(X \leq \Delta\lambda_{\mathcal{V}}) \geq 1-\gamma, \qquad \text{with} \quad X = \max_{i=1}^{m}\left[\frac{|\lambda_{a_i} - \lambda_{b_i}|}{\|a_i - b_i\|}\right] \tag{S22}$$

It trivially holds for $x \in \mathcal{B}_0$ and $\{y_1, \ldots, y_m\} \subset \mathcal{B}_0$ that:

$$\lambda_{y_i} = \lambda_x + \frac{\lambda_{y_i} - \lambda_x}{\|x - y_i\|} \cdot \|x - y_i\|$$

$$\max_{i=1}^{m} \lambda_{y_i} \leq \lambda_x + \max_{i=1}^{m} \frac{|\lambda_x - \lambda_y|}{\|x - y\|} \cdot r_x, \tag{S23}$$

for all y with $\|x - y_i\| \leq r_x$. From Eq. (S22) it follows that:

$$\Pr\Bigg(\lambda_x + \max_{i=1}^{m} \frac{|\lambda_x - \lambda_y|}{\|x - y\|} \cdot r_x \leq \tag{S24}$$

$$\leq \lambda_x + \Delta\lambda_{\mathcal{V}} \cdot r_x\Bigg) \geq 1-\gamma \tag{S25}$$

and using the monotony of the probability measure:

$$\Pr\left(\lambda_y \leq \lambda_x + \Delta\lambda_{\mathcal{V}} \cdot r_x\right) \geq 1-\gamma \tag{S26}$$

Using the mean value inequality for vector-valued functions it holds that:

$$\begin{aligned}
&|d_j(x) - d_j(y)| = |\,\|\chi(t_j,x) - \chi(t_j,x_0)\| - \\
&\quad - \|\chi(t_j,y) - \chi(t_j,x_0)\|\,| \quad \{\text{triangle inequality}\} \\
&\quad \leq \|\chi(t_j,x) - \chi(t_j,y)\| \quad \{\text{mean value theorem}\} \\
&\Rightarrow \exists z \in [x,y]\colon |d_j(x) - d_j(y)| \\
&\quad \leq \|\partial_x \chi(t_j,z)\| \|x - y\| = \lambda_z \cdot \|x - y\|
\end{aligned}$$

Combining this with Eq. (S26) and thus using $\lambda_x + \Delta\lambda_{\mathcal{V}} \cdot r_x$ as a probabilistic upper bound for λ_z, we obtain the following results:

$$\begin{aligned}
&\max_{i=1}^{m} |d_j(x) - d_j(y_i)| \leq \max_{i=1}^{m} \lambda_i \cdot r_x \\
&\Pr\big(\max_{i=1}^{m} |d_j(x) - d_j(y_i)| \leq \\
&\quad \leq (\lambda_x + \Delta\lambda_{\mathcal{V}} \cdot r_x) \cdot r_x\big) \geq 1-\gamma
\end{aligned} \tag{S27}$$

As r_x defined like in Eq. (S15) is the solution of the quadratic equation $\mu \cdot \bar{m}_{j,\mathcal{V}} - d_j(x) = \lambda_x r_x + \Delta\lambda_{\mathcal{V}} r_x^2$, it holds that:

$$\begin{aligned}
&\Pr\big(\max_{i=1}^{m} |d_j(x) - d_j(y_i)| \leq \mu \cdot \bar{m}_{j,\mathcal{V}} - d_j(x))\big) \geq \\
&\quad \geq 1-\gamma \quad \forall y_i \in B(x,r_x)^S
\end{aligned} \tag{S28}$$

We now distinguish between two cases for every y_i: (a) $d_j(y_i) \leq d_j(x)$ and (b) $d_j(y_i) \geq d_j(x)$. In case (a) it is trivial: $d_j(y_i) \leq d_j(x) \leq \mu \cdot \bar{m}_{j,\mathcal{V}}$. Having case (b):

$$\begin{gathered}
|d_j(x) - d_j(y_i)| \leq \mu \cdot \bar{m}_{j,\mathcal{V}} - d_j(x) \\
\Longleftrightarrow \\
d_j(y_i) - d_j(x) \leq \mu \cdot \bar{m}_{j,\mathcal{V}} - d_j(x) \\
\Longleftrightarrow \\
d_j(y_i) \leq \mu \cdot \bar{m}_{j,\mathcal{V}},
\end{gathered} \tag{S29}$$

thus from Eq. (S28) it follows that Eq. (S16) holds and $B(x,r_x)^S$ is a Lipschitz cap.

Theorem 2 (Convergence via Lipschitz Caps) *Given the tightness factor* $\mu > 1$*, the set of all sampled points* $\mathcal{V}$ *and the sample maximum* $\bar{m}_{j,\mathcal{V}} = \max_{x \in \mathcal{V}} d_j(x)$*. Let the initial ball maximum be defined by* $m_j^\star = \max_{\{x_1,\ldots,x_m\} \subset \mathcal{B}_0} d_j(x)$*. Then:*

$$\forall \gamma \in (0,1), \exists N \in \mathbb{N} \text{ s.t. } \Pr(\mu \cdot \bar{m}_{j,\mathcal{V}} \geq m_j^\star) \geq 1-\gamma \tag{S30}$$

where $N = |\mathcal{V}|$ *is the number of sampled points.*

Proof. Let $x_j^\star$ be a point such that $d_j(x_j^\star) = m_j^\star$. Given $\gamma \in (0,1)$ and cap radii r_x as defined in Eq. (S15), we know from the definition of a spherical cap that

$$p_{r_x} = \Pr(B(x,r_x)^S \ni x_j^\star) = \frac{\text{Area}(B(x,r_x)^S)}{\text{Area}\,(\mathcal{B}_0)} \quad \text{(S31)}$$

and thus it holds that:

$$\Pr(\exists y \in \mathcal{V}\colon B(y,r_y)^S \ni x_j^\star) = 1 - \prod_{x\in\mathcal{V}} (1-p_{r_x}) \quad \text{(S32)}$$

We derive a lower bound of r_x by using the first sample $x_{j,1}$ and replacing the values in Eq. (S15) as follows:

$$\mu \cdot \bar{m}_{j,\mathcal{V}} - d_j(x) \quad \text{(S33)}$$
$$\geq \mu \cdot \bar{m}_{j,\mathcal{V}} - \bar{m}_{j,\mathcal{V}} = (\mu-1)\cdot \bar{m}_{j,\mathcal{V}} \quad \text{(S34)}$$
$$\geq (\mu-1)\cdot d_j(x_{j,1}), \quad \text{(S35)}$$

thus a lower bound of all Lipschitz cap radii is given by

$$\begin{aligned} &r_{bound} = \\ &= \frac{-\lambda_x + \sqrt{\lambda_x^2 + 4\cdot\Delta\lambda_{\mathcal{V}}\cdot(\mu-1)\cdot d_j(x_{j,1})}}{2\cdot\Delta\lambda_{\mathcal{V}}} \leq \\ &\leq r_x \quad \forall x \in \mathcal{V} \\ &\Rightarrow \Pr(\exists y \in \mathcal{V}\colon B(y,r_y)^S \ni x_j^\star) \geq \\ &\geq 1-(1-p_{r_{bound}})^N \end{aligned} \quad \text{(S36)}$$

As in the limit of $N \to \infty$ the probability of Eq. (S36) is 1, it follows that $\forall \gamma \in (0,1)\ \exists N \in \mathbb{N}\colon \Pr(\exists x \in \mathcal{V}\colon B(x,r_x)^S \ni x_j^\star) \geq \sqrt{1-\gamma}$.

Using a set of sampled points $\mathcal{V}$ with cardinality N and using $\hat{\gamma} = 1-\sqrt{1-\gamma}$ as the error rate for the upper bound $\Delta\lambda_x$ of the confidence interval in Eq. (S14). Using the result of Theorem 1, the resulting probability $\forall y \in B(x,r_x)^S$ is:

$$\Pr\left(\max_{i=1}^{m} d_j(y_i) \leq \mu\cdot\bar{m}_{j,\mathcal{V}}\right) \geq 1-\hat{\gamma} = \sqrt{1-\gamma} \quad \text{(S37)}$$

If there is an $x \in \mathcal{V}$ such that $B(x,r_x)^S \ni x_j^\star$, then:

$$\Pr(d_j(x^\star) \leq \mu\cdot\bar{m}_{j,\mathcal{V}} | \exists x \in \mathcal{V}\colon B(x,r_x)^S \ni x^\star) \geq \sqrt{1-\gamma}$$

For any sets A, B it holds that $\Pr(A) \geq \Pr(A\cap B) = \Pr(A|B)\cdot\Pr(B)$, and using:

$$A = (\mu\cdot\bar{m}_{j,\mathcal{V}} \geq m_j^\star) \quad \text{(S38)}$$
$$B = (\exists x \in \mathcal{V}\colon B(x,r_x)^S \ni x_j^\star) \quad \text{(S39)}$$

it follows that $\Pr(\mu\cdot\bar{m}_{j,\mathcal{V}} \geq m_j^\star) \geq \Pr(A|B)\cdot\Pr(B) = 1-\gamma$ and therefore Eq. (S30) holds.

Reachability analysis of nonlinear ODEs

For linear ordinary differential equations (ODEs) there exists a closed-form solution, describing the behavior of the solution-traces over time, for every initial state. For nonlinear ODEs, there is no closed-form solution any more. One is able to calculate the solution for different initial states, but one does not know what happens in between of these already calculated traces.

The main goal in the reachability analysis of nonlinear ODEs, is to over-approximate the reachable states of the ODEs, starting from a set of initial states, such as, an interval, a ball, or an ellipsoid, in a way that one can guarantee that all traces are inside the over-approximation. We call such an over-approximation a *bounding tube*. Let us now define this mathematically:

Definition S1 (Initial value problem (IVP)) *We have a time-invariant ordinary differential equation $\partial_t x = f(x), f: \mathbb{R}^n \to \mathbb{R}^n$, a set of initial values defined by a ball $\mathcal{B}_0 = B(x_0,\delta_x)$ with center $x_0 \in \mathbb{R}^n$ and radius $\delta_0 \in \mathbb{R}$, the initial condition $x(t_0) \in \mathcal{B}_0$ and a sequence of k timesteps $\{t_j : j \in [1,\dots,k] \wedge (t_0 < t_1 < \dots < t_k)\}$. For every t_j, we want to know the solution of*

$$\partial_t x = f(x), \quad x(t_0) \in \mathcal{B}_0 = B(x_0,\delta_x). \quad \text{(S40)}$$

The definition can be generalized to time-variant ODEs, as time can be just seen as an additional variable x_{n+1} with $\partial_t x_{n+1} = 1$. Let $\chi_{t_0}^{t_j} x_0 = x(t_j)$ be the solution of Eq. (S40) at time t_j, for $x(t_0) = x_0$. In reachability analysis, the goal is to find for every time step t_j an overapproximation $\mathcal{B}_j \supseteq \{\chi_{t_0}^{t_j} x : x \in \mathcal{B}_0\}$, such that the set of these over-approximations build up a bounding tube, containing the reachable states. We define the bounding ball and bounding tube as follows:

Definition S2 (Bounding Ball) *Given an initial ball $\mathcal{B}_0 = B(x_0,\delta_0)$, we call $\mathcal{B}_j = B(\chi(t_j,x_0),\delta_j(\mathcal{B}_0))$ a* bounding ball *at time t_j, if it stochastically bounds the reachable states x at time t_j for all initial points around x_0 having the maximal initial perturbation δ_0.*

As we do not only want to bound the perturbation at one specific time, but on a time series, we define:

Definition S3 (Bounding Tube) *Given an initial ball $\mathcal{B}_0 = B(x_0,\delta_0)$ and bounding balls for $t_0 < \ldots < t_k = T$, we call the series of bounding balls $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k$ a* bounding tube.

When using reachability analysis to check for intersections with bad states, it is crucial to compute as tight as possible bounding tubes. Otherwise it would e.g. predict that a car driven by a controller would cause a crash even if the neural network controller is behaving perfectly and never causing a crash. Such wide bounding tubes are thus not useful for actually putting continuous depth-models into operation.