State-Dependent Actuator Saturation Ensures Adversarial Unreachability of High-Impact Inputs

Kaifan XIE
(0009-0005-6911-7295)

February 8, 2026

Abstract

We study a continuous-time control system subject to adversarial disturbances and direct actuator-side signal injection. A state-dependent actuator feasible set is imposed as a physical constraint. We prove that high-impact control inputs are physically unreachable outside a designated gated state region, independently of the controller, disturbance, or adversarial injection. The result establishes a safety property in the form of adversarial unreachability enforced at the physical layer.

Contents

1 Problem Formulation

We consider a safety verification problem for a continuous-time system with actuator saturation. The goal is to enforce a hard safety property that cannot be violated even by adversarial actuator-side input injection.

> High-impact actuator inputs must be physically impossible unless the system state lies in a designated admissible region.

The property is enforced independently of the control law.

2 System Model

2.1 State

Let

$$x(t) \in \mathcal{X} \subseteq \mathbb{R}^n \quad \text{and} \quad a(t) \in \mathcal{A} \subseteq \mathbb{R}^{n_a}$$

denote the system state and an auxiliary (anchor) state. Define the extended state

$$z(t) = \begin{bmatrix} x(t) \\ a(t) \end{bmatrix} \in \mathcal{Z} \subseteq \mathbb{R}^{n+n_a}.$$

2.2 Inputs

The system is subject to three classes of inputs:

- nominal control input $u(t) \in \mathcal{U} \subseteq \mathbb{R}^m$,
- disturbance input $d(t) \in \mathcal{D} \subseteq \mathbb{R}^p$,
- adversarial actuator injection $w(t) \in \mathcal{W} \subseteq \mathbb{R}^m$.

All signals are assumed measurable and essentially bounded on finite intervals.

2.3 Dynamics

The system evolves according to

$$\dot{z}(t) = F\big(z(t), u_{\text{phys}}(t), d(t)\big), \quad \text{a.e. } t. \tag{1}$$

Assumption 2.1 (Existence of trajectories)**.** For any initial condition and admissible inputs, (1) admits an absolutely continuous solution.

3 High-Impact Inputs

Definition 3.1 (High-impact input set)**.** Fix a norm $\|\cdot\|$ on $\mathbb{R}^m$ and a threshold $\tau > 0$. Define

$$\mathcal{U}_{\text{HI}} \coloneqq \{u \in \mathbb{R}^m : \ \|u\| \geq \tau\}.$$

Inputs in $\mathcal{U}_{\text{HI}}$ are considered unsafe or high-impact.

4 Admissible State Region

Definition 4.1 (Gated state set)**.** Let $h_i : \mathcal{Z} \to \mathbb{R}$, $i = 1, \ldots, r$, be given functions. Define the admissible (gated) state set

$$\mathcal{Z}_g = \{z \in \mathcal{Z} : \ h_i(z) \geq 0, \ \forall i\}.$$

Only states in $\mathcal{Z}_g$ are allowed to realise high-impact inputs.

5 Actuator Feasible Set

5.1 State-dependent saturation

Let constants satisfy

$$0 \le \bar{\tau} < \tau \le \bar{u}.$$

Definition 5.1 (Physical actuator feasible set)**.** Define

$$\mathcal{U}_{\text{phys}}(z) = \begin{cases} \{u : \ \|u\| \le \bar{\tau}\}, & z \notin \mathcal{Z}_g, \\ \{u : \ \|u\| \le \bar{u}\}, & z \in \mathcal{Z}_g. \end{cases}$$

This models a hard, state-dependent actuator authority limit.

5.2 Physical input realisation

The actual actuator input is given by

$$u_{\text{phys}}(t) = \Pi_{\mathcal{U}_{\text{phys}}(z(t))}\big(u(t) + w(t)\big). \tag{2}$$

Remark 5.1. This projection represents unavoidable physical saturation and is independent of software or control logic.

6 Adversarial Model

The adversary is assumed to have maximal power consistent with the physical interface:

- full knowledge of the system model,
- arbitrary disturbance $d(t)$,
- arbitrary nominal control $u(t)$,
- arbitrary actuator-side injection $w(t)$.

No restriction is imposed on coordination between these signals.

7 Safety Property

Definition 7.1 (Physical safety property)**.** The system is said to be *physically non-bypassable* if

$$u_{\text{phys}}(t) \in \mathcal{U}_{\text{HI}} \quad \Rightarrow \quad z(t) \in \mathcal{Z}_g \quad \text{for all } t.$$

8 Main Result

Lemma 8.1 (Exclusion of high-impact inputs outside the gate)**.** If $z(t) \notin \mathcal{Z}_g$, then

$$u_{\text{phys}}(t) \notin \mathcal{U}_{\text{HI}}.$$

Proof. If $z(t) \notin \mathcal{Z}_g$, then

$$\mathcal{U}_{\text{phys}}(z(t)) = \{u : \|u\| \le \bar{\tau}\}.$$

By (2),

$$u_{\text{phys}}(t) \in \mathcal{U}_{\text{phys}}(z(t)),$$

so $\|u_{\text{phys}}(t)\| \le \bar{\tau} < \tau$. Hence $u_{\text{phys}}(t) \notin \mathcal{U}_{\text{HI}}$. □

Theorem 8.1 (Adversarial unreachability of high-impact inputs)**.** For system (1)–(2), under any admissarial disturbance, control, and actuator injection, the physical safety property holds.

Proof. Immediate from the preceding lemma. □

9 Discussion

The safety guarantee is enforced at the actuator level and is therefore independent of controller correctness. Any violation would require altering the physical feasible set itself, not merely the control or software logic.