Data-Driven and Stealthy Deactivation of Safety Filters

Daniel Arnström DANIEL.ARNSTROM@IT.UU.SE
Department of Information Technology, Uppsala University

André M.H. Teixeira ANDRE.TEIXEIRA@IT.UU.SE
Department of Information Technology, Uppsala University

While learning-based controllers can improve performance over classical controllers (Coulson et al., 2019; Dörfler, 2023), they are seldom used in safety-critical applications due to their lack of *safety guarantees*. So-called *safety filters* (Wabersich et al., 2023; Tomlin et al., 2003; Ames et al., 2017; Wabersich and Zeilinger, 2018; Hobbs et al., 2023) can, in a modular fashion, augment any such unsafe learning-based controller with safety guarantees. A safety filter takes in a desired control action and the current state of the system, and outputs a filtered control action that guarantees a "safe" behaviour of the system. Since such filters separate safety from performance, any controller from the plethora of data-driven and learning-based controllers can be combined with a safety filter to give good performance together with safety guarantees.

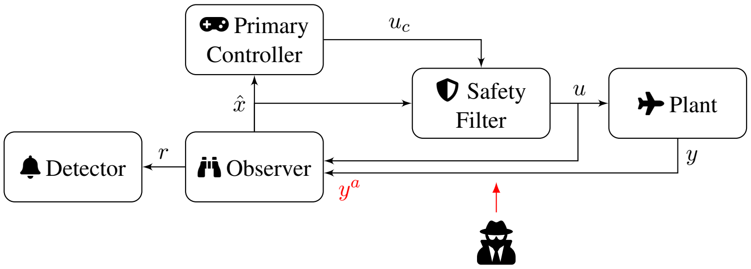
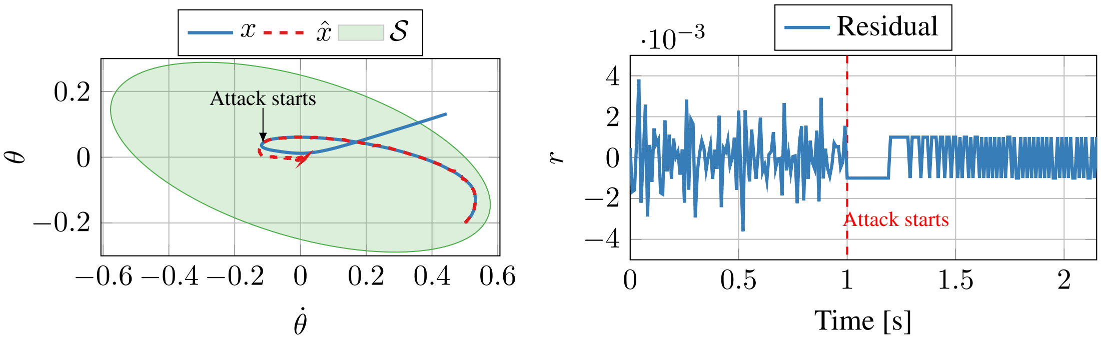


Figure 1: Overview of the system architecture considered. The safety filter produces a safe control action u given a desired control u_c and estimated state $\hat{x}$. An adversary tries to deactivate this filter through false-data injections on the communication channel between the sensors and the observer by replacing the true measurement y with a synthetic measurement y^a.

At the same time, cyber-physical systems (CPSs) are becoming more prevalent and enable advanced control systems that are efficient and resilient Dibaji et al. (2019). CPSs potential comes from their integration of communication, computation, and control technologies. The cyber component of CPSs do, however, open up for new vulnerabilities in the form of cyber attacks (Teixeira et al., 2015)(Annaswamy et al., 2023, §4.C). Since safety filters are the last layer before a control command is applied, they are prime targets for cyber attacks; moreover, if such an attack successfully compromises a safety-critical system, the consequence can be severe (Annaswamy et al., 2023, §4.B).

In this work, we consider cyber attacks that target safety filters. In particular, we consider an attack that injects false data on the communication channel from sensor measurements to a state observer, as illustrated in Figure 1. The goal of the attack is to produce synthetic measurements y^a that "deactivate" the safety filter, which in turn allows for dangerous control actions to be applied to the plant.



(a) Actual and percieved state trajectory

(b) Residual $r = y - \hat{y}$

Figure 2: The actual and perceived state trajectories, and the corresponding residual r, when the proposed false-data injection attack is initiated at $t = 1$. Consequently, the state leaves the safe set.

First, we consider the setting in Arnström and Teixeira (2024), where it is assumed that an adversary knows the model used in the safety filter and observer. Under these conditions, we show how the attacker can perform a false-data injection attack to deactivate the safety filter, leading to the states leaving the safe set (Figure 2a) while the attack remains undetected (Figure 2b). Moreover, we show to construct a detector that, in contrast to conventional detectors, detects these types of attacks.

Finally, we make the setting more realistic by relaxing the assumptions on that knowledge of the adversary; namely we assume that the adversary: (i) does not know the dynamics of the system; (ii) does not know the safety set that the safety filter uses; (iii) does not know the observer gain. Instead, the adversary learns all of these components by just observing data from the observer. Even in this more realistic scenario, we show that the adversary can perform an attack that deactivates the safety filter.

References

Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876, 2017. doi: 10.1109/TAC.2016.2638961.

Anuradha M. Annaswamy, Karl H. Johansson, and George J. Pappas. *Control for Societal-scale Challenges: Road Map 2030*. IEEE Control Systems Society, May 2023.

Daniel Arnström and André M.H. Teixeira. Stealthy deactivation of safety filters. In *2024 European Control Conference (ECC)*, pages 3077–3082, 2024. doi: 10.23919/ECC64448.2024.10590978.

Jeremy Coulson, John Lygeros, and Florian Dörfler. Data-enabled predictive control: In the shallows of the DeePC. In *2019 18th European Control Conference (ECC)*, pages 307–312. IEEE, 2019.

Seyed Mehran Dibaji, Mohammad Pirani, David Bezalel Flamholz, Anuradha M Annaswamy, Karl Henrik Johansson, and Aranya Chakrabortty. A systems and control perspective of CPS security. *Annual reviews in control*, 47:394–411, 2019.

Florian Dörfler. Data-driven control: Part two of two: Hot take: Why not go with models? *IEEE Control Systems Magazine*, 43(6):27–31, 2023. doi: 10.1109/MCS.2023.3310302.

Kerianne L Hobbs, Mark L Mote, Matthew CL Abate, Samuel D Coogan, and Eric M Feron. Runtime assurance for safety-critical systems: An introduction to safety filtering approaches for complex control systems. *IEEE Control Systems Magazine*, 43(2):28–65, 2023.

André Teixeira, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135–148, 2015.

Claire J Tomlin, Ian Mitchell, Alexandre M Bayen, and Meeko Oishi. Computational techniques for the verification of hybrid systems. *Proceedings of the IEEE*, 91(7):986–1001, 2003.

Kim P Wabersich and Melanie N Zeilinger. Linear model predictive safety certification for learning-based control. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 7130–7135. IEEE, 2018.

Kim P. Wabersich, Andrew J. Taylor, Jason J. Choi, Koushil Sreenath, Claire J. Tomlin, Aaron D. Ames, and Melanie N. Zeilinger. Data-driven safety filters: Hamilton-Jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, 43(5):137–177, 2023. doi: 10.1109/MCS.2023.3291885.