Fault detection via output-based barrier functions

Ballotta, Luca; Peruffo, Andrea; Ferrari, Riccardo M.G.; Mazo, Manuel

DOI
10.1016/j.ejcon.2025.101283

Publication date
2025

Document Version
Final published version

Published in
European Journal of Control

European Journal of Control 86 (2025) 101283

Contents lists available at ScienceDirect

European Journal of Control

journal homepage: www.sciencedirect.com/journal/european-journal-of-control

Fault detection via output-based barrier functions[☆]

Luca Ballotta[*], Andrea Peruffo, Riccardo M.G. Ferrari, Manuel Mazo Jr.

Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

ARTICLE INFO

Recommended by T. Parisini

Keywords:
Barrier function fault detection
Fault diagnosis
Fault-tolerant control
Nonlinear systems

ABSTRACT

Model-based fault detection identifies anomalies by comparing a system's output with the prediction from a model. Although such a technique can be very powerful, it may suffer from the computational complexity of its underlying models, especially for large systems. An alternative approach that circumvents this cost increase uses barrier functions, which abstract the system's behaviour into a single value. In this paper, we propose a fault detection mechanism via output-based barrier functions, that does not require to estimate the full state, copes with noisy processes, and is tailored to safety-critical faults as given by a user-defined safe region. We leverage such a mechanism by introducing so-called *p-fault tolerant sets*, which guarantee that a faulty system requires at least p time steps before reaching any unsafe state. Our approach is validated through numerical experiments on two systems with linear and nonlinear dynamics, along with the classic three-tank model.

1. Introduction

Safety is a crucial requirement in many cyber–physical systems (CPSs) such as industrial, automotive, or robotic applications, and requires keeping the CPS away from undesired configurations. A common way to do so is by imposing that its state remains within a predetermined safe region at all times. Ensuring safety, especially in presence of noise, uncertain or faulty dynamics, is thus a critical task of a system designer. A well established way to mitigate faults uses fault detection and isolation algorithms (Ding, 2008), in combination with a fault accommodation policy (Blanke, Kinnaert, Lunze, & Staroswiecki, 2015). Model-based approaches address the detection and isolation problem by continuously measuring the system's outputs, and comparing them to what is predicted by a model of the fault-free system. A powerful model-based scheme is the observer-based one, which employs one or more dynamical observers to detect and classify faults. A typical choice for the observer entails a Kalman filter (KF) for linear systems, or an extended KF or a particle filter for nonlinear systems — for a detailed discussion, see *e.g.,* Chen and Patton (2012), Ding (2008). However, these strategies may be computationally expensive, require ad-hoc designs for nonlinear models, and raise alarms whether or not the operating conditions are threatening the safety of the CPS. Most of all, they do not provide out of the box any formal guarantees that faults are detected early enough such that they can be accommodated before the system goes unsafe.

An alternative approach to CPS safety is constituted by barrier functions (BFs) or barrier certificates (Prajna, 2006; Prajna, Jadbabaie, & Pappas, 2007), which formally certify safety-by-design of a closed-loop system by forward invariance of a user-defined safe set (or avoidance of an unsafe set). This approach treats in a unified way a large class of nonlinear systems and provides a computationally tractable design, as BFs can be constructed offline, *e.g.,* via sum-of-squares decomposition (Papachristodoulou et al., 2021; Parrilo, 2000). Unfortunately, BFs are sensitive to uncertainty in the dynamical model used to synthesize them, including noise and delays, which requires extra care when using them for safety-critical operation.

Related Works. While both fault detection and (control) barrier functions (CBFs) are well established topics for CPS safety, their mutual connections have not been much explored. Early work (Prajna, 2006) proposes parametric model validation with BFs, where the parameters may account for uncertainty or faults. The work (Clark, Li, & Zhang, 2020) makes a CBF robust to sensor faults via reduced state estimators, each ignoring a (subset of) measurement(s). The computational cost of this approach grows proportionally to the combinations of faulty sensors. In a similar spirit, paper (Garg, Dawson, Xu, Ornik, & Fan, 2023) uses a bank of heuristic CBFs where each neglects a control input, to be robust to one faulty actuator. The authors in Wijk, Majji, and Hobbs (2024) use an extended KF to inform the CBF about abrupt changes in the dynamics, mitigating the effects of faults on the controller safety filter. Paper (Vyas, Roy, & Dey, 2022) uses a CBF to make a Lithium-ion

[☆] This work has been partially supported by the Horizon Europe program through the project TWAIN, grant id. 101122194, and by the Netherlands Enterprise Agency (RVO) through the TKI Offshore Energy program under the project PHYSIA, grant id. TKITOE_WOZ_2414.

[*] Corresponding author.

E-mail addresses: l.ballotta@tudelft.nl (L. Ballotta), r.ferrari@tudelft.nl (R.M.G. Ferrari), m.mazo@tudelft.nl (M. Mazo Jr.).

https://doi.org/10.1016/j.ejcon.2025.101283
Received 23 May 2025; Accepted 8 July 2025
Available online 29 July 2025

battery cell controller robust to thermal anomalies. These works assume that the faults can be precisely characterised or that safe behaviour can always be recovered through (a set of) modified CBF(s), which is a rather nontrivial requirement. Related works (Dean, Taylor, Cosner, Recht, & Ames, 2021; Garg & Panagou, 2021; Lindemann et al., 2024; Takano & Yamakita, 2018) robustify CBFs against model uncertainty or measurement noise but neglect faults.

Contributions. This paper proposes, to the best of our knowledge for the first time, *to use a BF for fault detection in safety-critical CPS*. We do this without estimating the inner state of the system by formalising a tailored output-based BF (Section 2). This approach *certainly* triggers alarms if the state trajectory enters a pre-defined unsafe set and tailors *critical faults*, remaining silent only if the fault does not impact the safety of the CPS. This is fundamentally different from common observer-based fault detectors, which assess the mismatch between estimation and measurements and cannot discern whether the actual dynamics is driving the system to unsafe configurations, resulting in potential conservatism. Further, we introduce the notion of *p-fault-tolerant* set, defining a portion of the state space from where trajectories enter the unsafe set in at least $p+1$ time steps under every faulty condition — this guarantees a minimum buffer time after detection, allowing to take action before the system's safety is compromised (Section 3). In particular, our design offers formal guarantees in a deterministic sense under bounded process and measurement noises, ensuring that any safety-critical fault is detected at least p time steps before the state becomes unsafe. Finally, we prove out method effective with numerical tests (Section 4).

2. System description and problem formulation

2.1. System model and fault detection task

System Model. Consider the discrete-time system

$$\begin{aligned} x(k+1) &= f(x(k), w(k), \phi(k)) \\ y(k) &= h(x(k), v(k)), \end{aligned} \tag{1}$$

where $k \in \mathbb{N}_+$ is the time step, the state $x(k) \in \mathcal{X} \subseteq \mathbb{R}^n$, and the measurements $y(k) \in \mathbb{R}^p$ are affected by measurement noise $v(k) \in \mathcal{V}$. The process noise $w(k) \in \mathcal{W}$ captures modelling uncertainties, while faults are represented by the unknown signal $\phi(k) \in \mathcal{F}$. We name "healthy" the fault-free dynamics $f(x,w,0)$ and "nominal" the noise- and fault-free dynamics $f_{\text{n}}(x) \doteq f(x,0,0)$. Let us introduce some definitions and assumptions needed for our problem definition.

Assumption 1. There exist constants $\bar{v}, \bar{w} \geq 0$ such that $\|v\| \leq \bar{v}\ \forall v \in \mathcal{V}$ and $\|w\| \leq \bar{w}\ \forall w \in \mathcal{W}$.

Fault Detection. To ensure safe operations, the state $x \in \mathcal{X}$ must remain outside an *unsafe set* $\mathcal{U} \subset \mathcal{X}$ that represents dangerous or undesired operation, *e.g.,* overheating or distance from an equilibrium. We name *admissible* the set $\mathcal{S} \doteq \mathcal{X} \setminus \mathcal{U}$.

Assumption 2. The admissible set $\mathcal{S}$ is compact.

We aim at detecting faults that drive the state of system (1) to the unsafe set $\mathcal{U}$. This is achieved through a fault detector that runs in real time and continually assesses if the state trajectory remains in the admissible set $\mathcal{S}$. Meanwhile, we require two features from the detector:

1. It must be computationally fast;
2. It should not trigger false alarms when faults that cannot drive the system unsafe occur.

These requirements are important to avoid interrupting operation when unnecessary and maintain high performance.

2.2. Output-based barrier function

Let us give the formal definition of barrier function.

Definition 1 (*Discrete-Time Barrier Function (Ahmadi, Singletary, Burdick, & Ames, 2019)*)**.** The continuous function $g : \mathbb{R}^n \to \mathbb{R}$ is a discrete-time barrier function for a system $x(k+1) = f(x(k))$, for the set

$$\mathcal{C} := \{x \in \mathcal{X} \mid g(x) \leq 0\}, \tag{2}$$

if there exists an extended class $\mathcal{K}$ function α satisfying $\alpha(r) < r$ for all $r > 0$ and a set $\mathcal{D} \supseteq \mathcal{C}$ such that

$$g(x(k+1)) - g(x(k)) \leq \alpha(-g(x(k))), \quad \forall x \in \mathcal{D}. \tag{3}$$

If a DTBF g exists, its sublevel set $\mathcal{C}$ in (2) is forward invariant (Ahmadi et al., 2019), which is a key feature we leverage later for fault detection. Definition 1 considers autonomous systems, corresponding to the nominal dynamics f_{n}, and full knowledge of the state. Let us now introduce an *output-based BF* that certifies that the state x never enters the unsafe set $\mathcal{U}$ based on output y. For clarity of presentation, we first present the case with healthy dynamics $f(x,w,0)$; this procedure is *not* robust to safety-critical faults. We present a robust barrier synthesis in Section 3, which we use for fault detection.

Hence, let us address the design of (output-based) BF $g : \mathbb{R}^p \to \mathbb{R}$ that takes a measurement y as argument. The relation between the output-based BF g and the state x is

$$\tilde{g}(x; v) \doteq g(h(x,v)) = g(y). \tag{4}$$

Poor sensing makes it impossible to infer safety based on the output. For instance, if a safety-critical state component x_i does not affect y, every $\tilde{g}$ parametrised as in (4) is useless for safety. To preclude such cases, we require a basic condition.

Assumption 3. The system (1) is detectable.

Assumption 3 is standard, *e.g.,* to make the covariance of a Kalman filter converge. We design the BF such that the admissible set $\mathcal{S}$ contains the sublevel set of $\tilde{g}$:[1]

$$\mathcal{S} \supseteq \mathcal{C}_{\tilde{g}}(0) \doteq \{x \in \mathcal{X} : \tilde{g}(x;0) \leq 0\}. \tag{5}$$

The BF is a safety certificate if $\mathcal{C}_{\tilde{g}}(0)$ is forward invariant, implying that the state trajectory never enters $\mathcal{U}$ if $x(0) \in \mathcal{C}_{\tilde{g}}(0)$. Under nominal dynamics f_{n}, this is guaranteed by condition (3), which in the space of measurements and without measurement noise, *i.e.,* with $y = h(x,0)$, is equivalent to

$$g(y(k+1)) - g(y(k)) \leq \alpha(-g(y(k))), \quad \forall x(k) \in \mathcal{S}. \tag{6}$$

Nonetheless, condition (6) holds also in presence of bounded process and measurement noises $w \neq 0, v \neq 0$ by setting suitable bounds on the functions f, h, g, and α.

Assumption 4. For all $x \in \mathcal{S}$, $w \in \mathcal{W}$, and $v \in \mathcal{V}$, it holds

$$|\tilde{g}(x;v) - \tilde{g}(x;0)| \leq \mathcal{L}_{\tilde{g}} \|v\| \tag{7}$$

$$|\tilde{g}(f_{\text{n}}(x);v) - \tilde{g}(f_{\text{n}}(x);0)| \leq \mathcal{L}_{\tilde{g}\circ f_{\text{n}}} \|v\| \tag{8}$$

$$|\alpha(-\tilde{g}(x;v)) - \alpha(-\tilde{g}(x;0))| \leq \mathcal{L}_{\alpha\circ\tilde{g}} \|v\| \tag{9}$$

$$|\tilde{g}(f(x,w,0);v) - \tilde{g}(f(x,0,0);v)| \leq \mathcal{L}_{\tilde{g}\circ f} \|w\| \tag{10}$$

The next result provides us with a sufficient condition for (3) to hold under bounded noises.

Proposition 1. *Let Assumptions 1 and 4 hold and define*

$$\beta \doteq (\mathcal{L}_{\tilde{g}} + \mathcal{L}_{\tilde{g}\circ f_{\text{n}}} + \mathcal{L}_{\alpha\circ\tilde{g}})\bar{v} + \mathcal{L}_{\tilde{g}\circ f}\bar{w}. \tag{11}$$

[1] Adapting the calculations and problem formulation to the superlevel set convention, *e.g.,* Ames, Xu, Grizzle, and Tabuada (2017), requires straightforward manipulations.

Then, $\mathcal{C}_{\tilde{g}}(0) \subseteq \mathcal{S}$ is forward invariant under $\phi = 0$ if there exists an extended class $\mathcal{K}$ function α such that, for all $x \in \mathcal{S}$,

$$\sup_{\substack{v,\xi\in\mathcal{V}\\ w\in\mathcal{W}}} \tilde{g}(f(x,w);\xi) - \tilde{g}(x;v) - \alpha(-\tilde{g}(x;v)) \leq -\beta. \tag{12}$$

Proof. For every noise values $v, \xi \in \mathcal{V}$ and $w \in \mathcal{W}$ it holds

$$\begin{aligned}\tilde{g}(f_{\mathrm{n}}(x);0) &\overset{(8)}{\leq} \tilde{g}(f_{\mathrm{n}}(x);\xi) + \mathcal{L}_{\tilde{g}\circ f_{\mathrm{n}}}\bar{v}\\ &\overset{(10)}{\leq} \tilde{g}(f(x,w);\xi) + \mathcal{L}_{\tilde{g}\circ f}\bar{w} + \mathcal{L}_{\tilde{g}\circ f_{\mathrm{n}}}\bar{v}\\ &\overset{(12)}{\leq} \alpha(-\tilde{g}(x;v)) + \tilde{g}(x;v) - \mathcal{L}_{\tilde{g}}\bar{v} - \mathcal{L}_{\alpha\circ\tilde{g}}\bar{v}\\ &\overset{(7)}{\leq} \alpha(-\tilde{g}(x;v)) - \mathcal{L}_{\alpha\circ\tilde{g}}\bar{v} + \tilde{g}(x;0)\\ &\overset{(9)}{\leq} \alpha(-\tilde{g}(x;0)) + \tilde{g}(x;0).\end{aligned} \tag{13}$$

Hence, condition (12) implies (3). Because (12) holds for all $x \in \mathcal{S}$ and in view of (5), the sublevel set $\mathcal{C}_{\tilde{g}}(0)$ is forward invariant following standard arguments, *e.g.,* Ahmadi et al. (2019). □

The condition (12) has the nuance of depending on both the state x and the noises w, v, which complicates numerically computing an output-based BF $\tilde{g}$. However, re-applying the smoothness arguments in (13) yields the following condition, which is more conservative but involves only the state x:

$$\Delta_{\mathrm{n}}\tilde{g}(x;0) \leq \alpha(-\tilde{g}(x;0)) - 2\beta \quad \forall x \in \mathcal{S}, \tag{14}$$

where we define $\Delta_{\mathrm{n}}\tilde{g}(x;0) \doteq \tilde{g}(f_{\mathrm{n}}(x);0) - \tilde{g}(x;0)$ for the sake of brevity. If condition (14) holds, then also (12) holds. The latter condition, which comprises all possible noisy measurements $y(k) = h(x(k), v(k))$ of a state $x(k)$ and all uncertain, fault-free transitions $x(k+1) = f(x(k), w(k), 0)$, translates into a fault detector test during operation as

$$\alpha(-g(y(k))) - (g(y(k+1)) - g(y(k))) - \beta \geq 0. \tag{15}$$

In particular, if condition (15) fails at some time k, then $\tilde{g}$ is not a valid barrier function for the system at that time and safety is not formally certified, hinting at possible faults.

The constant β makes (15) more conservative than (6) according to the smoothness of the dynamics and sensing models and of the output-based BF. Without noises, it holds $\bar{w} = \bar{v} = \beta = 0$ and condition (15) reduces back to (6).

Remark 1. The inequalities in (3) and (14) must theoretically hold in the sublevel set $\mathcal{C}_{\tilde{g}}(0)$ to make this forward invariant, instead of the entire set $\mathcal{S}$ as we impose, which is more conservative. However, our choice improves the synthesis of $\tilde{g}$ via convex optimisation, as discussed later, which is not possible if one enforces the more relaxed condition $x \in \mathcal{C}_{\tilde{g}}(0)$.

3. Barrier function for fault detection

Thus far, the presented barrier function considers solely healthy, fault-free dynamics. While robustness to bounded measurement noise v and process noise w is accommodated via β in the condition (15), there is no safety guarantee about faults ϕ. In this Section, we formalise the design of a robust output-based BF tailored to fault detection.

3.1. Fault-aware barrier function design

Since we are interested in faults that drive the state unsafe, we leverage the property that the sublevel set $\mathcal{C}_{\tilde{g}}(0)$ is invariant under healthy dynamics. Indeed, if $x(k^*) \in \mathcal{U}$, the trajectory must violate the condition (15) at some time $k \leq k^*$. This observation offers a simple way to detect faults that can potentially drive the state trajectory unsafe.

Remark 2. Our technique detects *only* faults that can drive the state towards an unsafe region. If a fault occurs but safety is *not* impaired, the condition (15) prevents the detection.

However, violation of (15) may occur exactly at time k^*, when the state enters the unsafe set, providing a tardive detection. To prevent this from happening, we add a robust condition that accounts for faults in a safe region of the state space. To this aim, let us introduce the notion of *p-fault-tolerant set*, a key concept to fault-aware design.

Definition 2. Given a parameter $p \in \mathbb{N}$, a set $\mathcal{S}_0^p \subseteq \mathcal{S}$ is a *p-fault-tolerant set* for (1) if

$$f^k(x, w, \phi) \in \mathcal{S} \quad \forall (w, \phi) \in \mathcal{Z}_k^p, \forall k \in [p], \tag{16}$$

where $f^k = f \circ \ldots \circ f$ denotes k iterates of f, and we define $[p] \doteq \{1, \ldots, p\}$ and the set $\mathcal{Z}_k^p \doteq \mathcal{S}_0^p \times \mathcal{W}^k \times \mathcal{F}^k$. The *buffer set* associated with $\mathcal{S}_0^p$ is $\mathcal{S}_{\mathrm{b}}^p \doteq \mathcal{S} \setminus \mathcal{S}_0^p$.

In words, every state $x \in \mathcal{S}_0^p$ takes at least $p+1$ steps to enter the unsafe set $\mathcal{U}$ under all faulty dynamics. This means that a fault detected within $\mathcal{S}_0^p$ guarantees a buffer of (at least) p time steps before the state actually enters the unsafe set, during which it evolves within the buffer set $\mathcal{S}_{\mathrm{b}}^p$.

In practice, the parameter p represents the time necessary to take action if a fault is detected, *e.g.,* to switch to a robust controller.

Computing a nontrivial p-fault-tolerant set directly in state space may be challenging. Further, we must check if $x \in \mathcal{S}_0^p$ from measurements y. The output-based BF offers a way to circumvent both issues. Given a BF $\tilde{g}$, we instantiate $\mathcal{S}_0^p$ as

$$\mathcal{S}_0^p = \mathcal{C}_{\tilde{g}}(-\gamma) = \{x \in \mathcal{X} : \tilde{g}(x;0) \leq -\gamma\} \tag{17}$$

where $\gamma \geq 0$ is computed as the minimum value that verifies

$$\tilde{g}(f^k(x, w, \phi); 0) \leq 0 \quad \forall (x, w, \phi) \in \mathcal{Z}_k^p, \forall k \in [p]. \tag{18}$$

The choice (17)–(18) guarantees that all states in the fault-tolerant set $\mathcal{S}_0^p$ evolve within the sublevel set $\mathcal{C}_{\tilde{g}}(0)$, and thus do not enter $\mathcal{U}$, for at least p steps. Further, it holds

$$\tilde{g}(x;v) \leq -\gamma - \mathcal{L}_{\tilde{g}}\bar{v} \implies \tilde{g}(x;0) \leq -\gamma. \tag{19}$$

Thus, if the output-based condition $g(y) \leq -\gamma - \mathcal{L}_{\tilde{g}}\bar{v}$ holds, the state x is in the fault-tolerant set $\mathcal{S}_0^p$ and faults are detected at least p steps before the state trajectory enters the unsafe set $\mathcal{U}$. Naturally, the existence of γ and $\mathcal{S}_0^p$ depends on the actual dynamics (1) and the parameter p; for some pathological dynamical flows, or if $\bar{w}$ is very large, $\mathcal{S}_0^p$ may be empty. This scenario corresponds to systems where the unsafe set $\mathcal{U}$ can be reached in p time steps from any point of the domain.

While faults detected within the p-fault tolerant set $\mathcal{S}_0^p$ will not drive the system unsafe for at least p steps, this is not guaranteed if detection occurs outside $\mathcal{S}_0^p$. Hence, we aim at maximising the volume of $\mathcal{S}_0^p$. This finally leads us to formulating our design problem, as follows.

Problem 1 (*Fault-Aware Barrier Function Design*). Given system (1), an extended class $\mathcal{K}$ function α, and a safety parameter $p \geq 1$, find an output-based BF g with maximal p-fault-tolerant set $\mathcal{S}_0^p$ designed according to (17):

$$\max_{\substack{g\in\mathcal{C}_0\\ \gamma\geq 0}} \quad \mathrm{vol}(\mathcal{S}_0^p) \tag{20a}$$

$$\text{s.t.} \qquad \tilde{g}(x;0) > 0 \qquad \forall x \in \mathcal{U} \tag{20b}$$

$$\Delta_{\mathrm{n}}\tilde{g}(x;0) \leq \alpha(-\tilde{g}(x;0)) - 2\beta \;\; \forall x \in \mathcal{S} \tag{20c}$$

$$\max_{k\in[p]} \tilde{g}(f^k(x, w, \phi); 0) \leq 0 \qquad \forall (x, w, \phi) \in \mathcal{Z}_k^p. \tag{20d}$$

In Problem 1, the constraints (20b)–(20c) certify that the set $\mathcal{C}_{\tilde{g}}(0)$ is safe (forward invariant) under healthy dynamics, while constraint (20d) ensures that the set $\mathcal{S}_0^p = \mathcal{C}_{\tilde{g}}(-\gamma)$ is p-fault-tolerant and guarantees timely detection of faults.

The following straightforward result states that Problem 1 is generally well posed.

Proposition 2. *Problem (20) is feasible if and only if the sub-problem (20a)–(20c) is feasible.*

***Proof*.** If $\gamma > -\min_{x\in\mathcal{S}} \tilde{g}(x;0)$, then $\mathcal{S}_0^p = \emptyset$ and the constraint (20d) is always inactive. □

Proposition 2 asserts that a trivial solution to Problem 1 is an output-based BF $\tilde{g}$ that satisfies (20b)–(20c) with $\mathcal{S}_0^p = \emptyset$. However, this is undesired for fault detection: it means that no output-based BF certifies safety under faulty dynamics for any point in the admissible region, namely the trajectory may go unsafe in less than p steps. Nonetheless, Problem 1 admits nontrivial solutions by setting suitable bounds on the faulty dynamics, which we formally state in the next Section.

Remark 3. A common theme when computing a BF for a user-defined admissible set $\mathcal{S}$ is to find the largest certifiably safe invariant set $\mathcal{C}_{\tilde{g}}(0) \subseteq \mathcal{S}$, which has been addressed in, *e.g.,* Clark (2021), Wang, Han, and Egerstedt (2018). While adapting the algorithms proposed therein to our setup is nontrivial, Problem 1 indeed relates to this theme because it seeks the largest sublevel set $\mathcal{C}_{\tilde{g}}(-\gamma)$ that is certifiably robust to faulty dynamics.

3.2. Maximising set volume

Computing $\mathrm{vol}(\mathcal{S}_0^p)$ in (20a) is challenging. To simplify Problem 1, we pick the largest ball in the fault-tolerant set:

$$\mathcal{B}_0^p \doteq \mathcal{B}_{r_0}(c), \quad r_0 \doteq \max r : \mathcal{B}_r(c) \subseteq \mathcal{S}_0^p, \tag{21}$$

where $\mathcal{B}_r(c) \doteq \left\{x : \|x-c\|_2 \le r\right\}$. The centre c may represent a reference operating configuration, *e.g.,* an equilibrium. This expedient allows us to (under-)approximate the cost $\mathrm{vol}(\mathcal{S}_0^p)$ with $\mathrm{vol}(\mathcal{B}_0^p)$, which is just proportional to r_0, and simplify Problem 1 as follows.

Problem 2. Given system (1), an extended class $\mathcal{K}$ function α, and a safety parameter $p \ge 0$, find an output-based BF g with maximal fault-tolerant ball $\mathcal{B}_0^p$ according to (17) and (21):

$$\max_{\substack{g\in\mathcal{C}_0\\ \gamma\ge 0}} \quad r_0 \tag{22a}$$

$$\text{s.t.} \qquad \tilde{g}(x;0) > 0 \qquad \forall x \in \mathcal{U} \tag{22b}$$

$$\Delta_{\mathrm{n}}\tilde{g}(x;0) \le \alpha(-\tilde{g}(x;0)) - 2\beta \;\; \forall x \in \mathcal{S}^p \tag{22c}$$

$$\max_{k\in[p]} \tilde{g}(f^k(x,w,\phi);0) \le 0 \qquad \forall (x,w,\phi) \in \mathcal{Z}_k^p. \tag{22d}$$

The bound (22d) remains the most complex element of Problem 2 as it must be satisfied jointly considering the states in $\mathcal{S}_0^p$, the noises, and the faults. Trivially, if faults can arbitrarily degrade the healthy system, the set $\mathcal{S}_0^p$ is always empty. A structural necessary condition to ensure $\mathcal{S}_0^p \neq \emptyset$ is the existence of a state such that every p faulty transitions stay in the admissible set, namely $\exists x^* \in \mathcal{S}$ such that $f^k(x^*,w,\phi) \in \mathcal{S}\ \forall w \in \mathcal{W}^k, \forall \phi \in \mathcal{F}^k, \forall k \in [p]$. Therefore, we state the following assumption.

Assumption 5. There exists $\bar{\phi} \ge 0$ such that $\|\phi\| \le \bar{\phi}\ \forall \phi \in \mathcal{F}$. Further, for $x \in \mathcal{S}$, $w \in \mathcal{W}$, and $k \in [p]$ it holds

$$|\tilde{g}(f^k(x,w,\phi);0) - \tilde{g}(f^k(x,w,0);0)| \le \mathcal{L}_{\phi k} \|\phi\|. \tag{23}$$

An upper bound on the constant $\mathcal{L}_{\phi k}$ can be found as follows,

$$\mathcal{L}_{\phi k} \le \max_{x\in\mathcal{S}, w\in\mathcal{W}^k} \left\|\nabla_\phi \tilde{g}(f^k(x,w,\phi);0)\right\|. \tag{24}$$

Then, if $I-\alpha$ is class $\mathcal{K}$ and the following condition holds,

$$(I-\alpha)^p(\gamma) \ge \max_{\substack{(x,w,\phi)\in\mathcal{Z}_k^p\\ k\in[p]}} \left\|\nabla_\phi \tilde{g}(f^k(x,w,\phi);0)\right\| \|\phi\|, \tag{25}$$

also (22d) holds, since, for every $x \in \mathcal{S}_0^p$ and $k \in [p]$,

$$\begin{aligned}
\tilde{g}(f^k(x,w,\phi);0) &\overset{(23)}{\le} \tilde{g}(f^k(x,w,0);0) + \mathcal{L}_{\phi k}\|\phi\| \\
&\overset{(10)}{\le} \tilde{g}(f^k(x,0,0);0) + \mathcal{L}_{\tilde{g}\circ f}\|w\| + \mathcal{L}_{\phi k}\|\phi\| \\
&\overset{(22c)}{\le} -(I-\alpha)^k(-\tilde{g}(x;0)) + \mathcal{L}_{\phi k}\|\phi\| \\
&\overset{(17)}{\le} -(I-\alpha)^k(\gamma) + \mathcal{L}_{\phi k}\|\phi\| \\
&\overset{(24)–(25)}{\le} -(I-\alpha)^p(\gamma) + (I-\alpha)^p(\gamma) = 0.
\end{aligned} \tag{26}$$

Hence, we replace (22d) with the more conservative (25).

Remark 4. This synthesis technique and particularly Assumptions 1 and 5 ensure the construction of a non-empty p-fault-tolerant set $\mathcal{S}_0^p$, where faults are detected at least p time steps before the system goes unsafe. Note that we do not need an explicit characterisation of $\mathcal{S}_0^p$, as it is sufficient to check if $g(y) \le -\gamma - \mathcal{L}_{\tilde{g}}\bar{v}$. By computing $\mathcal{S}_0^p$ for several values of p, we can instantly know at which "safety level" γ the system currently is. If Assumptions 1 and 5 do not hold, the fault detection based on condition (15) still detects faults by construction, but there is no formal guarantee that these are always detected sufficiently in advance before the state x enters the unsafe set $\mathcal{U}$. However, we argue that statistical or numerical analysis can be carried out in this case to assess the chance of missed detection due to noise bounds not holding true, for instance under Gaussian noises. This is a compelling direction of research that we will explore in future work.

Remark 5. An arbitrary template for g (*e.g.,* d-degree polynomial) might fail to solve Problem 2. However, this holds even for the standard barrier conditions (22b)–(22c) and sets a minimal requirement in terms of computational complexity of the design (Prajna et al., 2007).

3.3. Improving computational aspects

A possible approach to solve Problem 2 involves sum-of-squares (SoS) optimisation. In short, if f and h are polynomials and $\mathcal{U}$ is semi-algebraic, namely it is defined by a polynomial inequality, then we can choose a polynomial template for g and rewrite program (22) as SDP constraints (Papachristodoulou et al., 2021; Prajna et al., 2007), which is convex and can be solved efficiently. However, constraint (22d) and its conservative alternative (25) include the simultaneous optimisation of g and γ, making it nonconvex. To overcome this issue, we use alternations.

Our procedure is summarised in Algorithm 1. The first step fixes an initial (possibly small) radius r_0 and level γ and finds a function g that satisfies (22b)–(22c) – thus solving a feasibility problem. The second step uses the newly found g and finds the maximum radius r_0 for which (21) is valid. Finally, the third steps adjusts γ according to (25). If the result is unsatisfactory, we may iterate the procedure. For computational convenience, during the iterations we update γ replacing $x \in \mathcal{S}_0^p$ with $x \in \mathcal{B}_0^p$ in (25) (Line 5), and after the iterations we adjust γ as per (25). The alternation heuristics exploits the convexity of the first two optimisation steps, but it returns a local solution to the joint nonconvex problem.

On the other hand, we expect monotonic improvement of the objective function, and we can stop the procedure when some pre-defined convergence criteria are met. Other solution methods are also possible (e.g. stochastic gradient descent, sequential quadratic programming) but they all come with different downsides. In practice, the alternation represents a reliable and simple approach.

Algorithm 1: BF design with alternations.

```
Input: Functions $f$, $h$, $\alpha$, sets $\mathcal{U}$, $\mathcal{V}$, $\mathcal{W}$, $\mathcal{F}$, parameters $\beta$, $\gamma^0 > 0$,
       $r^0 > 0$.
Output: $\gamma$.
$\gamma \leftarrow \gamma^0$; $r_0 \leftarrow r^0$;
find $g$ s.t. (22b)–(22c) and $\tilde{g}(x;0) \leq -\gamma \ \forall x \in \mathcal{B}_0^p$;
repeat
    $r_0 \leftarrow \max r$ s.t. $\tilde{g}(x;0) \leq -\gamma \ \forall x \in \mathcal{B}_r(c)$;
    $\mathcal{L}_\phi \leftarrow \max_{(x,w,\phi)\in\mathcal{Z}_k^p, k\in[p]} \left\| \nabla_\phi \tilde{g}(f^k(x,w,\phi);0) \right\|$;
    $\gamma \leftarrow (I-\alpha)^{-p}(\mathcal{L}_\phi \max_{\phi\in\mathcal{F}^p} \|\phi\|)$;
    find $g$ s.t. (22b)–(22c) and $\tilde{g}(x;0) \leq -\gamma \ \forall x \in \mathcal{B}_0^p$;
until termination condition;
$\mathcal{L}_\phi \leftarrow \max_{(x,w,\phi)\in\mathcal{Z}_k^p, k\in[p]} \left\| \nabla_\phi \tilde{g}(f^k(x,w,\phi);0) \right\|$;
while $(I-\alpha)^p(\gamma) < \mathcal{L}_\phi \max_{\phi\in\mathcal{F}^p} \|\phi\|$ do
    increase $\gamma$;
    $\mathcal{L}_\phi \leftarrow \max_{(x,w,\phi)\in\mathcal{Z}_k^p, k\in[p]} \left\| \nabla_\phi \tilde{g}(f^k(x,\ \phi);0) \right\|$;
return $g, \gamma$.
```

Table 1
False detection rate (FDR), missed detection rate (MDR), and time to detection (TTD) with the linear system (27) (top) and nonlinear system (28) (bottom). The TTD is the average number of time steps between fault and detection.

	Noise	FDR	MDR	TTD
KF + χ^2-test	Gaussian Uniform	1.5% 97.5%	27.5% 0.04%	18.4 1.0
Output-based BF (proposed)	Gaussian Uniform	12.9% 10.1%	9.8% 0.25%	2.8 1.1
KF + χ^2-test	Gaussian Uniform	0.5% 9.3%	47.9% 57.8%	29.9 25.3
Output-based BF (proposed)	Gaussian Uniform	25.1% 9.6%	35.1% 43.7%	1.9 3.1

4. Case studies

4.1. Case study 1: A linear system

Let us consider the following linear system,

$$
\begin{aligned}
x(k+1) &= \begin{bmatrix} 0.8 & -0.5 \\ 0 & 0.9 \end{bmatrix} x(k) + w(k) + \phi(k) \\
y(k) &= \begin{bmatrix} 0.1 & -1.7 \\ -1.9 & 0.8 \end{bmatrix} x(k) + v(k).
\end{aligned}
\tag{27}
$$

We choose the unsafe set $\mathcal{U} = \mathbb{R}^2 \setminus \mathcal{B}_3(0)$ and instantiate our barrier function synthesis with fault bound $\|\phi\| \leq 0.2$. We solve Problem 2 with the alternating procedure in Algorithm 1, setting the degree of the barrier equal to 2, α simply as a linear function with constant slope $\bar{\alpha} = 10^{-4}$, and parameters $\beta = 0.1$, $p = 1$. The program returns the safety margin $\gamma = -1200$, corresponding to the dashed level set in Fig. 1.

We then simulate system (27) for 100 time steps, starting from $N = 150$ different initial states, with two types of noise statistics: (i) Gaussian process and observation noises both with zero mean and covariance matrices $10^{-4}I$ and $10^{-6}I$, respectively; and (ii) uniform distribution on $[-10^{-2}, 10^{-2}]^2$ and $[-10^{-3}, 10^{-3}]^2$ for the process and measurement noises, respectively. The simulations are fault-free for $k_f = 30$ time steps, and then we inject a fault as $\phi_2(k) = 0.2x_2(k) \ \forall k \geq k_f$.

Our technique simply asserts the condition (15): if the assertion fails, an alarm is triggered. We compare our detection technique against a classic fault detection technique: the analysis of residuals of a Kalman filter through the χ^2-test (Ding, 2008). The KF uses the *real* covariance matrices of the process and observation noises, and we instantiate the χ^2-test with a sliding window of length 10 setting significance levels 0.01 both for Gaussian and uniform noises.

Fig. 2 shows the distribution of detection times and Table 1 reports the aggregate results. Our approach is overall comparable to KF-based detection under Gaussian noises, with more false alarms but more attentive detection, and performs much better with uniform noises, which make the KF almost always raise alarms. Missed detection occurs with our detector when the system trajectory diverges slowly, and the state never enters the unsafe set through the simulations.

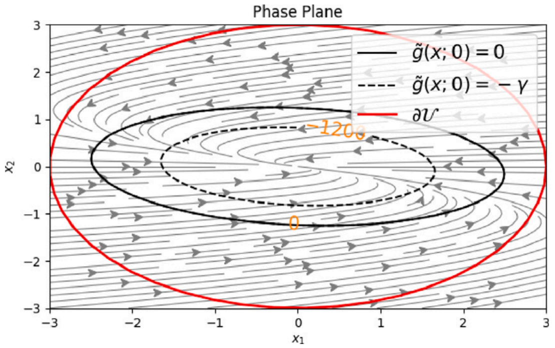

Fig. 1. Phase plane of system (27) and level sets of the output-based barrier function $\tilde{g}$ for levels 0 (solid black) and $-\gamma$ (dashed). The solid red line marks the boundary between admissible set $\mathcal{S} = \mathcal{B}_3(0)$ and unsafe set $\mathcal{U}$.

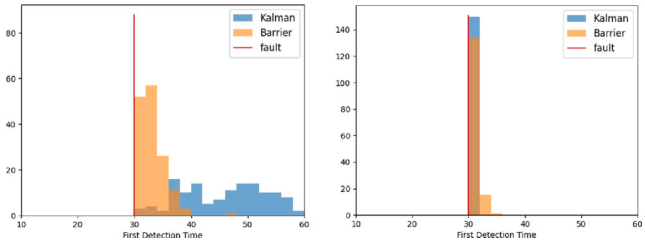

Fig. 2. Histogram of detection times for 150 different trajectories with Gaussian (left) and uniform (right) disturbance. The fault occurs at $k_\mathrm{f} = 30$.

4.2. Case study 2: A nonlinear system

Let us introduce the following nonlinear system,

$$
\begin{aligned}
x(k+1) &= \begin{bmatrix} -0.9x_1(k) + x_2^2(k) \\ -0.75x_2(k) + 0.3x_2^3(k) \end{bmatrix} + w(k) + \phi(k) \\
y(k) &= \begin{bmatrix} x_1(k) - x_2(k) \\ 0.1x_2^2(k) \end{bmatrix} + v(k).
\end{aligned}
\tag{28}
$$

We compute a BF of degree 4 with $\bar{\alpha} = \beta = 10^{-4}$ and the other parameters set as in the previous case study, and again we compare against the KF residual analysis. For the KF, we linearise model (28) about the origin. We simulate again a fault impacting the second state variable as $\phi_2(k) = -0.3x_2(k)$. Results are shown in Table 1 (bottom): our method performs better than the KF in terms of time to detection and missed alarms, whilst it produces more false alarms with Gaussian noise and comparable with uniform noise.

4.3. Case study 3: The three-tank model

We now consider the three-tank benchmark model

$$
\begin{aligned}
\dot{x}_1 &= A_1^{-1}(u_1 - Q_{13}) \\
\dot{x}_2 &= A_2^{-1}(u_2 + Q_{32} - Q_2) \\
\dot{x}_3 &= A_3^{-1}(Q_{13} - Q_{32}),
\end{aligned}
\tag{29}
$$

with the nonlinear autonomous components of the dynamics

$$
\begin{aligned}
Q_{ij} &= a_{ij}\,\mathrm{sgn}(x_i - x_j)\sqrt{2G|x_i - x_j|} \\
Q_i &= a_i\sqrt{2Gx_i}.
\end{aligned}
\tag{30}
$$

For tank i, x_i is the water level, $u_i \in [0, u_{\max}]$ is the inflow rate, A_i is the cross-sectional area. The parameters a_i (or a_{ij}) are coefficients for the outflow of pipe i (or between pipe i and j), and G is the acceleration of gravity. Parameters are set according to Kortela (2022). As usually assumed, all water levels are directly measured, *i.e.,* $y = x + v$. System (29) is non-polynomial and not suited to SoS optimisation. To compute a polynomial BF g, we derive a discrete-time linear approximation, which we do not use for simulation but only to compute the BF. We linearise system (29) about the operating point $\bar{x}_1 = 0.39$, $\bar{x}_2 = 0.23$, $\bar{x}_3 = 0.31, \bar{u}_1 = 5.26 \times 10^{-5}, \bar{u}_2 = 3.66 \times 10^{-5}$, use LQ control with $Q = R = I$, and apply exact discretisation with sampling period $T_s = 0.01$ s. This provides us with the autonomous dynamics $\tilde{x}(k+1) = A\tilde{x}(k)$ that describes the evolution of (small) deviations $\tilde{x} = x - \bar{x}$ about the operating point $\bar{x}$, with state matrix

$$A = \begin{bmatrix} 0.5223 & 0 & -0.1072 \\ 0 & 0.5223 & -0.1072 \\ 0.0002 & 0.0002 & 0.9995 \end{bmatrix}. \tag{31}$$

Numerical tests show that this approximation is quite accurate for a broad range of deviations, as compared to the heights of tanks. The input is computed as $u = \bar{u} - K\tilde{x}$ with feedback gain matrix

$$K = \begin{bmatrix} 0.9997 & 0.0001 & 0.2249 \\ 0.0001 & 0.9996 & 0.2249 \end{bmatrix}. \tag{32}$$

We then compute a barrier function tailored to faults affecting the inflows to the tanks. To compute the BF, we consider the disturbed version of the autonomous dynamics so obtained $\tilde{x}(k+1) = A\tilde{x}(k) + \phi(k)$, where $\|\phi\| \le u_{\max} A_i^{-1} T_s$, corresponding to both inflows either saturated or zeroed. We set $\mathcal{U} = \mathbb{R}^3 \setminus \mathcal{B}_{0.1}(0)$, *i.e.,* the maximum deviation from the operating point is 10 cm in any tank. We compute a BF with degree 2 and $\bar{\alpha} = 0.01$, $\beta = 0.1$, $p = 1$.

We simulate the original model (29) starting near the operating point, using forward Euler discretisation with step T_s, and inject the fault $u_1(k) \equiv u_{\max}$ after 500 steps. The measurement noise v is drawn from the uniform distribution supported on $\mathcal{B}_{0.005}(0)$, meaning a sensing accuracy of 5 mm. The trajectories of water levels x and deviations $\tilde{x}$ are shown in the top and middle boxes of Fig. 3, respectively, while the bottom box shows the LHS of (15), which is positive under healthy dynamics. Detection is marked by the dotted vertical line and occurs at time $k = 600$, namely 0.1 s after the fault, which is shown by the dashed vertical line. When the alarm is raised, the water levels are still near the operating point, as shown by the negligible deviations in the middle box.

5. Concluding remarks

This work proposes a barrier function based approach to fault detection. The faults are partially known, as we solely require their boundedness. The synthesis of the barrier function is achieved through an optimisation program, exploiting SoS constraints, hence our approach enjoys the speed and reliability of convex optimisation. The fault monitoring is achieved by simply tracking the value of the barrier function, without the need for state estimation. Numerical experiments show that our technique copes well with any kind of noise and disturbance for linear and nonlinear systems alike.

Future work includes several compelling directions of research. It is desirable to improve detection performance and scalability of our approach, possibly resorting to a different problem formulation or alternative BF design techniques. Extensions to stochastic formulations are also relevant to enhance the formal guarantees on false positive and missed detection rates and add flexibility to the design, possibly resorting to stochastic variants of the BF. A thorough numerical evaluation with real-world datasets or experimental setups would be insightful to more concretely assess performance, possibly tuning the algorithm according to the specific system, including multiple noise distributions and classes of faults.

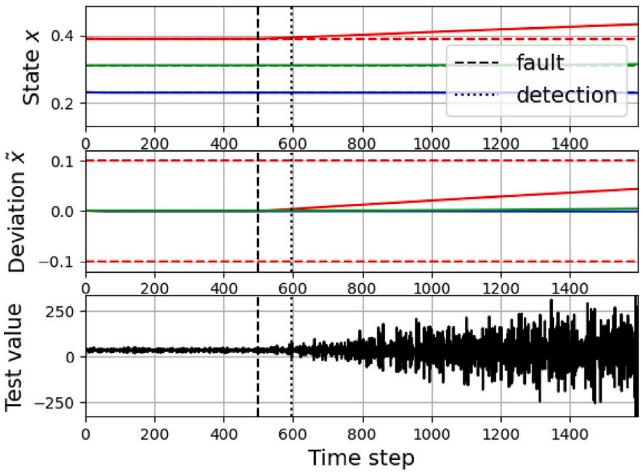

Fig. 3. State (top) and deviation (middle) trajectories along with the barrier function values (bottom). The coloured dashed horizontal lines in the top plot show the operating point and the two horizontal red dashed lines in the middle plot bound the maximum admissible deviation.

CRediT authorship contribution statement

Luca Ballotta: Formal analysis, Validation, Writing – original draft, Conceptualization, Methodology, Visualization, Writing – review & editing. **Andrea Peruffo:** Formal analysis, Validation, Writing – original draft, Conceptualization, Methodology, Visualization, Writing – review & editing. **Riccardo M.G. Ferrari:** Funding acquisition, Supervision, Conceptualization, Resources, Writing – review & editing. **Manuel Mazo:** Methodology, Writing – review & editing, Conceptualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmadi, M., Singletary, A., Burdick, J. W., & Ames, A. D. (2019). Safe policy synthesis in multi-agent POMDPs via discrete-time barrier functions. In *Proc. IEEE CDC* (pp. 4797–4803).

Ames, A. D., Xu, X., Grizzle, J. W., & Tabuada, P. (2017). Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, *62*(8), 3861–3876.

Blanke, M., Kinnaert, M., Lunze, J., & Staroswiecki, M. (2015). *Diagnosis and fault-tolerant control, 3rd Edition* (third ed.). Springer.

Chen, J., & Patton, R. J. (2012). *Robust model-based fault diagnosis for dynamic systems: Vol. 3*, Springer Science & Business Media.

Clark, A. (2021). Verification and synthesis of control barrier functions. In *Proc. IEEE CDC* (pp. 6105–6112).

Clark, A., Li, Z., & Zhang, H. (2020). Control barrier functions for safe CPS under sensor faults and attacks. In *Proc. IEEE CDC* (pp. 796–803).

Dean, S., Taylor, A., Cosner, R., Recht, B., & Ames, A. (2021). Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. In *Proc. CoRL* (pp. 654–670). PMLR.

Ding, S. X. (2008). *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Science & Business Media.

Garg, K., Dawson, C., Xu, K., Ornik, M., & Fan, C. (2023). Model-free neural fault detection and isolation for safe control. *IEEE Control Sys. Lett.*, *7*, 3169–3174.

Garg, K., & Panagou, D. (2021). Robust control barrier and control lyapunov functions with fixed-time convergence guarantees. In *Proc. ACC* (pp. 2292–2297).

Kortela, J. (2022). Model-predictive control for the three-tank system utilizing an industrial automation system. *ACS Omega*, *7*(22), 18605–18611.

Lindemann, L., Robey, A., Jiang, L., Das, S., Tu, S., & Matni, N. (2024). Learning robust output control barrier functions from safe expert demonstrations. *IEEE Open J. Control Syst.*, *3*, 158–172.

Papachristodoulou, A., Anderson, J., Valmorbida, G., Prajna, S., Seiler, P., Parrilo, P. A., et al. (2021). SOSTOOLS: Sum of squares optimization toolbox for MATLAB. ArXiv E-Prints arXiv:1310.4716.

Parrilo, P. A. (2000). *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. California Institute of Technology.

Prajna, S. (2006). Barrier certificates for nonlinear model validation. *Automatica*, *42*(1), 117–126.

Prajna, S., Jadbabaie, A., & Pappas, G. J. (2007). A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Transactions on Automatic Control*, *52*(8), 1415–1428.

Takano, R., & Yamakita, M. (2018). Robust constrained stabilization control using control Lyapunov and control barrier function in the presence of measurement noises. In *Proc. IEEE CCTA* (pp. 300–305).

Vyas, S. D., Roy, T., & Dey, S. (2022). Thermal fault-tolerance in lithium-ion battery cells: a barrier function based input-to-state safety framework. In *Proc. IEEE CCTA* (pp. 1178–1183).

Wang, L., Han, D., & Egerstedt, M. (2018). Permissive barrier certificates for safe stabilization using sum-of-squares. In *Proc. ACC* (pp. 585–590).

Wijk, D. v., Majji, M., & Hobbs, K. L. (2024). Fault tolerant run time assurance with control barrier functions for rigid body spacecraft rotation. In *AIAA SCITECH forum*.