Fundamental limitation of IIG gain for fault detection in LTI systems

Jingwei Dong and André M. H. Teixeira

I. INTRODUCTION

A critical issue in fault detection is the selection of appropriate metrics to assess diagnostic performance. For instance, in robust fault detection, mixing metric $\mathcal{H}_\infty/\mathcal{H}_-$ has been proposed to balance disturbance rejection and fault sensitivity through optimization approaches [1]. However, this metric suffers from conservatism because it separately evaluates the worst-case disturbance robustness through $\mathcal{H}_\infty$ norm and the worst-case fault sensitivity through $\mathcal{H}_-$ index. Moreover, the value of the $\mathcal{H}_-$ index becomes zero for faulty systems with non-minimum phase (NMP) zeros across the full frequency domain. To address these limitations, the authors in [2] introduced output-to-output (OOG) gain, a novel metric for quantifying the performance loss under stealthy injection attacks. Unlike traditional metrics that impose constraints on transfer functions, OOG directly computes the maximum performance loss in the output for all stealthy attacks. This integrates both attack impact and detectability while accounting for NMP zeros.

Given that the OOG method is designed for attack strategy only, this study explores its possible extension to robust fault detection. We develop an input-to-input (IIG) gain to measure fault sensitivity and disturbance robustness, offering a valuable alternative to conventional metrics. However, the conflicting nature of design constraints in IIG, coupled with some intrinsic system properties (zeros, poles, etc.), impose fundamental limitations on achievable performance of IIG. By leveraging the Poisson integral relation [3], this study further investigates the fundamental limitations of the proposed diagnostic metric .

II. PROBLEM FORMULATION

Let us consider a continuous-time LTI system

$$\Sigma : \begin{cases} \dot{x}(t) = Ax(t) + B_d d(t) + B_f f(t) \\ r(t) = Cx(t) + D_d d(t) + D_f f(t), \end{cases} \tag{1}$$

where $x(t) \in \mathbb{R}^{n_x}$, $d(t) \in \mathbb{R}^{n_d}$, $f(t) \in \mathbb{R}^{n_f}$, and $r(t) \in \mathbb{R}^{n_r}$ denote the state, disturbance, fault, and residual, respectively. The system (1) can be interpreted as a general form of a closed-loop system incorporating a fault detection observer [1]. The pair (A, C) is assumed to be observable. The residual r is used to indicate the occurrence of faults, which can be written as

$$r = \mathbb{T}_{dr}[d] + \mathbb{T}_{fr}[f], \tag{2}$$

This work is supported by the Swedish Research Council under the grant 2021-06316 and by the Swedish Foundation for Strategic Research.

Jingwei Dong and André M. H. Teixeira are with with the Department of Information Technology, Uppsala University, SE-75105 Uppsala, Sweden {jingwei.dong; andre.teixeira}@it.uu.se.

where $\mathbb{T}_{dr}$ and $\mathbb{T}_{fr}$ are the transfer functions from d to r and from f to r, respectively.

Given the effect of disturbances, we define a fault to be undetectable if the ℓ_2 norm of r satisfies:

$$\|r\|_{\ell_2}^2 = \|\mathbb{T}_{dr}[d] + \mathbb{T}_{fr}[f]\|_{\ell_2}^2 \leq \|\mathbb{T}_{dr}[d]\|_{\ell_2}^2. \tag{3}$$

This implies that the effect of f on r is masked by that of d. Moreover, the energy of d is assumed to be bounded, i.e., $\|d\|_{\ell_2}^2 \leq 1$, with the upper bound set to 1 without loss of generality.

Note that a larger undetectable fault indicates lower fault sensitivity. Thus, by calculating the maximum energy of undetectable faults characterized by (3), one can evaluate the fault sensitivity of (1). We then formulate the following optimization problem to find the largest fault that remains undetectable with respect to a given disturbance intensity:

$$\begin{aligned} \|\Sigma\|_{\ell_{2e}, f\leftarrow d}^2 \triangleq \sup_{f, d\in\ell_{2e}} \ & \|f\|_{\ell_2}^2 \\ \text{s.t. } & (1),\ x(0) = 0, \\ & \|\mathbb{T}_{dr}[d] + \mathbb{T}_{fr}[f]\|_{\ell_2}^2 \leq \|\mathbb{T}_{dr}[d]\|_{\ell_2}^2, \\ & \|d\|_{\ell_2}^2 \leq 1. \end{aligned} \tag{4}$$

The optimal value $\|\Sigma\|_{\ell_{2e}, f\leftarrow d}^2$ serves as a novel metric for fault sensitivity and is defined as IIG of Σ.

To simplify the subsequent analysis, we restrict our study to a special class of systems subject to periodic trajectories and consider that all inputs and outputs in (1) are one-dimensional signals. Based on the above setting, this study primarily focuses on identifying the fundamental limitations of IIG proposed in (4).

III. MAIN RESULTS

First, the undetectability condition (3), which characterizes all undetectable faults, is modified in the following lemma to derive a frequency-domain condition for IIG.

Lemma 1 (Reformulated undetectability condition)**.** *Consider the LTI system in* (1)*. The undetectability condition* (3) *together with* $\|d\|_{\ell_2}^2 \leq 1$ *can be realized through:*

$$\mathbb{T}_{dr}[\tilde{d}] = \mathbb{T}_{fr}[f],\ \|\tilde{d}\|_{\ell_2}^2 \leq \xi^2,\ \xi \in [0, 2]. \tag{5}$$

Proof. The proof is omitted here due to space limitation. □

While the condition (5) may be restrictive, requiring d and f to have the same output (up to a scale), we note that the condition naturally arises in the context of indistinguishability between fault and disturbances [1, Chapter 4].

We further apply the left co-prime factorization to $\mathbb{T}_{fr}$ and $\mathbb{T}_{dr}$:

$$\mathbb{T}_{fr} = M_I^{-1} N_f, \quad \mathbb{T}_{dr} = M_I^{-1} N_d. \tag{6}$$

Note that zeros in $\mathbb{T}_{fr}$ coincide with those in N_f. In the subsequent analysis, we assume that N_f have no NMP zeros as IIG value becomes infinite with certain inputs aligned with these NMP zeros, as demonstrated in [2, Theorem 2].

To simplify the notation, let us define $\gamma \triangleq \|\Sigma\|^2_{\ell_{2e}, f \leftarrow \tilde{d}}$. Now, we are in the position to present the frequency-domain inequality of the IIG in (4) with the undetectability condition replaced by (5).

Theorem 1 (Frequency domain inequality of IIG)**.** *Assume that the LTI systems in* (1) *is subject to periodic trajectories and consider that all the input and output signals are one-dimensional. Given the co-prime factorization in* (6) *with stable N_f, the optimal performance metrics introduced in* (4) *with the replaced condition* (5) *satisfy:*

$$\gamma \geq \xi^2 \left\| \frac{N_d}{N_f} \right\|^2_{\mathcal{H}_\infty}. \tag{7}$$

Proof. The proof is omitted due to space limitation. □

Furthermore, we establish relations between the developed IIG metric with the existing diagnostic metrics, namely, $\mathcal{H}_\infty/\mathcal{H}_\infty$ and $\mathcal{H}_\infty/\mathcal{H}_-$ employed in robust detection filter design, in the following corollary.

Corollary 1 (Relation with existing metrics)**.** *The IIG metric developed in* (7) *satisfies the following bounds:*

$$\frac{\|\mathbb{T}_{dr}\|^2_{\mathcal{H}_\infty}}{\|\mathbb{T}_{fr}\|^2_{\mathcal{H}_\infty}} \leq \left\| \frac{N_d}{N_f} \right\|^2_{\mathcal{H}_\infty} \leq \frac{\|\mathbb{T}_{dr}\|^2_{\mathcal{H}_\infty}}{\|\mathbb{T}_{fr}\|^2_{\mathcal{H}_-}}. \tag{8}$$

Proof. The proof is omitted here due to space limitation. □

To facilitate the derivation of the performance limitations in (7), we introduce the following notations. Let us define:

$$S_I = \frac{N_d}{N_f}, \quad P_I = 1 - \frac{N_d}{N_f}, \tag{9}$$

The NMP zeros of S_I are denoted as $\mu_i \in \mathcal{Z}_{S_I}, i = \{1, \ldots, n_\mu\}$, and those of P_I are denoted as $\nu_i \in \mathcal{Z}_{P_I}, i = \{1, \ldots, n_\nu\}$. Then, S_I and P_I can be factorized as

$$S_I = \tilde{S}_I \mathcal{B}_{S_I}, \quad P_I = \tilde{P}_I \mathcal{B}_{P_I}, \tag{10}$$

where $\tilde{S}_I$ and $\tilde{P}_I$ are the corresponding minimum-phase parts. The Blaschke products of NMP zeros in S_I and P_I are denoted as $\mathcal{B}_{S_I}$ and $\mathcal{B}_{P_I}$, which are given by

$$\mathcal{B}_{S_I}(s) = \prod_{i=1}^{n_\mu} \frac{s - \mu_i}{s + \bar{\mu}_i}, \quad \mathcal{B}_{P_I}(s) = \prod_{i=1}^{n_\nu} \frac{s - \nu_i}{s + \bar{\nu}_i}.$$

Note that $\mathcal{B}_{S_I}$ and $\mathcal{B}_{P_I}$ are all-pass filters, i.e., $|\mathcal{B}_{S_I}(j\omega)| = |\mathcal{B}_{P_I}(j\omega)| = 1$. If the set of NMP zeros is empty, we define the corresponding Blaschke product to be 1.

We then provide the performance limitations of IIG in the following theorem.

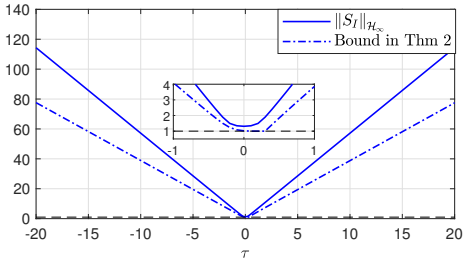

Fig. 1. Performance bound for $\|S_I\|_{\mathcal{H}_\infty}$.

Theorem 2 (Performance limitations of IIG)**.** *Consider the inequality developed for IIG in* (7) *and the factorization results in* (10)*. The performance limitations for S_I satisfy*

$$||S_I||_{\mathcal{H}_\infty} \geq \max_{\mu_h \in \mathcal{Z}_{S_I}, \nu_k \in \mathcal{Z}_{P_I}} \left\{ |\mathcal{B}^{-1}_{S_I}(\nu_k)|, |\mathcal{B}^{-1}_{P_I}(\mu_h)| - 1 \right\}.$$

Proof. The proof is omitted here due to space limitation. □

Theorem 2 relates the performance limitations of IIG to its NMP zeros. The lower bound is strictly larger than 1 if both S_I and P_I have NMP zeros, with the value determined by the distance between their NMP zeros: (1) The bound approaches 1 when the NMP zeros μ_h and ν_k are far apart. In this case, the performance bound does not closely reflect the actual value of $||S_I||_{\mathcal{H}_\infty}$; (2) The bound increases significantly when the NMP zeros μ_h and ν_k are close to each other, indicating the existence of large undetectable faults. Moreover, it is worth emphasizing that N_d and $N_f - N_d$ cannot share common NMP zeros because such a zero would also be an NMP zero of N_r, which, by assumption, has no NMP zeros in this study. The obtained results are validated through numerical examples in the following section.

IV. Numerical Examples

Consider transfer functions in (1) as follows:

$$\mathbb{T}_{fr}(s) = \frac{(s+0.1)(s+0.2)(s+0.6)}{(s+0.3)(s+0.4)(s+0.5)},$$
$$\mathbb{T}_{dr}(s) = \frac{(s+1)(s-0.04)(s-\tau)}{(s+0.3)(s+0.4)(s+0.5)},$$

where $\mathbb{T}_{dr}$ contains a NMP zero at 0.04 and a varying zero τ. We examine how the performance bound varies with respect to τ. Specifically, as the magnitude of τ increases in both directions, the NMP zero of $N_f - N_d$ gets close to that of N_d at 0.04. As a result, the performance bound of $\|S_I\|_{\mathcal{H}_\infty}$ increases when τ deviates from 0, as depicted in Fig. 1. This aligns with the theoretical results.

References

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

[2] A. Teixeira, H. Sandberg, and K. H. Johansson, "Strategic stealthy attacks: the output-to-output ℓ_2-gain", *in 54th IEEE Conference on Decision and Control (CDC)*, pp. 2582–2587, Osaka, Japan, 2015.

[3] J. S. Freudenberg and D. P. Looze, "Right half plane poles and zeros and design trade-offs in feedback systems," *IEEE Transactions on Automatic Control*, vol. 30, no. 6, pp. 555–565, 1985.

[4] A. Teixeira, "Optimal stealthy attacks on actuators for strictly proper systems", *in 58th IEEE Conference on Decision and Control (CDC)*, pp. 4385–4390, Nice, France, 2019.