New Bounds for State Transition Matrices

Frédéric Mazenc and Michael Malisoff, *Senior Member, IEEE*

***Abstract*—We address the problem of constructing matrix-valued interval observers for estimating state transition matrices for time-varying systems. We provide less conservative estimators than those in recent literature. We cover continuous- and discrete-time linear systems, under Metzler or nonnegativity conditions on the coefficient matrices. We show how to satisfy our Metzler conditions after simple changes of coordinates. We illustrate our method using a feedback stabilized underwater marine robotic dynamics with unknown control gains.**

***Index Terms*—Linear systems, estimation.**

I. Introduction

CONTROL theory for time-varying systems is complicated by the fact that it calls for replacing the matrix exponential by fundamental (i.e., state transition) matrices that are usually not available in explicit closed form. This problem is made even more difficult for time-varying linear systems that contain time-varying uncertainty in their coefficient matrices. This challenge arises even if the system is time invariant, e.g., for LTI systems $\dot{x} = Ax + Bu$ with globally asymptotically stabilizing feedback controls $u = Kx$ for constant matrices A, B, and K. This is because implementations naturally lead to uncertain controls gains, so for piecewise continuous bounded functions δ_K, the control acting on the system will be $u = (K + \delta_K(t))x$, which produces

$$\dot{x} = (A_{\rm cl} + \Delta_K(t))x \tag{1}$$

with $A_{\rm cl} = A + BK$ and $\Delta_K = B\delta_K$. Since Δ_K is unknown, the exponential representation $x(t) = e^{A_{\rm cl}t}x(0)$ for solutions of the unperturbed system $\dot{x} = A_{\rm cl}x$ does not apply to (1). Instead, a natural approach to addressing this problem is to use bounds on Δ_K to estimate solutions of (1).

This motivated [9] and [10], which derived matrix-valued interval observers for fundamental matrices of time-varying linear systems with unknown coefficient matrices, and used the observers in feedback control and observer designs. More precisely, for matrix-valued functions $M : [0, +\infty) \to \mathbb{R}^{n\times n}$ and $\Delta : [0, +\infty) \to \mathbb{R}^{n\times n}$ whose entries are bounded and piecewise continuous and such that $M(t)$ is Metzler and $0 \leq \underline{\Delta}(t) \leq \overline{\Delta}$ hold entrywise for all $t \geq 0$ for a known matrix $\overline{\Delta}$, [10] proved that the state transition matrix $\Phi_{M-\Delta}$ of

$$\dot{X}(t) = [M(t) - \Delta(t)]X(t) \tag{2}$$

is such that

$$\underline{\Phi}_{M,\overline{\Delta}}(t,s) \leq \Phi_{M-\Delta}(t,s) \leq \overline{\Phi}_{M,\overline{\Delta}}(t,s) \tag{3}$$

for all $s \geq 0$ and $t \geq s$, where

$$\underline{\Phi}_{M,\overline{\Delta}}(t,s) = \Phi_M(t,s) + \frac{\Phi_{M-\overline{\Delta}}(t,s) - \Phi_{M+\overline{\Delta}}(t,s)}{2}$$
$$\text{and } \overline{\Phi}_{M,\overline{\Delta}}(t,s) = \frac{\Phi_{M-\overline{\Delta}}(t,s) + \Phi_{M+\overline{\Delta}}(t,s)}{2}; \tag{4}$$

see our definitions and notation below.

Many systems can be transformed into the form (2) after a change of variables; see [10, Remark 2]. For instance, if $A_{\rm cl}$ in (1) is Metzler, and if each entry Δ_{Kij} of the matrix $\Delta_K = [\Delta_{Kij}]$ is known to be bounded by some constant $\Delta_{*ij} \geq 0$, then we obtain (2) with the choices $M = A_{\rm cl} + \Delta_*$ and $\Delta(t) = \Delta_* - \Delta_K(t)$, where Δ_* is the constant matrix $[\Delta_{*ij}]$ for all i and j. The Metzler requirement on $A_{\rm cl}$ can be met by a change of coordinates that transforms $A_{\rm cl}$ into its Jordan canonical form when all of its eigenvalues are real.

This motivates the important question of whether tighter bounds than (3) be determined. Its interest comes from the fact that the tighter the bounds are, the better the control laws or observers or stability analyses are which use them. In many cases, tighter bounds of this type can indeed by found. To understand why, let us consider the special case of (2) where $M(t) - \overline{\Delta}$ is Metzler for all $t \geq 0$. Then

$$\Phi_{M-\overline{\Delta}}(t,s) \leq \Phi_{M-\Delta}(t,s) \leq \Phi_M(t,s) \tag{5}$$

when $t \geq s \geq 0$; this monotonicity property follows, e.g., by the proof of [6, Lemma 2], with its matrix exponentials replaced by transition matrices. Since $\Phi_{M-\Delta}(t,s) = \Phi_{\underline{M}}(t,s)$ when $\Delta = 0$ and $\Phi_{M-\Delta}(t,s) = \Phi_{M-\overline{\Delta}}(t,s)$ when $\Delta = \overline{\Delta}$, (5) are the best possible bounds for $\Phi_{M-\Delta}(t,s)$. They are tighter than the bounds in (3) when $\Phi_M(t,s) < \frac{1}{2}(\Phi_{M+\overline{\Delta}}(t,s) + \Phi_{M-\overline{\Delta}}(t,s))$; see Section II-C below for an illustration where (5) provides tighter bounds than (3).

This motivates this letter. We revisit [9] and [10] by providing two results. First, in Section II, and in both continuous- and discrete-time cases, we obtain tighter bounds than those of [9] and [10], by decomposing the disturbances Δ in (2) into two parts Δ_1 and Δ_2 such that $M(t) + \Delta_1(t)$ is Metzler for all $t \geq 0$ and $\Delta_2 = \Delta - \Delta_1$. Second, in Section III, we present a family of matrices that are similar to full Metzler matrices, making it possible to obtain bounds of the type (5)

Manuscript received March 21, 2022; revised May 4, 2022; accepted May 6, 2022. Date of publication May 9, 2022; date of current version May 17, 2022. This work was supported by the U.S. National Science Foundation under Grant 1711299 and Grant 2009659. Recommended by Senior Editor V. Ugrinovskii. *(Corresponding author: Frédéric Mazenc.)*

Frédéric Mazenc is with Inria EPI DISCO, L2S-CNRS-CentraleSupélec, 91192 Gif-sur-Yvette, France (e-mail: frederic.mazenc@l2s.centralesupelec.fr).

Michael Malisoff is with the Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 USA (e-mail: malisoff@lsu.edu).

Digital Object Identifier 10.1109/LCSYS.2022.3173816

when $\overline{\Delta}$ is sufficiently small. This follows from the fact that if a matrix M is full Metzler, then $M-\overline{\Delta}$ is Metzler when $\overline{\Delta}$ is sufficiently small. This result completes [8], whose conditions ensure that a matrix of dimension 3 is similar to a full Metzler matrix. We illustrate this benefit in Section IV, using a marine robotic dynamics that is perturbed by uncertain control gains, where we find novel bounds for the transition matrix of the perturbed coefficient matrix.

The notation will be simplified when no confusion could arise given the context, and all matrices in this letter are assumed to have only real entries. We set $\mathbb{Z}_{\geq 0}=\{0,1,\ldots\}$ and $\mathbb{N}=\mathbb{Z}_{\geq 0}\backslash\{0\}$. We use I to denote the identity matrix, 0 to denote the zero matrix, and $\mathbf{1}$ to denote the matrix whose entries are all 1's, for any dimensions. A square matrix is called Metzler provided its off-diagonal entries are all nonnegative. A matrix is called nonpositive (resp., positive) provided all its entries are nonpositive (resp., positive). A square matrix is called full Metzler provided all its off-diagonal entries are positive. For vectors $V_1=(v_{1,1}\ \cdots\ v_{1,n})^\top$ and $V_2=(v_{2,1}\ \cdots\ v_{2,n})^\top$, we write $V_1<V_2$ provided that $v_{1,i}<v_{2,i}$ for $i=1,\ldots,n$; and $V_1\leq V_2$ provided for $v_{1,i}\leq v_{2,i}$ for $i=1,\ldots,n$. We use analogous componentwise inequalities for matrices.

Square matrices M_1 and M_2 of the same size are called similar provided there is an invertible matrix P so that $M_1=P^{-1}M_2P$.

For any matrix-valued function $\mathcal{F}:[0,+\infty)\to\mathbb{R}^{n\times n}$ whose entries are locally bounded and piecewise continuous, the fundamental (or state transition) matrix solution $\Phi_{\mathcal{F}}$ is defined to be the unique matrix-valued function satisfying

$$\frac{\partial\Phi_{\mathcal{F}}}{\partial t}(t,t_0)=\mathcal{F}(t)\Phi_{\mathcal{F}}(t,t_0),\quad \Phi_{\mathcal{F}}(t_0,t_0)=I \tag{6}$$

for all $t_0\geq 0$ and $t\geq t_0$. Note for later use that this uniqueness property gives $\Phi_{R\mathcal{F}R^{-1}}=R\Phi_{\mathcal{F}}R^{-1}$ for all invertible $n\times n$ matrices R [10].

II. Bounds for Fundamental Matrices

We derive our bounds for fundamental matrices of both continuous- and discrete-time systems, and we illustrate how they can produce tighter bounds than those in the literature.

A. Continuous-Time system

Let us consider the system

$$\dot{x}(t)=[M(t)-\Delta(t)]x(t) \tag{7}$$

where $M:[0,+\infty)\to\mathbb{R}^{n\times n}$ and $\Delta:[0,+\infty)\to\mathbb{R}^{n\times n}$ are locally bounded and piecewise continuous functions. Assume the following (but see Remark 1 for extensions where the signs of the additive uncertainties on M have no restrictions).

Assumption 1: The matrix $M(t)$ is Metzler for all $t\geq 0$.

Assumption 2: There is a known matrix $\overline{\Delta}\geq 0$ such that

$$0\leq\Delta(t)\leq\overline{\Delta} \tag{8}$$

for all $t\geq 0$.

We let $m_{i,j}$ denote the entry of M in its ith row and jth column for all i and j. Similarly, $\delta_{i,j}$ (resp. $\overline{\delta}_{i,j}$) is the ij entry of Δ (resp. $\overline{\Delta}$). We define the function $\Delta_1=[\delta^1_{i,j}]$ by:

$$\delta^1_{i,j}(t)=\begin{cases}\delta_{i,j}(t), & \text{if } i=j \text{ or if}\\ & m_{ij}(t)-\overline{\delta}_{i,j}\geq 0\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

We also define the function $\Delta_2(t)=\Delta(t)-\Delta_1(t)$. Let $\overline{\Delta}_1=[\overline{\delta}^1_{i,j}]$ be defined by

$$\overline{\delta}^1_{i,j}(t)=\begin{cases}\overline{\delta}_{i,j}(t), & \text{if } i=j \text{ or if}\\ & m_{ij}(t)-\overline{\delta}_{i,j}\geq 0\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

and $\overline{\Delta}_2$ is defined by $\overline{\Delta}_2=\overline{\Delta}-\overline{\Delta}_1$. Our first theorem is:

Theorem 1: Let Assumptions 1-2 be satisfied. Then, for all $t_0\geq 0$ and $t\geq t_0$, the inequalities

$$\begin{aligned}&\Phi_{M-\overline{\Delta}_1}(t,t_0)+\frac{\Phi_{M-\overline{\Delta}_2}(t,t_0)-\Phi_{M+\overline{\Delta}_2}(t,t_0)}{2}\\ &\leq\Phi_{M-\Delta}(t,t_0)\leq\frac{\Phi_{M-\overline{\Delta}_2}(t,t_0)+\Phi_{M+\overline{\Delta}_2}(t,t_0)}{2}\end{aligned} \tag{11}$$

are satisfied.

Proof: Let us introduce the matrix

$$\Omega(t)=\begin{bmatrix}M(t)-\Delta_1(t) & \Delta_2(t)\\ \Delta_2(t) & M(t)-\Delta_1(t)\end{bmatrix}, \tag{12}$$

which is Metzler for all $t\geq 0$, because $M(t)-\Delta_1(t)$ is Metzler and $\Delta_2(t)\geq 0$ for all $t\geq 0$. Also, with the choices

$$\begin{aligned}\overline{\Omega}(t)&=\begin{bmatrix}M(t) & \overline{\Delta}_2\\ \overline{\Delta}_2 & M(t)\end{bmatrix}\text{ and}\\ \underline{\Omega}(t)&=\begin{bmatrix}M(t)-\overline{\Delta}_1 & 0\\ 0 & M(t)-\overline{\Delta}_1\end{bmatrix},\end{aligned} \tag{13}$$

the inequalities

$$\underline{\Omega}(t)\leq\Omega(t)\leq\overline{\Omega}(t) \tag{14}$$

hold for all $t\geq 0$. Also, using the choice

$$\mu=\begin{bmatrix}\frac{I}{\sqrt{2}} & -\frac{I}{\sqrt{2}}\\ \frac{I}{\sqrt{2}} & \frac{I}{\sqrt{2}}\end{bmatrix}, \tag{15}$$

we have

$$\mu^{-1}=\mu^\top. \tag{16}$$

Also, one can readily check that

$$\begin{aligned}&\mu\Omega(t)\mu^{-1}\\ &=\begin{bmatrix}M(t)-\Delta_1(t)-\Delta_2(t) & 0\\ 0 & M(t)-\Delta_1(t)+\Delta_2(t)\end{bmatrix}\end{aligned} \tag{17}$$

and

$$\mu\overline{\Omega}(t)\mu^{-1}=\begin{bmatrix}M(t)-\overline{\Delta}_2 & 0\\ 0 & M(t)+\overline{\Delta}_2\end{bmatrix} \tag{18}$$

hold for all $t\geq 0$. From (14), (17) and (18), we deduce that

$$\begin{aligned}&\underline{\Omega}(t)\\ &\leq\mu^{-1}\begin{bmatrix}M(t)-\Delta_1(t)-\Delta_2(t) & 0\\ 0 & M(t)-\Delta_1(t)+\Delta_2(t)\end{bmatrix}\mu\\ &\leq\mu^{-1}\begin{bmatrix}M(t)-\overline{\Delta}_2 & 0\\ 0 & M(t)+\overline{\Delta}_2\end{bmatrix}\mu.\end{aligned} \tag{19}$$

Since $\underline{\Omega}(t)$, $\Omega(t)$ and $\overline{\Omega}(t)$ are Metzler, the reasoning that gave (5) gives

$$\begin{aligned}&\Phi_{\underline{\Omega}}(t,t_0)\\ &\leq\mu^{-1}\begin{bmatrix}\Phi_{M-\Delta}(t,t_0) & 0\\ 0 & \Phi_{M-\Delta_1+\Delta_2}(t,t_0)\end{bmatrix}\mu\end{aligned}$$

$$\leq \mu^{-1}\begin{bmatrix}\Phi_{M-\overline{\Delta}_2}(t,t_0) & 0\\ 0 & \Phi_{M+\overline{\Delta}_2}(t,t_0)\end{bmatrix}\mu \tag{20}$$

for all $t_0 \geq 0$ and $t \geq t_0$, where

$$\Phi_{\underline{\Omega}}(t,t_0) = \begin{bmatrix}\Phi_{M-\overline{\Delta}_1}(t,t_0) & 0\\ 0 & \Phi_{M-\overline{\Delta}_1}(t,t_0)\end{bmatrix} \tag{21}$$

for all $t \geq t_0$. These equalities are equivalent to

$$\Phi_{\underline{\Omega}} \leq \begin{bmatrix}\frac{\Phi_{M-\Delta}+\Phi_{M-\Delta_1+\Delta_2}}{2} & \frac{\Phi_{M-\Delta_1+\Delta_2}-\Phi_{M-\Delta}}{2}\\ \frac{\Phi_{M-\Delta_1+\Delta_2}-\Phi_{M-\Delta}}{2} & \frac{\Phi_{M-\Delta}+\Phi_{M-\Delta_1+\Delta_2}}{2}\end{bmatrix}$$
$$\leq \begin{bmatrix}\frac{\Phi_{M-\overline{\Delta}_2}+\Phi_{M+\overline{\Delta}_2}}{2} & \frac{\Phi_{M+\overline{\Delta}_2}-\Phi_{M-\overline{\Delta}_2}}{2}\\ \frac{\Phi_{M+\overline{\Delta}_2}-\Phi_{M-\overline{\Delta}_2}}{2} & \frac{\Phi_{M-\overline{\Delta}_2}+\Phi_{M+\overline{\Delta}_2}}{2}\end{bmatrix}. \tag{22}$$

Comparing the upper left and upper right blocks in (22) gives

$$\Phi_{M-\overline{\Delta}_1}(t,t_0) \leq \frac{\Phi_{M-\Delta}(t,t_0)+\Phi_{M-\Delta_1+\Delta_2}(t,t_0)}{2}$$
$$\leq \frac{\Phi_{M-\overline{\Delta}_2}(t,t_0)+\Phi_{M+\overline{\Delta}_2}(t,t_0)}{2} \text{ and} \tag{23}$$
$$0 \geq -\frac{\Phi_{M-\Delta_1+\Delta_2}(t,t_0)-\Phi_{M-\Delta}(t,t_0)}{2}$$
$$\geq -\frac{\Phi_{M+\overline{\Delta}_2}(t,t_0)-\Phi_{M-\overline{\Delta}_2}(t,t_0)}{2} \tag{24}$$

for all $t \geq t_0$. Adding these inequalities yields (11). ■

Remark 1: The nonnegativeness of Δ in Assumption 2 is not restrictive at all. Indeed, if we know constant matrices $\Delta_s \leq 0$ and $\Delta_l \geq 0$ such that $\Delta_s \leq \Delta(t) \leq \Delta_l$ for all $t \geq 0$, then (7) can be rewritten as $\dot{x}(t) = [M_\star(t) - \Delta_\star(t)]x(t)$ with $M_\star(t) = M(t) - \Delta_s$ and $\Delta_\star(t) = \Delta(t) - \Delta_s$, and $M_\star$(t) is Metzler (if $M(t)$ is Metzler) and $0 \leq \Delta_\star(t) \leq \overline{\Delta}$ with $\overline{\Delta} = \Delta_l - \Delta_s \geq 0$ for all t. Hence, when the inequalities $\Delta_s \leq \Delta(t) \leq \Delta_l$ hold for all $t \geq 0$ and $\Delta_s \leq 0$, we can always find a new decomposition for $M - \Delta$ as the difference between a Metzler and a nonnegative matrix for which Assumptions 1-2 hold. For simplicity, we use Assumption 2. Let us add that any square matrix can be decomposed as the sum of a Metzler matrix and a nonpositive matrix. The smaller the nonpositive matrix is in the decomposition, the closer to the fundamental solutions are the bounds obtained from Theorem 1. Thus, to obtain useful bounds, it may be worthwhile to first transform a system through a time-varying change of coordinates to get a new system for which, roughly speaking, a small function $\Delta_2(t)$ can be obtained. See [7] for such a change of coordinates.

B. Discrete-Time system

We consider the discrete-time system

$$X_{k+1} = (M_k - \Delta_k)X_k \tag{25}$$

where $M_k \in \mathbb{R}^{n\times n}$ and $\Delta_k \in \mathbb{R}^{n\times n}$ for $k \in \mathbb{Z}_{\geq 0}$. Assume:

Assumption 3: For all $k \in \mathbb{Z}_{\geq 0}$, we have $M_k \geq 0$.

Assumption 4: There is a matrix $\overline{\Delta} \geq 0$ such that for all $k \in \mathbb{Z}_{\geq 0}$, the inequalities

$$0 \leq \Delta_k \leq \overline{\Delta} \tag{26}$$

are satisfied.

We let $\delta_{k,i,j}$ (resp. $\overline{\delta}_{i,j}$) denote the entry of Δ_k (resp. $\overline{\Delta}$) in its ith row and jth column for all i and j. We also use

$$E_{(i,j)} = (M_{j-1+i} - \Delta_{j-1+i})\cdots(M_{i+1} - \Delta_{i+1})(M_i - \Delta_i) \tag{27}$$

for all $i \in \mathbb{Z}_{\geq 0}$ and $j \in \mathbb{N}$, where in (27) and some of what follows, we place parentheses around the (i,j) (which are used to indicate a range of indices that are used in the product of matrices on the right side) to distinguish it from the subscripts without parentheses that indicate entries in the ith row and jth column of a matrix. We also define the entries $\delta^1_{k,i,j}$ of the matrix $\Delta^1_k = [\delta^1_{k,i,j}]$ for all $k \in \mathbb{Z}_{\geq 0}$ by

$$\delta^1_{k,i,j} = \begin{cases}\delta_{k,i,j}, & \text{if } m_{k,i,j} - \overline{\delta}_{i,j} \geq 0\\ 0, & \text{otherwise,}\end{cases} \tag{28}$$

where $m_{k,i,j}$ is the entry of M_k in row i and column j. Similarly, we define $\overline{\Delta}^1_k = [\overline{\delta}^1_{k,i,j}]$ for all $k \in \mathbb{Z}_{\geq 0}$ by

$$\overline{\delta}^1_{k,i,j} = \begin{cases}\overline{\delta}_{i,j}, & \text{if } m_{k,i,j} - \overline{\delta}_{i,j} \geq 0\\ 0, & \text{otherwise.}\end{cases} \tag{29}$$

Similarly, we define the sequences

$$\overline{\Delta}^2_k = \overline{\Delta} - \overline{\Delta}^1_k, \tag{30}$$
$$\overline{R}_{(i,j)} = (M_{j-1+i} + \overline{\Delta}^2_{j-1+i})\cdots(M_{i+1} + \overline{\Delta}^2_{i+1})(M_i + \overline{\Delta}^2_i), \tag{31}$$

and

$$\overline{S}_{(i,j)} = (M_{j-1+i} - \overline{\Delta}^2_{j-1+i})\cdots(M_{i+1} - \overline{\Delta}^2_{i+1})(M_i - \overline{\Delta}^2_i). \tag{32}$$

Our main result in the discrete-time case is then as follows.

Theorem 2: For all $i \in \mathbb{Z}_{\geq 0}$ and all $j \in \mathbb{N}$, the inequalities

$$(M_{j-1+i} - \overline{\Delta}^1_{j-1+i})\cdots(M_i - \overline{\Delta}^1_i) + \frac{\overline{S}_{(i,j)} - \overline{R}_{(i,j)}}{2}$$
$$\leq E_{(i,j)} \leq \frac{\overline{R}_{(i,j)} + \overline{S}_{(i,j)}}{2} \tag{33}$$

are satisfied.

Proof: We use the matrices $\Delta^2_k = \Delta_k - \Delta^1_k$,

$$\Omega_k = \begin{bmatrix}M_k - \Delta^1_k & \Delta^2_k\\ \Delta^2_k & M_k - \Delta^1_k\end{bmatrix}, \tag{34}$$
$$\overline{\Omega}_k = \begin{bmatrix}M_k & \overline{\Delta}^2_k\\ \overline{\Delta}^2_k & M_k\end{bmatrix}, \tag{35}$$
$$\text{and } \underline{\Omega}_k = \begin{bmatrix}M_k - \overline{\Delta}^1_k & 0\\ 0 & M_k - \overline{\Delta}^1_k\end{bmatrix}. \tag{36}$$

Then $\Omega_k \geq 0$, $\underline{\Omega}_k \geq 0$, $\overline{\Omega}_k \geq 0$, and

$$\underline{\Omega}_k \leq \Omega_k \leq \overline{\Omega}_k \tag{37}$$

hold for all $k \in \mathbb{Z}_{\geq 0}$.

We also use the matrices (15) and the j-fold products

$$\kappa_{(i,j)} = \Omega_{j-1+i}\cdots\Omega_{i+1}\Omega_i,\ \overline{\kappa}_{(i,j)} = \overline{\Omega}_{j-1+i}\cdots\overline{\Omega}_{i+1}\overline{\Omega}_i, \tag{38}$$
$$\mathcal{G}_{(i,j)} = (M_{j-1+i} - \overline{\Delta}^1_{j-1+i})\cdots(M_i - \overline{\Delta}^1_i), \tag{39}$$
$$\text{and } \underline{\kappa}_{(i,j)} = \begin{bmatrix}\mathcal{G}_{(i,j)} & 0\\ 0 & \mathcal{G}_{(i,j)}\end{bmatrix}. \tag{40}$$

Then we deduce from the inequalities (37) that

$$\underline{\kappa}_{(i,j)} \leq \kappa_{(i,j)} \leq \overline{\kappa}_{(i,j)} \tag{41}$$

for all $i \in \mathbb{Z}_{\geq 0}$ and $j \in \mathbb{N}$. One can also readily check that

$$\mu\kappa_{(i,j)}\mu^{-1} = \begin{bmatrix}E_{(i,j)} & 0\\ 0 & S_{(i,j)}\end{bmatrix}, \tag{42}$$

where $S_{(i,j)}$ is the j-fold product of matrices

$$S_{(i,j)} = (M_{j-1+i} - \Delta^1_{j-1+i} + \Delta^2_{j-1+i}) \cdots (M_{i+1} - \Delta^1_{i+1} + \Delta^2_{i+1})(M_i - \Delta^1_i + \Delta^2_i) \tag{43}$$

and $$\mu\overline{\kappa}_{(i,j)}\mu^{-1} = \begin{bmatrix} \overline{S}_{(i,j)} & 0 \\ 0 & \overline{R}_{(i,j)} \end{bmatrix} \tag{44}$$

with $\overline{R}_{(i,j)}$ and $\overline{S}_{(i,j)}$ defined respectively in (31) and (32) and μ defined in (15). Consequently, the inequalities

$$\begin{aligned} \underline{\kappa}_{(i,j)} &\le \mu^{-1}\begin{bmatrix} E_{(i,j)} & 0 \\ 0 & S_{(i,j)} \end{bmatrix}\mu \\ &\le \mu^{-1}\begin{bmatrix} \overline{S}_{(i,j)} & 0 \\ 0 & \overline{R}_{(i,j)} \end{bmatrix}\mu \end{aligned} \tag{45}$$

are satisfied. They can be rewritten as:

$$\begin{aligned} \underline{\kappa}_{(i,j)} &\le \begin{bmatrix} \frac{E_{(i,j)}+S_{(i,j)}}{2} & \frac{-E_{(i,j)}+S_{(i,j)}}{2} \\ \frac{-E_{(i,j)}+S_{(i,j)}}{2} & \frac{E_{(i,j)}+S_{(i,j)}}{2} \end{bmatrix} \\ &\le \begin{bmatrix} \frac{\overline{R}_{(i,j)}+\overline{S}_{(i,j)}}{2} & \frac{-\overline{S}_{(i,j)}+\overline{R}_{(i,j)}}{2} \\ \frac{-\overline{S}_{(i,j)}+\overline{R}_{(i,j)}}{2} & \frac{\overline{R}_{(i,j)}+\overline{S}_{(i,j)}}{2} \end{bmatrix}. \end{aligned} \tag{46}$$

Comparing the upper left and upper right blocks in (46) gives

$$\begin{aligned} (M_{j-1+i} - \overline{\Delta}^1_{j-1+i}) \cdots (M_i - \overline{\Delta}^1_i) &\le \frac{E_{(i,j)} + S_{(i,j)}}{2} \\ &\le \frac{\overline{R}_{(i,j)} + \overline{S}_{(i,j)}}{2} \end{aligned} \tag{47}$$

and

$$0 \ge -\frac{-E_{(i,j)} + S_{(i,j)}}{2} \ge \frac{\overline{S}_{(i,j)} - \overline{R}_{(i,j)}}{2} \tag{48}$$

for all i and j. By adding (47)-(48), we obtain (33). ■

C. Comparison

Through a simple example, we show that the bounds from Theorem 1 can be better than those of the inequalities (3).

Consider the one dimensional system

$$\dot{X}(t) = [-1 - \Delta(t)]X(t) \tag{49}$$

where Δ is a piecewise continuous function that admits a known constant $\overline{\Delta}$ such that $0 \le \Delta(t) \le \overline{\Delta}$ for all $t \ge 0$.

For this system, the inequalities (3) are

$$\underline{\Phi}_{-1,\overline{\Delta}}(t,s) \le \Phi_{-1-\Delta}(t,s) \le \overline{\Phi}_{-1,\overline{\Delta}}(t,s) \tag{50}$$

for all $t \ge s \ge 0$ with the following choices:

$$\begin{aligned} \underline{\Phi}_{-1,\overline{\Delta}}(t,s) &= e^{-(t-s)} + \frac{e^{(-1-\overline{\Delta})(t-s)} - e^{(-1+\overline{\Delta})(t-s)}}{2} \\ \overline{\Phi}_{-1,\overline{\Delta}}(t,s) &= \frac{e^{(-1+\overline{\Delta})(t-s)} + e^{(-1-\overline{\Delta})(t-s)}}{2}. \end{aligned} \tag{51}$$

By the definitions of Δ_1 and Δ_2, we have $\overline{\Delta}_1(t) = \overline{\Delta}(t)$ and $\overline{\Delta}_2(t) = 0$. Then Theorem 1 gives

$$e^{-(1+\overline{\Delta})(t-s)} \le \Phi_{M-\Delta}(t,s) \le e^{-(t-s)} \tag{52}$$

for all $t \ge s$. Then we observe that

$$e^{-m} < \frac{e^{(-1+\overline{\Delta})m} + e^{(-1-\overline{\Delta})m}}{2} \tag{53}$$

for all $m > 0$ because this inequality is equivalent to

$$1 < \frac{e^{\overline{\Delta}m} + e^{-\overline{\Delta}m}}{2} \tag{54}$$

Similarly, for all $m > 0$, we have

$$e^{-m} + \frac{e^{(-1-\overline{\Delta})m} - e^{(-1+\overline{\Delta})m}}{2} < e^{-(1+\overline{\Delta})m} \tag{55}$$

because (55) is also equivalent to (54). We conclude that the bounds in (52) are tighter than those of (50).

III. Full Metzler Matrices

In this section, we exhibit important families of matrices which are similar to full Metzler matrices. Starting from a constant matrix M that is similar to full Metzler matrix F, this will make it possible to use the methods of the previous section to provide bounds on fundamental solutions for the system $\dot{x} = (M + \Delta(t))x$ after a change of coordinates when the absolute values of the entries of the piecewise continuous unknown matrix Δ are small enough. To see why, notice that if $F = Q^{-1}MQ$ for some invertible matrix Q, then the dynamics of the new variable $z = Q^{-1}x$ are $\dot{z} = (F+Q^{-1}\Delta(t)Q)z$, which in conjunction with the methods from the previous section will allow us to bound the state transition matrix for $F+Q^{-1}\Delta(t)Q$ when $Q^{-1}\Delta(t)Q$ is treated as a (small) perturbation; see our illustration below.

A. Block Triangular Matrices

This lemma is key to proving our first theorem of this section, by providing full Metzler matrices F_ϵ, where $0_{n_1\times n_2}$ denotes the $n_1 \times n_2$ zero matrix for integers n_1 and n_2:

Lemma 1: Consider the constant matrix

$$N = \begin{bmatrix} d & 0_{1\times n} \\ 0_{n\times 1} & p \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} \tag{56}$$

where $d > 0$, $p = [p_{i,j}] \in \mathbb{R}^{n\times n}$, $p_{i,j} > 0$ for all i and j, and $d < p_{1,1}$. Then, with the choices

$$F_\epsilon = \begin{bmatrix} \frac{d+\epsilon^2 p_{11}}{1+\epsilon^2} & \frac{\epsilon(p_{1,1}-d)}{1+\epsilon^2} & \epsilon p_{1,2} & \cdots & \epsilon p_{1,n} \\ \frac{\epsilon(p_{1,1}-d)}{1+\epsilon^2} & \frac{d\epsilon^2+p_{1,1}}{1+\epsilon^2} & p_{1,2} & \cdots & p_{1,n} \\ \frac{\epsilon p_{2,1}}{1+\epsilon^2} & \frac{p_{2,1}}{1+\epsilon^2} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\epsilon p_{n,1}}{1+\epsilon^2} & \frac{p_{n,1}}{1+\epsilon^2} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix} \tag{57}$$

and

$$R_\epsilon = \begin{bmatrix} 1 & \epsilon & 0 & \dots & 0 \\ -\epsilon & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix} \tag{58}$$

for any constant $\epsilon > 0$, we have $R_\epsilon N R_\epsilon^{-1} = F_\epsilon$.

Proof: The result follows from a block multiplication of matrices. In fact, using the fact that

$$R_\epsilon^{-1} = \begin{bmatrix} \frac{1}{1+\epsilon^2} & -\frac{\epsilon}{1+\epsilon^2} & 0 & \dots & 0 \\ \frac{\epsilon}{1+\epsilon^2} & \frac{1}{1+\epsilon^2} & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix}, \tag{59}$$

one can prove that the conclusion holds. ∎

Remark 2: One can easily extend Lemma 1 to matrices with nonpositive diagonal entries. Indeed, if matrix M is similar to a full Metzler matrix, then a matrix $M + \lambda I$ where λ is any real number is similar to a full Metzler matrix.

B. Diagonal Matrices With Distinct Diagonal Entries

Using Lemma 1, we prove the following, which we will use below to estimate the effects of uncertainties on systems whose coefficient matrices are not necessarily diagonal:

Theorem 3: Consider the diagonal matrix

$$D = \text{diag}\{d_m, \dots, d_1\} \in \mathbb{R}^{m\times m} \tag{60}$$

where $d_i > d_j$ when $i < j$. Then there are a matrix $\theta \in \mathbb{R}^{m\times m}$ and a full Metzler matrix F such that $\theta D\theta^{-1} = F$.

Proof: We can assume that $d_m > 0$. Indeed, if this were not satisfied, then we can first study $D + \nu I$, where $\nu \in \mathbb{R}$ is such that $d_m + \nu > 0$. We proceed by induction.

Induction Hypothesis: There is an invertible matrix $L \in \mathbb{R}^{k\times k}$ such that $L\text{diag}\{d_k, \dots, d_1\}L^{-1} = G$, where $G = [g_{ij}] \in \mathbb{R}^{k\times k}$ is positive. Moreover, $g_{11} > d_{k+1}$.

Step 1: The induction hypothesis is satisfied at the step 1 for $k = 1$, with $L = 1$, because $d_1 > d_2$.

Step $k+1$: Let us assume that the induction hypothesis is satisfied at the step k for some choices of G and L. Then let

$$L_\star = \begin{bmatrix} 1 & 0 \\ 0 & L \end{bmatrix} \tag{61}$$

Then $L_\star\text{diag}\{d_{k+1}, \dots, d_1\}L_\star^{-1} = G_\star$, where

$$G_\star = \begin{bmatrix} d_{k+1} & 0 \\ 0 & G \end{bmatrix}. \tag{62}$$

By Lemma 1, the matrix $G_\star$ is similar to

$$H_\epsilon = \begin{bmatrix} \frac{d_{k+1}+\epsilon^2 g_{1,1}}{1+\epsilon^2} & \frac{\epsilon(g_{1,1}-d_{k+1})}{1+\epsilon^2} & \epsilon g_{1,2} & \dots & \epsilon g_{1,k} \\ \frac{\epsilon(g_{1,1}-d_{k+1})}{1+\epsilon^2} & \frac{d_{k+1}\epsilon^2+g_{1,1}}{1+\epsilon^2} & g_{1,2} & \dots & g_{1,k} \\ \frac{\epsilon g_{2,1}}{1+\epsilon^2} & \frac{g_{2,1}}{1+\epsilon^2} & g_{2,2} & \dots & g_{2,k} \\ \vdots & & \vdots & & \vdots \\ \frac{\epsilon g_{k,1}}{1+\epsilon^2} & \frac{g_{k,1}}{1+\epsilon^2} & g_{k,2} & \dots & g_{k,k} \end{bmatrix}. \tag{63}$$

for any $\epsilon > 0$. The matrix H_ϵ is positive. Also, $d_{k+2} < d_{k+1}$. It follows that there is a constant $\epsilon > 0$ such that

$$\frac{d_{k+1} + \epsilon^2 g_{11}}{1+\epsilon^2} > d_{k+2} \tag{64}$$

Hence, the induction assumption holds at step $k+1$. ∎

Remark 3: The preceding theorem can be extended easily to any diagonal matrix with distinct diagonal entries, because then the matrix is similar to a matrix $D = \text{diag}\{d_m, \dots, d_1\}$ such that $d_i > d_j$ when $i < j$.

C. Diagonal Matrices With Repeated Diagonal Entries

Let us establish the following result.

Theorem 4: For any $n \in \mathbb{N}$, the matrix

$$D = \text{diag}\{d_1, d_2, \dots, d_2\} \in \mathbb{R}^{n\times n} \tag{65}$$

for values $d_1 > d_2$ is similar to the full Metzler matrix $F = d_2 I + \frac{d_1-d_2}{n}\mathbf{1}$ where $\mathbf{1} \in \mathbb{R}^{n\times n}$ is the matrix whose entries are all 1's. When $n \geq 3$ and $d_1 \leq d_2$, D is not similar to a full Metzler matrix.

Proof (First Part): Consider the case where $d_1 > d_2$. Set

$$Q = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & -1 & 1 & & \vdots \\ 1 & 0 & -1 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 1 & 0 & 0 & \dots & -1 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{66}$$

which is invertible (e.g., since any $x \in \mathbb{R}^n$ such that $x^\top Q = 0$ would need to satisfy $x_1 + \dots x_n = 0$ and $x_i - x_{i+1} = 0$ for $i = 1, \dots, n-1$ and so is $x = 0$). Then

$$Q\text{diag}\{1, 0, \dots, 0\} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \tag{67}$$

and

$$\mathbf{1}Q = n\begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \tag{68}$$

It follows that $Q\text{diag}\{1, 0, \dots, 0\}Q^{-1} = \frac{1}{n}\mathbf{1}$. Since

$$D = d_2 I + (d_1 - d_2)\text{diag}\{1, 0, \dots, 0\}, \tag{69}$$

it now follows that $QDQ^{-1} = d_2 I + \frac{d_1-d_2}{n}\mathbf{1}$. This concludes the first part of the proof.

Second Part: We consider the case where $n \geq 3$ and $d_1 \leq d_2$.

The case $d_1 = d_2$ is trivial because of (69). Therefore, let us focus our attention to the case where $d_1 < d_2$. Since $D = d_2 I + (d_2 - d_1)D_1$ with $D_1 = \text{diag}\{-1, 0, \dots, 0\}$, the matrix D is similar to a full Metzler matrix if and only if D_1 is similar to a full Metzler matrix.

To show that D_1 is not similar to a full Metzler matrix, let us proceed by contradiction. Suppose that there were an invertible matrix $R \in \mathbb{R}^{n\times n}$ such that $RD_1R^{-1} = F$, where F is full Metzler. Let $\mu > 1$ be a large enough constant such that $F + \mu I$ is positive. By the Perron–Frobenius theorem, $F + \mu I$ has a simple dominant eigenvalue. By similarity, the characteristic polynomial of $F + \mu I$ equals that of $D_1 + \mu I$ and is given by $(x - [\mu - 1])(x - \mu)^{n-1}$, where $n - 1 \geq 2$. In particular, μ is the dominant eigenvalue and is not simple, which yields a contradiction. This concludes the proof. ∎

IV. Marine Robotic Application

We revisit the dynamics from [12] for controlling the depth and pitch degrees-of-freedom (or DOF) for an autonomous underwater vehicle (or AUV), which was shown to be suitable for studying the BlueROV2 vehicle that is widely used for environmental surveys such as the study of corals. As in [12], we assume that the AUV is equipped with a Doppler Velocity Logger (or DVL), which is used to estimate its velocity. For marine surveys close to the sea floor, the DVL can experience bottom lock, making it impractical to ensure consistent thrust controls or to continuously change the control values. In [11] and [12], event-triggered controls (as defined, e.g., in [3] and [4]) were designed for this model, to reduce the numbers of time instances when the control values are changed. Here, we study a complementary problem of inconsistent thrust controls by estimating the transition matrix of the closed loop AUV dynamics when it is affected by unknown control gains. We use Theorem 3.

Assuming that the vehicle is neutrally buoyant, we can linearize the dynamics in the depth plane to obtain [12]

$$\begin{aligned}&(m - X_{\dot{w}(t)})\dot{w}(t) - (mx_g + Z_{\dot{q}})\dot{q}(t)\\ &\quad - Z_w w(t) - (mU + z_q)q(t) = Z_{\gamma_s} u_Z\\ &\text{and } (mx_g + M_{\dot{w}}(t))\dot{w}(t) + (I_{yy} - M_{\dot{q}})\dot{q}(t)\\ &\quad - M_w w(t) + (mx_g U - M_q)q(t) - M_\theta \theta = M_{\gamma_s} u_M \end{aligned} \tag{70}$$

whose parameters were obtained experimentally in [13]. Its states are the depth and pitch velocity $x = [w,\ q]^\top$, and its control inputs u_Z and u_M are the force and moment required to produce motion of the vehicle, respectively, where we assume that the surge nominal velocity is $U = 0.1$m/s. With the controller and parameters from [12] and [13], the dynamics (70) become $\dot{x}(t) = Ax(t) + Bu$ with

$$A = \begin{bmatrix} -0.17742 & -0.3027 \\ 0.5394 & -1.4685 \end{bmatrix} \text{ and } B = \begin{bmatrix} -0.2063 \\ -0.7629 \end{bmatrix}. \tag{71}$$

Our strategy for studying this system can be summarized in several steps. First, we choose K_0 such that $A + BK_0$ has distinct negative real eigenvalues. Hence, using Remark 3, we can find a matrix P such that $F = P(A + BK_0)P^{-1}$ is a full Metzler matrix, by first diagonalizing $A+BK_0$ to obtain the required structure (60). Next, we consider the case where the control $u = (K_0 + \delta_K(t))x$ is perturbed by a piecewise continuous bounded uncertainty $\delta_K(t)$ (representing uncertain control gains) having known bounds. Then the new state variable $z = Px$ satisfies $\dot{z} = (F + P\Delta_K(t)P^{-1})z$, where $\Delta_K = B\delta_K$. One can find constant matrices $\underline{\kappa}$ and $\overline{\kappa}$ such that $\underline{\kappa} \le P\Delta_K(t)P^{-1} \le \overline{\kappa}$, so the tight bounds

$$e^{(F+\underline{\kappa})(t-s)} \le \Phi_{F+P\Delta_K P^{-1}}(t, s) \le e^{(F+\overline{\kappa})(t-s)} \tag{72}$$

hold for all $s \ge 0$ and $t \ge s$, provided $F + \underline{\kappa}$ is still Metzler (again by the monotonicity condition from [6, Lemma 2]). This uses the fullness of the Metzler condition on F in an essential way, because if F is a full Metzler matrix, then $F + \underline{\kappa}$ will be Metzler provided all entries $\underline{\kappa}$ are small enough.

To see how this can be done in the special case where we use the choice $K_0 = [0.59, 0.23]$ from [11], notice that this choice of K_0 produces a matrix $A+BK_0$ having eigenvalues -1.6203 and -0.322801. Using a diagonalizing matrix R_0 for $A+BK_0$, we can now apply Theorem 3 to $D = 1.7I+R_0(A+BK_0)R_0^{-1} = \text{diag}\{d_2, d_1\} = \text{diag}\{0.0796973, 1.3772\}$, which by the proof of Theorem 3 (applied with $n = 2$ and $\epsilon = 1$) satisfies

$$\theta D\theta^{-1} = \begin{bmatrix} \frac{d_1+d_2}{2} & \frac{d_1-d_2}{2} \\ \frac{d_1-d_2}{2} & \frac{d_1+d_2}{2} \end{bmatrix} \text{ where } \theta = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}. \tag{73}$$

Then we can satisfy the preceding requirements with $P = \theta R_0$ and $F = \theta D\theta^{-1} - 1.7I$. This gives the requirement $\min\{\underline{\kappa}_{1,2}, \underline{\kappa}_{2,1}\} \ge \frac{d_2-d_1}{2}$ on the off-diagonal elements of $\underline{\kappa}$, to ensure that $F + \underline{\kappa}$ is still Metzler and to obtain (72).

V. Conclusion

We provided new matrix-valued interval observers for transition matrices for continuous-time and discrete-time time-varying systems. We illustrated how they can be less conservative than those that were provided by [9] and [10]. A key ingredient was our novel transformations of diagonalized matrices into full Metzler matrices. Our estimations are of independent interest, owing to the need to find nonconservative interval observers for fundamental matrices under unknown control gains. Also, as noted in [9], transition matrix estimations are useful for stabilization and observer design [1], [2], [5]. We aim to study how our new transition matrix estimates may improve on the convergence rates for the stabilization and observer results from [9] and [10].

References

[1] P. Bernard, V. Andrieu, and D. Astolfi, "Observer design for continuous-time dynamical systems," *Annu. Rev. Control*, to be published.

[2] A. Borri *et al.*, "Luenberger-like observers for nonlinear time-delay systems with application to the artificial pancreas: The attainment of good performance," *IEEE Control Syst. Mag.*, vol. 37, no. 4, pp. 33–49, Aug. 2017.

[3] A. Borri and P. Pepe, "Event-triggered control of nonlinear systems with time-varying state delays," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2846–2853, Jun. 2021.

[4] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *Proc. 51st IEEE Conf. Decis. Control*, 2012, pp. 3270–3285.

[5] R. Katz, E. Fridman, and A. Selivanov, "Boundary delayed observer-controller design for reaction–diffusion systems," *IEEE Trans. Autom. Control*, vol. 66, no. 1, pp. 275–282, Jan. 2021.

[6] F. Mazenc, V. Andrieu, and M. Malisoff, "Design of continuous–discrete observers for time-varying nonlinear systems," *Automatica*, vol. 57, pp. 135–144, Jul. 2015.

[7] F. Mazenc and O. Bernard, "Interval observers for linear time-invariant systems with disturbances," *Automatica*, vol. 47, no. 1, pp. 140–147, 2011.

[8] F. Mazenc and O. Bernard, "When is a matrix of dimension 3 similar to a Metzler matrix application to interval observer design," *IEEE Trans. Autom. Control*, early access, Oct. 27, 2021, doi: 10.1109/TAC.2021.3123245.

[9] F. Mazenc and M. Malisoff, "Feedback stabilization with discrete measurements using bounds on fundamental matrices," in *Proc. 60th IEEE Conf. Decis. Control*, 2021, pp. 1814–1819.

[10] F. Mazenc and M. Malisoff, "Feedback stabilization and robustness analysis using bounds on fundamental matrices," *Syst. Control Lett.*, vol. 164, 2022, Art. no. 105212.

[11] F. Mazenc, M. Malisoff, C. Barbalata, and Z.-P. Jiang, "Event-triggered control for systems with state delays using a positive systems approach," in *Proc. 60th IEEE Conf. Decis. Control*, 2021, pp. 552–557.

[12] F. Mazenc, M. Malisoff, C. Barbalata, and Z.-P. Jiang, "Event-triggered control for discrete-time systems using a positive systems approach," *IEEE Control Syst. Lett.*, vol. 6, pp. 1843–1848, 2022.

[13] T. Prestero, "Verification of a six-degree of freedom simulation model for the REMUS autonomous underwater vehicle," M.S. thesis, Dept. Ocean Eng., MIT, Cambridge, MA, USA, 2001.