Linear Algebra and its Applications 679 (2023) 275–304

Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

A greedy Galerkin method to efficiently select sensors for linear dynamical systems

Drew P. Kouri [a,*,1], Zuhao Hua [b,2], Madeleine Udell [c,2]

[a] *Optimization and Uncertainty Quantification, Sandia National Laboratories, Albuquerque, NM 87185, USA*
[b] *Physics, Cornell University, Ithaca, NY 14850, USA*
[c] *Management Science and Engineering, Stanford University, Stanford, CA 94305, USA*

ARTICLE INFO

Article history:
Received 22 September 2022
Received in revised form 26 July 2023
Accepted 4 September 2023
Available online 18 September 2023
Submitted by V. Mehrmann

MSC:
62K05

ABSTRACT

A key challenge in inverse problems is the selection of sensors to gather the most effective data. In this paper, we consider the problem of inferring the initial condition to a linear dynamical system and develop an efficient control-theoretical approach for greedily selecting sensors. Our method employs a Galerkin projection to reduce the size of the inverse problem, resulting in a computationally efficient algorithm for sensor selection. As a byproduct of our algorithm, we obtain a preconditioner for the inverse problem that enables the rapid

* Corresponding author.
E-mail addresses: dpkouri@sandia.gov (D.P. Kouri), zh296@cornell.edu (Z. Hua), udell@stanford.edu (M. Udell).

[1] DPK was partially supported by the DOE ASCR Early Career Research Project "Adaptive and Fault-Tolerant Algorithms for Data-Driven Optimization, Design, and Learning", U.S. Air Force Office of Scientific Research award F4FGA09135G001, and the Sandia Laboratory Directed Research and Development (LDRD) program. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.
[2] ZH and MU were partially supported by NSF Award IIS-1943131, the ONR Young Investigator Program, and the Alfred P. Sloan Foundation.

https://doi.org/10.1016/j.laa.2023.09.003

34H05
93B07
65F45
49N05

Keywords:
Sensor placement
Linear dynamics
Gramian
Observability
Lyapunov equation
Preconditioning
Galerkin method
Greedy algorithm
Inverse problems
Optimal control
Submodularity

recovery of the initial condition. We analyze the theoretical performance of our greedy sensor selection algorithm as well as the performance of the associated preconditioner. Finally, we verify our theoretical results on various inverse problems involving partial differential equations.

1. Introduction

In this paper, we consider the task of inferring the initial condition $\mathbf{x}_0$ to a linear time-invariant dynamical system from measured data, which we formulate as the optimization problem

$$\min_{\mathbf{x}(t),\,\mathbf{x}_0} \ \frac{1}{2}\int_0^{t_f} \|\mathbf{C}\mathbf{x}(t) - \mathbf{y}_d(t)\|_2^2 \,\mathrm{d}t + \frac{1}{2}\mathbf{x}_0^\top \mathbf{R}\mathbf{x}_0 + \Phi(\mathbf{x}_0) \tag{1a}$$

$$\text{subject to} \quad \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{f}(t) \quad \text{for } t \in (0, t_f] \\ \mathbf{x}(0) = \mathbf{x}_0 \end{cases}. \tag{1b}$$

Here, $t_f > 0$ denotes the final time, $\mathbf{E}$ and $\mathbf{A}$ are m-by-m matrices, $\mathbf{f}(t) \in \mathbb{R}^m$ is a load or force, $\mathbf{C}$ is an S-by-m rectangular matrix that produces measurements of the state $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y}_d(t) \in \mathbb{R}^S$ is observed data, $\mathbf{R}$ is a symmetric m-by-m matrix representing a quadratic regularization function, and Φ is an extended real-valued, convex function that encapsulates auxiliary convex constraints and nonsmooth regularization terms. We interpret each row of $\mathbf{C}$ to be a sensor. In an attempt to reduce the recovery error for the initial condition, we develop an efficient numerical method to select sensors from a finite set of possible sensors. As a byproduct of our algorithm, we produce a good preconditioner to accelerate the iterative solution of the optimization problem (1). Optimization problems similar to (1) arise in numerous applications including data assimilation for weather forecasting and climate modeling [1,2], detecting sources of contaminants such as pollution, radiation or contagions [3,4], and calibrating financial models [5] and experimental facilities [6].

A common statistical approach for selecting sensors, and more generally designing experiments, is based on the Fisher information matrix or the covariance matrix of the estimated initial condition $\mathbf{x}_0$ [7,8]. In design of experiments, it is common to maximize some metric of the Fisher information matrix such as the determinant, the minimum eigenvalue, or the trace. In the context of (1), [9,10] choose sensors using the determinant

of the Fisher information matrix. A related approach is based on frame potential, which measures the orthogonality of the rows of the Fisher information matrix [11,12].

Our approach is closely related to these statistical approaches. However, instead of using the Fisher information matrix, we select sensors using the observability Gramian of the dynamical system. As with design of experiments, several scalarized version of the observability Gramian have been investigated including the minimum eigenvalue [13], the trace [14], and the determinant [15]. This approach attempts to maximize the output energy of the dynamical system and thereby minimize the worst-case recovery error for any initial condition $\mathbf{x}_0$. In the context of (1), this approach assumes an infinite time horizon $t_f = +\infty$ and consequently only provides an approximation when t_f is finite—an approximation that must be accounted for to guarantee that the computed sensors are near optimal for (1).

Traditional Gramian-based sensor selection has been successfully applied to stable linear time-invariant dynamical systems with the form (1b). However, the typical stability assumptions are too restrictive for many practical applications [16]. To circumvent this issue, the Gramian-based approach has been generalized to handle nonlinear dynamical systems [14,17], unstable dynamics [18], infinite-dimensional systems of partial differential equations (PDEs) [19,20], differential algebraic equations [21], and closed-loop control systems [22].

In general, we formulate sensor selection as the binary optimization problem

$$\max_{w} \ \Psi\left(\mathbf{R} + \sum_{s=1}^{\bar{S}} w_s \mathbf{Q}_s\right) \quad \text{subject to} \quad \sum_{s=1}^{\bar{S}} w_s \leqslant S, \quad w_s \in \{0,1\}, \tag{2}$$

where $\mathbf{Q}_s$ is the observability Gramian for sensor s, S is the sensor budget, $\bar{S}$ is the number of possible sensors and the functional Ψ acts on square matrices to measure the quality of the selected sensors. For example, Ψ is the determinant, minimum eigenvalue, or trace. The solution of (2) is NP-hard. Even evaluating the objective function for a given set of sensors can scale with the cube of the problem dimension.

To address these challenges, we develop a greedy sensor placement algorithm that offers guaranteed solution quality. To ensure that our approach is computationally tractable, we employ a Galerkin projection of the observability Gramian to reduce the overall dimensionality of the problem. Although we consider the finite time horizon case $t_f < \infty$, we employ the infinite time horizon observability Gramian, which introduces an error. We quantify this error, showing that the observability Gramian produces a sufficiently accurate approximation of the reduced Hessian matrix from (1). As a consequence, we can employ the observability Gramian to place sensors. Moreover, we can use the observability Gramian as a preconditioner for the reduced Hessian to accelerate the iterative solution of (1).

The paper is organized as follows. In Section 2, we introduce the notation and standing assumptions used throughout. In Section 3, we review the optimality conditions for (1) and derive a bound on the recovery error. We use this bound to motivate our sensor

selection algorithm. In Section 4, we prove that the observability Gramian provides an accurate approximation to the reduced Hessian matrix from (1) and therefore it serves well as a preconditioner and as a proxy for selecting sensors. In Section 5, we introduce our Galerkin approximation to the observability Gramian and demonstrate that this approximation is exact under certain assumptions. Finally, in Section 6 we describe our greedy sensory selection algorithm and demonstrate its performance on various PDE examples in Section 7.

2. Notation and standing assumptions

Given a symmetric positive definite matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$, we denote the $\mathbf{M}$-inner product by

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} := \mathbf{x}^\top \mathbf{M}\mathbf{y}$$

and the associated norm by $\|\mathbf{x}\|_{\mathbf{M}}^2 := \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{M}}$. Given another symmetric positive definite matrix $\mathbf{N} \in \mathbb{R}^{m\times m}$, we denote the induced matrix norm associated with $\mathbf{M}$ and $\mathbf{N}$ by

$$\|\mathbf{B}\|_{\mathbf{M},\mathbf{N}} := \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{B}\mathbf{x}\|_{\mathbf{M}}}{\|\mathbf{x}\|_{\mathbf{N}}}.$$

When $\mathbf{M} = \mathbf{N}$, we denote the norm $\|\cdot\|_{\mathbf{M},\mathbf{M}} = \|\cdot\|_{\mathbf{M}}$ and when $\mathbf{M} = \mathbf{N} = \mathbf{I}$, where $\mathbf{I}$ is the appropriately sized identity matrix, the induced matrix norm is the usual matrix 2-norm, i.e., $\|\cdot\|_{\mathbf{I}} = \|\cdot\|_2$. If either $\mathbf{M} = \mathbf{I}$ or $\mathbf{N} = \mathbf{I}$, we replace the associated subscript with the number 2. In addition, for any square matrices $\mathbf{B}$ and $\mathbf{M}$ of the same size, we denote the maximum and minimum generalized eigenvalues of the pair $(\mathbf{B}, \mathbf{M})$ by $\lambda_{\max}(\mathbf{B}, \mathbf{M})$ and $\lambda_{\min}(\mathbf{B}, \mathbf{M})$, respectively. Recall that λ is a generalized eigenvalue of the pair $(\mathbf{B}, \mathbf{M})$ if there exists a generalized eigenvector $\mathbf{v}$ satisfying

$$\mathbf{B}\mathbf{v} = \lambda \mathbf{M}\mathbf{v}.$$

When $\mathbf{M} = \mathbf{I}$, we simplify this notation to $\lambda_{\max}(\mathbf{B})$ and $\lambda_{\min}(\mathbf{B})$. Finally, for an arbitrary square matrix $\mathbf{B}$, we denote the maximum real part of the eigenvalues of $\mathbf{B}$ by $\alpha(\mathbf{B})$ and refer to this quantity as the *spectral abscissa*.

Throughout, $\mathbf{E}$, $\mathbf{A}$ and $\mathbf{R}$ denote m-by-m matrices with real entries, $\mathbf{f} : (0, t_f] \to \mathbb{R}^m$, and $\Phi : \mathbb{R}^m \to (-\infty, +\infty]$. We assume that $\mathbf{E}$ is symmetric positive definite, $\mathbf{R}$ is symmetric positive semidefinite, and Φ is given by

$$\Phi(\mathbf{x}) = \begin{cases} \Phi_0(\mathbf{x}) & \text{if } \mathbf{x} \in \mathcal{F}, \\ +\infty & \text{otherwise,} \end{cases} \tag{3}$$

where $\Phi_0 : \mathbb{R}^m \to (-\infty, +\infty]$ is convex and Lipschitz continuous with respect to the $\mathbf{E}$-norm on a neighborhood of the nonempty, closed and convex set $\mathcal{F} \subseteq \mathbb{R}^m$. We denote the Lipschitz modulus of Φ_0 by $L \geqslant 0$. As (3) suggests, (1) includes the constraints $\mathbf{x}_0 \in \mathcal{F}$.

We denote by $\mathcal{X}$ the vector space $\mathbb{R}^m$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathbf{E}}$. We associate with $\mathcal{X}$ the dual space $\mathcal{X}'$, which is the vector space $\mathbb{R}^m$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathbf{E}^{-1}}$. We recall that for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}'$, the usual continuity bound holds, i.e.,

$$|\mathbf{x}^\top \mathbf{y}| \leqslant \|\mathbf{x}\|_{\mathbf{E}} \, \|\mathbf{y}\|_{\mathbf{E}^{-1}} . \tag{4}$$

We further note that for each $\bar{\mathbf{x}} \in \mathcal{X}'$, the vector $\mathbf{x} = \mathbf{E}^{-1}\bar{\mathbf{x}} \in \mathcal{X}$ is the Riesz representer of $\bar{\mathbf{x}}$. We view the matrices $\mathbf{E}$, $\mathbf{A}$ and $\mathbf{R}$ as linear operators from $\mathcal{X}$ to $\mathcal{X}'$ and the vector $\mathbf{f}(t)$ as an element in $\mathcal{X}'$ for $t > 0$. In particular, $\mathbf{E}^{-1}\mathbf{A}$ and $\exp(\mathbf{E}^{-1}\mathbf{A}t)$ for $t > 0$ are linear operators from $\mathcal{X}$ to $\mathcal{X}$. We make the following assumption on $\mathbf{E}^{-1}\mathbf{A}$ to ensure stability of the dynamical system (1b).

Assumption 1 *(Stability).* The logarithmic norm, with respect to the $\mathbf{E}$-norm, of the matrix $\mathbf{E}^{-1}\mathbf{A}$ satisfies

$$\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A}) < 0,$$

where the logarithmic norm of a square matrix $\mathbf{B}$ with respect to the $\mathbf{E}$-norm is defined as

$$\mu_{\mathbf{E}}(\mathbf{B}) := \lim_{h \downarrow 0} \frac{\|\mathbf{I} + h\mathbf{B}\|_{\mathbf{E}} - 1}{h} = \alpha(\tfrac{1}{2}(\mathbf{B}^\top + \mathbf{E}\mathbf{B}\mathbf{E}^{-1})).$$

One consequence of Assumption 1 is that

$$\left\|\exp(\mathbf{E}^{-1}\mathbf{A}t)\right\|_{\mathbf{E}} \leqslant \exp(\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t) \leqslant 1 \quad \forall\, t \geqslant 0.$$

Another consequence is that the spectral abscissa of $\mathbf{E}^{-1}\mathbf{A}$ is negative and therefore the linear time-invariant system

$$\begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{f}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \end{cases} \tag{5}$$

is *stable*. Owing to a similarity transformation, we have that $\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})$ is the maximal eigenvalue of $\frac{1}{2}\mathbf{E}^{-1}(\mathbf{A} + \mathbf{A}^\top)$, or equivalently the maximal generalized eigenvalue of the matrix pair $(\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top), \mathbf{E})$. In addition, we denote the logarithmic norm, with respect to the matrix 2-norm, of a square matrix $\mathbf{B}$ by $\mu_2(\mathbf{B})$ and note that this quantity is the spectral abscissa of $\frac{1}{2}(\mathbf{B} + \mathbf{B}^\top)$. For more information on the logarithmic norm, see [23–25].

Finally, we denote the set of possible sensors by $\{\mathbf{c}_s\}_{s=1}^{\bar{S}}$, where $\mathbf{c}_s \in \mathcal{X}'$. We construct the observation matrix $\mathbf{C}_\sigma$ using a subset $\sigma \subseteq \{1, \ldots, \bar{S}\}$ of possible sensors. In particular, the observation matrix $\mathbf{C}_\sigma \in \mathbb{R}^{S \times m}$, with $S = |\sigma|$, has rows given by $\{\mathbf{c}_s^\top\}_{s \in \sigma}$. We

interpret $\mathbf{C}_\sigma$ as a linear operator from $\mathcal{X}$ to $\mathbb{R}^S$, where the range space is endowed with the Euclidean inner product. We recall that the matrix-matrix product $\mathbf{C}_\sigma^\top \mathbf{C}_\sigma$ satisfies

$$\mathbf{C}_\sigma^\top \mathbf{C}_\sigma = \sum_{s\in\sigma} \mathbf{c}_s \mathbf{c}_s^\top \tag{6}$$

for all subsets $\sigma \subseteq \{1,\dots,\bar{S}\}$. For notational convenience, we often denote $\mathbf{C} = \mathbf{C}_\sigma$, where σ is a fixed subset of $\{1,\dots,\bar{S}\}$. We further denote the data associated with the observations $\{\mathbf{c}_s\}_{s\in\sigma}$ by $\mathbf{y}_d : (0,t_f] \to \mathbb{R}^S$.

3. Optimality conditions and recovery error

In this section, we review some basic results regarding the optimization problem (1). First, we recall that the solution $\mathbf{x}(t) = [\mathbf{S}(\mathbf{x}_0)](t)$ to the linear dynamical system (1b) is affine in the initial condition $\mathbf{x}_0$ and is given by

$$\mathbf{x}(t) = \exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{x}_0 + \exp(\mathbf{E}^{-1}\mathbf{A}t)\int_0^t \exp(-\mathbf{E}^{-1}\mathbf{A}\tau)\mathbf{E}^{-1}\mathbf{f}(\tau)\,\mathrm{d}\tau.$$

We denote by $\bar{\mathbf{y}} : [0,t_f] \to \mathbb{R}^S$ the function

$$\bar{\mathbf{y}}(t) := \mathbf{y}_d(t) - \mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t)\int_0^t \exp(-\mathbf{E}^{-1}\mathbf{A}\tau)\mathbf{E}^{-1}\mathbf{f}(\tau)\,\mathrm{d}\tau,$$

which allows us to rewrite the optimization problem (1) in reduced form as

$$\min_{\mathbf{x}_0\in\mathbb{R}^m} \frac{1}{2}\int_0^{t_f} \|\mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{x}_0 - \bar{\mathbf{y}}(t)\|_2^2\,\mathrm{d}t + \frac{1}{2}\mathbf{x}_0^\top\mathbf{R}\mathbf{x}_0 + \Phi(\mathbf{x}_0). \tag{7}$$

Any solution to (7) satisfies the first-order necessary and sufficient optimality condition

$$\left[\int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top\mathbf{C}^\top\bar{\mathbf{y}}(t)\,\mathrm{d}t - (\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{x}_0\right]^\top (\mathbf{x}_0' - \mathbf{x}_0) \leqslant \Phi(\mathbf{x}_0') - \Phi(\mathbf{x}_0) \tag{8}$$

for all $\mathbf{x}_0' \in \mathcal{X}$, where the Hessian matrix $\mathbf{H}_0^{t_f}$ is given by

$$\mathbf{H}_0^{t_f} := \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top\mathbf{C}^\top\mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t)\,\mathrm{d}t. \tag{9}$$

Note that the sub and superscript on the Hessian matrix refer to the initial and final times. Consequently, $\mathbf{H}_0^\infty$ is the Hessian matrix for the infinite time horizon problem (i.e., $t_f = +\infty$) and is the so-called observability Gramian. We also note that $\mathbf{H}_0^{t_f}$ is a linear operator from $\mathcal{X}$ into $\mathcal{X}'$ and the Hessian of the objective function in (7) with respect to the $\mathbf{E}$-inner product, omitting the regularization terms, is $\mathbf{E}^{-1}\mathbf{H}_0^{t_f}$. In particular,

$\mathbf{E}^{-1}\mathbf{H}_0^{t_f}$ is self-adjoint with respect to the $\mathbf{E}$-inner product, i.e., for all $\mathbf{u}, \mathbf{v} \in \mathcal{X}$, we have that

$$\langle \mathbf{u}, \mathbf{E}^{-1}\mathbf{H}_0^{t_f}\mathbf{v}\rangle_{\mathbf{E}} = \mathbf{u}^\top \mathbf{E}\mathbf{E}^{-1}\mathbf{H}_0^{t_f}\mathbf{v} = \mathbf{u}^\top \mathbf{H}_0^{t_f}\mathbf{v} = \mathbf{v}^\top \mathbf{H}_0^{t_f}\mathbf{u} = \langle \mathbf{v}, \mathbf{E}^{-1}\mathbf{H}_0^{t_f}\mathbf{u}\rangle_{\mathbf{E}}.$$

By Assumption 1, the spectral abscissa of $\mathbf{E}^{-1}\mathbf{A}$ is negative and consequently $\mathbf{H}_0^\infty$ solves the Lyapunov equation

$$(\mathbf{E}^{-1}\mathbf{A})^\top \mathbf{H}_0^\infty + \mathbf{H}_0^\infty(\mathbf{E}^{-1}\mathbf{A}) + \mathbf{C}^\top\mathbf{C} = \mathbf{0}. \tag{10}$$

When $\mathbf{H}_0^\infty$ is positive definite, the system (5) is *observable*. Additionally, making the substitution $\mathbf{H}_0^\infty = \mathbf{E}\overline{\mathbf{H}}_0^\infty\mathbf{E}$ in (10), demonstrates that $\overline{\mathbf{H}}_0^\infty$ solves the generalized Lyapunov equation

$$\mathbf{A}^\top\overline{\mathbf{H}}_0^\infty\mathbf{E} + \mathbf{E}\overline{\mathbf{H}}_0^\infty\mathbf{A} + \mathbf{C}^\top\mathbf{C} = \mathbf{0}. \tag{11}$$

Solving (11) can have computational advantages over solving (10) because it does not require the inverse matrix $\mathbf{E}^{-1}$.

Owing to (9) and the linearity of the integral, we can write the finite-time Hessian $\mathbf{H}_0^{t_f}$ in terms of the observability Gramian $\mathbf{H}_0^\infty$. In particular,

$$\mathbf{H}_0^{t_f} = \mathbf{H}_0^\infty - \exp(\mathbf{E}^{-1}\mathbf{A}t_f)^\top\mathbf{H}_0^\infty\exp(\mathbf{E}^{-1}\mathbf{A}t_f).$$

From this, we see that $\mathbf{H}_0^{t_f}$ solves the Lyapunov equation

$$(\mathbf{E}^{-1}\mathbf{A})^\top\mathbf{H}_0^{t_f} + \mathbf{H}_0^{t_f}(\mathbf{E}^{-1}\mathbf{A}) + \mathbf{C}^\top\mathbf{C} - \exp(\mathbf{E}^{-1}\mathbf{A}t_f)^\top\mathbf{C}^\top\mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t_f) = \mathbf{0}$$

(cf. [26,27] for additional details). In particular, we can compute $\mathbf{H}_0^{t_f}$ by solving the generalized Lyapunov equation

$$\mathbf{A}^\top\overline{\mathbf{H}}_0^{t_f}\mathbf{E} + \mathbf{E}\overline{\mathbf{H}}_0^{t_f}\mathbf{A} + \mathbf{C}^\top\mathbf{C} - \exp(\mathbf{E}^{-1}\mathbf{A}t_f)^\top\mathbf{C}^\top\mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t_f) = \mathbf{0},$$

and setting $\mathbf{H}_0^{t_f} = \mathbf{E}\overline{\mathbf{H}}_0^{t_f}\mathbf{E}$. Notice that the computation of $\mathbf{H}_0^{t_f}$ requires multiple applications of the matrix exponential $\exp(\mathbf{E}^{-1}\mathbf{A}t_f)$, which can be performed with $\mathcal{O}(m^3)$ complexity using, e.g., the scaling-and-squaring method [28].

To conclude this discussion, suppose that the target data $\mathbf{y}_d$ is given by the additive noise relationship

$$\mathbf{y}_d(t) = \widehat{\mathbf{C}}[\widehat{\mathbf{S}}(\mathbf{x}_0^\star)](t) + \eta_t,$$

where η_t is an S-dimensional random vector of measurement noise, $\widehat{\mathbf{C}}$ is the true observation operator, $\widehat{\mathbf{S}}$ is the true state map, and $\mathbf{x}_0^\star \in \mathbb{R}^m$ is the unknown initial condition. The true observation operator $\widehat{\mathbf{C}}$ and state map $\widehat{\mathbf{S}}$ may differ from $\mathbf{C}$ and $\mathbf{S}$, respectively,

because of incorrectly specified parameters, inaccurate modeling assumptions, and unknown environmental conditions. These true operators may even be nonlinear. Let $\mathbf{p}_0^\star$ denote the metric projection in $\mathcal{X}$ of $\mathbf{x}_0^\star$ onto the feasible set $\mathcal{F}$. Substituting the assumed form of $\mathbf{y}_d$ into the bracketed term on the left-hand side of the optimality conditions (8), adding and subtracting $\mathbf{CS}(\mathbf{p}_0^\star)$ to $\mathbf{y}_d$, and adding and subtracting $\mathbf{R}\mathbf{p}_0^\star$ yields

$$\begin{aligned}\int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \bar{\mathbf{y}}(t)\,\mathrm{d}t - (\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{x}_0 \\ = (\mathbf{H}_0^{t_f} + \mathbf{R})(\mathbf{p}_0^\star - \mathbf{x}_0) + \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \eta_t\,\mathrm{d}t - \mathbf{R}\mathbf{p}_0^\star \\ + \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top (\widehat{\mathbf{C}}[\widehat{\mathbf{S}}(\mathbf{x}_0^\star)] - \mathbf{C}[\mathbf{S}(\mathbf{p}_0^\star)])(t)\,\mathrm{d}t.\end{aligned}$$

A consequence of this equation, the optimality conditions (8), the Lipschitz continuity of Φ_0, and the triangle inequality is that the error committed between the solution $\mathbf{x}_0$ to (7) and the unknown $\mathbf{x}_0^\star$ is bounded by

$$\|\mathbf{x}_0 - \mathbf{x}_0^\star\|_{\mathbf{E}} \leqslant \|\mathbf{p}_0^\star - \mathbf{x}_0^\star\|_{\mathbf{E}} + \frac{M}{\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})}, \tag{12}$$

where $M \geqslant 0$ is bounded above by the error associated with the measurement noise η_t as well as the model error and regularization biases, i.e.,

$$\begin{aligned}M \leqslant \left\| \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top (\widehat{\mathbf{C}}[\widehat{\mathbf{S}}(\mathbf{x}_0^\star)] - \mathbf{C}[\mathbf{S}(\mathbf{p}_0^\star)])(t)\,\mathrm{d}t \right\|_{\mathbf{E}^{-1}} \\ + \left\| \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \eta_t\,\mathrm{d}t \right\|_{\mathbf{E}^{-1}} + \|\mathbf{R}\mathbf{p}_0^\star\|_{\mathbf{E}^{-1}} + L.\end{aligned}$$

In general, the noise and biases that make up M are difficult or impossible to reduce. Motivated by this, our goal is to choose the observation matrix $\mathbf{C}$ so that $\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})$ is large. By increasing $\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})$, we reduce the second term in the recovery error bound (12).

4. Theoretical results

In this section, we discuss the main theoretical results used to demonstrate that choosing the observation matrix $\mathbf{C}$ based on $\mathbf{H}_0^\infty$ provides a good proxy for increasing the minimum eigenvalue of $\mathbf{H}_0^{t_f}$. A fortuitous consequence of this analysis is that $\mathbf{H}_0^\infty + \mathbf{R}$ is a good preconditioner for $\mathbf{H}_0^{t_f} + \mathbf{R}$ and can be used to accelerate the iterative solution of (7). We begin with a technical lemma regarding the eigenvalues of perturbed matrices.

Lemma 1. *Suppose* $\mathbf{H} \in \mathbb{R}^{m\times m}$ *is symmetric positive definite,* $\mathbf{D} \in \mathbb{R}^{m\times m}$ *is symmetric positive semidefinite, and define* $\mathbf{P} = \mathbf{H} + \mathbf{D}$*. We consider* $\mathbf{H}$*,* $\mathbf{P}$*, and* $\mathbf{D}$ *to be linear operators from* $\mathcal{X}$ *into* $\mathcal{X}'$*. Then the eigenvalues of* $\mathbf{P}^{-1}\mathbf{H}$ *lie in the interval*

$$\left[\frac{\lambda_{\min}(\mathbf{H},\mathbf{E})}{\lambda_{\min}(\mathbf{H},\mathbf{E}) + \lambda_{\max}(\mathbf{D},\mathbf{E})}, \frac{\lambda_{\max}(\mathbf{H},\mathbf{E})}{\lambda_{\max}(\mathbf{H},\mathbf{E}) + \lambda_{\min}(\mathbf{D},\mathbf{E})}\right] \subseteq (0,1].$$

Moreover, the minimum eigenvalues of $\mathbf{H}$ *and* $\mathbf{P}$ *satisfy*

$$\lambda_{\min}(\mathbf{D},\mathbf{E}) \leqslant \lambda_{\min}(\mathbf{P},\mathbf{E}) - \lambda_{\min}(\mathbf{H},\mathbf{E}) \leqslant \lambda_{\max}(\mathbf{D},\mathbf{E}).$$

Proof. First, recall that the eigenvalues of $\mathbf{P}^{-1}\mathbf{H}$ are the generalized eigenvalues of $(\mathbf{H},\mathbf{P})$. Suppose μ is a generalized eigenvalue of $(\mathbf{H},\mathbf{P})$ with associated eigenvector $\mathbf{v}$, i.e., $\mathbf{H}\mathbf{v} = \mu\mathbf{P}\mathbf{v}$. Then we have that

$$(1-\mu)\mathbf{H}\mathbf{v} = \mu\mathbf{D}\mathbf{v}.$$

If $\mu > 1$, then $(1-\mu)\mathbf{v}^\top\mathbf{H}\mathbf{v} < 0$ and $\mu\mathbf{v}^\top\mathbf{D}\mathbf{v} \geqslant 0$, which cannot happen. On the other hand, if $\mu \leqslant 0$, then $(1-\mu)\mathbf{v}^\top\mathbf{H}\mathbf{v} > 0$ and $\mu\mathbf{v}^\top\mathbf{D}\mathbf{v} \leqslant 0$, which again cannot happen. Consequently, $\mu \in (0,1]$ and we have that

$$(1-\mu)\lambda_{\min}(\mathbf{H},\mathbf{E})\|\mathbf{v}\|_{\mathbf{E}}^2 \leqslant (1-\mu)\mathbf{v}^\top\mathbf{H}\mathbf{v} = \mu\mathbf{v}^\top\mathbf{D}\mathbf{v} \leqslant \mu\lambda_{\max}(\mathbf{D},\mathbf{E})\|\mathbf{v}\|_{\mathbf{E}}^2.$$

Rearranging this inequality gives the bound

$$\mu \geqslant \frac{\lambda_{\min}(\mathbf{H},\mathbf{E})}{\lambda_{\min}(\mathbf{H},\mathbf{E}) + \lambda_{\max}(\mathbf{D},\mathbf{E})}.$$

Analogously, we have that

$$(1-\mu)\lambda_{\max}(\mathbf{H},\mathbf{E})\|\mathbf{v}\|_{\mathbf{E}}^2 \geqslant (1-\mu)\mathbf{v}^\top\mathbf{H}\mathbf{v} = \mu\mathbf{v}^\top\mathbf{D}\mathbf{v} \geqslant \mu\lambda_{\min}(\mathbf{D},\mathbf{E})\|\mathbf{v}\|_{\mathbf{E}}^2,$$

which produces the stated upper bound. To conclude the proof, we note that

$$\lambda_{\min}(\mathbf{P},\mathbf{E}) \geqslant \lambda_{\min}(\mathbf{H},\mathbf{E}) + \lambda_{\min}(\mathbf{D},\mathbf{E}).$$

In addition, we have that

$$\lambda_{\min}(\mathbf{H},\mathbf{E}) \geqslant \lambda_{\min}(\mathbf{P},\mathbf{E}) + \lambda_{\min}(-\mathbf{D},\mathbf{E}) \geqslant \lambda_{\min}(\mathbf{P},\mathbf{E}) - \lambda_{\max}(\mathbf{D},\mathbf{E}),$$

which concludes the proof. □

Our intention is to use $\mathbf{P} = \mathbf{H}_0^\infty + \mathbf{R}$ as a surrogate and preconditioner for the Hessian of the quadratic objective function in (7), $\mathbf{H} = \mathbf{H}_0^{t_f} + \mathbf{R}$. As suggested by Lemma 1, we

must demonstrate that the eigenvalues of $\mathbf{D} = \mathbf{H}_0^\infty - \mathbf{H}_0^{t_f}$ are sufficiently small to ensure that $\mathbf{P}$ is a good approximation to $\mathbf{H}$. In the next result, we use Assumption 1 to bound the eigenvalues of $\mathbf{D}$.

Proposition 1. *The matrix* $\mathbf{D} = \mathbf{H}_0^\infty - \mathbf{H}_0^{t_f}$ *is symmetric positive semidefinite and satisfies the bound*

$$\lambda_{\max}(\mathbf{D}, \mathbf{E}) = \|\mathbf{D}\|_{\mathbf{E}^{-1},\mathbf{E}} \leqslant \frac{\|\mathbf{C}\|_{2,\mathbf{E}}^2}{2|\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})|} \exp(2\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t_f).$$

In particular, $\lambda_{\max}(\mathbf{D}, \mathbf{E})$ *decays exponentially as* t_f *increases. Furthermore, the following bound holds*

$$\begin{aligned} 0 \leqslant \lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E}) &\leqslant \lambda_{\max}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E}) \\ &\leqslant \|\mathbf{H}_0^\infty + \mathbf{R}\|_{\mathbf{E}^{-1},\mathbf{E}} + \frac{\|\mathbf{C}\|_{2,\mathbf{E}}^2}{2|\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})|} \exp(2\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A}t_f). \end{aligned}$$

Proof. First, we note that the matrix $\mathbf{D}$ can be rewritten as

$$\mathbf{D} = \int_{t_f}^\infty \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \mathbf{C} \exp(\mathbf{E}^{-1}\mathbf{A}t)\, \mathrm{d}t.$$

Clearly, $\mathbf{D}$ is symmetric since $\mathbf{H}_0^\infty$ and $\mathbf{H}_0^{t_f}$ are. To show that $\mathbf{D}$ is positive semidefinite, we set $\mathbf{M}_{t_f} := \exp(\mathbf{E}^{-1}\mathbf{A}t_f)$ and rewrite $\mathbf{D}$ as

$$\begin{aligned} \mathbf{D} &= \mathbf{M}_{t_f}^\top \int_{t_f}^\infty \exp(\mathbf{E}^{-1}\mathbf{A}(t-t_f))^\top \mathbf{C}^\top \mathbf{C} \exp(\mathbf{E}^{-1}\mathbf{A}(t-t_f))\, \mathrm{d}t \mathbf{M}_{t_f} \\ &= \mathbf{M}_{t_f}^\top \int_0^\infty \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \mathbf{C} \exp(\mathbf{E}^{-1}\mathbf{A}t)\, \mathrm{d}t \mathbf{M}_{t_f} \\ &= \mathbf{M}_{t_f}^\top \mathbf{H}_0^\infty \mathbf{M}_{t_f}. \end{aligned}$$

Since $\mathbf{M}_{t_f}$ is invertible and $\mathbf{H}_0^\infty$ is positive semidefinite, $\mathbf{D}$ is also positive semidefinite. We now show that $\|\mathbf{D}\|_{\mathbf{E}^{-1},\mathbf{E}} = \lambda_{\max}(\mathbf{D}, \mathbf{E})$. For any $\mathbf{x} \in \mathcal{X}$ with $\mathbf{x} \neq 0$, we have that

$$\frac{\mathbf{x}^\top \mathbf{D}\mathbf{E}^{-1}\mathbf{D}\mathbf{x}}{\mathbf{x}^\top \mathbf{E}\mathbf{x}} = \frac{\|\mathbf{E}^{-\frac{1}{2}}\mathbf{D}\mathbf{E}^{-\frac{1}{2}}\mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2},$$

where $\mathbf{y} = \mathbf{E}^{\frac{1}{2}}\mathbf{x}$ and $\mathbf{E}^{\frac{1}{2}}$ is the square root of the symmetric positive definite matrix $\mathbf{E}$. By taking $\mathbf{y}$ to be the eigenvector corresponding to the maximum eigenvalue of $\mathbf{E}^{-\frac{1}{2}}\mathbf{D}\mathbf{E}^{-\frac{1}{2}}$, we see that $\|\mathbf{D}\|_{\mathbf{E}^{-1},\mathbf{E}} = \lambda_{\max}(\mathbf{E}^{-\frac{1}{2}}\mathbf{D}\mathbf{E}^{-\frac{1}{2}})$. Moreover, we have that

$$\lambda_{\max}(\mathbf{E}^{-\frac{1}{2}}\mathbf{D}\mathbf{E}^{-\frac{1}{2}}) = \frac{\mathbf{y}^\top \mathbf{E}^{-\frac{1}{2}}\mathbf{D}\mathbf{E}^{-\frac{1}{2}}\mathbf{y}}{\mathbf{y}^\top \mathbf{y}} = \frac{\mathbf{x}^\top \mathbf{D}\mathbf{x}}{\mathbf{x}^\top \mathbf{E}\mathbf{x}} = \lambda_{\max}(\mathbf{D}, \mathbf{E}),$$

where $\mathbf{y}$ is again the eigenvector associated with the maximum eigenvalue of $\mathbf{E}^{-\frac{1}{2}}\mathbf{D}\mathbf{E}^{-\frac{1}{2}}$ and $\mathbf{x} = \mathbf{E}^{-\frac{1}{2}}\mathbf{y}$. Now, using this $\mathbf{x}$ and the properties of the logarithmic norm, we arrive at the following bound

$$\begin{aligned}\lambda_{\max}(\mathbf{D},\mathbf{E}) &= \|\mathbf{x}\|_{\mathbf{E}}^{-2}\int_{t_f}^{\infty}\|\mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{x}\|_2^2\,\mathrm{d}t\\ &\leqslant \|\mathbf{C}\|_{2,\mathbf{E}}^2\int_{t_f}^{\infty}\exp(2\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t)\,\mathrm{d}t.\end{aligned}$$

Evaluating the integral on the right-hand side gives the desired bound. The final bound is a consequence of the previous bound and the triangle inequality,

$$\begin{aligned}\left\|\mathbf{H}_0^{t_f}+\mathbf{R}\right\|_{\mathbf{E}^{-1},\mathbf{E}} &\leqslant \|\mathbf{D}\|_{\mathbf{E}^{-1},\mathbf{E}} + \|\mathbf{H}_0^{\infty}+\mathbf{R}\|_{\mathbf{E}^{-1},\mathbf{E}}\\ &= \|\mathbf{H}_0^{\infty}+\mathbf{R}\|_{\mathbf{E}^{-1},\mathbf{E}} + \lambda_{\max}(\mathbf{D},\mathbf{E}),\end{aligned}$$

which yields the desired result. □

Using Proposition 1, the following result is a direct corollary of Lemma 1 and demonstrates that the observability Gramian $\mathbf{H}_0^{\infty}$ is an increasingly accurate approximation of the Hessian matrix $\mathbf{H}_0^{t_f}$ as t_f increases.

Corollary 1. *Suppose $\mathbf{H}_0^{\infty}+\mathbf{R}$ is positive definite. In addition, let the assumptions of Proposition 1 hold and define*

$$\beta := 2\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A}) \qquad \text{and} \qquad c := |\beta|^{-1}\|\mathbf{C}\|_{2,\mathbf{E}}.$$

Then the eigenvalues of $(\mathbf{H}_0^{\infty}+\mathbf{R})^{-1}(\mathbf{H}_0^{t_f}+\mathbf{R})$ lie in the interval

$$\left[\frac{\lambda_{\min}(\mathbf{H}_0^{t_f}+\mathbf{R},\mathbf{E})}{\lambda_{\min}(\mathbf{H}_0^{t_f}+\mathbf{R},\mathbf{E})+c\exp(\beta t_f)},1\right]$$

and the minimum eigenvalues of $\mathbf{H}_0^{\infty}+\mathbf{R}$ and $\mathbf{H}_0^{t_f}+\mathbf{R}$ satisfy

$$0 \leqslant \lambda_{\min}(\mathbf{H}_0^{\infty}+\mathbf{R},\mathbf{E}) - \lambda_{\min}(\mathbf{H}_0^{t_f}+\mathbf{R},\mathbf{E}) \leqslant c\exp(\beta t_f). \tag{13}$$

As a consequence of Corollary 1, the spectral diameter of the preconditioned Hessian is bounded above by

$$\begin{aligned}\lambda_{\max}((\mathbf{H}_0^{\infty}+\mathbf{R})^{-1}(\mathbf{H}_0^{t_f}+\mathbf{R})) - \lambda_{\min}((\mathbf{H}_0^{\infty}+\mathbf{R})^{-1}(\mathbf{H}_0^{t_f}+\mathbf{R}))&\\ \leqslant \frac{c\exp(\beta t_f)}{\lambda_{\min}(\mathbf{H}_0^{t_f}+\mathbf{R},\mathbf{E})+c\exp(\beta t_f)}&\\ \leqslant \frac{c\exp(\beta t_f)}{\lambda_{\min}(\mathbf{H}_0^{\infty}+\mathbf{R},\mathbf{E})}&\end{aligned}$$

and therefore approaches zero as t_f increases. Moreover, let $0 < \varepsilon < \lambda_{\min}(\mathbf{H}_0^\infty + \mathbf{R}, \mathbf{E})$ be arbitrary. Then for all $t_f \geqslant \beta^{-1}\log(c^{-1}\varepsilon)$, we have that $c\exp(\beta t_f) \leqslant \varepsilon$ and

$$0 < \lambda_{\min}(\mathbf{H}_0^\infty + \mathbf{R}, \mathbf{E}) - \varepsilon \leqslant \lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})$$

by (13). Consequently, the condition number of $(\mathbf{H}_0^\infty + \mathbf{R})^{-1}(\mathbf{H}_0^{t_f} + \mathbf{R})$ is bounded above by

$$\begin{aligned}\frac{\lambda_{\max}((\mathbf{H}_0^\infty + \mathbf{R})^{-1}(\mathbf{H}_0^{t_f} + \mathbf{R}))}{\lambda_{\min}((\mathbf{H}_0^\infty + \mathbf{R})^{-1}(\mathbf{H}_0^{t_f} + \mathbf{R}))} &\leqslant \frac{\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E}) + c\exp(\beta t_f)}{\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})} \\ &\leqslant \frac{\lambda_{\min}(\mathbf{H}_0^\infty + \mathbf{R}, \mathbf{E}) + c\exp(\beta t_f)}{\lambda_{\min}(\mathbf{H}_0^\infty + \mathbf{R}, \mathbf{E}) - \varepsilon}\end{aligned}$$

and approaches one as t_f increases since ε is arbitrary.

Remark 1 *(Positive definite $\mathbf{H}_0^\infty + \mathbf{R}$).* Since $\mathbf{H}_0^\infty$ and $\mathbf{R}$ are positive semidefinite, the null space of $\mathbf{H}_0^\infty + \mathbf{R}$ satisfies

$$\ker(\mathbf{H}_0^\infty + \mathbf{R}) = \ker(\mathbf{H}_0^\infty) \cap \ker(\mathbf{R}).$$

Consequently, as long as $\ker(\mathbf{H}_0^\infty) \cap \ker(\mathbf{R}) = \varnothing$, the matrix $\mathbf{H}_0^\infty + \mathbf{R}$ is positive definite as required by Proposition 1. This is the case if either $\mathbf{H}_0^\infty$ or $\mathbf{R}$ are nonsingular. Recall that $\mathbf{H}_0^\infty$ is nonsingular if and only if the system (5) is observable, which by Hautus' lemma, is equivalent to the condition: $\mathbf{C}\mathbf{v} \neq \mathbf{0}$ for all eigenvectors $\mathbf{v}$ of $\mathbf{E}^{-1}\mathbf{A}$ [29, Th. 4.26]. As Hautus' lemma suggests, the system (5) may not be observable if $\mathbf{C}$ has a large null space. When (5) is unobservable, we rely on the regularization matrix $\mathbf{R}$ to ensure that $\mathbf{H}_0^\infty + \mathbf{R}$ is positive definite.

Remark 2 *(Stability).* Assumption 1 is too strong for many applications. In particular, it can happen that the spectral abscissa of $\mathbf{E}^{-1}\mathbf{A}$ is negative (i.e., the dynamical system (1b) is stable), yet the logarithmic norm $\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})$ is positive. If this is the case, then the bounds in Proposition 1 and Corollary 1 do not decay as t_f increases. However, if the spectral abscissa of $\mathbf{E}^{-1}\mathbf{A}$ is negative, then we have that

$$\lim_{t\to+\infty} \left\|\exp(\mathbf{E}^{-1}\mathbf{A}t)\right\|_{\mathbf{E}} = 0.$$

Unfortunately, this property does not provide a convergence rate as in the bounds of Proposition 1 and Corollary 1. To obtain a convergence rate, we can use the results in [30], which guarantee the existence of a symmetric positive definite matrix $\mathbf{M}$ for which $\mu_{\mathbf{M}}(\mathbf{E}^{-1}\mathbf{A})$ is negative. In fact, $\mathbf{M}$ solves the Lyapunov equation

$$(\mathbf{E}^{-1}\mathbf{A})^\top\mathbf{M} + \mathbf{M}(\mathbf{E}^{-1}\mathbf{A}) + \mathbf{I} = 0$$

or equivalently, $\mathbf{M} = \mathbf{E}\overline{\mathbf{M}}\mathbf{E}$, where $\overline{\mathbf{M}}$ solves the generalized Lyapunov equation

$$\mathbf{A}^\top\overline{\mathbf{M}}\mathbf{E} + \mathbf{E}\overline{\mathbf{M}}\mathbf{A} + \mathbf{I} = 0.$$

Moreover, since all norms are equivalent in finite dimensions, there exists a positive constant $C > 0$ such that

$$\|\exp(\mathbf{E}^{-1}\mathbf{A}t)\|_{\mathbf{E}} \leqslant C\|\exp(\mathbf{E}^{-1}\mathbf{A}t)\|_{\mathbf{M}} \leqslant C\exp(\mu_{\mathbf{M}}(\mathbf{E}^{-1}\mathbf{A})t). \tag{14}$$

Using these facts, we can replace $\exp(\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t)$ in Proposition 1 and Corollary 1 with the upper bound in (14), producing a bound that decays as t_f increases. Given the finite-dimensional nature of this argument, when $\mathbf{E}$ and $\mathbf{A}$ arise from the spatial discretization of a system of PDEs, the constant C will generally depend on the problem size m, leading to a bound that changes as the discretization is refined. We provide a numerical example in Section 7.2 that confirms this discussion, but also demonstrates that our results and methods work even when Assumption 1 fails to hold.

5. Galerkin approximation

As the results in Section 4 suggest, $\mathbf{H}_0^\infty + \mathbf{R}$ is a good preconditioner for the finite-time horizon Hessian $\mathbf{H}_0^{t_f} + \mathbf{R}$. Unfortunately, the Lyapunov equation that defines the observability Gramian $\mathbf{H}_0^\infty$ can be difficult to solve numerically when the problem size m is large. In this case, we can reduce the computational cost by approximating the Gramian $\mathbf{H}_0^\infty$ using a Galerkin projection. Let the columns of $\mathbf{Y} \in \mathbb{R}^{m\times d}$ be orthonormal with respect to the $\mathbf{E}$-inner product, i.e., $\mathbf{Y}^\top\mathbf{E}\mathbf{Y} = \mathbf{I}$. We treat $\mathbf{Y}$ as a linear operator from $\mathbb{R}^d$, endowed with the Euclidean inner product, into $\mathcal{X}$. We denote by $\mathbf{Y}_0$ a matrix whose columns span the orthogonal complement of $\mathbf{Y}$, so that the columns of the square matrix $\bar{\mathbf{Y}} := [\mathbf{Y}|\mathbf{Y}_0]$ are $\mathbf{E}$-orthonormal, i.e., $\mathbf{Y}^\top\mathbf{E}\mathbf{Y}_0 = 0$ and $\mathbf{Y}_0^\top\mathbf{E}\mathbf{Y}_0 = \mathbf{I}$. Throughout this section, we assume that the set $\widehat{\mathcal{F}} := \{\mathbf{p} \in \mathbb{R}^d \,|\, \mathbf{Y}\mathbf{p} \in \mathcal{F}\}$ is nonempty.

Before introducing the Galerkin approximation, we make the following observations. First, we have that $\|\mathbf{Y}\|_{\mathbf{E},2} = 1$ since $\mathbf{Y}$ is $\mathbf{E}$-orthonormal. Second, we have that $\|\mathbf{Y}^\top\|_{2,\mathbf{E}^{-1}} = \|\mathbf{Y}^\top\mathbf{E}\|_{2,\mathbf{E}} = 1$. To see this, take any $\mathbf{x} \in \mathcal{X}$ with $\mathbf{x} \neq 0$. We can decompose $\mathbf{x}$ as $\mathbf{x} = \mathbf{Y}\mathbf{v} + \mathbf{Y}_0\mathbf{v}_0$ since $[\mathbf{Y}|\mathbf{Y}_0]$ forms a basis of $\mathcal{X}$. From this, we see that

$$\|\mathbf{Y}^\top\mathbf{E}\mathbf{x}\|_2^2 = \|\mathbf{v}\|_2^2 \qquad \text{and} \qquad \|\mathbf{x}\|_{\mathbf{E}}^2 = \|\mathbf{v}\|_2^2 + \|\mathbf{v}_0\|_2^2.$$

Consequently, we have that

$$\frac{\|\mathbf{Y}^\top\mathbf{E}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_{\mathbf{E}}^2} = \frac{\|\mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2 + \|\mathbf{v}_0\|_2^2} \leqslant 1 \qquad \Longrightarrow \qquad \|\mathbf{Y}^\top\mathbf{E}\|_{2,\mathbf{E}} \leqslant 1,$$

and the upper bound is attained if $\mathbf{v}_0 = 0$. Combining these facts with the definition of the norm $\|\mathbf{Y}^\top\mathbf{E}\|_{2,\mathbf{E}}$ proves that the norm is equal to one.

The Galerkin approximation of $\mathbf{H}_0^\infty$ is given by $\mathbf{E}\mathbf{Y}\mathbf{G}_0^\infty\mathbf{Y}^\top\mathbf{E}$, where $\mathbf{G}_0^\infty \in \mathbb{R}^{d\times d}$ solves the reduced Lyapunov equation

$$(\mathbf{Y}^\top\mathbf{A}\mathbf{Y})^\top\mathbf{G}_0^\infty + \mathbf{G}_0^\infty(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}) + \mathbf{Y}^\top\mathbf{C}^\top\mathbf{C}\mathbf{Y} = \mathbf{0}. \tag{15}$$

For notational convenience, we denote $\widehat{\mathbf{A}} := \mathbf{Y}^\top\mathbf{A}\mathbf{Y}$, $\widehat{\mathbf{C}} := \mathbf{C}\mathbf{Y}$, and $\widehat{\mathbf{R}} := \mathbf{Y}^\top\mathbf{R}\mathbf{Y}$. We define $\mathbf{G}_0^{t_f}$ analogously to $\mathbf{H}_0^{t_f}$, i.e.,

$$\mathbf{G}_0^{t_f} := \int_0^{t_f} \exp(\widehat{\mathbf{A}}t)^\top\widehat{\mathbf{C}}^\top\widehat{\mathbf{C}}\exp(\widehat{\mathbf{A}}t)\,\mathrm{d}t \tag{16}$$

One overwhelming benefit of the Galerkin approximation is that if $d \ll m$, then the reduced Lyapunov equation (15) can be solved efficiently using direct or iterative methods [31]. For example, one possible direct method to compute $\mathbf{G}_0^\infty$, when d is small, is to solve the linear system of equations

$$(\mathbf{I}\otimes\widehat{\mathbf{A}}^\top + \widehat{\mathbf{A}}\otimes\mathbf{I})\mathrm{vec}(\mathbf{G}_0^\infty) = -\mathrm{vec}(\widehat{\mathbf{C}}^\top\widehat{\mathbf{C}})$$

using a matrix factorization. Here, $\mathrm{vec}(\mathbf{M})$ is the vector of stacked columns of $\mathbf{M}$ and $\otimes$ denotes the Kronecker product. For additional information on Galerkin methods for solving large Lyapunov equations, see [31–33].

To connect the Galerkin approximation $\mathbf{G}_0^\infty$ to (7), we note that $\mathbf{G}_0^{t_f} + \widehat{\mathbf{R}}$ is the Hessian to the Galerkin projected optimization problem

$$\min_{\widehat{\mathbf{x}}_0\in\mathbb{R}^d} \frac{1}{2}\int_0^{t_f} \|\widehat{\mathbf{C}}\exp(\widehat{\mathbf{A}}t)\widehat{\mathbf{x}}_0 - \bar{\mathbf{y}}(t)\|_2^2\,\mathrm{d}t + \frac{1}{2}\widehat{\mathbf{x}}_0^\top\widehat{\mathbf{R}}\widehat{\mathbf{x}}_0 + \Phi(\mathbf{Y}\widehat{\mathbf{x}}_0). \tag{17}$$

Optimization problem (17) arises by first requiring that solutions to the differential equation (1b) be in the subspace spanned by the columns of $\mathbf{Y}$. Since the solution to (1b) is generally not a linear combination of the columns of $\mathbf{Y}$, we instead require that the residual of (1b) be orthogonal to the columns of $\mathbf{Y}$. At the expense of this approximation, we reduce the dimensionality of (1) from m to d. Galerkin projections are commonly used in model reduction such as proper orthogonal decomposition and reduced basis methods (cf. [34] and the references therein). In contrast, by replacing the initial condition $\mathbf{x}_0$ in (1b) with $\mathbf{Y}\overline{\mathbf{x}}_0$ for some $\overline{\mathbf{x}}_0 \in \mathbb{R}^d$, we see that (17) is closely related to the transformed optimization problem

$$\min_{\overline{\mathbf{x}}_0\in\mathbb{R}^d} \frac{1}{2}\int_0^{t_f} \|\mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{Y}\overline{\mathbf{x}}_0 - \bar{\mathbf{y}}(t)\|_2^2\,\mathrm{d}t + \frac{1}{2}\overline{\mathbf{x}}_0^\top\widehat{\mathbf{R}}\overline{\mathbf{x}}_0 + \Phi(\mathbf{Y}\overline{\mathbf{x}}_0). \tag{18}$$

In particular, the error between the optimal solution $\widehat{\mathbf{x}}_0$ for (17) and the optimal solution $\overline{\mathbf{x}}_0$ for (18) is controlled by the difference of the matrix functions

$$\mathbf{h}(t) := \mathbf{C}\exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{Y} \qquad \text{and} \qquad \mathbf{g}(t) := \widehat{\mathbf{C}}\exp(\widehat{\mathbf{A}}t).$$

Additionally, the optimal solutions $\mathbf{Y}\hat{\mathbf{x}}_0$ and $\mathbf{Y}\overline{\mathbf{x}}_0$ are generally approximations to the optimal solution, $\mathbf{x}_0$, to (7), unless $\mathbf{x}_0 \in \text{Range}(\mathbf{Y})$. To quantify the error between $\mathbf{Y}\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$, we note that

$$\begin{aligned}\|\mathbf{Y}\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_{\mathbf{E}} &\leqslant \|\mathbf{Y}(\hat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0)\|_{\mathbf{E}} + \|\mathbf{Y}\overline{\mathbf{x}}_0 - \mathbf{x}_0\|_{\mathbf{E}} \\ &= \|\hat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0\|_2 + \|\mathbf{Y}\overline{\mathbf{x}}_0 - \mathbf{x}_0\|_{\mathbf{E}}.\end{aligned}$$

Consequently, it is important to understand the error between $\hat{\mathbf{x}}_0$ and $\overline{\mathbf{x}}_0$ as well as the error between $\mathbf{Y}\overline{\mathbf{x}}_0$ and $\mathbf{x}_0$.

The error between $\mathbf{Y}\overline{\mathbf{x}}_0$ and $\mathbf{x}_0$ is controlled by the approximation quality of the subspace spanned by the columns of $\mathbf{Y}$. The following proposition provides a bound for the resulting recovery error. For this result, we require that $\mathbf{H}_0^{t_f} + \mathbf{R}$ is positive definite, which in turn implies that $\mathbf{H}_0^{\infty} + \mathbf{R}$ is positive definite since

$$\mathbf{H}_0^{\infty} + \mathbf{R} = (\mathbf{H}_0^{t_f} + \mathbf{R}) + (\mathbf{H}_0^{\infty} - \mathbf{H}_0^{t_f}),$$

where the first matrix in parentheses on the right-hand side is positive definite and the second is positive semidefinite by Proposition 1.

Proposition 2. *Suppose that* $\mathbf{H}_0^{t_f} + \mathbf{R}$ *is positive definite,* $\mathbf{x}_0 \in \mathbb{R}^m$ *solves* (7) *and* $\overline{\mathbf{x}}_0 \in \mathbb{R}^d$ *solves* (18)*. Then*

$$\|\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0\|_{\mathbf{E}}^2 \leqslant \frac{K}{\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})} \min_{\mathbf{p} \in \hat{\mathcal{F}}} \|\mathbf{Y}\mathbf{p} - \mathbf{x}_0\|_{\mathbf{E}}, \tag{19}$$

where the constant $K > 0$ *is given by*

$$K := L + \left\| (\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{Y}\overline{\mathbf{x}}_0 - \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^{\top}\mathbf{C}^{\top}\bar{\mathbf{y}}(t)\,\mathrm{d}t \right\|_{\mathbf{E}^{-1}}.$$

In particular, if $\mathbf{x}_0 \in \text{Range}(\mathbf{Y})$*, then* $\|\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0\|_{\mathbf{E}} = 0$*. Moreover, if* $\mathbf{x}_0$ *is in the interior of* $\mathcal{F}$*, then*

$$K \leqslant 2L + \left\| \mathbf{H}_0^{t_f} + \mathbf{R} \right\|_{\mathbf{E}^{-1},\mathbf{E}} \|\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0\|_{\mathbf{E}}.$$

If, in addition, $\Phi_0 \equiv 0$*, then*

$$\|\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0\|_{\mathbf{E}} \leqslant \frac{\left\| \mathbf{H}_0^{t_f} + \mathbf{R} \right\|_{\mathbf{E}^{-1},\mathbf{E}}}{\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})} \min_{\mathbf{p} \in \hat{\mathcal{F}}} \|\mathbf{x}_0 - \mathbf{Y}\mathbf{p}\|_{\mathbf{E}}.$$

Proof. The optimality conditions for (18) are

$$\begin{aligned}\left[\mathbf{Y}^{\top}\int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^{\top}\mathbf{C}^{\top}\bar{\mathbf{y}}(t)\,\mathrm{d}t - \left(\mathbf{Y}^{\top}\mathbf{H}_0^{t_f}\mathbf{Y} + \mathbf{Y}^{\top}\mathbf{R}\mathbf{Y}\right)\overline{\mathbf{x}}_0\right]^{\top}(\mathbf{p} - \overline{\mathbf{x}}_0) \\ \leqslant \Phi(\mathbf{Y}\mathbf{p}) - \Phi(\mathbf{Y}\overline{\mathbf{x}}_0)\end{aligned} \tag{20}$$

for all $\mathbf{p} \in \mathbb{R}^d$. Adding (20) to the optimality conditions (8) for (7) with $\mathbf{x}_0' = \mathbf{Y}\overline{\mathbf{x}}_0$ results in the bound

$$(\mathbf{Y}\mathbf{p} - \mathbf{x}_0)^\top \left(\int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \bar{\mathbf{y}}(t)\,\mathrm{d}t - (\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{Y}\overline{\mathbf{x}}_0 \right)$$
$$+ (\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0)^\top (\mathbf{H}_0^{t_f} + \mathbf{R})(\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0) \leqslant \Phi(\mathbf{Y}\mathbf{p}) - \Phi(\mathbf{x}_0).$$

Rearranging terms, applying the bound

$$\lambda_{\min}(\mathbf{H}_0^{t_f} + \mathbf{R}, \mathbf{E})\, \|\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0\|_{\mathbf{E}}^2 \leqslant (\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0)^\top (\mathbf{H}_0^{t_f} + \mathbf{R})(\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0)$$

on the left-hand side, applying the Cauchy-Schwarz inequality on the right, employing the Lipschitz continuity of Φ_0, and passing to the infimum over $\widehat{\mathcal{F}}$ yields (19).

Now, suppose that $\mathbf{x}_0$ is in the interior of $\mathcal{F}$. To obtain the bound on K, we add and subtract $(\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{x}_0$ in the norm and apply the triangle inequality to obtain

$$\left\| (\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{Y}\overline{\mathbf{x}}_0 - \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \bar{\mathbf{y}}(t)\,\mathrm{d}t \right\|_{\mathbf{E}^{-1}}$$
$$\leqslant \left\| (\mathbf{H}_0^{t_f} + \mathbf{R})(\mathbf{Y}\overline{\mathbf{x}}_0 - \mathbf{x}_0) \right\|_{\mathbf{E}^{-1}}$$
$$+ \left\| \int_0^{t_f} \exp(\mathbf{E}^{-1}\mathbf{A}t)^\top \mathbf{C}^\top \bar{\mathbf{y}}(t)\,\mathrm{d}t - (\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{x}_0 \right\|_{\mathbf{E}^{-1}}.$$

The expression in the second term on the right-hand side of the previous inequality is a subgradient of Φ_0 at $\mathbf{x}_0$ by the optimality of $\mathbf{x}_0$ and the assumption that $\mathbf{x}_0$ is an interior point. Consequently, this term is bounded above by the Lipschitz constant L. This fact, combined with the submultiplicativity of the matrix norm, proves the desired bound. Finally, suppose $\Phi_0 \equiv 0$. Then, $L = 0$ and dividing both sides of (19), using the upper bound on K, by $\|\mathbf{x}_0 - \mathbf{Y}\overline{\mathbf{x}}_0\|_{\mathbf{E}}$ results in the desired bound. □

Remark 3 *(Nonsmoothness and the projected problem).* As Proposition 2 demonstrates, the presence of nonsmoothness, via a nonsmooth regularizer Φ_0 and active constraints $\mathbf{x}_0 \in \mathcal{F}$, results in a larger error between $\mathbf{x}_0$ and $\mathbf{Y}\overline{\mathbf{x}}_0$. In this case, the error between $\mathbf{x}_0$ and $\mathbf{Y}\overline{\mathbf{x}}_0$ is on the order of the square root of the best approximation error in the subspace spanned by the columns on $\mathbf{Y}$. In contrast, the error is proportional to the best approximation error when the problem is smooth and the constraints are not active at $\mathbf{x}_0$.

The Hessian of the transformed problem (18), omitting $\widehat{\mathbf{R}}$, is given by

$$\widehat{\mathbf{H}}_0^{t_f} := \mathbf{Y}^\top \mathbf{H}_0^{t_f} \mathbf{Y},$$

which can be expressed in terms of $\mathbf{H}_0^{\infty}$ as

$$\widehat{\mathbf{H}}_0^{t_f} = \mathbf{Y}^\top \mathbf{H}_0^\infty \mathbf{Y} - (\exp(\mathbf{E}^{-1}\mathbf{A}t_f)\mathbf{Y})^\top \mathbf{H}_0^\infty (\exp(\mathbf{E}^{-1}\mathbf{A}t_f)\mathbf{Y}).$$

Consequently, the evaluation of $\widehat{\mathbf{H}}_0^{t_f}$ in this way requires the solution of an m-dimensional Lyapunov equation to compute $\mathbf{H}_0^\infty$ and d applications of the matrix exponential $\exp(\mathbf{E}^{-1}\mathbf{A}t_f)$. The Galerkin Gramian $\mathbf{G}_0^\infty$ is an approximation of $\widehat{\mathbf{H}}_0^\infty$ and plays an important role in bounding the error between $\widehat{\mathbf{x}}_0$ and $\overline{\mathbf{x}}_0$. We provide this error bound in the following proposition.

Proposition 3. *Suppose that* $\mathbf{Y}^\top(\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{Y}$ *is positive definite,* $\widehat{\mathbf{x}}_0 \in \mathbb{R}^d$ *solves* (17) *and* $\overline{\mathbf{x}}_0 \in \mathbb{R}^d$ *solves* (18)*. Then*

$$\begin{aligned} &\lambda_{\min}(\mathbf{Y}^\top(\mathbf{H}_0^{t_f} + \mathbf{R})\mathbf{Y}) \, \|\widehat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0\|_2 \\ &\leqslant \|\mathbf{G}_0^{t_f} - \widehat{\mathbf{H}}_0^{t_f}\|_2 \, \|\widehat{\mathbf{x}}_0\|_2 + \int_0^{t_f} \|\mathbf{h}(t) - \mathbf{g}(t)\|_2 \, \|\bar{\mathbf{y}}(t)\|_2 \, \mathrm{d}t, \end{aligned} \tag{21}$$

where $\|\mathbf{G}_0^{t_f} - \widehat{\mathbf{H}}_0^{t_f}\|_2$ *is bounded by*

$$\|\mathbf{G}_0^{t_f} - \widehat{\mathbf{H}}_0^{t_f}\|_2 \leqslant 2\,\|\mathbf{C}\|_{2,\mathbf{E}} \int_0^{t_f} \exp(\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t)\|\mathbf{h}(t) - \mathbf{g}(t)\|_2 \, \mathrm{d}t. \tag{22}$$

Proof. Let $\widehat{J}(\mathbf{x}) + \frac{1}{2}\mathbf{x}^\top\widehat{\mathbf{R}}\mathbf{x} + \Phi(\mathbf{Y}\mathbf{x})$ denote the objective function from (17) and let $\overline{J}(\mathbf{x}) + \frac{1}{2}\mathbf{x}^\top\widehat{\mathbf{R}}\mathbf{x} + \Phi(\mathbf{Y}\mathbf{x})$ be the objective function from (18). Both $\widehat{J}$ and $\overline{J}$ are continuously differentiable and convex. Moreover, $\widehat{\mathbf{x}}_0$ and $\overline{\mathbf{x}}_0$ satisfy the optimality conditions

$$(\widehat{\mathbf{x}}_0 - \mathbf{p})^\top(\nabla\widehat{J}(\widehat{\mathbf{x}}_0) + \widehat{\mathbf{R}}\widehat{\mathbf{x}}_0) \leqslant \Phi(\mathbf{Y}\mathbf{p}) - \Phi(\mathbf{Y}\widehat{\mathbf{x}}_0)$$

and

$$(\overline{\mathbf{x}}_0 - \mathbf{p})^\top(\nabla\overline{J}(\overline{\mathbf{x}}_0) + \widehat{\mathbf{R}}\overline{\mathbf{x}}_0) \leqslant \Phi(\mathbf{Y}\mathbf{p}) - \Phi(\mathbf{Y}\overline{\mathbf{x}}_0)$$

for all $\mathbf{p} \in \mathbb{R}^d$. Setting $\mathbf{p} = \overline{\mathbf{x}}_0$ in the first variational inequality and $\mathbf{p} = \widehat{\mathbf{x}}_0$ in the second yields

$$(\widehat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0)^\top\widehat{\mathbf{R}}(\widehat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0) + (\widehat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0)^\top(\nabla\widehat{J}(\widehat{\mathbf{x}}_0) - \nabla\overline{J}(\overline{\mathbf{x}}_0)) \leqslant 0. \tag{23}$$

The gradients of $\widehat{J}$ and $\overline{J}$ are given by

$$\begin{aligned} \nabla\widehat{J}(\widehat{\mathbf{x}}_0) &= \mathbf{G}_0^{t_f}\widehat{\mathbf{x}}_0 - \int_0^{t_f} \exp(\widehat{\mathbf{A}}t)^\top\mathbf{Y}^\top\mathbf{C}^\top\bar{\mathbf{y}}(t)\,\mathrm{d}t \\ \nabla\overline{J}(\overline{\mathbf{x}}_0) &= \widehat{\mathbf{H}}_0^{t_f}\overline{\mathbf{x}}_0 - \int_0^{t_f} \mathbf{Y}^\top\exp(\widehat{\mathbf{E}}^{-1}\mathbf{A}t)^\top\mathbf{C}^\top\bar{\mathbf{y}}(t)\,\mathrm{d}t, \end{aligned}$$

respectively. Substituting these expressions into (23) and adding and subtracting $\widehat{\mathbf{H}}_0^{t_f}\widehat{\mathbf{x}}_0$ yields

$$(\hat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0)^\top (\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}})(\hat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0)$$
$$\leqslant (\hat{\mathbf{x}}_0 - \overline{\mathbf{x}}_0)^\top \left((\widehat{\mathbf{H}}_0^{t_f} - \mathbf{G}_0^{t_f})\hat{\mathbf{x}}_0 + \int_0^{t_f} (\mathbf{h}(t) - \mathbf{g}(t))^\top \bar{\mathbf{y}}(t)\,\mathrm{d}t \right)$$

Consequently, the Cauchy-Schwarz and triangle inequalities produce (21). We now bound the error $\|\mathbf{G}_0^{t_f} - \widehat{\mathbf{H}}_0^{t_f}\|_2$. Using the definition of $\mathbf{h}$ and $\mathbf{g}$, we can rewrite the projected and Galerkin Gramians as

$$\widehat{\mathbf{H}}_0^{t_f} = \int_0^{t_f} \mathbf{h}(t)^\top \mathbf{h}(t)\,\mathrm{d}t \quad \text{and} \quad \mathbf{G}_0^{t_f} = \int_0^{t_f} \mathbf{g}(t)^\top \mathbf{g}(t)\,\mathrm{d}t,$$

respectively. The submultiplicativity of the matrix 2-norm produces the upper bound

$$\|\widehat{\mathbf{H}}_0^{t_f} - \mathbf{G}_0^{t_f}\|_2 = \int_0^{t_f} (\|\mathbf{h}(t)^\top (\mathbf{h}(t) - \mathbf{g}(t)) + \mathbf{g}(t)^\top (\mathbf{h}(t) - \mathbf{g}(t))\|_2\,\mathrm{d}t$$
$$\leqslant \int_0^{t_f} (\|\mathbf{h}(t)\|_2 + \|\mathbf{g}(t)\|_2)\|\mathbf{h}(t) - \mathbf{g}(t)\|_2\,\mathrm{d}t$$

and by the arguments in the proof of Proposition 1, we can bound $\|\mathbf{h}(t)\|_2$ as

$$\|\mathbf{h}(t)\|_2 \leqslant \|\mathbf{C}\|_{2,\mathbf{E}} \left\|\exp(\mathbf{E}^{-1}\mathbf{A}t)\right\|_{\mathbf{E}} \|\mathbf{Y}\|_{\mathbf{E},2} \leqslant \|\mathbf{C}\|_{2,\mathbf{E}} \exp(\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t).$$

Similarly, we can bound $\|\mathbf{g}(t)\|_2$ as

$$\|\mathbf{g}(t)\|_2 \leqslant \|\mathbf{C}\|_{2,\mathbf{E}} \|\exp(\widehat{\mathbf{A}}t)\|_2 \leqslant \|\mathbf{C}\|_{2,\mathbf{E}} \exp(\mu_2(\widehat{\mathbf{A}})t).$$

To estimate the logarithmic norm $\mu_2(\widehat{\mathbf{A}})$, we note that

$$\mu_2(\widehat{\mathbf{A}}) = \lim_{h \downarrow 0} \frac{\|\mathbf{I} + h\widehat{\mathbf{A}}\|_2 - 1}{h} = \lim_{h \downarrow 0} \frac{\|\mathbf{Y}^\top(\mathbf{E} + h\mathbf{A})\mathbf{Y}\|_2 - 1}{h}$$
$$\leqslant \lim_{h \downarrow 0} \frac{\|\mathbf{Y}\|_{\mathbf{E},2}^2 \|\mathbf{E} + h\mathbf{A}\|_{\mathbf{E}^{-1},\mathbf{E}} - 1}{h} = \lim_{h \downarrow 0} \frac{\left\|\mathbf{I} + h\mathbf{E}^{-1}\mathbf{A}\right\|_{\mathbf{E}} - 1}{h}$$
$$= \mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})$$

and consequently, we have that

$$\|\mathbf{g}(t)\|_2 \leqslant \|\mathbf{C}\|_{2,\mathbf{E}} \exp(\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})t).$$

Combining these results, produces the desired error bound (22). □

The bound in (22) demonstrates that $\mathbf{G}_0^{t_f}$ is typically a good preconditioner for $\widehat{\mathbf{H}}_0^{t_f}$ and can be used for sensor selection. Owing to (16), we can write $\mathbf{G}_0^{t_f}$ in terms of $\mathbf{G}_0^{\infty}$ as

$$\mathbf{G}_0^{t_f} = \mathbf{G}_0^{\infty} - \exp(\widehat{\mathbf{A}} t_f)^{\top} \mathbf{G}_0^{\infty} \exp(\widehat{\mathbf{A}} t_f)$$

or by solving the reduced Lyapunov equation

$$\widehat{\mathbf{A}}^{\top} \mathbf{G}_0^{t_f} + \mathbf{G}_0^{t_f} \widehat{\mathbf{A}} + \widehat{\mathbf{C}}^{\top} \widehat{\mathbf{C}} - \exp(\widehat{\mathbf{A}} t_f)^{\top} \widehat{\mathbf{C}}^{\top} \widehat{\mathbf{C}} \exp(\widehat{\mathbf{A}} t_f) = \mathbf{0}.$$

Consequently, the computation of $\mathbf{G}_0^{t_f}$ requires roughly $\min\{d, S\}$ applications of the matrix exponential $\exp(\widehat{\mathbf{A}} t_f)^{\top}$ (i.e., either to the left and right of $\mathbf{G}_0^{\infty}$ or to $\mathbf{Y}^{\top}\mathbf{c}_s$ for $s = 1, \ldots, S$). When it is not possible to apply $\mathbf{G}_0^{t_f}$, one can instead employ $\mathbf{G}_0^{\infty}$ to select sensors and to precondition $\widehat{\mathbf{H}}_0^{t_f}$. Similar to the discussion in Section 4, we can use $\mathbf{G}_0^{\infty}$ as a preconditioner for $\widehat{\mathbf{H}}_0^{t_f}$. In order to gauge the quality of this preconditioner, we must bound the eigenvalues of the matrix $\mathbf{D} = \mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{t_f}$.

Proposition 4. *Let the assumptions of Proposition 1 hold. Then*

$$\|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{t_f}\|_2 \leqslant \|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}\|_2 + c \exp(\beta t_f),$$

where $\|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}\|_2$ *is bounded above by* (22) *with* t_f *replaced by* $+\infty$.

Proof. This result follows directly from the triangle inequality and Proposition 1. In particular, we note that

$$\begin{aligned} \mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{t_f} &= (\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}) + (\widehat{\mathbf{H}}_0^{\infty} - \widehat{\mathbf{H}}_0^{t_f}) \\ &= (\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}) + \mathbf{Y}^{\top}(\mathbf{H}_0^{\infty} - \mathbf{H}_0^{t_f})\mathbf{Y}. \end{aligned}$$

Using the triangle inequality, we obtain

$$\|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{t_f}\|_2 \leqslant \|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}\|_2 + \|\mathbf{Y}\|_{\mathbf{E},2}^2 \|\mathbf{H}_0^{\infty} - \mathbf{H}_0^{t_f}\|_{\mathbf{E}^{-1},\mathbf{E}}$$

and Proposition 1 yields the desired bound. □

As a consequence of Proposition 4, if $\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{t_f}$ is symmetric positive semidefinite and $\lambda_{\min}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}}) > 0$, then the condition number of the preconditioned Hessian $(\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})^{-1}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}})$ is bounded above by

$$\frac{\lambda_{\max}((\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})^{-1}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}}))}{\lambda_{\min}((\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})^{-1}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}}))} \leqslant \frac{\lambda_{\min}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}}) + \|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}\|_2 + c \exp(\beta t_f)}{\lambda_{\min}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}})}$$

and its spectral diameter is bounded above by

$$\begin{aligned} &\lambda_{\max}((\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})^{-1}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}})) - \lambda_{\min}((\mathbf{G}_0^{\infty} + \mathbf{R})^{-1}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}})) \\ &\qquad \leqslant \frac{\|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}\|_2 + c \exp(\beta t_f)}{\lambda_{\min}(\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}}) + \|\mathbf{G}_0^{\infty} - \widehat{\mathbf{H}}_0^{\infty}\|_2 + c \exp(\beta t_f)}. \end{aligned}$$

Therefore, the quality of the preconditioner is controlled by t_f and the Galerkin approximation error, which is bounded as in (22).

Remark 4 *(Petrov-Galerkin approximation).* In principal, it is possible to employ a Petrov-Galerkin approximation to reduce the dimensionality of the observability Gramian. That is, given a d-by-m matrix $\mathbf{X}$ such that the columns of $\mathbf{X}$ and $\mathbf{Y}$ are biorthogonal with respect to the $\mathbf{E}$-inner product, i.e., $\mathbf{X}^\top\mathbf{E}\mathbf{Y} = \mathbf{I}$, we can approximate the Gramian $\mathbf{H}_0^\infty$ by $\mathbf{E}\mathbf{X}\mathbf{G}_0^\infty\mathbf{X}^\top\mathbf{E}$, where $\mathbf{G}_0^\infty$ solves the reduced Lyapunov equation

$$(\mathbf{X}^\top\mathbf{A}\mathbf{Y})^\top\mathbf{G}_0^\infty + \mathbf{G}_0^\infty(\mathbf{X}^\top\mathbf{A}\mathbf{Y}) + \mathbf{Y}^\top\mathbf{C}^\top\mathbf{C}\mathbf{Y} = \mathbf{0}.$$

Petrov-Galerkin approximations are common in control theory where one can perform a balanced reduction to reduce the dimensionality of the underlying control system [29]. Unfortunately, it is unclear how to produce a bound of the form

$$\mu_2(\mathbf{X}^\top\mathbf{A}\mathbf{Y}) \leqslant \kappa\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A}), \quad \kappa > 0,$$

as required by the proof of Proposition 3.

Remark 5 *(Polynomial approximation).* In many applications, the differential equation (1b) arises from the discretization of a system of PDEs, in which case the initial condition is a function defined on some domain Ω. For this class of problems, it is often practical to approximate the initial condition using a set of orthogonal polynomials. For example, if $\Omega = (0,1)^d$, $d = 1,2,3$, and the $\mathbf{E}$-inner product is a discretization of the $L^2(\Omega)$ inner product, then we can define the columns of $\mathbf{Y}$ as the values of the Legendre polynomials at predefined mesh vertices (reorthogonalized with respect to the $\mathbf{E}$-inner product). When $d > 1$, this set of polynomials could consist of polynomials with total or maximum degree less than a prescribed order p. Since the polynomials are dense in $L^2(\Omega)$, we would expect to reduce the recovery error by increasing p.

Example 1 *(Spectral approximation for symmetric* $\mathbf{A}$*).* In the following results, we assume that $\mathbf{A}$ is symmetric. Under this assumption, we can show that the Galerkin error is zero when using eigenvectors as the columns of $\mathbf{Y}$.

Proposition 5. *Suppose* $\mathbf{A}$ *is symmetric and let the columns of* $\mathbf{Y}$ *be* d *eigenvectors of* $\mathbf{E}^{-1}\mathbf{A}$*. Then the Galerkin error is zero, i.e.,* $\|\mathbf{h}(t) - \mathbf{g}(t)\|_2 = 0$.

Proof. We first note that $\mathbf{E}^{-1}\mathbf{A}$ is the similarity transformation of a symmetric matrix. In particular, let $\mathbf{E}^{\frac{1}{2}}$ denote the square root of the symmetric positive definite matrix $\mathbf{E}$, then

$$\mathbf{E}^{-1}\mathbf{A} = \mathbf{E}^{-\frac{1}{2}}\mathbf{E}^{-\frac{1}{2}}\mathbf{A}\mathbf{E}^{-\frac{1}{2}}\mathbf{E}^{\frac{1}{2}} = \mathbf{E}^{-\frac{1}{2}}\bar{\mathbf{A}}\mathbf{E}^{\frac{1}{2}},$$

where $\bar{\mathbf{A}} = \mathbf{E}^{-\frac{1}{2}}\mathbf{A}\mathbf{E}^{-\frac{1}{2}}$. We denote the eigenvectors of $\mathbf{E}^{-1}\mathbf{A}$ as the columns of the matrix $\bar{\mathbf{Y}}$, and the eigenvectors of $\bar{\mathbf{A}}$ as the columns of $\bar{\mathbf{U}}$. Since they are related by a similarity transformation, the eigenvalues of $\mathbf{E}^{-1}\mathbf{A}$ and $\bar{\mathbf{A}}$ coincide and are negative by Assumption 1. We denote the diagonal matrix of eigenvalues by $\bar{\mathbf{L}}$ and employ the decompositions

$$\bar{\mathbf{Y}} = [\mathbf{Y}|\mathbf{Y}_0], \quad \bar{\mathbf{U}} = [\mathbf{U}|\mathbf{U}_0], \quad \bar{\mathbf{L}} = \begin{bmatrix} \mathbf{L} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_0 \end{bmatrix}.$$

We note that $\bar{\mathbf{U}} = \mathbf{E}^{\frac{1}{2}}\bar{\mathbf{Y}}$ and $\mathbf{U} = \mathbf{E}^{\frac{1}{2}}\mathbf{Y}$. By definition, we have that

$$\widehat{\mathbf{A}} = \mathbf{Y}^\top\mathbf{A}\mathbf{Y} = \mathbf{Y}^\top\mathbf{E}^{\frac{1}{2}}\bar{\mathbf{A}}\mathbf{E}^{\frac{1}{2}}\mathbf{Y} = \mathbf{U}^\top\bar{\mathbf{A}}\mathbf{U} = \mathbf{L}$$

and $\exp(\widehat{\mathbf{A}}t) = \exp(\mathbf{L}t)$. To bound the Galerkin error, we must bound the quantity $\|\mathbf{h}(t) - \mathbf{g}(t)\|_2$. To this end, we notice that

$$\begin{aligned}
\exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{Y} - \mathbf{Y}\exp(\widehat{\mathbf{A}}t) &= \exp(\mathbf{E}^{-\frac{1}{2}}\bar{\mathbf{A}}\mathbf{E}^{\frac{1}{2}}t)\mathbf{Y} - \mathbf{Y}\exp(\mathbf{L}t) \\
&= \mathbf{E}^{-\frac{1}{2}}\bar{\mathbf{U}}\exp(\bar{\mathbf{L}}t)\bar{\mathbf{U}}^\top\mathbf{E}^{\frac{1}{2}}\mathbf{Y} - \mathbf{Y}\exp(\mathbf{L}t) \\
&= \bar{\mathbf{Y}}\exp(\bar{\mathbf{L}}t)\bar{\mathbf{U}}^\top\mathbf{U} - \mathbf{Y}\exp(\mathbf{L}t) \\
&= \mathbf{Y}\exp(\mathbf{L}t) - \mathbf{Y}\exp(\mathbf{L}t) = 0.
\end{aligned}$$

Here, we have used the fact that the columns of $\bar{\mathbf{U}}$ are orthonormal. Consequently, $\mathbf{h}(t) = \mathbf{g}(t)$ for all $t \geqslant 0$ as desired. □

In the setting of Proposition 5, the state trajectory satisfies

$$\begin{aligned}
\mathbf{x}(t) &= \exp(\mathbf{E}^{-1}\mathbf{A}t)\mathbf{Y}\hat{\mathbf{x}}_0 + \exp(\mathbf{E}^{-1}\mathbf{A}t)\int_0^t \exp(-\mathbf{E}^{-1}\mathbf{A}\tau)\mathbf{E}^{-1}\mathbf{f}(\tau)\,\mathrm{d}\tau \\
&= \mathbf{Y}\exp(\mathbf{L}t)\hat{\mathbf{x}}_0 + \bar{\mathbf{Y}}\exp(\bar{\mathbf{L}}t)\int_0^t \exp(-\bar{\mathbf{L}}\tau)\bar{\mathbf{Y}}^\top\mathbf{f}(\tau)\,\mathrm{d}\tau.
\end{aligned}$$

Consequently, if $\mathbf{f}(t) = \mathbf{E}\mathbf{Y}\widehat{\mathbf{f}}(t)$ for some function $\widehat{\mathbf{f}} : (0, t_f] \to \mathbb{R}^d$, then

$$\mathbf{x}(t) = \mathbf{Y}\exp(\mathbf{L}t)\hat{\mathbf{x}}_0 + \mathbf{Y}\exp(\mathbf{L}t)\int_0^t \exp(-\mathbf{L}\tau)\widehat{\mathbf{f}}(t)(\tau)\,\mathrm{d}\tau.$$

In particular, $\mathbf{x}(t) = \mathbf{Y}\hat{\mathbf{x}}(t)$, where $\hat{\mathbf{x}}(t)$ solves the reduced system of differential equations

$$\begin{cases} \dot{\hat{\mathbf{x}}} = \mathbf{L}\hat{\mathbf{x}} + \widehat{\mathbf{f}}, & \text{in } (0, t_f] \\ \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0 \end{cases},$$

which is equal to the Galerkin approximation of the differential equation (1b). However, if $\mathbf{f}$ does not have this form, then the Galerkin projection of the dynamical system (1b) is no longer exact and is given by

$$\begin{cases} \dot{\mathbf{w}} = \mathbf{L}\mathbf{w} + \mathbf{Y}^\top \mathbf{f}, & \text{in } (0, t_f] \\ \mathbf{w}(0) = \widehat{\mathbf{x}}_0 \end{cases}. \quad \square$$

6. Greedy sensor placement

As the results in Sections 4 and 5 suggest, $\mathbf{G}_0^\infty + \widehat{\mathbf{R}}$ provides a good surrogate for $\widehat{\mathbf{H}}_0^{t_f} + \widehat{\mathbf{R}}$ when selecting sensors to improve the recovery error (12). Owing to (6) and the linearity of the Lyapunov equation, we have that for any set of sensors $\sigma \subseteq \{1, \ldots, \bar{S}\}$, the Galerkin Gramian associated with $\mathbf{C} = \mathbf{C}_\sigma$ can be written as a sum over the sensors,

$$\mathbf{G}_0^\infty = \sum_{s \in \sigma} \mathbf{Q}_s,$$

where the Gramian $\mathbf{Q}_s$ for sensor $s \in \{1, \ldots, \bar{S}\}$ solves the reduced Lyapunov equation

$$\widehat{\mathbf{A}}^\top \mathbf{Q}_s + \mathbf{Q}_s \widehat{\mathbf{A}} + \mathbf{Y}^\top \mathbf{c}_s \mathbf{c}_s^\top \mathbf{Y} = 0. \tag{24}$$

As a consequence, we formulate the objective function for our optimal sensor placement problem as

$$g(\sigma) := \log\det\left(\sum_{s \in \sigma} \mathbf{Q}_s + \widehat{\mathbf{R}}\right).$$

We employ the following greedy algorithm to determine a sensor set σ that approximately maximizes g.

Algorithm 1 Greedy sensor placement.

Require: The desired number of sensors $S_0 \in \mathbb{N}$.
1: Set $S = 0$, $\mathbf{G}_0^\infty = 0$ and $\sigma = \varnothing$.
2: Compute $\mathbf{Q}_s$ for each $s \in \{1, \ldots, \bar{S}\}$ by solving the Lyapunov equation (24).
3: **while** $S < S_0$ **do**
4: Compute $\gamma_s = \log\det(\mathbf{G}_0^\infty + \mathbf{Q}_s + \widehat{\mathbf{R}})$ for $s \in \{1, \ldots, \bar{S}\} \backslash \sigma$.
5: Choose $s^\star = \arg\max_s \gamma_s$.
6: Update $S \leftarrow S + 1$, $\sigma \leftarrow \sigma \cup \{s^\star\}$, and $\mathbf{G}_0^\infty \leftarrow \mathbf{G}_0^\infty + \mathbf{Q}_{s^\star}$.
7: **end while**

The main computational effort required by Algorithm 1 is the computation of the single measurement Gramians $\mathbf{Q}_s$, which is performed offline—a computation that can easily be done in parallel. If $\bar{S}$ and d are small, then this step can be performed efficiently using direct or iterative solvers. Once these are computed, one only needs to compute the determinants of $\mathbf{G}_0^\infty + \mathbf{Q}_s + \widehat{\mathbf{R}}$ for $s \in \{1, \ldots, \bar{S}\} \backslash \sigma$ at each iteration of the algorithm.

This cost decreases at each iteration because the set of remaining sensors, $\{1,\dots,\bar{S}\}\backslash\sigma$, decreases in size. One interesting direction for future work would identify safe selection rules that can reject sensors that will not be used without computing the determinant.

To analyze Algorithm 1, we note that the objective function g is *submodular* as a function of the set σ: for any sets $\tau \subseteq \sigma \subset \{1,\dots,\bar{S}\}$ and any sensor $s \in \{1,\dots,\bar{S}\}\backslash\sigma$,

$$g(\tau \cup \{s\}) - g(\tau) \geqslant g(\sigma \cup \{s\}) - g(\sigma).$$

In words, the improvement from adding an additional sensor is always larger when the sensor is added to the smaller set $\tau \subseteq \sigma$. For a submodular objective, the greedy algorithm, Algorithm 1, provably yields a solution within $1 - 1/e$ of the optimum [35] (and often even better [36]). Formally, let σ be the set of sensors returned by Algorithm 1. Then the submodularity of g ensures that

$$g(\sigma) \geqslant \left(1 - \frac{1}{e}\right) \max_{|\tau| \leqslant S_0} g(\tau).$$

Consequently, the greedy algorithm is inexpensive to run and produces a solution that is nearly optimal.

In comparison, we can reformulate the problem of maximizing g over the sets $\sigma \subset \{1,\dots,\bar{S}\}$ as the binary optimization problem

$$\max_{w} \ \log\det\left(\sum_{s=1}^{\bar{S}} w_s \mathbf{Q}_s + \widehat{\mathbf{R}}\right) \quad \text{subject to} \quad \sum_{s=1}^{\bar{S}} w_s \leqslant S_0, \quad w_s \in \{0,1\}. \tag{25}$$

Given an optimal binary vector w, the associated measurement set is $\sigma = \{s \,|\, w_s = 1\}$. Although the objective function for the binary program (25) is concave for $w \in [0,1]^{\bar{S}}$, its numerical solution can be quite expensive. For example, a branch-and-bound method [16] might require exponentially many evaluations of the determinant.

7. Numerical results

In this section, we investigate the performance of Algorithm 1 for a two-dimensional advection-diffusion-reaction example and a one-dimensional wave example. We discretize the PDEs in space using continuous piecewise (bi)linear finite elements on a uniform mesh (quadrilateral elements in 2d) and the backward Euler method in time (500 time steps for $t_f = 5$). Upon discretizing in space, we arrive at an optimization problem with the form (18). We set the measurement vectors to be point measurements collocated at the mesh vertices, i.e., $\mathbf{c}_s = \mathbf{e}_s$, where $\mathbf{e}_s$ has the value one as its s-th component and zero for all other components, and the desired number of sensors to $S_0 = 16$. We further generate the data $\mathbf{y}_d$ by solving the discretized PDE to obtain the "true" signal u_{true}, applying the current observation matrix $\mathbf{C}$, and adding normally distributed noise of

zero mean and standard deviation σ chosen so that the resulting data has signal-to-noise ratio equal to a prescribed value. We define the signal-to-noise ratio to be the quantity

$$\mathrm{SNR} = \frac{\sigma^{-2}}{t_f} \int_0^{t_f} \int_\Omega |u_{\mathrm{true}}|^2 \, \mathrm{d}\Omega \mathrm{d}t.$$

For each example, we set the regularization matrix $\mathbf{R}$ to be the associated discretization of the squared H^1-norm scaled by $\delta = 10^{-2}$ and set the nonsmooth term Φ to zero. Finally, for the Galerkin approximation, we choose the columns of $\mathbf{Y}$ to be the Legendre polynomials of degree less than or equal to six, $\{p_0, \ldots, p_6\}$, evaluated at the mesh vertices in 1d. In 2d, we employ the total degree Legendre polynomials

$$p_{i,j}(x, y) = p_i(x) p_j(y) \qquad \text{for} \qquad i + j \leqslant 6, \qquad i, j \in \{0, \ldots, 6\}.$$

The subspace dimension is $d = 7$ in 1d and $d = 21$ in 2d.

To describe the discretization, we restrict the presentation to the advection-diffusion-reaction equation and then repurpose the notation for the wave equation. Let $\Omega = (0, 1)^p$, $p = 1, 2$, with boundary $\Gamma = \partial\Omega$ and consider the parabolic PDE

$$\begin{aligned} \dot{u} - \nabla \cdot (\kappa \nabla u) + \mathbf{v} \cdot \nabla u + r u &= 0 \qquad \text{in } (0, t_f) \times \Omega \\ \kappa \nabla u \cdot n &= \beta u \quad \text{on } (0, t_f) \times \Gamma \\ u(0, \cdot) &= u_0 \quad \text{in } \Omega. \end{aligned} \tag{26}$$

Let $U_m = \mathrm{span}\{\varphi_1, \ldots, \varphi_m\}$ denote a linear subspace of the Sobolev space $H^1(\Omega)$. We approximate the PDE solution u in space as

$$u(t, x) \approx u_m(t, x) := \sum_{i=1}^{m} \mathbf{x}_i(t) \varphi_i(x)$$

and discretize the PDE (26) by enforcing that the PDE residual at u_m is orthogonal to U_m resulting in a problem of the form (1b). We will use the following finite-element matrix notation in the upcoming sections:

$$\mathbf{K}_{ij} = \int_\Omega \{(\kappa \nabla \varphi_i) \cdot \nabla \varphi_j + (\mathbf{v} \cdot \nabla \varphi_i) \varphi_j + r \varphi_i \varphi_j\} \, \mathrm{d}\Omega + \beta \int_\Gamma \varphi_i \varphi_j \, \mathrm{d}\Gamma \tag{27a}$$

$$\mathbf{M}_{ij} = \int_\Omega \varphi_i \varphi_j \, \mathrm{d}\Omega \tag{27b}$$

$$\mathbf{R}_{ij} = \delta \int_\Omega \{\nabla \varphi_i \cdot \nabla \varphi_j + \varphi_i \varphi_j\} \, \mathrm{d}\Omega \tag{27c}$$

To solve the resulting reduced, discretized optimization problem (18), we employ the preconditioned conjugate gradient (CG) method. We use the same algorithm applied to (17) to generate the initial guess for (18).

Table 1
First Row: Logarithmic norm for the two-dimensional advection-diffusion-reaction example with varying diffusivities $\kappa \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. Second Row: Average number of conjugate gradient iterations using $\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}}$ as a preconditioner. Third Row: Average number of unpreconditioned conjugate gradient iterations.

κ	10^{-3}	10^{-2}	10^{-1}	10^{0}
$\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})$	-0.019735	-0.196642	-1.897475	-13.801099
PCG iters	21.0000	16.2500	8.9375	5.4375
CG iters	26.0625	24.5200	26.6250	28.6250

7.1. Advection-diffusion-reaction equation

For this example, we set $\mathbf{A} = -\mathbf{K}$, $\mathbf{E} = \mathbf{M}$, and $\mathbf{f} \equiv 0$. We discretize in space on a uniform mesh of 25,600 quadrilateral elements. We choose the coefficients $\beta = 10$, $\kappa(x) \equiv 10^a$ for $a \in \{-3, -2, -1, 0\}$, $\mathbf{v}(x) \equiv (1, 1)^{\top}$ and $r(x) \equiv 0$. In addition, we choose the true initial condition to be $u_0(x) = \sin(\pi x_1)\sin(\pi x_2)$ and set the signal-to-noise ratio to SNR = 5. The logarithmic norms of $\mathbf{E}^{-1}\mathbf{A}$ with respect to the $\mathbf{E}$-norm for the varying κ are listed in the first row of Table 1. In the second and third rows, we list the average number of $(\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})$-preconditioned CG iterations and unpreconditioned CG iterations for solving (1) with $\mathbf{C}$ corresponding to the greedy sensor locations depicted in the left column of Fig. 1. As demonstrated in Section 4, the quality of the preconditioner improves as $|\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})|$ increases, which is confirmed by our numerical results. For example, when $\kappa = 1$ the preconditioner $(\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})$ reduces the average number of iterations by a factor of roughly 5.2644 when compared with no preconditioning. In contrast, when $\kappa = 10^{-3}$, $|\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A})|$ is relatively small and CG performs comparably with and without preconditioning. In particular, the preconditioner $(\mathbf{G}_0^{\infty} + \widehat{\mathbf{R}})$ reduces the average number of iterations only slightly compared to unpreconditioned CG. When using $(\mathbf{G}_0^{t_f} + \widehat{\mathbf{R}})$ as a preconditioner, the average number of CG iterations was nearly identical to the values in the second row of Table 1.

In the right column of Fig. 1, we depict the discretized L^2-recovery error. For comparison, we solved each instance of (1) with 100 realizations of S randomly selected sensors for $S = 1, \ldots, S_0$. The solid red lines in the right images are the median of the error for the random sensors, while the dashed red lines correspond to the 10% and 90% quantiles of the error. The black line corresponds to the recovery error for the greedy sensors using $\mathbf{G}_0^{\infty}$. As depicted, the greedy sensor selection outperforms the randomly selected sensors by a considerable margin for all κ. We note that more sensors are placed near the boundary of Ω as κ decreases in size, with increasing numbers of sensors in the direction of the advection. This is expected since for small κ the advection dominates the diffusion, pushing the state towards the boundary at a higher rate. This feature of the greedy algorithm is impossible to recover with randomly placed sensors, resulting in poor recovery errors. We note that the greedy sensors selected when using the finite-time Gramian $\mathbf{G}_0^{t_f}$ were identical to those selected using $\mathbf{G}_0^{\infty}$.

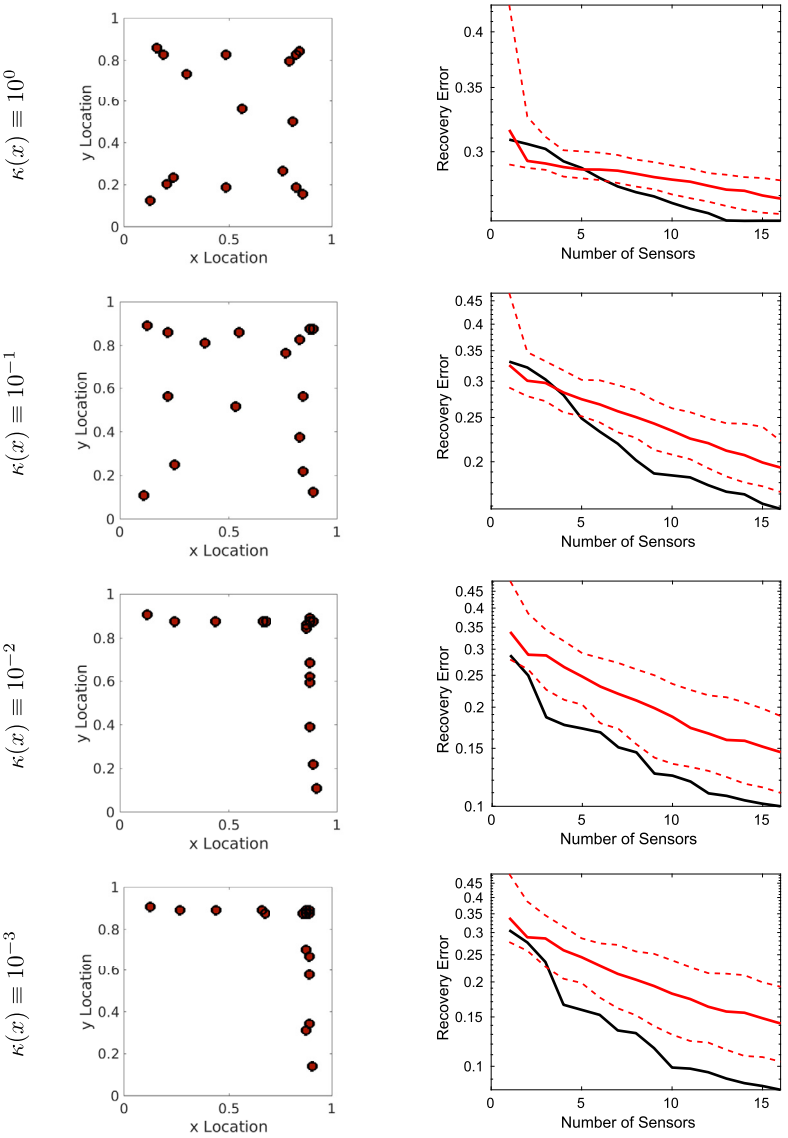

Fig. 1. Two-dimensional advection-diffusion-reaction results. Left: The greedy sensor locations. Right: The recovery error for the greedy and randomly selected sensors measured in the discretized L^2 norm.

7.2. Wave equation

For our second example, we consider the one-dimensional wave equation

$$\begin{aligned} \ddot{u} + \gamma \dot{u} - c^2 \Delta u &= 0 \quad \text{in } (0, t_f) \times \Omega \\ c^2 \nabla u \cdot n &= \beta u \quad \text{on } (0, t_f) \times \partial\Omega \\ u(0, \cdot) = u_0, \quad \dot{u}(0, \cdot) &= v_0 \quad \text{in } \Omega. \end{aligned} \tag{28}$$

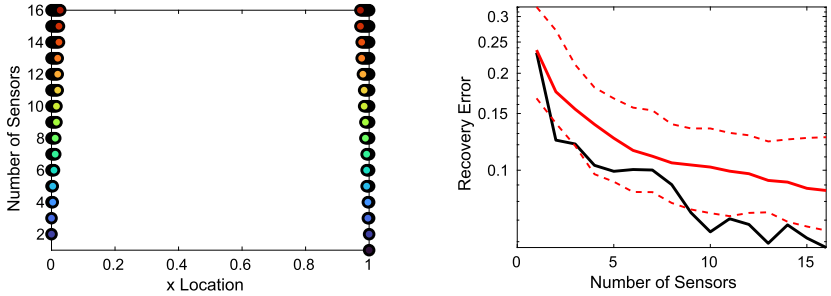

Fig. 2. One-dimensional wave equation results. Left: The greedy sensor locations. Right: The recover error for the greedy and random sensors measured in the discretized L^2 norm.

We discretize (28) on a uniform mesh of 256 intervals and we choose the wave speed $c = 2$, the damping factor $\gamma = 10^{-1}$, and the Robin coefficient $\beta = 10$. We transform (28) into a first-order equation in time by adding the velocity variable $v = \dot{u}$, producing the PDE

$$\begin{aligned} \dot{u} &= v && \text{in } (0,t_f)\times\Omega \\ \dot{v} + \gamma v - c^2\Delta u &= 0 && \text{in } (0,t_f)\times\Omega \\ c^2\nabla u\cdot n &= \beta u && \text{on } (0,t_f)\times\partial\Omega. \end{aligned} \tag{29}$$

In addition, we choose the true initial conditions to be $u_0(x) = \sin(\pi x)$ and $v_0(x) \equiv 0$, and set the signal-to-noise ratio to SNR = 10. Upon discretization, the system matrices are

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{M} \\ -\mathbf{K} & -\gamma\mathbf{M} \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} \mathbf{M} & 0 \\ 0 & \mathbf{M} \end{pmatrix}, \qquad \text{and} \qquad \mathbf{f} \equiv 0,$$

where $\mathbf{K}$ is defined as in (27a) with $\kappa \equiv c^2$, $\mathbf{v} \equiv 0$, and $r \equiv 0$. In addition, we restrict the set of possible sensors to only measure the u-variable. The logarithmic norm of $\mathbf{E}^{-1}\mathbf{A}$, with respect to the $\mathbf{E}$-norm, is

$$\mu_{\mathbf{E}}(\mathbf{E}^{-1}\mathbf{A}) \approx 1{,}572{,}898.12.$$

Clearly, Assumption 1 is violated. However, the spectral abscissa associated with $\mathbf{E}^{-1}\mathbf{A}$ is negative (i.e., $\alpha(\mathbf{E}^{-1}\mathbf{A}) = -\frac{\gamma}{2} = -0.05$). Consequently, the discussion in Remark 2 applies to this example. A simple computation shows that the constant C in (14) decreases by $\sqrt{2}$ as the finite-element mesh size is doubled, confirming that the $\mathbf{E}$ and $\mathbf{P}$ norms are not equivalent in the limit of the spatial discretization. However, this does suggest that our results and method are applicable even though Assumption 1 fails to hold. In Fig. 2, we depict the greedy sensor locations, using $\mathbf{G}_0^\infty$, in the left image and the discretized L^2-recovery error in the right image. For comparison, we solved (1) with 100 realizations of S randomly selected sensors for $S = 1,\ldots,S_0$. The solid red line in the right image

Table 2
Greedy sensor locations for the one-dimensional wave equation. The first column indicates the order in which each sensor was selected, the second column are the sensors selected using $\mathbf{G}_0^{\infty}$ and the third column are the sensors selected using $\mathbf{G}_0^{t_f}$.

Sensor	$\mathbf{G}_0^{\infty}$	$\mathbf{G}_0^{t_f}$
1	1.00000000	1.00000000
2	0.00000000	0.00000000
3	0.99609375	0.99609375
4	0.00390625	0.00390625
5	0.99218750	0.00781250
6	0.00781250	0.99218750
7	0.01171875	0.98828125
8	0.98828125	0.01171875
9	0.01562500	0.98437500
10	0.98437500	0.01562500
11	0.01953125	0.01953125
12	0.98046875	0.98046875
13	0.97656250	0.97656250
14	0.02343750	0.02343750
15	0.97265625	0.97265625
16	0.02734375	0.02734375

is the median of the error for the random sensors, while the dashed red lines correspond to the 10% and 90% quantiles of the error. The black line corresponds to the recovery error for the greedy sensors. As seen in this image, greedily selected sensors tend to outperform the randomly placed sensors. As with the advection-diffusion-reaction example, the greedy sensors are all clustered near the boundary of Ω—a physically intuitive feature that cannot be replicated using randomly placed sensors. In Table 2, we list the greedy sensor locations, in the order that they were selected. The second column lists the sensor locations selected using $\mathbf{G}_0^{\infty}$, while the third column lists the sensor locations selected using $\mathbf{G}_0^{t_f}$. In comparison, the sensors selected using $\mathbf{G}_0^{t_f}$ and using $\mathbf{G}_0^{\infty}$ differ only slightly in the order in which they were selected, suggesting that both $\mathbf{G}_0^{t_f}$ and $\mathbf{G}_0^{\infty}$ perform comparably for sensor selection.

8. Conclusion

The quality of the solution recovered by solving an inverse problem depends on the quality of the data. The location of sensors is one important parameter determining data quality. We have demonstrated a method to select sensors with guarantees on the quality of the resulting solution. Choosing the optimal location of sensors is computationally challenging; our method uses a Galerkin projection to reduce the size of the problem (which also generates a preconditioner for the inverse problem) and uses a greedy method to select each subsequent sensor. Empirically, we have seen that our sensor selection method yields improved accuracy compared to selecting random sensors in most cases.

The following important questions remain open. For which dynamical systems do random sensors perform well, and for which dynamical systems are optimized, not greedy, sensors important? Can we adapt our Galerkin method for problems with time-varying, nonlinear, or unstable dynamics? Can we use similar ideas to design a method for actuator placement using the controllability Gramian? We leave these as future work.

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.

Data availability

No data was used for the research described in the article.

References

[1] M. Ghil, P. Malanotte-Rizzoli, Data Assimilation in Meteorology and Oceanography, Advances in Geophysics, vol. 33, Elsevier, 1991, pp. 141–266.
[2] H. Goosse, E. Crespin, A. de Montety, M.E. Mann, H. Renssen, A. Timmermann, Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation, J. Geophys. Res., Atmos. 115 (D9) (2010), https://doi.org/10.1029/2009JD012737.
[3] A.O. Zaporozhets, V.V. Khaidurov, Mathematical models of inverse problems for finding the main characteristics of air pollution sources, Water Air Soil Pollut. 231 (12) (2020) 1–13.
[4] A.Y. Sun, S.L. Painter, G.W. Wittmeyer, A constrained robust least squares approach for contaminant release history identification, Water Resour. Res. 42 (4) (2006).
[5] T. Tian, X. Ge, Calibration of stochastic differential equation models using implicit numerical methods and particle swarm optimization, in: 2012 Proceedings of International Conference on Modelling, Identification and Control, IEEE, 2012, pp. 1049–1054.
[6] J. Frankel, M. Keyhani, B. Elkins, Surface heat flux prediction through physics-based calibration, part 1: theory, J. Thermophys. Heat Transf. 27 (2) (2013) 189–205.
[7] F. Pukelsheim, Optimal Design of Experiments, SIAM, 2006.
[8] S. Silvey, Optimal Design: an Introduction to the Theory for Parameter Estimation, vol. 1, Springer Science & Business Media, 2013.
[9] S. Martinez, F. Bullo, Optimal sensor placement and motion coordination for target tracking, Automatica 42 (4) (2006) 661–668, https://doi.org/10.1016/j.automatica.2005.12.018.
[10] D.C. Kammer, Optimal sensor placement for modal identification using system-realization methods, J. Guid. Control Dyn. 19 (3) (1996) 729–731, https://doi.org/10.2514/3.21688.
[11] C. Jiang, Y.C. Soh, H. Li, Sensor placement by maximal projection on minimum eigenspace for linear inverse problems, IEEE Trans. Signal Process. 64 (21) (2016) 5595–5610, https://doi.org/10.1109/TSP.2016.2573767.
[12] J. Ranieri, A. Chebira, M. Vetterli, Near-optimal sensor placement for linear inverse problems, IEEE Trans. Signal Process. 62 (5) (2014) 1135–1146, https://doi.org/10.1109/TSP.2014.2299518.
[13] W. Kang, L. Xu, Optimal placement of mobile sensors for data assimilations, Tellus A, Dyn. Meterol. Oceanogr. 64 (1) (2012), https://doi.org/10.3402/tellusa.v64i0.17133.
[14] A. Singh, J. Hahn, Sensor location for stable nonlinear dynamic system: multiple sensor case, Ind. Eng. Chem. Res. 45 (10) (2006) 3615–3623, https://doi.org/10.1021/ie0511175.
[15] J. Qi, K. Sun, W. Kang, Optimal PMU placement for power system dynamic state estimation by using empirical observability Gramian, IEEE Trans. Power Syst. 30 (4) (2015) 2041–2054, https://doi.org/10.1109/TPWRS.2014.2356797.
[16] D. Georges, The use of observability and controllability Gramians or functions for optimal sensor and actuator location in finite-dimensional systems, in: Proceedings of 1995 34th IEEE Conference on Decision and Control, vol. 4, IEEE, 1995, pp. 3319–3324.

[17] A.K. Singh, J. Hahn, Determining optimal sensor locations for state and parameter estimation for stable nonlinear systems, Ind. Eng. Chem. Res. 44 (15) (2005) 5645–5659.
[18] H.R. Shaker, M. Tahavori, Optimal sensor and actuator location for unstable systems, J. Vib. Control 19 (12) (2013) 1915–1920.
[19] U. Vaidya, R. Rajaram, S. Dasgupta, Actuator and sensor placement in linear advection PDE with building system application, J. Math. Anal. Appl. 394 (1) (2012) 213–224, https://doi.org/10.1016/j.jmaa.2012.03.046.
[20] H. Fang, R. Sharma, R. Patil, Optimal sensor and actuator deployment for HVAC control system design, in: 2014 American Control Conference, IEEE, 2014, pp. 2240–2246.
[21] B. Marx, D. Koenig, D. Georges, Optimal sensor and actuator location for descriptor systems using generalized Gramians and balanced realizations, in: Proceedings of the 2004 American Control Conference, Vol. 3, IEEE, 2004, pp. 2729–2734.
[22] M. Güney, E. Eşkinat, Optimal actuator and sensor placement in flexible structures using closed-loop criteria, J. Sound Vib. 312 (1–2) (2008) 210–233.
[23] E. Deutsch, On matrix norms and logarithmic norms, Numer. Math. 24 (1) (1975) 49–51.
[24] G. Söderlind, The logarithmic norm: history and modern theory, BIT Numer. Math. 46 (3) (2006) 631–652.
[25] T. Ström, On logarithmic norms, SIAM J. Numer. Anal. 12 (5) (1975) 741–753, https://doi.org/10.1137/0712055.
[26] W. Gawronski, J.-N. Juang, Model reduction in limited time and frequency intervals, Int. J. Syst. Sci. 21 (2) (1990) 349–376.
[27] P. Kürschner, Balanced truncation model order reduction in limited time intervals for large systems, Adv. Comput. Math. 44 (6) (2018) 1821–1844.
[28] N. Higham, Functions of Matrices: Theory and Computation, Other Titles in Applied Mathematics, Society for Industrial and Applied Mathematics, 2008, https://books.google.com/books?id=2Wz_zVUEwPkC.
[29] A.C. Antoulas, Approximation of Large-Scale Dynamical Systems, SIAM, 2005.
[30] C.V. Pao, Logarithmic derivates of a square matrix, Linear Algebra Appl. 6 (1973) 159–164.
[31] V. Simoncini, Computational methods for linear matrix equations, SIAM Rev. 58 (3) (2016) 377–441.
[32] D. Palitta, V. Simoncini, Optimality properties of Galerkin and Petrov–Galerkin methods for linear matrix equations, Vietnam J. Math. 48 (4) (2020) 791–807.
[33] V. Simoncini, V. Druskin, Convergence analysis of projection methods for the numerical solution of large Lyapunov equations, SIAM J. Numer. Anal. 47 (2) (2009) 828–843.
[34] P. Benner, E. Sachs, S. Volkwein, Model order reduction for PDE constrained optimization, in: Trends in PDE Constrained Optimization, Springer, 2014, pp. 303–326.
[35] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, An analysis of approximations for maximizing submodular set functions—I, Math. Program. 14 (1) (1978) 265–294.
[36] D. Sharma, A. Kapoor, A. Deshpande, On greedy maximization of entropy, in: International Conference on Machine Learning, PMLR, 2015, pp. 1330–1338.