Mathematics of Control, Signals, and Systems (2020) 32:19–41
https://doi.org/10.1007/s00498-020-00253-z

ORIGINAL ARTICLE

# Approximability models and optimal system identification

**Mahmood Ettehad**[1] · **Simon Foucart**[1]

Received: 6 December 2018 / Accepted: 4 February 2020 / Published online: 13 February 2020


**Abstract**
This article considers the problem of optimally recovering stable linear time-invariant systems observed via linear measurements made on their transfer functions. A common modeling assumption is replaced here by the related assumption that the transfer functions belong to a model set described by approximation capabilities. Capitalizing on recent optimal recovery results relative to such approximability models, we construct some optimal algorithms and characterize the optimal performance for the identification and evaluation of transfer functions in the framework of the Hardy Hilbert space and of the disk algebra. In particular, we determine explicitly the optimal recovery performance for frequency measurements taken at equispaced points on an inner circle or on the torus.

**Keywords** Optimal recovery · System identification · Hardy spaces · Disk algebra

**Mathematics Subject Classification** 93B30 · 46J10 · 30H10

## 1 Background and motivation

System identification can be viewed as the learning task of inferring a model to describe the behavior of a system observed via input–output data. As such, it influences control theory as a whole, since the model selection dictates the subsequent design of efficient controllers for the system. Linear dynamical systems, due to their appearance in many applications, have been the focus of many investigations. They are often handled by moving to the frequency domain, where one studies their transfer functions, typically assumed to be elements of the Hardy space $\mathcal{H}_2$ (see e.g., [18, Chapter 13]) or $\mathcal{H}_\infty$ (see e.g., [18, Chapter 14]). Then, based on *a priori* information given as an hypothesized

Simon Foucart is partially supported by NSF Grants DMS-1622134 and DMS-1664803, and also acknowledges the NSF Grant CCF-1934904.

✉ Simon Foucart
simon.foucart@centraliens.net

[1] Texas A&M University, College Station, USA

model and on *a posteriori* information given as frequency observations, one seeks to identify the transfer function in a way that minimizes the error with respect to the given norm.

For the model put forward in [11], various algorithms for the identification of transfer functions in $\mathcal{H}_\infty$ have already been proposed: In [7], interpolatory algorithms are constructed by solving a Nevanlinna–Pick problem, [8] presents some closely related algorithms with explicit bounds in $\mathcal{H}_\infty$ and $\ell_1$ norms, and in [6] interpolatory algorithms are generalized to time-domain data by solving a Carathéodory–Fejér problem via standard convex optimization methods. More recent works focus on the trade-off between the number of samples required to accurately build models of dynamical systems and the degradation of performance in various control objectives. For instance, the article [16] derived bounds on the number of noisy input–output samples from a stable linear time-invariant system that are sufficient to guarantee closeness of the finite impulse response approximation to the true system in $\mathcal{H}_\infty$-norm. In [15], linear system identification from pointwise noisy measurements in the frequency domain was formulated as a convex minimization problem involving an $\ell_1$-penalty and error estimates in $\mathcal{H}_2$-norm were provided. There has also been progress in quantifying the sample complexity of dynamical system identification using statistical estimation theory, see [4,17].

The article [11] was one of the first to suggest adopting a perspective from optimal recovery [12] in system identification, leading to a framework that is now textbook material (see e.g., [13, Section 4.4] or [5, Section 3.2]). The purpose of this note is to revisit this classical framework in light of recent optimal recovery results involving approximability models, see [2,9,10]. Our setting is closely related to the one from [11], in that the *a priori* information is encapsulated by a model set $\mathcal{K}$ and the *a posteriori* information consists of frequency response measurements (possibly nonuniformly spaced, as in [1]), while the objective is to establish performance bounds and devise algorithms for the recovery of transfer functions $F$. By recovering $F$, we mean here approximating it in full (i.e., identifying) in the context of $\mathcal{H}_2$ or evaluating a point value $F(\zeta_0)$ (i.e., estimating) in the context of $\mathcal{H}_\infty$—in fact, as in [3], we replace $\mathcal{H}_\infty$ by a subspace known as the disk algebra to avoid certain technical difficulties. The novelty of our setting lies in the model set $\mathcal{K}$: Whereas the model set of [11] can be viewed as an intersection $\mathcal{K} = \cap_{n\geq 0}\mathcal{K}_n$ of approximation sets, we focus here on a single approximation set $\mathcal{K}_n$ chosen as model set. This slight modification of the framework allows us to construct identification algorithms that are *optimal*, not just near-optimal. In addition, these optimal algorithms turn out to also be *linear* algorithms.

The structure of this article is as follows. In Sect. 2, we formulate precisely the problem we are considering, we introduce the approximability model, and we describe some optimal recovery results that have recently emerged. In Sect. 3, we zoom in on results about Hilbert spaces and adapt them to the identification of transfer functions in $\mathcal{H}_2$, leading to a matrix-analytic method for the construction of an optimal algorithm and the determination of the optimal performance. In the case of data gathered at equispaced points on inner circles, we also find the exact value of the optimal performance in terms of the number $m$ of observations and we remark that it is independent of the dimension $n$ of the polynomial space underlying the approximability model.

In Sect. 4, we turn to results about quantities of interest in Banach spaces and adapt them to the optimal evaluation of transfer functions in the disk algebra, leading to a convex optimization method for the construction of an optimal algorithm and the determination of the optimal performance. In the case of data gathered at equispaced points on the torus, we also show how the optimal performance behaves in terms of $m$ and $n$. Section 5 concludes with some perspective on further research. Finally, an appendix collects some that have been delayed to keep the flow of the main text going.

## 2 Problem formulation

In its most abstract form, the scenario we are considering in the rest of this article involves unknown objects $F$ from a normed space $\mathcal{X}$ which are acquired through observational data

$$y_k = \ell_k(F), \qquad k = 1, \dots, m, \tag{1}$$

where $\ell_1, \dots, \ell_m \in \mathcal{X}^*$ are known linear functionals. The perfect accuracy of the acquisition process is of course an idealization. For brevity, we shall write $y = \mathcal{L}(F) \in \mathbb{C}^m$. Discarding some of the $\ell_k$'s if necessary, we may assume that the operator $\mathcal{L} : \mathcal{X} \to \mathbb{C}^m$ has full range. The task at hand consists in making use of the data $y$ to approximate $F$, or merely to evaluate a quantity of interest $Q(F)$, where $Q$ is a linear map from $\mathcal{X}$ into another normed space $\mathcal{Y}$—typically $\mathcal{Y} = \mathbb{C}$. There is also an *a priori* knowledge about $F$, often conveyed by the assumption that $F$ belongs to a certain model set $\mathcal{K} \subseteq \mathcal{X}$. The feasibility of the task is assessed in a worst-case setting via the quantity

$$E^{\mathrm{opt}}(\mathcal{K}, \mathcal{L}, Q) := \inf_{A:\mathbb{C}^m \to \mathcal{Y}} \sup_{F \in \mathcal{K}} \|Q(F) - A(\mathcal{L}(F))\|_{\mathcal{Y}}. \tag{2}$$

An optimal algorithm (relative to the model set $\mathcal{K}$) is a map $A^{\mathrm{opt}}$ from $\mathbb{C}^m$ into $\mathcal{Y}$ for which the infimum is achieved.

In system identification, the standard objects are transfer functions belonging to the Hardy space $\mathcal{X} = \mathcal{H}_2(\mathbb{D})$ or $\mathcal{X} = \mathcal{H}_\infty(\mathbb{D})$ relative to the open unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$. We recall that the Hardy spaces $\mathcal{H}_p(\mathbb{D})$ are defined for $1 \le p \le \infty$ by

$$\mathcal{H}_p(\mathbb{D}) := \left\{ F \text{ is analytic on } \mathbb{D} \text{ and } \|F\|_{\mathcal{H}_p} < +\infty \right\}, \tag{3}$$

where the $\mathcal{H}_p$-norms of $F(z) = \sum_{n=0}^{\infty} f_n z^n$ are given for $p = 2$ and $p = \infty$ by

$$\|F\|_{\mathcal{H}_2} := \sup_{r \in [0,1)} \left[ \frac{1}{2\pi} \int_0^{2\pi} |F(re^{i\theta})|^2 d\theta \right]^{1/2} = \left[ \sum_{n=0}^{\infty} |f_n|^2 \right]^{1/2}, \tag{4}$$

$$\|F\|_{\mathcal{H}_\infty} := \sup_{r \in [0,1)} \sup_{\theta \in [-\pi,\pi]} |F(re^{i\theta})| = \sup_{|z|=1} |F(z)|. \tag{5}$$

The last equality in (5) does not imply that functions in $\mathcal{H}_\infty(\mathbb{D})$ are well defined on the torus $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$. To bypass this peculiarity, instead of the whole Hardy space $\mathcal{H}_\infty(\mathbb{D})$, we shall work within the subspace consisting of functions that are continuous on $\mathbb{T}$. This space is called the disk algebra and is denoted $\mathcal{A}(\mathbb{D})$.

Concerning the acquisition process, although our considerations are valid for any linear functionals $\ell_1, \ldots, \ell_m$, we shall concentrate in this initial work on the situation commonly encountered in system identification where $\ell_1, \ldots, \ell_m$ are point evaluations at some $\zeta_1, \ldots, \zeta_m$ located in the disk or on the torus. It is worth recalling that the point evaluation defined at some $\zeta \in \mathbb{C}$ by

$$e_\zeta(F) := F(\zeta), \qquad F \in \mathcal{X}, \tag{6}$$

is well defined for $|\zeta| \leq 1$ when $\mathcal{X} = \mathcal{A}(\mathbb{D})$, but only for $|\zeta| < 1$ when $\mathcal{X} = \mathcal{H}_2(\mathbb{D})$. Its Riesz representer $E_\zeta \in \mathcal{H}_2(\mathbb{D})$, characterized by the identity $e_\zeta(F) = \langle F, E_\zeta \rangle_{\mathcal{H}_2}$ for all $F \in \mathcal{H}_2(\mathbb{D})$, is the Cauchy kernel given by

$$E_\zeta(z) := \frac{1}{1 - \overline{\zeta} z}, \qquad |z| < 1. \tag{7}$$

We also point out that, in $\mathcal{A}(\mathbb{D})$, the dual norm of a point evaluation at $\zeta \in \mathbb{C}$, $|\zeta| \leq 1$, is

$$\|e_\zeta\|_{\mathcal{A}^*} = 1, \tag{8}$$

where $\|e_\zeta\|_{\mathcal{A}^*} \geq 1$ follows, e.g., from $e_\zeta(1) = 1$ and where $\|e_\zeta\|_{\mathcal{A}^*} \leq 1$ holds because, for any $F \in \mathcal{A}(\mathbb{D})$, one has $|F(\zeta)| \leq \|F\|_{\mathcal{H}_\infty}$ by the maximum modulus theorem.

As for the model, a popular representation of the *a priori* information proposed in [11] prevails. Given $\rho > 1$ and $M > 0$, it in fact involves two model sets, $\mathcal{K}_{\mathcal{H}_2}$ and $\mathcal{K}_{\mathcal{H}_\infty}$, both of them consisting of functions $F$ that are analytic on the disk $\mathbb{D}_\rho := \{z \in \mathbb{C} : |z| < \rho\}$ and that satisfy, respectively,

$$\|F(\rho\,\cdot)\|_{\mathcal{H}_2} = \left[\sum_{n=0}^{\infty} |f_n|^2 \rho^{2n}\right]^{1/2} \leq M, \tag{9}$$

$$\|F(\rho\,\cdot)\|_{\mathcal{H}_\infty} = \sup_{|z|=\rho} |F(z)| \leq M. \tag{10}$$

These model sets are closely related (see the appendix for the precise statement and its justification) to the model sets given, for $\mathcal{X} = \mathcal{H}_2(\mathbb{D})$ or $\mathcal{X} = \mathcal{H}_\infty(\mathbb{D})$, by

$$\bigcap_{n \geq 0} \{F \in \mathcal{X} : \mathrm{dist}_{\mathcal{X}}(F, \mathcal{P}_n) \leq \varepsilon_n\}, \tag{11}$$

where $\mathcal{P}_n$ denotes the space of polynomials of degree at most $n-1$ and where $\varepsilon_n = M\rho^{-n}$ [with constants $\rho > 1$ and $M > 0$ differing from the ones in (9)–(10)]. In the

rest of this article, we consider the model sets obtained by overlooking the intersection and focusing on single sets associated with fixed $n$'s. In this situation, we will be able to produce optimal algorithms in the sense of (2). As we will realize, these optimal algorithms are in fact linear maps.[1] Our method leverages recent results described next. Strictly speaking, they were established in [2,9,10] for the real setting only—their validity for the complex setting is justified in the Appendix.

Inspired by parametric PDEs, the authors of the articles [2,9,10] emphasized approximation properties and considered the model sets

$$\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V}) := \{F \in \mathcal{X} : \operatorname{dist}_{\mathcal{X}}(F, \mathcal{V}) \leq \varepsilon\} \tag{12}$$

relative to a normed space $\mathcal{X}$, a subspace $\mathcal{V}$ of $\mathcal{X}$, and a parameter $\varepsilon > 0$. As a summary of general results for the full approximation problem where $Q = \mathrm{Id}$ (see [10] for details), we underline that

(i) an important role is played by an indicator of the compatibility of the acquisition process and the model (as represented by the spaces $\ker(\mathcal{L})$ and $\mathcal{V}$, respectively), namely by

$$\mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V}) := \sup_{F \in \ker(\mathcal{L})} \frac{\|F\|_{\mathcal{X}}}{\operatorname{dist}_{\mathcal{X}}(F, \mathcal{V})}; \tag{13}$$

(ii) the optimal performance over the model set $\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V})$ is essentially characterized by this quantity, since

$$\mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V})\, \varepsilon \leq E^{\mathrm{opt}}(\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V}), \mathcal{L}, I) \leq 2\, \mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V})\, \varepsilon; \tag{14}$$

(iii) the map $A' : \mathbb{C}^m \to \mathcal{X}$ defined independently of $\varepsilon > 0$ by

$$A'(y) := \operatorname*{argmin}_{F \in \mathcal{X}} \operatorname{dist}_{\mathcal{X}}(F, \mathcal{V}) \qquad \text{subject to } \mathcal{L}(F) = y \tag{15}$$

provides a near-optimal algorithm (though not necessarily a practical one) in the sense that

$$\sup_{F \in \mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V})} \|F - A'(\mathcal{L}(F))\|_{\mathcal{X}} \leq 2\, \mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V})\, \varepsilon. \tag{16}$$

It is worth noting that $\mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V})$ decreases when observations are added, as $\ker(\mathcal{L})$ shrinks, and increases when the space $\mathcal{V}$ is enlarged, as $\operatorname{dist}_{\mathcal{X}}(F, \mathcal{V})$ becomes smaller. In fact, $\mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V}) = \infty$ when $n := \dim(\mathcal{V}) > m$ since then one can find $F \in \ker(\mathcal{L}) \cap \mathcal{V}$. Thus, we assume that $n \leq m$ throughout. The articles [2] and [9] improved the results (i)–(ii)–(iii) in two specific situations. Precisely, [2] considered the full approximation problem when $\mathcal{X}$ is a Hilbert space, while [9] dealt with an arbitrary normed space $\mathcal{X}$ but placed the focus on quantities of interest $Q$ that are

[1] In the case $\mathcal{K} = \mathcal{H}_2(\mathbb{D})$, the optimal algorithm over an ellipsoidal model set such as the one described by (9) is linear, too. It is a variant of the minimal-norm interpolation presented in Section 5.3. of [13].

linear functionals. Our objective consists in adapting and supplementing these contributions for the spaces $\mathcal{X} = \mathcal{H}_2(\mathbb{D})$ and $\mathcal{X} = \mathcal{A}(\mathbb{D})$, which are done in Sects. 3 and 4, respectively. More precisely, our novel contribution consists, for the case of $\mathcal{H}_2(\mathbb{D})$, in a slight but computation-ready variation on the result of [2] (Theorem 1) and in the exact determination of the indicator $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ when $\zeta_1, \ldots, \zeta_m$ are equispaced points on an inner circle (Proposition 2), and for the case of $\mathcal{A}(\mathbb{D})$, in a slight but appropriate extension of the result of [9] (Theorem 3) and in the asymptotic determination of the indicator $\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ when $\zeta_1, \ldots, \zeta_m$ are equispaced points on the torus (Proposition 5).

## 3 Optimal identification in $\mathcal{H}_2(\mathbb{D})$

As was just mentioned, the results (i)–(ii)–(iii) can be enhanced when $\mathcal{X}$ is a Hilbert space, which is the case of the Hardy space $\mathcal{H}_2(\mathbb{D})$. To present the enhancement provided by [2], we let $(V_1, \ldots, V_n)$ denote an orthonormal basis for the subspace $\mathcal{V}$ of the Hilbert space $\mathcal{X}$ and $L_1, \ldots, L_m \in \mathcal{H}_2(\mathbb{D})$ denote the Riesz representers of the linear functionals $\ell_1, \ldots, \ell_m \in \mathcal{X}^*$—recall that they are characterized by

$$\ell_k(F) = \langle F, L_k \rangle_{\mathcal{H}_2}, \qquad k = 1, \ldots, m. \tag{17}$$

Concerning (iii), it was shown that the algorithm of (15) is not just near optimal, but genuinely *optimal*. It can in fact be written as the *linear* map $A^{\mathrm{opt}} : \mathbb{C}^m \to \mathcal{X}$ given by

$$A^{\mathrm{opt}}(y) = V_y^\star + W_y - P_{\mathcal{W}}(V_y^\star), \qquad V_y^\star = \operatorname*{argmin}_{V \in \mathcal{V}} \|W_y - P_{\mathcal{W}}(V)\|_{\mathcal{X}}, \tag{18}$$

where $P_{\mathcal{W}}$ is the orthogonal projector onto the space $\mathcal{W} := \mathrm{span}\{L_1, \ldots, L_m\}$ and $W_y$ denotes the element of $\mathcal{W}$ satisfying $\ell_k(W_y) = y_k$ for all $k = 1, \ldots, m$ (so that $W_y = P_{\mathcal{W}}(F)$ if $y = \mathcal{L}(F)$ for some $F \in \mathcal{X}$). Concerning (ii), the optimal performance over $\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V})$ is then completely characterized by a sharpening of (14) to

$$E^{\mathrm{opt}}(\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V}), \mathcal{L}, I) = \mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V})\, \varepsilon. \tag{19}$$

Concerning (i), the compatibility indicator introduced in (13) becomes fully computable as

$$\mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V}) = \frac{1}{\sigma_{\min}(\widetilde{G})}, \tag{20}$$

where $\sigma_{\min}(\widetilde{G})$ is the smallest positive singular value of the cross-Gramian matrix $\widetilde{G} \in \mathbb{C}^{m \times n}$ relative to the orthonormal basis $(V_1, \ldots, V_n)$ for $\mathcal{V}$ and to an orthonormal basis $(\widetilde{L}_1, \ldots, \widetilde{L}_m)$ for $\mathcal{W}$. The entries of this matrix are

$$\widetilde{G}_{k,j} := \langle V_j, \widetilde{L}_k \rangle, \qquad k = 1, \ldots, m, \quad j = 1, \ldots, n. \tag{21}$$

### 3.1 Optimal performance and algorithm

When turning to the implementation of these results for the Hardy space $\mathcal{H}_2(\mathbb{D})$, a difficulty arises from the fact that orthonormal bases for $\mathcal{W}$ are not automatically available. A Gram–Schmidt orthonormalization of $(L_1, \ldots, L_m)$ is not ideal, because calculating inner products in $\mathcal{H}_2$ is not exact (if performed either as contour integrals or as inner products of infinite sequences). We are going to present a more reliable method of implementation, based only on matrix computations and relying on data directly available to the users. These data consist of two matrices $G \in \mathbb{C}^{m\times n}$ and $H \in \mathbb{C}^{m\times m}$. The first one, with entries

$$G_{k,j} := \ell_k(V_j), \qquad k = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{22}$$

is simply the cross-Gramian matrix relative to $(V_1, \ldots, V_n)$ and to $(L_1, \ldots, L_m)$, in view of the identity $\ell_k(V_j) = \langle V_j, L_k\rangle_{\mathcal{H}_2}$. The second one, with entries

$$H_{k,j} := \ell_k(L_j), \qquad k, j = 1, \ldots, m, \tag{23}$$

is the Gramian matrix relative to $(L_1, \ldots, L_m)$, in view of the identity $\ell_k(L_j) = \langle L_j, L_k\rangle_{\mathcal{H}_2}$.

Once the matrices $G$ and $H$ are set, we can make very explicit the optimal performance of system identification in $\mathcal{H}_2$ for the approximability model relative to a subspace $\mathcal{V}$, as well as an algorithm achieving this optimal performance. This is the object of the theorem below—to reiterate, it is a variation on a result of [2] which advantageously lends itself more easily to practical implementation.

**Theorem 1** *With $G \in \mathbb{C}^{m\times n}$ and $H \in \mathbb{C}^{m\times m}$ defined in (22) and (23), one has*

$$\mu_{\mathcal{H}_2}(\ker(\mathcal{L}), \mathcal{V}) = \frac{1}{\lambda_{\min}(G^* H^{-1} G)^{1/2}}. \tag{24}$$

*Moreover, with $c^\star = (G^* H^{-1} G)^{-1} G^* H^{-1} y$ and $d^\star = H^{-1}(y - Gc^\star)$, the map $A^{\mathrm{opt}} : \mathbb{C}^m \to \mathcal{H}_2(\mathbb{D})$ defined by*

$$A^{\mathrm{opt}}(y) = \sum_{j=1}^{n} c_j^\star V_j + \sum_{k=1}^{m} d_k^\star L_k \tag{25}$$

*is an optimal algorithm in the sense that*

$$\begin{aligned} &\sup_{F\in\mathcal{K}_{\mathcal{H}_2}(\varepsilon,\mathcal{V})} \|F - A^{\mathrm{opt}}(\mathcal{L}(F))\|_{\mathcal{H}_2} \\ &\quad = \inf_{A:\mathbb{C}^m\to\mathcal{H}_2} \sup_{F\in\mathcal{K}_{\mathcal{H}_2}(\varepsilon,\mathcal{V})} \|F - A(\mathcal{L}(F))\|_{\mathcal{H}_2}, \end{aligned} \tag{26}$$

*with the value of the latter being $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}), \mathcal{V})\,\varepsilon$.*

***Proof*** Since the Gramian matrix $H$ is positive definite, it has an eigenvalue decomposition

$$H = UDU^*, \tag{27}$$

where $U \in \mathbb{C}^{m\times m}$ is a unitary matrix and $D = \mathrm{diag}[d_1, \ldots, d_m] \in \mathbb{C}^{m\times m}$ is a diagonal matrix with positive entries. It is routine to verify that the functions $\widetilde{L}_1, \ldots, \widetilde{L}_m \in \mathcal{H}_2(\mathbb{D})$ defined by

$$\widetilde{L}_k = \frac{1}{\sqrt{d_k}} \sum_{j=1}^{m} U_{j,k} L_j, \qquad k = 1, \ldots, m, \tag{28}$$

form an orthonormal basis for $\mathcal{W}$. It is also easy to verify that the cross-Gramian matrix $\widetilde{G} \in \mathbb{C}^{m\times n}$ relative to $(V_1, \ldots, V_n)$ and to $(\widetilde{L}_1, \ldots, \widetilde{L}_m)$ is then expressed as

$$\widetilde{G} = D^{-1/2} U^* G. \tag{29}$$

In turn, we derive that

$$\widetilde{G}^* \widetilde{G} = G^* H^{-1} G. \tag{30}$$

The first part of the theorem, namely (24), is now a consequence of (20). For the second part of the theorem, by virtue of (18), our strategy is simply to determine in turn $W_y$, $V_y^\star$, and $P_{\mathcal{W}}(V_y^\star)$. To start with, it is readily checked, taking inner products with $L_1, \ldots, L_k$, that

$$W_y = \sum_{k=1}^{m} (H^{-1} y)_k L_k. \tag{31}$$

With the orthonormal basis $(\widetilde{L}_1, \ldots, \widetilde{L}_m)$ for $\mathcal{W}$ introduced in (28), the easily verifiable change-of-basis formula

$$W = \sum_{k=1}^{m} c_k L_k \iff W = \sum_{k=1}^{m} (D^{1/2} U^* c)_k \widetilde{L}_k \tag{32}$$

yields the alternative representation

$$W_y = \sum_{k=1}^{m} (D^{-1/2} U^* y)_k \widetilde{L}_k. \tag{33}$$

Next, for an arbitrary function $V = \sum_{j=1}^{n} c_j V_j \in \mathcal{V}$, we observe that

$$P_{\mathcal{W}}(V) = \sum_{k=1}^{m} \langle V, \widetilde{L}_k \rangle \widetilde{L}_k = \sum_{k=1}^{m} \sum_{j=1}^{n} c_j \langle V_j, \widetilde{L}_k \rangle \widetilde{L}_k = \sum_{k=1}^{m} (\widetilde{G} c)_k \widetilde{L}_k. \tag{34}$$

It follows from (33) and (34) that

$$\begin{aligned}\|W_y - P_{\mathcal{W}}(V)\|_{\mathcal{H}_2}^2 &= \sum_{k=1}^{m} |(D^{-1/2}U^*y)_k - (\widetilde{G}c)_k|^2 \\ &= \|D^{-1/2}U^*y - \widetilde{G}c\|_2^2. \end{aligned} \tag{35}$$

Therefore, the minimizer $V_y^\star \in \mathcal{V}$ of this expression takes the form

$$V_y^\star = \sum_{j=1}^{n} c_j^\star V_j, \qquad c^\star = \widetilde{G}^\dagger D^{-1/2}U^*y, \tag{36}$$

where $\widetilde{G}^\dagger = (\widetilde{G}^*\widetilde{G})^{-1}\widetilde{G}^*$ is the Moore–Penrose pseudo-inverse of $\widetilde{G}$. According to (29), we easily derive that the coefficient vector $c^\star$ is indeed

$$c^\star = (G^*H^{-1}G)^{-1}G^*H^{-1}y. \tag{37}$$

Finally, in view of (34) and (32), we obtain

$$P_{\mathcal{W}}(V_y^\star) = \sum_{k=1}^{m} (\widetilde{G}c^\star)_k \widetilde{L}_k = \sum_{k=1}^{m} (U D^{-1/2}\widetilde{G}c^\star)_k L_k = \sum_{k=1}^{m} (H^{-1}Gc^\star)_k L_k. \tag{38}$$

Putting (36)–(37), (33), and (38) together in (18), we arrive at the expression announced in (25). This completes the proof of the theorem. □

### 3.2 Evaluations at points on an inner circle

It is time to specify the results to our scenario of interest. In particular, keeping (11) in mind, we fix $\mathcal{V}$ to be the space $\mathcal{P}_n$ of polynomials of degree at most $n-1$. It is equipped with the orthonormal basis $(V_1, \ldots, V_n)$ given by $V_j(z) = z^{j-1}$, $j = 1, \ldots, n$. Let us also consider a radius $r < 1$ and suppose that the linear functionals $\ell_k$, $k = 1, \ldots, m$, take the form

$$\ell_k(F) = F(\zeta_k) \qquad \text{for some } \zeta_k \in \mathbb{C} \text{ with } |\zeta_k| = r. \tag{39}$$

We are going to compare two situations, one where $\zeta_1, \ldots, \zeta_m$ are randomly selected on the circle of radius $r$ and one where they are equispaced on this circle. In the latter situation, the optimal performance of system identification in $\mathcal{H}_2$ can be determined explicitly.

**Proposition 2** *Given $0 < r < 1$ and $\zeta_1, \ldots, \zeta_m \in \mathbb{D}$ defined by*

$$\zeta_k = r \exp\left(i\frac{2\pi}{m}(k-1)\right), \qquad k = 1, \ldots, m, \tag{40}$$

*the indicator of compatibility of the acquisition process $\mathcal{L}_\zeta$ and the approximability model relative to $\mathcal{P}_n$ is, for any $n \leq m$,*

$$\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n) = \frac{1}{\sqrt{1 - r^{2m}}}. \tag{41}$$

***Proof*** We shall prove the stronger statement that

$$G^* H^{-1} G = (1 - r^{2m}) I_n, \tag{42}$$

which clearly implies the announced result by taking (24) into account. To this end, we notice that the entries of the matrix $G \in \mathbb{C}^{m \times n}$ are

$$\begin{aligned} G_{k,j} = V_j(\zeta_k) = r^j \exp\left(i \frac{2\pi}{m}(k-1)(j-1)\right), \\ k = 1, \ldots, m, \quad j = 1, \ldots, n, \end{aligned} \tag{43}$$

and, in view of the form (7) of the representer $L_j = E_{\zeta_j}$, that the entries of the matrix $H \in \mathbb{C}^{m \times m}$ are

$$\begin{aligned} H_{k,j} = L_j(\zeta_k) = \frac{1}{1 - \overline{\zeta_j} \zeta_k} = \frac{1}{1 - r^2 \exp\left(i \frac{2\pi(k-j)}{m}\right)}, \\ k = 1, \ldots, m, \quad j = 1, \ldots, m. \end{aligned} \tag{44}$$

Since $H$ is a circulant matrix, it 'diagonalizes in Fourier,' meaning that the matrix $U$ in the eigendecomposition (27) has columns $u^{(1)}, \ldots, u^{(m)} \in \mathbb{C}^m$ given by

$$u^{(k)} = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 \\ \exp(i 2\pi (k-1)/m) \\ \vdots \\ \exp(i 2\pi (k-1)(m-1)/m) \end{bmatrix}. \tag{45}$$

The eigenvalues $d_1, \ldots, d_m$ are found through the calculation

$$\begin{aligned} (H u^{(k)})_{k'} &= \sum_{j=1}^{m} \frac{1}{1 - r^2 \exp(i 2\pi (k' - j)/m)} \frac{\exp(i 2\pi (k-1)(j-1))}{\sqrt{m}} \\ &= \sum_{j=1}^{m} \frac{\exp(i 2\pi (k-1)(j-k')/m)}{1 - r^2 \exp(i 2\pi (k' - j)/m)} (u^{(k)})_{k'}, \end{aligned} \tag{46}$$

so that, after the change of summation index $h = k' - j$, the eigenvalue $d_k$ appears to be

$$d_k = \sum_{h=1}^{m} \frac{\exp(-i 2\pi (k-1)h/m)}{1 - r^2 \exp(i 2\pi h/m)}$$

$$= \sum_{h=1}^{m} \exp(-i2\pi(k-1)h/m) \sum_{t=0}^{\infty} r^{2t} \exp(i2\pi t h/m)$$
$$= \sum_{t=0}^{\infty} r^{2t} \sum_{h=1}^{m} \exp(i2\pi(t-k+1)h/m) = \sum_{t=0}^{\infty} r^{2t}\, m\, \mathbf{1}_{\{t \in k-1+m\mathbb{N}\}}$$
$$= m(r^{2(k-1)} + r^{2(k-1+m)} + r^{2(k-1+2m)} + \cdots) = \frac{m r^{2(k-1)}}{1-r^{2m}}. \tag{47}$$

Thus, we have now made explicit the eigendecomposition of $H$ as

$$H = \frac{m}{1-r^{2m}}\, U\, \mathrm{diag}[1, r^2, \ldots, r^{2(m-1)}]\, U^*, \qquad U = \left[u^{(1)} | \cdots | u^{(m)}\right]. \tag{48}$$

Besides, we easily observe that the expression (43) reads, in matrix form,

$$G = \sqrt{m}\, \widetilde{U}\, \mathrm{diag}[1, r, \ldots, r^{n-1}], \qquad \widetilde{U} = \left[u^{(1)} | \cdots | u^{(n)}\right]. \tag{49}$$

At this point, it is an easy matter to verify that (48) and (49) imply (42). □

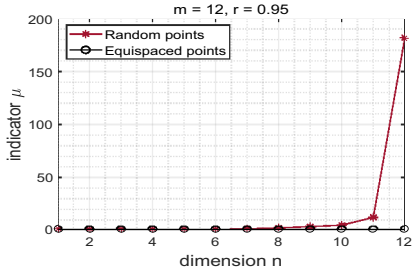
In the case of observations gathered at equispaced points, Proposition 2 shows that the indicator $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ is independent of the dimension $n$ of $\mathcal{V} = \mathcal{P}_n$. Keeping in mind that the value of $\varepsilon = \varepsilon_n$ appearing in (19) decreases with $n$ [e.g., as $\varepsilon_n = M\rho^{-n}$ for the model set imposed by (9)–(11)], the optimal performance becomes most favorable when $n$ as large as possible, i.e., when $n = m$. This differs from the typical situation where a trade-off is to be found between the increase of $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ and the decrease of $\varepsilon_n$. Figure 1 illustrates numerically the fact that $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ generally increases with $n$. The experiments that generated this figure can be reproduced by downloading from the authors' Web pages the MATLAB files accompanying this article.[2]

***Remark*** The choice $n = m$ implies the existence of some $A_m : \mathbb{C}^m \to \mathcal{H}_2(\mathbb{D})$ yielding the estimate

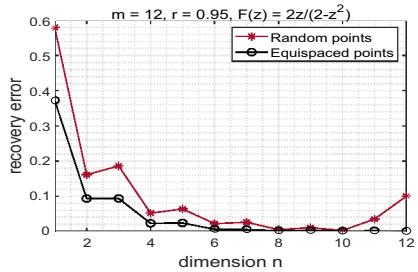
$$\|F - A_m(\mathcal{L}_\zeta(F))\|_{\mathcal{H}_2} \le \frac{1}{\sqrt{1-r^{2m}}}\, \mathrm{dist}_{\mathcal{H}_2}(F, \mathcal{P}_m) \qquad \text{for all } F \in \mathcal{H}_2(\mathbb{D}), \tag{50}$$

where the factor $1/\sqrt{1-r^{2m}}$ is bounded independently of $m$ by $1/\sqrt{1-r^2}$. As it turns out, the map $F \mapsto A_m(\mathcal{L}_\zeta(F))$ is just the operator $P_m : \mathcal{H}_2(\mathbb{D}) \to \mathcal{P}_m$ of polynomial interpolation at $\zeta_1, \ldots, \zeta_m$: Indeed, thanks to $\mathcal{L}_\zeta(P_m(F)) = \mathcal{L}_\zeta(F)$, (50) applied to $P_m(F) \in \mathcal{P}_m$ instead of $F$ reveals that $A_m(\mathcal{L}_\zeta(F)) = P_m(F)$. In short, the operator $P_m : \mathcal{H}_2(\mathbb{D}) \to \mathcal{P}_m$ of polynomial interpolation at $m$ equispaced points on the circle of radius $r < 1$ acts as an operator of near-best approximation from $\mathcal{P}_m$

[2] To compute the $\mathcal{H}_2$-error between functions $F = \sum_{j=1}^{\infty} b_j V_j$ and $\widetilde{F} = \sum_{j=1}^{n} c_j V_j + \sum_{k=1}^{m} d_k L_k$, we used the fact that $\|F - \widetilde{F}\|_{\mathcal{H}_2}^2 = \|F\|_{\mathcal{H}_2}^2 + \|\widetilde{F}\|_{\mathcal{H}_2}^2 - 2\,\mathrm{Re}\langle F, \widetilde{F}\rangle$, together with $\|\widetilde{F}\|_{\mathcal{H}_2}^2 = \|c\|_2^2 + \langle d, Hd\rangle + 2\,\mathrm{Re}\langle c, Gd\rangle$ and $\langle F, \widetilde{F}\rangle = \langle b_{1:n}, c\rangle + \sum_{k=1}^{m} \overline{d_k} \ell_k(F)$.




**(a)** The indicator $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ as a function of $n$ for equispaced points (no dependence) and for random points (fast increase).




**(b)** The identification error $\|F - A^{\text{opt}}(\mathcal{L}_\zeta(F))\|_{\mathcal{H}_2}$ as a function of $n$ for equispaced points and for random points.

**Fig. 1** Behavior of the indicator $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n)$ and of the identification error $\|F - A^{\text{opt}}(\mathcal{L}_\zeta(F))\|_{\mathcal{H}_2}$ when the points $\zeta_1, \ldots, \zeta_m$ are chosen on a circle of radius $r < 1$

relative to $\mathcal{H}_2$. As pointed out by a referee, this specific observation can be obtained without relying on the preceding machinery. Indeed, the expression

$$P_m(G)(z) = \sum_{\ell=0}^{m-1} \Big( \sum_{t=0}^{\infty} g_{\ell+tm} r^{tm} \Big) z^\ell \qquad \text{whenever } G(z) = \sum_{\ell=0}^{\infty} g_\ell z^\ell \tag{51}$$

is easily verified by checking that $P_m(G)(\zeta_k) = G(\zeta_k)$ for all $k = 1, \ldots, m$ and that $P_m(V_j) = V_j$ for all $j = 1, \ldots, m$. Then, an application of Cauchy–Schwarz inequality gives

$$\|P_m(G)\|_{\mathcal{H}_2} \le \frac{1}{\sqrt{1 - r^{2m}}} \|G\|_{\mathcal{H}_2}. \tag{52}$$

The inequality (50) finally follows by taking $G = F - V$ where $V \in \mathcal{P}_m$ is chosen such that $\|F - V\|_{\mathcal{H}_2} = \text{dist}_{\mathcal{H}_2}(F, \mathcal{P}_m)$. We emphasize that this short argument for (50) incidentally justifies that $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_m) \le 1/\sqrt{1 - r^{2m}}$, according to the leftmost inequality in (14), and in turn that $\mu_{\mathcal{H}_2}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n) \le 1/\sqrt{1 - r^{2m}}$ for all $n \le m$.

## 4 Optimal estimation and identification in $\mathcal{A}(\mathbb{D})$

As mentioned at the end of Sect. 2, the results (i)–(ii)–(iii) can also be enhanced when $\mathcal{X}$ is an arbitrary normed space and one does not target the full approximation of $F \in \mathcal{X}$ but only the estimation of a quantity of interest $Q(F)$ for some linear functional $Q \in \mathcal{X}^*$. To present the enhancement provided in [9], we let $(V_1, \ldots, V_n)$ denote a basis for the subspace $\mathcal{V}$ of $\mathcal{X}$—this time, it does not need to be an orthonormal basis. Concerning (iii), an algorithm which is *optimal* over the approximation set $\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V})$ was uncovered: It consists in outputting

$$A^{\mathrm{opt}}(y) = \sum_{k=1}^{m} a_k^\star y_k, \qquad y \in \mathbb{C}^m, \tag{53}$$

where the vector $a^\star \in \mathbb{C}^m$ is obtained as a solution of the optimization problem

$$\underset{a \in \mathbb{C}^m}{\text{minimize}} \left\| Q - \sum_{k=1}^{m} a_k \ell_k \right\|_{\mathcal{X}^*} \quad \text{subject to} \sum_{k=1}^{m} a_k \ell_k(V_j) = Q(V_j) \quad \text{for all } j = 1, \ldots, n. \tag{54}$$

Note that the vector $a^\star$ is computed offline once and for all and is subsequently used for each fresh occurrence of observational data $y$ via the rule (53), thus providing an optimal algorithm which is incidentally a *linear* functional. Note also that the knowledge of $\varepsilon$ is unnecessary to produce the vector $a^\star$. Concerning (ii), or rather its alteration to incorporate quantities of interest, the optimal performance over $\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V})$ is again completely characterized by

$$E^{\mathrm{opt}}(\mathcal{K}_{\mathcal{X}}(\varepsilon, \mathcal{V}), \mathcal{L}, Q) = \mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V}, Q)\, \varepsilon, \tag{55}$$

where the indicator of compatibility of the acquisition process and the model now becomes

$$\mu_{\mathcal{X}}(\ker(\mathcal{L}), \mathcal{V}, Q) := \sup_{F \in \ker(\mathcal{L})} \frac{|Q(F)|}{\mathrm{dist}_{\mathcal{X}}(F, \mathcal{V})}. \tag{56}$$

Concerning (i), the value of this indicator is in fact obtained as the minimum of the optimization program (54). At first sight, this program seems intractable because the dual norm is not automatically amenable to numerical computations. The purpose of this section is to show that the optimization can be performed effectively when the acquisition functionals are evaluations at points on the torus $\mathbb{T}$ and when $\mathcal{X}$ is the disk algebra $\mathcal{A}(\mathbb{D})$ endowed with the $\mathcal{H}_\infty$-norm. It could also be performed when $\mathcal{X} = \mathcal{H}_2(\mathbb{D})$—we do not give details about this case, since it is essentially covered in [9, Subsection 5.2], which is dedicated to reproducing kernel Hilbert spaces and hence applies to $\mathcal{H}_2(\mathbb{D})$.

### 4.1 Optimal performance and algorithm

The key to transforming (54) into a practically solvable optimization program when $\mathcal{X} = \mathcal{A}(\mathbb{D})$ consists in reformulating the $\mathcal{X}^*$-norm of a linear combination of linear functionals as a workable expression of its coefficients. With acquisition functionals being evaluations at points on the torus, this reformulation is made possible by Rudin–Carleson theorem. It provides access not only to the optimal performance of system identification in $\mathcal{A}(\mathbb{D})$ for the approximability model relative to a subspace $\mathcal{V}$, but also to an algorithm achieving this optimal performance. This is the object of the theorem below—to reiterate, it extends a result of [9] from the space of continuous functions to the more involved disk algebra.

**Theorem 3** *Given distinct points $\zeta_1, \dots, \zeta_m \in \mathbb{T}$ and a quantity of interest taking the form $Q(F) = F(\zeta_0)$ for some $\zeta_0 \in \mathbb{T} \setminus \{\zeta_1, \dots, \zeta_m\}$, consider the $\ell_1$-minimization problem*

$$\underset{a \in \mathbb{C}^m}{\text{minimize}} \sum_{k=1}^{m} |a_k| \quad \text{subject to} \sum_{k=1}^{m} a_k V_j(\zeta_k) = V_j(\zeta_0) \quad \text{for all } j = 1, \dots, n. \tag{57}$$

*If $a^\star \in \mathbb{C}^m$ denotes a solution to this problem, then*

$$\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}, Q) = 1 + \sum_{k=1}^{m} |a_k^\star|, \tag{58}$$

*and the linear functional $A^{\mathrm{opt}} : \mathbb{C}^m \to \mathbb{C}$ defined by*

$$A^{\mathrm{opt}}(y) = \sum_{k=1}^{m} a_k^\star y_k, \qquad y \in \mathbb{C}^m, \tag{59}$$

*is an optimal algorithm over $\mathcal{K}_{\mathcal{A}}(\varepsilon, \mathcal{V})$ in the sense that*

$$\sup_{F \in \mathcal{K}_{\mathcal{A}}(\varepsilon, \mathcal{V})} |Q(F) - A^{\mathrm{opt}}(\mathcal{L}_\zeta(F))| = \inf_{A : \mathbb{C}^m \to \mathbb{C}} \sup_{F \in \mathcal{K}_{\mathcal{A}}(\varepsilon, \mathcal{V})} |Q(F) - A(\mathcal{L}_\zeta(F))|, \tag{60}$$

*with the value of the latter being $\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}, Q)\, \varepsilon$.*

***Proof*** The argument is based on the remark that, for any $a \in \mathbb{C}^m$,

$$\left\| e_{\zeta_0} - \sum_{k=1}^{m} a_k e_{\zeta_k} \right\|_{\mathcal{A}^*} = 1 + \sum_{k=1}^{m} |a_k|. \tag{61}$$

The '$\le$'-part follows from the triangle inequality and the observation (8). As for the '$\ge$'-part, since the set $\{\zeta_0, \zeta_1, \dots, \zeta_m\} \subseteq \mathbb{T}$ is closed and have measure zero, Rudin–Carleson theorem (see e.g., [13, Theorem 2.3.2]) ensures that one can find $F \in \mathcal{A}(\mathbb{D})$ such that

$$\begin{aligned} &\|F\|_{\mathcal{H}_\infty} = 1, \qquad F(\zeta_0) = 1 =: w_0, \\ &F(\zeta_k) = -\frac{\overline{a_k}}{|a_k|} =: w_k, \quad k = 1, \dots, m, \end{aligned} \tag{62}$$

which in turn implies that

$$\left\| e_{\zeta_0} - \sum_{k=1}^{m} a_k e_{\zeta_k} \right\|_{\mathcal{A}^*} \ge F(\zeta_0) - \sum_{k=1}^{m} a_k F(\zeta_k) = 1 + \sum_{k=1}^{m} |a_k|. \tag{63}$$

It is then clear from (61) that the optimization program (54) reformulates as (57), omitting the additive constant 1. The announced results are now restatements of (53)–(54) and of the announced equality between the indicator (56) and the minimum of the optimization program. □

***Remark*** The argument (62) justifying the identity (61) would not hold if $\zeta_0, \zeta_1, \ldots, \zeta_m$ were all inside the unit circle. Indeed, by Nevanlinna–Pick theorem (see e.g., [13, Theorem 2.1.6] or [5, Theorem 2.3.4]), it would mean that the matrix $\mathbf{Q} \in \mathbb{C}^{(m+1)\times(m+1)}$ with entries

$$Q_{k,j} = \frac{1 - \overline{w_j} w_k}{1 - \overline{\zeta_j} \zeta_k}, \qquad k, j = 0, 1, \ldots, m, \tag{64}$$

is positive semidefinite. This cannot occur because its diagonal entries are all zero, hence its trace equals zero, so $\mathbf{Q}$ itself would be the zero matrix.

***Remark*** Theorem 3 can effortlessly be extended to quantities of interest taking the form of a weighted sum of evaluations at points on $\mathbb{T}$. With a little more work, it could also be extended to quantities of interest of the form $Q(F) = F^{(s)}(\zeta)$ for some $\zeta \in \mathbb{D}$. The details are left to the reader. They essentially amount to justifying the '$\geq$'-part of an analog of (61). This is done via a limiting argument which involves a discretization of the Cauchy formula for $F^{(s)}(\zeta)$ and allows for Rudin–Carleson theorem to be applied.

### 4.2 Evaluations at points on the unit circle

In this subsection, general considerations are again particularized to our situation of interest where we fix $\mathcal{V}$ to be the space of polynomials of degree at most $n-1$. In this case, and with acquisition functionals being evaluations at equispaced points on the torus $\mathbb{T}$, we are able to determine quite explicitly the optimal performance of system identification in $\mathcal{A}(\mathbb{D})$. Underlying the argument is a crucial observation about $\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V})$ valid regardless of the space $\mathcal{V}$ and of the evaluation points $\zeta_1, \ldots, \zeta_m \in \mathbb{T}$. This observation is isolated below.

**Lemma 4** *Given a finite-dimensional subspace $\mathcal{V}$ of $\mathcal{A}(\mathbb{D})$ and distinct points $\zeta_1, \ldots, \zeta_m \in \mathbb{T}$,*

$$\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}) = 1 + \sup_{V \in \mathcal{V}} \frac{\|V\|_{\mathcal{H}_\infty}}{\max\limits_{k=1,\ldots,m} |V(\zeta_k)|}. \tag{65}$$

***Proof*** In order to establish (65), we first recall that

$$\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}) = \sup_{F(\zeta_1)=\cdots=F(\zeta_m)=0} \frac{\|F\|_{\mathcal{H}_\infty}}{\mathrm{dist}_{\mathcal{H}_\infty}(F, \mathcal{V})}. \tag{66}$$

To prove the '$\leq$'-part of (65), we remark that, if $F \in \mathcal{A}(\mathbb{D})$ satisfies $F(\zeta_1) = \cdots = F(\zeta_m) = 0$ and $V \in \mathcal{V}$ satisfies $\|F - V\|_{\mathcal{H}_\infty} = \mathrm{dist}_{\mathcal{H}_\infty}(F, \mathcal{V})$, then

$$\begin{aligned} \frac{\|F\|_{\mathcal{H}_\infty}}{\mathrm{dist}_{\mathcal{H}_\infty}(F, \mathcal{V})} &= \frac{\|F\|_{\mathcal{H}_\infty}}{\|F - V\|_{\mathcal{H}_\infty}} \leq 1 + \frac{\|V\|_{\mathcal{H}_\infty}}{\|F - V\|_{\mathcal{H}_\infty}} \\ &\leq 1 + \frac{\|V\|_{\mathcal{H}_\infty}}{\max\limits_{k=1,\ldots,m} |(F - V)(\zeta_k)|} \\ &= 1 + \frac{\|V\|_{\mathcal{H}_\infty}}{\max\limits_{k=1,\ldots,m} |V(\zeta_k)|}, \end{aligned} \tag{67}$$

and the required inequality immediately follows. As for the '$\geq$'-part, let $\eta$ denote the maximum appearing in (65), and let us select $V \in \mathcal{V}$ with $\max\limits_{k=1,\ldots,m} |V(\zeta_k)| = 1$ and $\|V\|_{\mathcal{H}_\infty} = \eta$. We then pick $z \in \mathbb{T}$ such that $|V(z)| = \eta$—we may assume $z \notin \{\zeta_1, \ldots, \zeta_m\}$, otherwise we can slightly perturb it and replace $\eta$ by $\eta - \varepsilon$ for an arbitrary small $\varepsilon > 0$. By Rudin–Carleson theorem, we can find $H \in \mathcal{A}(\mathbb{D})$ such that

$$\|H\|_{\mathcal{H}_\infty} = 1, \qquad H(z) = -\frac{V(z)}{|V(z)|}, \qquad H(\zeta_k) = V(\zeta_k), \quad k = 1, \ldots, m. \tag{68}$$

Then, since $F = V - H \in \mathcal{A}(\mathbb{D})$ satisfies $F(\zeta_1) = \cdots = F(\zeta_m) = 0$, we have

$$\begin{aligned} \mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}) &\geq \frac{\|F\|_{\mathcal{H}_\infty}}{\mathrm{dist}_{\mathcal{H}_\infty}(F, \mathcal{V})} \geq \frac{\|V - H\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}} \\ &\geq |V(z) - H(z)| = |V(z)| \left(1 + \frac{1}{|V(z)|}\right) \\ &= 1 + \eta. \end{aligned} \tag{69}$$

This is the required inequality. □

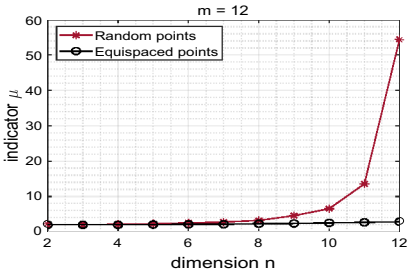
With Lemma 4 at our disposal, it becomes almost immediate to establish a result about optimal identification in $\mathcal{A}(\mathbb{D})$ for acquisition functionals being evaluations at equispaced points. There is a direct repercussion on optimal estimation in $\mathcal{A}(\mathbb{D})$—the setting of Theorem 3—since

$$\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}) = \sup_{\zeta_0 \in \mathbb{T}} \mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{V}, e_{\zeta_0}). \tag{70}$$

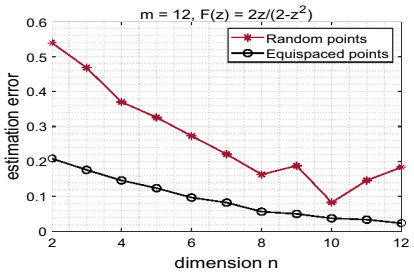
The awaited result about optimal identification reads as follows.

**Proposition 5** *Given points $\zeta_1, \ldots, \zeta_m \in \mathbb{T}$ defined by*

$$\zeta_k = \exp\left(i \frac{2\pi}{m}(k - 1)\right), \qquad k = 1, \ldots, m, \tag{71}$$




**(a)** The indicator $\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n, e_{\zeta_0})$ as a function of $n$ for equispaced points and for random points.




**(b)** The estimation error $|F(\zeta_0) - A^{\text{opt}}(\mathcal{L}_\zeta(F))|$ as a function of $n$ for equispaced points and for random points.

**Fig. 2** Behavior of the indicator $\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n, e_{\zeta_0})$ and of the estimation error $|F(\zeta_0) - A^{\text{opt}}(\mathcal{L}_\zeta(F))|$ when the points $\zeta_0, \zeta_1, \ldots, \zeta_m$ are chosen on the torus

*the indicator of compatibility of the acquisition process $\mathcal{L}_\zeta$ and the approximability model relative to $\mathcal{P}_n$ satisfies*

$$\mu_{\mathcal{A}}(\ker(\mathcal{L}_\zeta), \mathcal{P}_n) = 2 + \kappa(m, n), \qquad \textit{with} \quad \kappa(m, n) \asymp \ln\left(\frac{m}{m - n + 1}\right). \tag{72}$$

***Proof*** We invoke the result of [14], which precisely says that

$$\sup_{V \in \mathcal{P}_n} \frac{\|V\|_{\mathcal{H}_\infty}}{\max\limits_{k=1,\ldots,m} |V(\zeta_k)|} = 1 + \kappa(m, n), \qquad \text{with}$$
$$\kappa(m, n) \asymp \ln\left(\frac{m}{m - n + 1}\right). \tag{73}$$

It remains to take Lemma 4 into account for the proof to be complete. □

To close this subsection, we present in Fig. 2 a brief numerical illustration comparing the optimal estimation algorithms of Theorem 3 for equispaced points and random points on the torus. The experiment is also included in the reproducible MATLAB file.

## 5 Conclusion

In this article, we formulated an optimal system identification problem by expressing the *a priori* information via approximability properties. To the best of our knowledge, this is the first work in this direction. We partially solved our problem by devising methods for the construction of optimal algorithms, which turn out to be linear algorithms. Yet, our contribution is certainly an exploratory work that should be extended in several directions. We highlight a few possible directions below.

*Measurement types:* We have mostly considered frequency responses obtained as point evaluations of transfer functions, and we have obtained sharp results when the evaluation points are equispaced. Our underlying method, however, is valid for any

type of linear measurements, so it is worth studying its repercussions for other relevant measurement scenarios, including (i) impulse response samples, (ii) convolutions of the impulse response with (pseudorandom) input signals, and (iii) pairs of input–output time series.

*Measurements quality:* The initial results presented here were based on the assumption that the observational data can be acquired with perfect accuracy. In realistic situations, errors always occur in the acquisition process. It is possible to formulate an optimal recovery problem taking this inaccuracy into account. (In fact, such a problem can be transformed onto a standard optimal recovery problem, at least formally—see [12] for details.) In the context of system identification, it would be valuable to design implementable algorithms that are optimal in this inaccurate setting.

*MIMO systems:* The focus in this preliminary work was put on single-input single-output (SISO) systems. A step further would consist in treating multi-input multi-output (MIMO) systems without simply applying our techniques to each input–output pair separately, as this becomes inefficient for large-scale systems.

## Appendix: Proofs of essential results

In this section, we fully justify some statements appearing in the text but not yet established, namely the relation between (9)–(10) and (11), as well as the validity in the complex setting of results about optimal identification in Hilbert spaces [2] and optimal estimation in Banach spaces [9]. We start with how (11) connects to the descriptions (9)–(10) of the models put forward in [11].

**Proposition 6** *With $\mathcal{X}$ denoting either $\mathcal{H}_2(\mathbb{D})$ or $\mathcal{A}(\mathbb{D})$, the following properties are equivalent:*

$$\text{there exist } \rho > 1 \text{ and } M > 0 \text{ such that } \|F(\rho\,\cdot)\|_{\mathcal{X}} \le M; \tag{74}$$

$$\text{there exist } \rho > 1 \text{ and } M > 0 \text{ such that } \operatorname{dist}_{\mathcal{X}}(F, \mathcal{P}_n) \le M\rho^{-n} \text{ for all } n \ge 0. \tag{75}$$

***Proof*** We write $F(z) = \sum_{n=0}^{\infty} f_n z^n$ throughout the proof. We first establish the equivalence in the case $\mathcal{X} = \mathcal{H}_2(\mathbb{D})$. Let us assume that (74) holds, i.e., that $\sum_{n=0}^{\infty} |f_n|^2 \rho^{2n} \le M^2$ for some $\rho > 1$ and $M > 0$. In particular, we have $|f_n|^2 \le M^2 \rho^{-2n}$ for all $n \ge 0$. It follows that, for all $n \ge 0$,

$$\operatorname{dist}_{\mathcal{H}_2}(F, \mathcal{P}_n)^2 = \sum_{k=n}^{\infty} |f_k|^2 \le M^2 \sum_{k=n}^{\infty} \rho^{-2k} = \frac{M^2}{1-\rho^{-2}} \rho^{-2n}; \tag{76}$$

hence, (75) holds with a change in the constant $M$. Conversely, let us assume that (75) holds, i.e., that there are $\rho > 1$ and $M > 0$ such that $\sum_{k=n}^{\infty} |f_k|^2 \le M^2 \rho^{-2n}$ for all $n \ge 0$. In particular, we have $|f_n|^2 \le M^2 \rho^{-2n}$ for all $n \ge 0$. Then, picking $\widetilde{\rho} \in (1, \rho)$, we derive that

$$\sum_{n=0}^{\infty} |f_n|^2 \widetilde{\rho}^{\,2n} \le M^2 \sum_{n=0}^{\infty} (\widetilde{\rho}/\rho)^{2n} = \frac{M^2}{1-(\widetilde{\rho}/\rho)^2}; \tag{77}$$

hence, (74) holds with a change in both $\rho$ and $M$.

We now establish the equivalence in the case $\mathcal{X} = \mathcal{A}(\mathbb{D})$. Let us assume that (74) holds, i.e., that $\sup_{|z|=\rho} |F(z)| \leq M$ for some $\rho > 1$ and $M > 0$. This implies that the Taylor coefficients of $F$ satisfy, for any $k \geq 0$,

$$|f_k| = \left| \frac{1}{2\pi i} \int_{|z|=\rho} \frac{F(z)}{z^{k+1}} dz \right| \leq \frac{1}{2\pi} \times \frac{M}{\rho^{k+1}} \times 2\pi\rho = M\rho^{-k}. \tag{78}$$

Considering $P \in \mathcal{P}_n$ defined by $P(z) := \sum_{k=0}^{n-1} f_k z^k$, we obtain

$$\begin{aligned} \mathrm{dist}_{\mathcal{X}}(F, \mathcal{P}_n) \leq \|F - P\|_{\mathcal{H}_\infty} &= \sup_{|z|=1} \left| \sum_{k=n}^{\infty} f_k z^k \right| \leq \sum_{k=n}^{\infty} |f_k| \\ &\leq M \sum_{k=n}^{\infty} \rho^{-k} = \frac{M}{1-\rho} \rho^{-n}; \end{aligned} \tag{79}$$

hence, (75) holds with a change in the constant $M$. Conversely, let us assume that (75) holds, i.e., that there are $\rho > 1$ and $M > 0$ such that there exists, for each $n \geq 0$, a polynomial $P^{[n]} \in \mathcal{P}_n$ with $\|F - P^{[n]}\|_{\mathcal{H}_\infty} \leq M\rho^{-n}$. For all $n \geq 0$, since the coefficients in $z^n$ of $F$ are the same as that of $F - P^{[n]}$, we have

$$\begin{aligned} |f_n| &= \left| \frac{1}{2\pi i} \int_{|z|=1} \frac{(F - P^{[n]})(z)}{z^{n+1}} dz \right| \\ &\leq \frac{1}{2\pi} \times \|F - P^{[n]}\|_{\mathcal{H}_\infty} \times 2\pi \leq M\rho^{-n}. \end{aligned} \tag{80}$$

Then, picking $\widetilde{\rho} \in (1, \rho)$, we derive that

$$\sup_{|z|=\widetilde{\rho}} |F(z)| \leq \sum_{n=0}^{\infty} |f_n| \widetilde{\rho}^{\,n} \leq M \sum_{n=0}^{\infty} (\widetilde{\rho}/\rho)^n = \frac{M}{1 - \widetilde{\rho}/\rho}; \tag{81}$$

hence, (74) holds with a change in both $\rho$ and $M$. □

We now turn to the justification for the complex setting of the results from [2] about optimal identification in Hilbert spaces. As in [2, Theorem 2.8], these results are easy consequences of the following statement.

**Theorem 7** *Let $\mathcal{V}$ be a subspace of a Hilbert space $\mathcal{X}$, and let $\ell_1, \ldots, \ell_m$ be linear functionals defined on $\mathcal{X}$. With a model set given by*

$$\mathcal{K} = \{f \in \mathcal{X} : \mathrm{dist}_{\mathcal{X}}(f, \mathcal{V}) \leq \varepsilon\}, \tag{82}$$

*the performance of optimal identification from some $y \in \mathbb{C}^m$ satisfies*

$$\inf_{A:\mathbb{C}^m \to \mathcal{X}} \sup_{f \in \mathcal{K} \cap \mathcal{L}^{-1}(y)} \|f - A(y)\|_{\mathcal{X}} = \mu \left( \varepsilon^2 - \min_{f \in \mathcal{L}^{-1}(y)} \|f - P_{\mathcal{V}} f\|_{\mathcal{X}}^2 \right)^{1/2}, \tag{83}$$

*where the constant $\mu$ is defined by*

$$\mu = \sup_{u \in \ker(\mathcal{L})} \frac{\|u\|_{\mathcal{X}}}{\operatorname{dist}_{\mathcal{X}}(u, \mathcal{V})}. \tag{84}$$

***Proof*** Let $f^\star \in \mathcal{X}$ be constructed from $y \in \mathbb{C}^m$ via $f^\star := \operatorname*{argmin}_{f \in \mathcal{X}} \|f - P_{\mathcal{V}} f\|_{\mathcal{X}}$ subject to $\mathcal{L}(f) = y$. We shall prove on the one hand that

$$\sup_{f \in \mathcal{K} \cap \mathcal{L}^{-1}(y)} \|f - f^\star\|_{\mathcal{X}} \leq \mu \left(\varepsilon^2 - \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2\right)^{1/2} \tag{85}$$

and on the other hand that, for any $g \in \mathcal{X}$,

$$\sup_{f \in \mathcal{K} \cap \mathcal{L}^{-1}(y)} \|f - g\|_{\mathcal{X}} \geq \mu \left(\varepsilon^2 - \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2\right)^{1/2}. \tag{86}$$

Justification of (85): Let us point out that $f^\star - P_{\mathcal{V}} f^\star$ is orthogonal to both $\mathcal{V}$ and $\ker(\mathcal{L})$. To see this, given $v \in \mathcal{V}$, $u \in \ker(\mathcal{L})$, and $\theta \in [-\pi, \pi]$, we notice that, as functions of $t \in \mathbb{R}$, the expressions

$$\begin{aligned} \|f^\star - P_{\mathcal{V}} f^\star + t e^{i\theta} v\|_{\mathcal{X}}^2 &= \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2 \\ &\quad + 2t \operatorname{Re}(e^{-i\theta} \langle f^\star - P_{\mathcal{V}} f^\star, v \rangle) + \mathcal{O}(t^2), \end{aligned} \tag{87}$$

$$\begin{aligned} \|f^\star + t e^{i\theta} u - P_{\mathcal{V}}(f^\star + t e^{i\theta} u)\|_{\mathcal{X}}^2 &= \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2 \\ &\quad + 2t \operatorname{Re}(e^{-i\theta} \langle f^\star - P_{\mathcal{V}} f^\star, u - P_{\mathcal{V}} u \rangle) + \mathcal{O}(t^2), \end{aligned} \tag{88}$$

are minimized at $t = 0$. Therefore, $\operatorname{Re}(e^{-i\theta} \langle f^\star - P_{\mathcal{V}} f^\star, v \rangle) = 0$ and $\operatorname{Re}(e^{-i\theta} \langle f^\star - P_{\mathcal{V}} f^\star, u - P_{\mathcal{V}} u \rangle) = 0$ for all $\theta \in [-\pi, \pi]$. This implies that $\langle f^\star - P_{\mathcal{V}} f^\star, v \rangle = 0$ and $\langle f^\star - P_{\mathcal{V}} f^\star, u - P_{\mathcal{V}} u \rangle = 0$ for all $v \in \mathcal{V}$ and $u \in \ker(\mathcal{L})$, hence our claim. Now, consider $f \in \mathcal{K} \cap \mathcal{L}^{-1}(y)$. Since $\mathcal{L}(f) = y = \mathcal{L}(f^\star)$, we can write $f = f^\star + u$ for some $u \in \ker(\mathcal{L})$. The fact that $f \in \mathcal{K}$ then yields

$$\begin{aligned} \varepsilon^2 \geq \|f - P_{\mathcal{V}} f\|_{\mathcal{X}}^2 &= \|f^\star - P_{\mathcal{V}} f^\star + u - P_{\mathcal{V}} u\|_{\mathcal{X}}^2 \\ &= \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2 + \|u - P_{\mathcal{V}} u\|_{\mathcal{X}}^2, \end{aligned} \tag{89}$$

so that

$$\operatorname{dist}_{\mathcal{X}}(u, \mathcal{V}) = \|u - P_{\mathcal{V}} u\|_{\mathcal{X}} \leq \left(\varepsilon^2 - \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2\right)^{1/2}. \tag{90}$$

It remains to take the definition of $\mu$ into account to obtain

$$\|f - f^\star\|_{\mathcal{X}} = \|u\|_{\mathcal{X}} \leq \mu \operatorname{dist}_{\mathcal{X}}(u, \mathcal{V}) \leq \mu \left(\varepsilon^2 - \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2\right)^{1/2}. \tag{91}$$

Justification of (86): Let us select $u \in \ker(\mathcal{L})$ such that

$$\|u\|_{\mathcal{X}} = \mu \operatorname{dist}_{\mathcal{X}}(u, \mathcal{V}) \qquad \text{and} \qquad \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2 + \|u - P_{\mathcal{V}} u\|_{\mathcal{X}}^2 = \varepsilon^2. \quad (92)$$

We now consider $f^\pm := f^\star \pm u$. It is clear that $f^\pm \in \mathcal{L}^{-1}(y)$, and we also have $f^\pm \in \mathcal{K}$, since

$$\begin{aligned} \|f^\pm - P_{\mathcal{V}} f^\pm\|_{\mathcal{X}}^2 &= \|(f^\star - P_{\mathcal{V}} f^\star) \pm (u - P_{\mathcal{V}} u)\|_{\mathcal{X}}^2 \\ &= \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2 + \|u - P_{\mathcal{V}} u\|_{\mathcal{X}}^2 = \varepsilon^2. \end{aligned} \quad (93)$$

Then, for any $g \in \mathcal{X}$,

$$\begin{aligned} \sup_{f \in \mathcal{K} \cap \mathcal{L}^{-1}(y)} \|f - g\|_{\mathcal{X}} &\geq \max_{\pm} \|f^\pm - g\|_{\mathcal{X}} \geq \frac{1}{2} \left( \|f^+ - g\|_{\mathcal{X}} + \|f^- - g\|_{\mathcal{X}} \right) \\ &\geq \frac{1}{2} \|f^+ - f^-\|_{\mathcal{X}} \\ &= \|u\|_{\mathcal{X}} = \mu \operatorname{dist}_{\mathcal{X}}(u, \mathcal{V}) = \mu \left( \varepsilon^2 - \|f^\star - P_{\mathcal{V}} f^\star\|_{\mathcal{X}}^2 \right)^{1/2}. \end{aligned} \quad (94)$$

This completes the proof of the theorem. □

Finally, we justify below that the result from [9] about optimal estimation in Banach spaces holds in the complex setting, too.

**Theorem 8** *Let $\mathcal{V}$ be a subspace of a Banach space $\mathcal{X}$, let $\ell_1, \ldots, \ell_m$ be linear functionals defined on $\mathcal{X}$, and let $Q$ be another linear functional defined on $\mathcal{X}$. With a model set given by*

$$\mathcal{K} = \{f \in \mathcal{X} : \operatorname{dist}_{\mathcal{X}}(f, \mathcal{V}) \leq \varepsilon\}, \quad (95)$$

*the performance of optimal estimation of $Q$ satisfies*

$$\inf_{A : \mathbb{C}^m \to \mathbb{C}} \sup_{f \in \mathcal{K}} |Q(f) - A(\mathcal{L}(f))| = \mu \, \varepsilon, \quad (96)$$

*where the constant $\mu$ equals the minimum of the optimization problem*

$$\begin{aligned} &\underset{a \in \mathbb{C}^m}{\text{minimize}} \left\| Q - \sum_{k=1}^m a_k \ell_k \right\|_{\mathcal{X}^*} \qquad \textit{subject to} \qquad \sum_{k=1}^m a_k \ell_k(v) = Q(v) \\ &\quad \textit{for all } v \in \mathcal{V}. \end{aligned} \quad (97)$$

***Proof*** Let $a^\star \in \mathbb{C}^m$ be a minimizer of the optimization program (97), and let $\nu$ denote the value of the minimum. Let us also consider

$$\mu = \sup_{u \in \mathcal{U}} \frac{|Q(u)|}{\operatorname{dist}_{\mathcal{X}}(u, \mathcal{V})}. \quad (98)$$

We shall prove on the one hand that

$$\sup_{f\in\mathcal{K}}\left|Q(f)-\sum_{k=1}^{m}a_k^\star\ell_k(f)\right|\le\nu\,\varepsilon,\tag{99}$$

and on the other hand that, for any $A:\mathbb{C}^m\to\mathbb{C}$,

$$\sup_{f\in\mathcal{K}}|Q(f)-A(\mathcal{L}(f))|\ge\mu\,\varepsilon,\tag{100}$$

and we shall show as a last step that

$$\nu\le\mu.\tag{101}$$

Justification of (99): Given $f\in\mathcal{K}$, we select $v\in\mathcal{V}$ such that $\|f-v\|_{\mathcal{X}}=\operatorname{dist}_{\mathcal{X}}(f,\mathcal{V})$. The required inequality follows by noticing that

$$\begin{aligned}\left|Q(f)-\sum_{k=1}^{m}a_k^\star\ell_k(f)\right|&=\left|Q(f-v)-\sum_{k=1}^{m}a_k^\star\ell_k(f-v)\right|\\&\le\left\|Q-\sum_{k=1}^{m}a_k^\star\ell_k\right\|_{\mathcal{X}^*}\|f-v\|_{\mathcal{X}}\\&=\nu\operatorname{dist}_{\mathcal{X}}(f,\mathcal{V})\le\nu\,\varepsilon.\end{aligned}\tag{102}$$

Justification of (100): Let us select $u\in\ker(\mathcal{L})$ such that

$$|Q(u)|=\mu\operatorname{dist}_{\mathcal{X}}(u,\mathcal{V})\qquad\text{and}\qquad\operatorname{dist}_{\mathcal{X}}(u,\mathcal{V})=\varepsilon.\tag{103}$$

Then, for any $A:\mathbb{C}^m\to\mathbb{C}$, we have

$$\begin{aligned}\sup_{f\in\mathcal{K}}|Q(f)-A(\mathcal{L}(f))|&\ge\max_{\pm}|Q(\pm u)-A(0)|\\&\ge\frac{1}{2}\left(|Q(u)-A(0)|+|Q(-u)-A(0)|\right)\\&\ge\frac{1}{2}|Q(u)-Q(-u)|=|Q(u)|=\mu\,\varepsilon.\end{aligned}\tag{104}$$

Justification of (101): We assume that $\ker(\mathcal{L})\cap\mathcal{V}=\{0\}$, otherwise $\mu=\infty$ and there is nothing to prove. We consider a linear functional $\lambda$ defined on $\ker(\mathcal{L})\oplus\mathcal{V}$ by

$$\lambda(u)=Q(u)\quad\text{for all }u\in\ker(\mathcal{L})\qquad\text{and}\qquad\lambda(v)=0\quad\text{for all }v\in\mathcal{V}.\tag{105}$$

We then consider a Hahn–Banach extension $\widetilde{\lambda}$ of $\lambda$ defined on $\mathcal{X}$. Because $Q-\widetilde{\lambda}$ vanishes on $\ker(\mathcal{L})$, we can write $Q-\widetilde{\lambda}=\sum_{k=1}^{n}\widetilde{a}_k\ell_k$ for some $\widetilde{a}\in\mathbb{C}^m$, and because

$\widetilde{\lambda}$ vanishes on $\mathcal{V}$, we have $\sum_{k=1}^{m} \widetilde{a}_k \ell_k(v) = Q(v)$ for all $v \in \mathcal{V}$. We therefore derive that

$$\nu \le \left\| Q - \sum_{k=1}^{n} \widetilde{a}_k \ell_k \right\|_{\mathcal{X}^*} = \left\| \widetilde{\lambda} \right\|_{\mathcal{X}^*} = \|\lambda\|_{(\ker(\mathcal{L})\oplus\mathcal{V})^*} = \sup_{\substack{u\in\ker(\mathcal{L}) \\ v\in\mathcal{V}}} \frac{|\lambda(u-v)|}{\|u-v\|_{\mathcal{X}}}$$

$$= \sup_{\substack{u\in\ker(\mathcal{L}) \\ v\in\mathcal{V}}} \frac{|Q(u)|}{\|u-v\|_{\mathcal{X}}} = \sup_{u\in\ker(\mathcal{L})} \frac{|Q(u)|}{\operatorname{dist}_{\mathcal{X}}(u,\mathcal{V})} = \mu. \tag{106}$$

This concludes the proof of the theorem. □

## References


1. Akcay H, Gu G, Khargonekar PP (1994) Identification in $\mathcal{H}_\infty$ with nonuniformly spaced frequency response measurements. Int J Robust Nonlinear Control 4(4):613–629
2. Binev P, Cohen A, Dahmen W, DeVore R, Petrova G, Wojtaszczyk P (2017) Data assimilation in reduced modeling. SIAM/ASA J Uncertain Quantif 5(1):1–29
3. Bokor J, Schipp F, Gianone L (1995) Approximate $\mathcal{H}_\infty$ identification using partial sum operators in disc algebra basis. In: Proceedings of American control conference, pp 1981–1985
4. Campi MC, Weyer E (2002) Finite sample properties of system identification methods. IEEE Trans Autom Control 47(8):1329–1334
5. Chen J, Gu G (2000) Control-oriented system identification: an $\mathcal{H}_\infty$ approach, vol 19. Wiley, Hoboken
6. Chen J, Nett CN (1993) The Caratheodory–Fejer problem and $\mathcal{H}_\infty$ identification: a time domain approach. In: Proceedings of the 32nd IEEE conference on decision and control, pp 68–73
7. Chen J, Nett CN, Fan MKH (1992) Worst-case system identification in $\mathcal{H}_\infty$: validation of a priori information, essentially optimal algorithms, and error bounds. In: Proceedings of American control conference, pp 251–257
8. Chen J, Nett CN, Fan MKH (1992) Optimal non-parametric system identification from arbitrary corrupt finite time series: a control-oriented approach. In: Proceedings of American control conference, pp 279–285
9. DeVore R, Foucart S, Petrova G, Wojtaszczyk P (2019) Computing a quantity of interest from observational data. Constr Approx 49(3):461–508
10. DeVore R, Petrova G, Wojtaszczyk P (2017) Data assimilation and sampling in Banach spaces. Calcolo 54(3):963–1007
11. Helmicki AJ, Jacobson CA, Nett CN (1991) Control oriented system identification: a worst-case/deterministic approach in $\mathcal{H}_\infty$. IEEE Trans Autom Control 36(10):1163–1176
12. Micchelli C, Rivlin T (1977) A survey of optimal recovery. Optimal estimation in approximation theory. Springer, Boston, pp 1–54
13. Partington JR (1997) Interpolation, identification, and sampling. Oxford University Press, Oxford
14. Rakhmanov E, Shekhtman B (2006) On discrete norms of polynomials. J Approx Theory 139(1–2):2–7
15. Shah P, Bhaskar BN, Tang G, Recht B (2012) Linear system identification via atomic norm regularization. arXiv preprint arXiv:1204.0590
16. Tu S, Boczar R, Packard A, Recht B (2017) Non-asymptotic analysis of robust control from coarse-grained identification. arXiv preprint arXiv:1707.04791
17. Vidyasagar M, Karandikar RL (2008) A learning theory approach to system identification and stochastic adaptive control. J Process Control 18(3):421–430
18. Zhou K, Doyle JC (1998) Essentials of robust control. Prentice Hall, Upper Saddle River