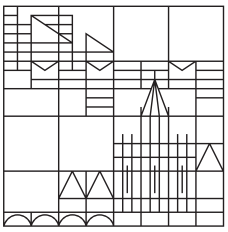
On the Convergence of a Greedy Algorithm for Operator Reconstruction

Master's Thesis

by
Simon Buchwald

at the

Universität Konstanz



Working Group Numerical Optimization
Department of Mathematics and Statistics

1. Evaluated by	Jun.-Prof. Dr. Gabriele Ciaramella
2. Evaluated by	Prof. Dr. Stefan Volkwein

Konstanz, June 2020

Abstract

In [MS09], Yvon Maday and Julien Salomon introduce an algorithm for the reconstruction of operators in the context of quantum systems. This algorithm shows good results in the numerical application. However, no convergence theory has been developed so far. In this work we investigate the algorithm in the setting of a linear ordinary differential equation, present problematic cases and establish assumptions under which convergence is guaranteed. We also develop working improvements to the algorithm, demonstrate their performance in numerical examples and discuss numerical instabilities.
Hereafter, we give a detailed description of the algorithm in its original form and setting. In this context we introduce monotonic schemes as an efficient method to solve optimal control problems for quantum systems.
Finally, we apply one of the improvements from the linear setting and show, with the aid of numerical experiments, that it also has a positive effect on the algorithm's performance for the solution of reconstruction problems governed by quantum systems.

Zusammenfassung

In [MS09] entwickeln Yvon Maday und Julien Salomon einen Algorithmus zur Rekonstruktion von Operatoren im Kontext von Quantensystemen, der gute Ergebnisse in der numerischen Anwendung liefert. Bislang wurde hierfür jedoch noch keinerlei Konvergenztheorie vorgestellt. In dieser Arbeit betrachten wir den Algorithmus im Rahmen einer linearen gewöhnlichen Differentialgleichung, zeigen problematische Fälle auf und erarbeiten Annahmen, unter denen die Konvergenz garantiert werden kann. Zudem entwickeln wir Verbesserungen für den Algorithmus, demonstrieren ihre Performance in numerischen Experimenten und diskutieren numerische Instabilitäten.
Anschließend geben wir eine detaillierte Beschreibung des Algorithmus in seiner ursprünglichen Form und seinem ursprünglichen Rahmen. In diesem Zusammenhang führen wir sogenannte "monotonic schemes" ein, als effiziente Methode zur Lösung von Optimalsteuerungsproblemen für Quantensystemen.
Schließlich übertragen wir eine der Verbesserungen aus dem linearen Fall und zeigen anhand eines numerischen Beispiels, dass es auch im Fall von Rekonstruktionsproblemen mit Bezug zu Quantensystemen einen positiven Effekt auf die Performance des Algorithmus hat.

Contents

Introduction

Operator identification is a challenging topic in many fields of mathematical application such as physics in general, chemistry and engineering. Especially for the topic of Hamiltonian identification in the field of quantum mechanics, the physics literature has constantly been enriched with new results over the last years (see e.g. [DR14, WDQ+18, XWLJ19, ZS14]). Correspondingly different numerical methods have been developed from the mathematical side. Some of these strategies are applied more in the *online* phase of the procedure, meaning that they make use of laboratory data given by the experiment (see e.g. [BMR09, LBMRT07]). The strategy that we will investigate in our work (see [MS09]) does most of the calculations during the *offline* phase, meaning it does not require any laboratory data to be provided on the way.

The algorithm developed in [MS09] is a greedy-type method for the reconstruction of an unknown dipole moment operator in a Schrödinger-type equation. By *reconstruction* we do not only mean the process of solving an inverse problem (the identification) but the entire process of finding the respective operator. In this particular case the first and main part of the strategy consists in the generation of selective laser fields such that the final identification has a better chance of finding the true dipole moment operator. In the second phase these fields are then applied to the true model in order to generate the laboratory data. These observations are then used in the third and final phase to identify the unknown Hamiltonian by fitting the data from the experiment with theoretical simulations for the same laser fields. Although we solve many inverse problems during the first step and one in the last step, the full procedure of this algorithm is, in some sense, more related to the idea of a *design of experiment* (see [Atk11]).

To acquire the selective laser fields, we also have to solve a series of optimal control problems. In fact, the first step of the algorithm is a constant interplay between solving inverse problems and optimal control problems. In the context of quantum systems, the idea of *monotonic schemes* has shown to be an efficient strategy in order to solve these kind of optimal control problems (see [MST06, Sal07, TKO92, ZR98]).

The thesis is structured as follows:
In Chapter 1 the algorithm is analysed in the setting of a linear ordinary differential equation. The relation between existence of feasible controls and observability of the system is investigated, leading to the definition of sufficient conditions for the convergence of the algorithm. Based on the findings regarding the observability of our system, improvements will be introduced and tested in numerical experiments. As a last step some numerical instabilities are being discussed.

In Chapter 2 the algorithm is described in its original setting of a Schrödinger-type equation. The main focus of this chapter will be the introduction of monotonic schemes. We will show two different approaches and prove convergence for the one used in the numerical experiments. In the last part, an improvement from Chapter 1 is reintroduced and investigated on the basis of numerical experiments

CHAPTER 1

Linear system

1.1 The identification problem and a greedy identification algorithm

1.1.1 Problem formulation

Consider a state $\boldsymbol{y}$ that evolves according to the linear ordinary differential equation (ODE)

$$\begin{cases} \dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + B^\star \boldsymbol{\epsilon}(t), \quad t \in (0, T] \\ \boldsymbol{y}(0) = \boldsymbol{y}_0, \end{cases} \tag{1}$$

for an initial condition $\boldsymbol{y}_0 \in \mathbb{R}^N$ and a control function $\boldsymbol{\epsilon} \in E_{ad}$, where E_{ad} is a non-empty, closed, convex and bounded subset of $L^2(0, T; \mathbb{R}^M)$. The system matrix $A \in \mathbb{R}^{N\times N}$ is assumed to be known. The goal is to reconstruct the unknown control matrix $B^\star$ that lies in the space spanned by a basis $\mathcal{B} = \{B_1, \dots, B_K\} \subset \mathbb{R}^{N\times M}$ with $1 \le K \le NM$. In order to find $B^\star$ one uses experimentally some control functions $\boldsymbol{\epsilon}$ and measures at some fixed time $T > 0$ the quantities $C\boldsymbol{y}(T)$, obtained by the different control functions. Here C is some known observer matrix $C \in \mathbb{R}^{P\times N}$ ($P \le N$). Moreover, we assume that the measurements are not affected by any type of errors (like noise). Using the measured data $C\boldsymbol{y}(T)$, one can also simulate numerically (1) at time T for any control matrix $B \in \mathcal{B}$, using the same controls $\boldsymbol{\epsilon}$, and compare the numerical and experimental results.
In this setting the goal is to find a matrix $B \in \mathcal{B}$ for which the difference at time T between the experimental data $C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon})$ and the numerical data $C\boldsymbol{y}_T(B, \boldsymbol{\epsilon})$ is the smallest for any control $\boldsymbol{\epsilon} \in E_{ad}$. In other words we want to find the matrix B that solves

$$\min_{B\in\mathcal{B}} \max_{\boldsymbol{\epsilon}\in E_{ad}} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B, \boldsymbol{\epsilon}) \right\|^2, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\boldsymbol{y}_T(B, \boldsymbol{\epsilon})$ is the solution of (1) at time T for the control matrix B and the control $\boldsymbol{\epsilon}$:

$$\boldsymbol{y}_T(B, \boldsymbol{\epsilon}) = e^{TA}\boldsymbol{y}_0 + \int_0^T e^{(T-s)A} B\boldsymbol{\epsilon}(s) ds.$$

Notice that we implicitly assume that the linear system (1) corresponds to the true laboratory model. For this reason the experimental data are denoted by $C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon})$, where $B^\star$ is the true matrix of the laboratory system.
Since we know that $B^\star = \sum_{j=1}^K \boldsymbol{\alpha}_j^\star B_j =: B(\boldsymbol{\alpha}^\star)$ for some $\boldsymbol{\alpha}^\star = [\boldsymbol{\alpha}_1^\star, \dots, \boldsymbol{\alpha}_K^\star]^T \in \mathbb{R}^K$ we can write the numerically guessed matrix B in the same form $B(\boldsymbol{\alpha}) = \sum_{j=1}^K \boldsymbol{\alpha}_j B_j$ and hence rewrite problem (2) as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \max_{\boldsymbol{\epsilon}\in E_{ad}} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}) \right\|^2. \tag{3}$$

However if we wanted to directly solve this problem we would have to try many different linear combinations $B(\boldsymbol{\alpha})$ and therefore perform a lot of experiments for many different controls.

The idea presented in [MS09] is to look at the problem from a different point of view. The

algorithm introduced in [MS09] does not consider (at first) any experimental data from the system with the true control matrix $B^\star$. Instead it tries to iteratively find a set of controls $(\boldsymbol{\epsilon}^1, \dots, \boldsymbol{\epsilon}^K) \subset E_{ad}$ that attempt to maximize the identifiability of the system and separate the observed numerical solution at time T for any two different linear combinations $B(\tilde{\boldsymbol{\alpha}}), B(\hat{\boldsymbol{\alpha}})$. With this set of controls one then solves the least-squares problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \sum_{j=1}^{K} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2, \tag{4}$$

where K laboratory experiments have to be performed.

Remark 1
The problem (4) *can also be formulated more generally as a linear problem, see [MS09, Section 6].*

Obviously $\boldsymbol{\alpha}^\star$ solves problem (4). The question remaining is whether all solutions of (4) also solve (3). One case where this would be true is when $B^\star$ is the unique solution to (4). We will discuss assumptions for this and other scenarios in Sections 1.2.4 and 1.3.

Since a priori we might not know a basis $\mathcal{B}$ with dimension $K < NM$ such that $B^\star \in \text{span}(\mathcal{B})$, we will in a first stage consider a full basis of the matrix space $Mat_{N\times M}(\mathbb{R})$.

Assumtion 1
Let $K = NM$.

1.1.2 Introduction to the algorithm

In the case of our linear ODE, the algorithm in [MS09] can be written as:

Algorithm 1 Greedy Reconstruction Algorithm

Require: A basis $\mathcal{B} = (B_1, \dots, B_{NM})$.

1: Solve the initialization problem

$$\max_{\boldsymbol{\epsilon} \in E_{ad}} \left\| C\boldsymbol{y}_T(B_1, \boldsymbol{\epsilon}) \right\|^2. \tag{5}$$

which gives the field $\boldsymbol{\epsilon}^1$ and set $k = 1$.

2: **while** $k \leq NM - 1$ **do**

3: Fitting step: Find $(\boldsymbol{\alpha}_j^k)_{j=1,\dots,k}$ that solve the problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \sum_{j=1}^{k} \left\| C\boldsymbol{y}_T(B_{k+1}, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2. \tag{6}$$

4: Discriminatory step: Find $\boldsymbol{\epsilon}^{k+1}$ that solves the problem:

$$\max_{\boldsymbol{\epsilon} \in E_{ad}} \left\| C\boldsymbol{y}_T(B_{k+1}, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}^k), \boldsymbol{\epsilon}) \right\|^2. \tag{7}$$

5: Update $k \leftarrow k + 1$.

6: **end while**

Let us briefly explain Algorithm 1. The idea is to generate controls that separate the observation of system (1) at time T for the different elements $B_1, \dots, B_{NM}$, making possible the identification of their respective coefficients $\boldsymbol{\alpha}_1^\star, \dots, \boldsymbol{\alpha}_{NM}^\star$ when solving problem (4). The initialization is performed by solving the optimal control problem (5). Here we maximize what can be observed at time T for the system

$$\begin{cases} \dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + B_1\boldsymbol{\epsilon}(t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0. \end{cases}$$

Remark 2
This initialization is somewhat arbitrary as mentioned in [MS09, Section 3.1]. However, this is generally not true: an arbitrary initialization can lead to some problems concerning the convergence of the algorithm. We will discuss the relevancy of the initialization for the convergence in Section 1.2.3.1 and present an improved choice in Section 1.3.1.1.

In each following iteration we consider the next B_{k+1}. First we perform a fitting-step, meaning that we solve the finite-dimensional minimization problem (6). Here we want to find a linear combination of the previous $B_1, \dots, B_k$ that yields the same or a similar solution at time T for the system for all previously computed controls $\boldsymbol{\epsilon}^1, \dots, \boldsymbol{\epsilon}^k$.
In other words, we compare the two systems

$$\begin{cases} \dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + B_{k+1}\boldsymbol{\epsilon}^j(t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0, \end{cases} \qquad \begin{cases} \dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + \sum_{i=1}^k \boldsymbol{\alpha}_i B_i\boldsymbol{\epsilon}^j(t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0. \end{cases}$$

for $j \in \{1, \dots, k\}$ and try to find an $\boldsymbol{\alpha} \in \mathbb{R}^k$ for which their observed solution at time T are as similar as possible. We denote the computed vector by $\boldsymbol{\alpha}^k$. The fitting step is followed by a discriminatory-step that uses the computed $\boldsymbol{\alpha}^k$. This step consists in solving the optimal control problem (7). Here we compute a control $\boldsymbol{\epsilon}^{k+1}$ that splits the solutions obtained at time T with B_{k+1} and the linear combination obtained in the previous fitting step. In other words, we again look at the two systems

$$\begin{cases} \dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + B_{k+1}\boldsymbol{\epsilon}(t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0, \end{cases} \qquad \begin{cases} \dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + \sum_{i=1}^k \boldsymbol{\alpha}_i^k B_i\boldsymbol{\epsilon}(t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0, \end{cases}$$

this time with fixed coefficients $\boldsymbol{\alpha}_i^k$ and a variable control $\boldsymbol{\epsilon}$. We try to find an $\boldsymbol{\epsilon} \in E_{ad}$ for which the observed solutions at time T of both systems are most distinct from one another.

Remark 3
The fact that each generated control is a solution to a maximization problem is the greedy character of the algorithm.

1.1.3 Well-posedeness

In this section the goal is to show that problem (4) and the three problems in the algorithm, namely (5), (6) and (7), are well posed.

1.1.3.1 Main identification problem

Proposition 1
The least-squares problem (4) *is well posed.*

proof. The problem is finite dimensional and the cost functional is bounded from below (by 0). The problem is in fact equivalent to a quadratic optimization problem $\min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \left\langle \boldsymbol{\alpha}^\star - \boldsymbol{\alpha}, \widehat{W}(\boldsymbol{\alpha}^\star - \boldsymbol{\alpha}) \right\rangle$, where $\widehat{W}$ is a symmetric positive semi-definite matrix (see Section 1.2). Therefore a minimizer exists. □

1.1.3.2 Initialization problem

Proposition 2
The initialization problem (5) *is well posed.*

proof. We denote by J_I the cost functional of (5), that is

$$J_I(\boldsymbol{\epsilon}) = \|C\boldsymbol{y}_T(B_1, \boldsymbol{\epsilon})\|^2 .$$

First we show that J_I is bounded from above. We have

$$\begin{aligned}
\|C\boldsymbol{y}_T(B_1,\boldsymbol{\epsilon})\| &= \left\| Ce^{TA}\boldsymbol{y}_0 + \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds \right\| \\
&\leq \underbrace{\left\|Ce^{TA}\boldsymbol{y}_0\right\|}_{=:c_0<\infty} + \left\| \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds \right\| \\
&\leq c_0 + \underbrace{\|C\|}_{<\infty} \left\| \int_0^T e^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds \right\| \\
&\leq c_0 + c_1 \int_0^T \left\| e^{(T-s)A}B_1\boldsymbol{\epsilon}(s) \right\| ds \\
&\leq c_0 + c_1 \int_0^T \underbrace{\|e^{(T-s)A}\|}_{<\infty} \underbrace{\|B_1\|}_{<\infty} \|\boldsymbol{\epsilon}(s)\| ds
\end{aligned}$$

$$
\begin{aligned}
&\leq c_0 + c_2 \|\boldsymbol{\epsilon}\|_{L^1(0,T;\mathbb{R}^M)} \\
&\leq c_0 + c_3 \underbrace{\|\boldsymbol{\epsilon}\|_{L^2(0,T;\mathbb{R}^M)}}_{\overset{(\star)}{\leq} c_4 < \infty} \qquad (\star) \text{ since } E_{ad} \subset L^2(0,T;\mathbb{R}^M) \text{ is bounded} \\
&\leq c_5,
\end{aligned}
$$

for some $c_5 \geq 0$, which is independent of $\boldsymbol{\epsilon}$. Therefore J_I is bounded from above.

Since E_{ad} is non-empty, we can find a maximizing sequence $\{\boldsymbol{\epsilon}^n\}_{n\in\mathbb{N}} \subset E_{ad}$,

$$
\lim_{n\to\infty} J_I(\boldsymbol{\epsilon}^n) = \sup_{\boldsymbol{\epsilon}\in E_{ad}} J_I(\boldsymbol{\epsilon}) =: \widetilde{J}.
$$

E_{ad} is also weakly compact in $L^2(0,T;\mathbb{R}^M)$, since it is a closed, convex and bounded subset of $L^2(0,T;\mathbb{R}^M)$. Hence there exists a weakly in $L^2(0,T;\mathbb{R}^M)$ convergent subsequence $\boldsymbol{\epsilon}^{\tilde{n}} \rightharpoonup \tilde{\boldsymbol{\epsilon}} \in E_{ad}$.

Notice that $\boldsymbol{y}_T(B_1, \boldsymbol{\epsilon})$ is the solution at final time T of the differential equation

$$
\begin{cases}
\dot{\boldsymbol{y}}(t) = A\boldsymbol{y}(t) + B_1\boldsymbol{\epsilon}(t), \quad t \in (0,T] \\
\boldsymbol{y}(0) = \boldsymbol{y}_0,
\end{cases} \tag{8}
$$

By Carathéodory's Existence Theorem, for each $\boldsymbol{\epsilon} \in E_{ad}$ there exists a unique solution $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{\epsilon}) \in AC([0,T];\mathbb{R}^N)$. With this we define the sequence $\{\boldsymbol{y}_{\tilde{n}}\}_{\tilde{n}\in\mathbb{N}}$ as $\boldsymbol{y}_{\tilde{n}} = \boldsymbol{y}(\boldsymbol{\epsilon}^{\tilde{n}}) \subset AC([0,T];\mathbb{R}^N)$. By similar calculations as at the beginning of the proof we can show that $\{\boldsymbol{y}_{\tilde{n}}\}_{\tilde{n}\in\mathbb{N}}$ is (uniformly) bounded in L^2. Recalling 8 and that E_{ad} is bounded, we obtain that $\{\boldsymbol{y}_{\tilde{n}}\}_{\tilde{n}\in\mathbb{N}}$ is also bounded in $H^1(0,T,\mathbb{R}^{\mathbb{N}})$. Since $H^1(0,T;\mathbb{R}^N)$ is reflexive there exists a subsequence $\{\boldsymbol{y}_{\tilde{n}_k}\}_{k\in\mathbb{N}}$ that converges weakly in $H^1(0,T;\mathbb{R}^N)$ and, by the Sobolev compact embedding $H^1(0,T;\mathbb{R}^N) \Subset C([0,T];\mathbb{R}^N)$, uniformly in $C([0,T];\mathbb{R}^N)$ with limit $\tilde{\boldsymbol{y}} \in H^1(0,T;\mathbb{R}^N) \cap C([0,T];\mathbb{R}^N)$.

Since $\boldsymbol{\epsilon}^{\tilde{n}} \rightharpoonup \tilde{\boldsymbol{\epsilon}}$ in $L^2(0,T;\mathbb{R}^M)$ and $\boldsymbol{y}_{\tilde{n}_k} \to \tilde{\boldsymbol{y}}$ uniformly in $C([0,T];\mathbb{R}^N)$ we get

$$
e^{tA}\boldsymbol{y}_{\tilde{n}_k}(0) + \int_0^t e^{(t-s)A} B\boldsymbol{\epsilon}^{\tilde{n}_k}(s)ds \quad \to \quad e^{tA}\tilde{\boldsymbol{y}}(0) + \int_0^t e^{(t-s)A} B\tilde{\boldsymbol{\epsilon}}(s)ds, \tag{9}
$$

where we use that the map $\epsilon \mapsto \int_0^t e^{(t-s)A} B\boldsymbol{\epsilon}(s)ds$ which means that $\dot{\tilde{\boldsymbol{y}}}(t) = A\tilde{\boldsymbol{y}}(t) + B\tilde{\boldsymbol{\epsilon}}(t)$ a.e. in $(0,T)$ and $\tilde{\boldsymbol{y}}(0) = \boldsymbol{y}_0$. Therefore we obtain $\tilde{\boldsymbol{y}} = \boldsymbol{y}(\tilde{\boldsymbol{\epsilon}})$. Now by continuity of the norm and matrix-vector multiplication we have

$$
\widetilde{J} = \lim_{k\to\infty} J_I(\boldsymbol{\epsilon}^{\tilde{n}_k}) = \lim_{k\to\infty} \left\| C(\boldsymbol{y}(\boldsymbol{\epsilon}^{\tilde{n}_k}))(T) \right\|^2 = \|C(\boldsymbol{y}(\tilde{\boldsymbol{\epsilon}}))(T)\|^2 = J(\tilde{\boldsymbol{\epsilon}}).
$$

This implies that $\tilde{\boldsymbol{\epsilon}}$ is a minimizer for the initialization problem (5). □

1.1.3.3 Fitting-step problem

Proposition 3
The fitting-step problem (6) *is well posed.*

proof. As the least-squares problem (4), this one is also finite dimensional, bounded from below and can be written as a quadratic optimization problem $\min_{\boldsymbol{\alpha}\in\mathbb{R}^k} \left\langle \boldsymbol{\alpha}, W\boldsymbol{\alpha} \right\rangle + \boldsymbol{b}^\top\boldsymbol{\alpha}$ for some symmetric, positive semidefinite matrix W and some vector $\boldsymbol{b}$ (see Section 1.2). □

1.1.3.4 Discriminatory-step problem

Denote by $J_D(\boldsymbol{\epsilon})$ the cost functional in (7), that is

$$J_D(\boldsymbol{\epsilon}) = \left\| C\boldsymbol{y}_T(B_{k+1}, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}^k), \boldsymbol{\epsilon}) \right\|^2. \tag{10}$$

We get

$$\begin{aligned} J_D(\boldsymbol{\epsilon}) &= \left\| C\Big(\int_0^T e^{(T-s)A} B_{k+1}\boldsymbol{\epsilon}(s)ds - \int_0^T e^{(T-s)A} \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l \boldsymbol{\epsilon}(s)ds \Big) \right\|^2 \\ &= \left\| C\Big(\int_0^T e^{(T-s)A} \Big(B_{k+1} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l \Big) \boldsymbol{\epsilon}(s)ds \Big) \right\|^2. \end{aligned}$$

This is equivalent to the initialization problem, if one replaces B_1 in (5) with $B_{k+1} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l$. Therefore the proof of the well posedeness is analogous to the well posedeness of the initialization problem, and the following result holds.

Proposition 4
The discrimiatory-step problem (7) *is well posed.*

1.2 Convergence analysis of the greedy identification algorithm

First we need to introduce a new matrix form notation for our optimization problems.

1.2.1 Algorithm in matrix form

1.2.1.1 Main identification problem

We start with the main least-squares problem (4). We define

$$J(\boldsymbol{\alpha}) := \sum_{j=1}^{NM} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2. \tag{11}$$

One can write this as

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \sum_{j=1}^{NM} \left\| \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^{NM} \boldsymbol{\alpha}_l^\star B_l\Big)\boldsymbol{\epsilon}^j(s)ds - \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^{NM} \boldsymbol{\alpha}_l B_l\Big)\boldsymbol{\epsilon}^j(s)ds \right\|^2 \\ &= \sum_{j=1}^{NM} \left\| \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^{NM} (\boldsymbol{\alpha}_l^\star - \boldsymbol{\alpha}_l) B_l\Big)\boldsymbol{\epsilon}^j(s)ds \right\|^2 \\ &= \sum_{j=1}^{NM}\sum_{l=1}^{NM}\sum_{m=1}^{NM} (\boldsymbol{\alpha}_l^\star - \boldsymbol{\alpha}_l)(\boldsymbol{\alpha}_m^\star - \boldsymbol{\alpha}_m) \Big\langle \int_0^T Ce^{(T-s)A}B_l\boldsymbol{\epsilon}^j(s)ds, \int_0^T Ce^{(T-s)A}B_m\boldsymbol{\epsilon}^j(s)ds \Big\rangle. \end{aligned}$$

We define the matrix

$$\widehat{W} := \sum_{j=1}^{NM} W_j \in \mathbb{R}^{NM\times NM}, \tag{12}$$

where W_j has the entries

$$[W(\boldsymbol{\epsilon}^j)]_{l,m} := [W_j]_{l,m} := \Big\langle \boldsymbol{\gamma}_{l,j}, \boldsymbol{\gamma}_{m,j} \Big\rangle, \quad 1 \le l, m, j \le NM, \tag{13}$$

with

$$\boldsymbol{\gamma}_l(\boldsymbol{\epsilon}^j) := \boldsymbol{\gamma}_{l,j} := \int_0^T Ce^{(T-s)A}B_l\boldsymbol{\epsilon}^j(s)ds. \tag{14}$$

We get

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \sum_{l=1}^{NM}\sum_{m=1}^{NM} (\boldsymbol{\alpha}_l^\star - \boldsymbol{\alpha}_l)(\boldsymbol{\alpha}_m^\star - \boldsymbol{\alpha}_m) \sum_{j=1}^{NM} \Big\langle \boldsymbol{\gamma}_{l,j}, \boldsymbol{\gamma}_{m,j} \Big\rangle \\ &= \Big\langle \boldsymbol{\alpha}^\star - \boldsymbol{\alpha}, \sum_{j=1}^{NM} W_j(\boldsymbol{\alpha}^\star - \boldsymbol{\alpha}) \Big\rangle = \Big\langle \boldsymbol{\alpha}^\star - \boldsymbol{\alpha}, \widehat{W}(\boldsymbol{\alpha}^\star - \boldsymbol{\alpha}) \Big\rangle. \end{aligned}$$

We have then obtained the following result.

Lemma 1
Problem (4) *is equivalent to*

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{NM}} \Big\langle \boldsymbol{\alpha}^\star - \boldsymbol{\alpha}, \widehat{W}(\boldsymbol{\alpha}^\star - \boldsymbol{\alpha}) \Big\rangle, \tag{15}$$

where $\widehat{W}$ *is defined in* (12).

1.2.1.2 Initialization problem

As in the last section we introduce

$$\boldsymbol{\gamma}_l(\boldsymbol{\epsilon}) := \int_0^T Ce^{(T-s)A}B_l\boldsymbol{\epsilon}(s)ds \tag{16}$$

and

$$[W(\boldsymbol{\epsilon})]_{l,m} := \Big\langle \boldsymbol{\gamma}_l(\boldsymbol{\epsilon}), \boldsymbol{\gamma}_m(\boldsymbol{\epsilon}) \Big\rangle, \quad 1 \le l, m \le NM, \tag{17}$$

for any $\boldsymbol{\epsilon} \in E_{ad}$. Using this notation we can write the initialization problem (5) in terms of W:

$$\begin{aligned} J_l(\boldsymbol{\epsilon}) &= \left\| e^{TA}\boldsymbol{y}_0 + \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds \right\|^2 \\ &= \left\| e^{TA}\boldsymbol{y}_0 \right\|^2 + 2\Big\langle e^{TA}y_0, \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds \Big\rangle \\ &\quad + \Big\langle \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds, \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds \Big\rangle \\ &\overset{(14)}{=} \left\| e^{TA}\boldsymbol{y}_0 \right\|^2 + 2\Big\langle e^{TA}y_0, \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}) \Big\rangle + \Big\langle \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}), \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}) \Big\rangle \\ &\overset{(13)}{=} \left\| e^{TA}\boldsymbol{y}_0 \right\|^2 + 2\Big\langle e^{TA}y_0, \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}) \Big\rangle + [W(\boldsymbol{\epsilon})]_{1,1}. \end{aligned}$$

Hence, we obtain the following result

Lemma 2
Problem (5) *is equivalent to*

$$\max_{\boldsymbol{\epsilon} \in E_{ad}} \left\| e^{TA}\boldsymbol{y}_0 \right\|^2 + 2\Big\langle e^{TA}y_0, \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}) \Big\rangle + [W(\boldsymbol{\epsilon})]_{1,1}, \tag{18}$$

where $\boldsymbol{\gamma}_1(\boldsymbol{\epsilon})$ *and* $[W(\boldsymbol{\epsilon})]_{1,1}$ *are defined in* (16) *and* (17)*, respectively.*

1.2.1.3 Fitting-step problem

Let us denote by $J_F(\boldsymbol{\alpha})$ the cost functional in (6) at step k, that is

$$J_F(\boldsymbol{\alpha}) = \sum_{j=1}^{k} \left\| C\boldsymbol{y}_T(B_{k+1}, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2. \tag{19}$$

This can be written as

$$\begin{aligned}
J_F(\boldsymbol{\alpha}) &= \sum_{j=1}^{k} \left\| \int_0^T Ce^{(T-s)A} B_{k+1} \boldsymbol{\epsilon}^j(s) ds - \int_0^T Ce^{(T-s)A} \Big(\sum_{l=1}^{k} \boldsymbol{\alpha}_l B_l \Big) \boldsymbol{\epsilon}^j(s) ds \right\|^2 \\
&\overset{(14)}{=} \sum_{j=1}^{k} \left\| \boldsymbol{\gamma}_{k+1,j} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l \boldsymbol{\gamma}_{l,j} \right\|^2 \\
&= \sum_{j=1}^{k} \left[\left\langle \boldsymbol{\gamma}_{k+1,j}, \boldsymbol{\gamma}_{k+1,j} \right\rangle - 2 \Big\langle \boldsymbol{\gamma}_{k+1,j}, \sum_{l=1}^{k} \boldsymbol{\alpha}_l \boldsymbol{\gamma}_{l,j} \Big\rangle + \Big\langle \sum_{l=1}^{k} \boldsymbol{\alpha}_l \boldsymbol{\gamma}_{l,j}, \sum_{l=1}^{k} \boldsymbol{\alpha}_l \boldsymbol{\gamma}_{l,j} \Big\rangle \right] \\
&= \sum_{j=1}^{k} \left\| \boldsymbol{\gamma}_{k+1,j} \right\|^2 - 2 \sum_{j=1}^{k} \sum_{l=1}^{k} \boldsymbol{\alpha}_l \Big\langle \boldsymbol{\gamma}_{k+1,j}, \boldsymbol{\gamma}_{l,j} \Big\rangle + \sum_{j=1}^{k} \sum_{l=1}^{k} \sum_{\mu=1}^{k} \boldsymbol{\alpha}_l \boldsymbol{\alpha}_\mu \Big\langle \boldsymbol{\gamma}_{l,j}, \boldsymbol{\gamma}_{\mu,j} \Big\rangle \\
&= \sum_{j=1}^{k} \left\| \boldsymbol{\gamma}_{k+1,j} \right\|^2 - 2 \Big\langle \boldsymbol{\alpha}, \widehat{W}^k_{[1:k,k+1]} \Big\rangle + \Big\langle \boldsymbol{\alpha}, \widehat{W}^k_{[1:k,1:k]} \boldsymbol{\alpha} \Big\rangle,
\end{aligned}$$

where $\widehat{W}^k$ is the k-th partial sum of $\widehat{W}$:

$$\widehat{W}^k = \sum_{j=1}^{k} W_j, \tag{20}$$

where W_j is defined in (13), and $\widehat{W}^k_{[1:j,1:l]}$ denotes the $j \times l$ upper left submatrix of $\widehat{W}^k$.

This leads to the following result.

Lemma 3
Problem (6) *is equivalent to*

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \Big\langle \boldsymbol{\alpha}, \widehat{W}^k_{[1:k,1:k]} \boldsymbol{\alpha} \Big\rangle - 2 \Big\langle \widehat{W}^k_{[1:k,k+1]}, \boldsymbol{\alpha} \Big\rangle, \tag{21}$$

where $\widehat{W}^k$ *is defined in* (20).

1.2.1.4 Discriminatory-step probem

Let $\boldsymbol{\alpha}^k$ be a solution to the fitting-step problem (21). Then we can write problem (7) as

$$J_D(\boldsymbol{\epsilon}) = \left\| \int_0^T Ce^{(T-s)A} B_{k+1}\boldsymbol{\epsilon}(s)ds - \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l\Big)\boldsymbol{\epsilon}(s)ds \right\|^2$$
$$\overset{(14)}{=} \left\| \boldsymbol{\gamma}_{k+1}(\boldsymbol{\epsilon}) - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k \boldsymbol{\gamma}_l(\boldsymbol{\epsilon}) \right\|^2 .$$

If we define $\boldsymbol{v} := [\boldsymbol{\alpha}^k,\ -1]^\top$, we can write

$$\begin{aligned} J_D(\boldsymbol{\epsilon}) &= \left\| \sum_{l=1}^{k+1} \boldsymbol{v}_l \boldsymbol{\gamma}_l(\boldsymbol{\epsilon}) \right\|^2 \\ &= \Big\langle \sum_{l=1}^{k+1} \boldsymbol{v}_l \boldsymbol{\gamma}_l(\boldsymbol{\epsilon}), \sum_{l=1}^{k+1} \boldsymbol{v}_l \boldsymbol{\gamma}_l(\boldsymbol{\epsilon}) \Big\rangle \\ &= \sum_{l=1}^{k+1}\sum_{m=1}^{k+1} \boldsymbol{v}_l \boldsymbol{v}_m \Big\langle \boldsymbol{\gamma}_l(\boldsymbol{\epsilon}), \boldsymbol{\gamma}_m(\boldsymbol{\epsilon}) \Big\rangle \\ &\overset{(13)}{=} \Big\langle \boldsymbol{v}, [W(\boldsymbol{\epsilon})]_{[1:k+1,1:k+1]}\boldsymbol{v} \Big\rangle, \end{aligned}$$

and obtain the following result.

Lemma 4
Problem (7) *is equivalent to*

$$\max_{\boldsymbol{\epsilon}\in E_{ad}} \Big\langle \boldsymbol{v}, [W(\boldsymbol{\epsilon})]_{[1:k+1,1:k+1]}\boldsymbol{v} \Big\rangle, \tag{22}$$

where $W(\boldsymbol{\epsilon})$ *is defined in* (13) *and* $\boldsymbol{v} := [\boldsymbol{\alpha}^k,\ -1]^\top$.

1.2.2 Outline of the convergence analysis

We want to motivate the convergence analysis with the results we obtained in the last section.

Our main goal is to show that the algorithm produces control functions such that problem (4) (which is equivalent to problem (15) by Lemma 1) is uniquely solvable.
Consider the matrix $\Gamma^j \in \mathbb{R}^{P\times NM}$ $(1 \le j \le NM)$ with columns

$$[\Gamma^j]_{\cdot,m} := \boldsymbol{\gamma}_{m,j} = \int_0^T Ce^{(T-s)A}B_m\boldsymbol{\epsilon}^j(s)ds, \qquad (1 \le m \le NM), \tag{23}$$

we get

$$[(\Gamma^j)^\top\Gamma^j]_{l,m} = \left\langle \boldsymbol{\gamma}_{l,j}, \boldsymbol{\gamma}_{m,j} \right\rangle = [W_j]_{l,m}.$$

This means that we can write $W_j = (\Gamma^j)^\top\Gamma^j$. Hence the matrices W_j are positive semi-definite and therefore the same is true for $\widehat{W} = \sum_{j=1}^{NM} W_j$.
We get the following result.

Proposition 5
Problem (15) *is uniquely solvable if and only if* $\widehat{W}$ *is positive definite.*

Therefore we have to show that the matrix $\widehat{W}$ that corresponds to the set of controls $\{\boldsymbol{\epsilon}^j\}_{j=1}^{NM}$, generated by the algorithm, is positive definite.

The idea is that at iteration k $(1 \le k \le (NM-1))$ we assume the k-by-k upper left block of $\widehat{W}^k$ (defined in (20)) to be positive definite. If the $(k+1)$-by-$(k+1)$ upper left block is also positive definite, it does not matter what control $\boldsymbol{\epsilon}^{k+1}$ the algorithm computes, the $(k+1)$-by-$(k+1)$ upper left block of $\widehat{W}^{k+1} = \widehat{W}^k + W_{k+1}$ will be positive definite as well. If the $(k+1)$-by-$(k+1)$ upper left block is only positive semi-definite (with a non-trivial kernel), then the following statement holds.

Proposition 6
Let $\widehat{W}^k_{[1:k,1:k]}$ *be positive definite. Then the fitting-step problem* (21) *is uniquely solvable with sufficient first-order optimality condition*

$$\widehat{W}^k_{[1:k,1:k]}\boldsymbol{\alpha}^k = \widehat{W}^k_{[1:k,k+1]}. \tag{24}$$

Remark 4
In the application Proposition 6 allows us to perform a fitting step (compute $\boldsymbol{\alpha}^k$*) by simply solving the linear system* (24).

Now one can show that the vector $\boldsymbol{v} := [\boldsymbol{\alpha}^k,\ -1]^\top$ spans the kernel of $\widehat{W}^k_{[1:k+1,1:k+1]}$. We use the following lemma for a more general case.

Lemma 5
Consider a symmetric, positive semi-definite matrix $\tilde{A} \in \mathbb{R}^{n\times n}$ *of the form*

$$\tilde{A} = \begin{pmatrix} A & \boldsymbol{b} \\ \boldsymbol{b}^\top & c \end{pmatrix},$$

where $A \in \mathbb{R}^{(n-1)\times(n-1)}$ *is symmetric, positive definite and* $\boldsymbol{b} \in \mathbb{R}^{n-1}$ *and* $c \in \mathbb{R}$ *are such that the kernel of* $\tilde{A}$ *is non-trivial.*
Then

$$ker(\tilde{A}) = span\Big\{ \begin{pmatrix} A^{-1}\boldsymbol{b} \\ -1 \end{pmatrix} \Big\}.$$

proof. Since the kernel of $\tilde{A}$ is non-trivial, there exists a non-zero vector $\boldsymbol{u} = \begin{pmatrix} \boldsymbol{v} \\ d \end{pmatrix} \in \mathbb{R}^n \setminus \{0\}$ (with $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and $d \in \mathbb{R}$) such that $\tilde{A}\boldsymbol{u} = 0$. Moreover, since A is positive definite, the kernel of $\tilde{A}$ must be one-dimensional and equal to the span of $\{\boldsymbol{u}\}$. Using the structure of $\boldsymbol{u}$, we write $\tilde{A}\boldsymbol{u}$ as

$$\begin{cases} A\boldsymbol{v} + d\,\boldsymbol{b} = 0, \\ \boldsymbol{b}^\top \boldsymbol{v} + dc = 0, \end{cases} \quad \overset{A \text{ invertible}}{\Leftrightarrow} \quad \begin{cases} \boldsymbol{v} = -dA^{-1}\boldsymbol{b}, \\ -d\,\boldsymbol{b}^\top A^{-1}\boldsymbol{b} + dc = 0. \end{cases} \tag{25}$$

Now, suppose that $d = 0$. This implies that $\boldsymbol{v} = -dA^{-1}\boldsymbol{b} = 0$, which in return implies that $\boldsymbol{u} = 0$. However, this is a contradiction to the fact that $\boldsymbol{u} \neq 0$. Hence $d \neq 0$ and the result follows by the right equations in (25) (divided by $-d$). □

In our case we have

$$\tilde{A} = \widehat{W}^k_{[1:k+1,1:k+1]},\ A = \widehat{W}^k_{[1:k,1:k]},\ \boldsymbol{b} = \widehat{W}^k_{[1:k,k+1]},\ c = \widehat{W}^k_{[k+1,k+1]},$$

where by assumption $\widehat{W}^k_{[1:k+1,1:k+1]}$ is positive semi-definite with a non-trivial kernel and $\widehat{W}^k_{[1:k,1:k]}$ is positive definite.

Therefore, we have

$$\ker(\widehat{W}^k_{[1:k+1,1:k+1]}) = \mathrm{span}\Big\{ \begin{pmatrix} (\widehat{W}^k_{[1:k,1:k]})^{-1}\widehat{W}^k_{[1:k,k+1]} \\ -1 \end{pmatrix} \Big\} = \mathrm{span}\Big\{ \begin{pmatrix} \boldsymbol{\alpha}^k \\ -1 \end{pmatrix} \Big\} = \mathrm{span}\Big\{ \boldsymbol{v} \Big\}.$$

Using this vector $\boldsymbol{v}$, the discriminatory step tries to find the next control $\boldsymbol{\epsilon}^{k+1}$ such that $[W_{k+1}]_{[1:k+1,1:k+1]}$ (defined in (13)) is strictly positive on $\mathrm{span}\Big\{ \boldsymbol{v} \Big\}$. Then the $(k+1)$-by-$(k+1)$ upper left block of $\widehat{W}^{k+1} = \widehat{W}^k + W_{k+1}$ is also strictly positive on this span and therefore positive definite.

Proceeding iteratively, we obtain that the matrix $\widehat{W}$ is positive definite after at most $(NM-1)$-iterations.

However there are two issues in these arguments that can pose problems: the initialization and the strict positivity of the discriminatory step. We will discuss both issues and present solutions in the following section.

1.2.3 Problems of the algorithm

1.2.3.1 Initialization

If we want to follow the induction argument from the previous section, we first need that the initial control $\boldsymbol{\epsilon}^1$ (solution to the initialization problem (18)) satisfies

$$[W_1]_{1,1} = \left\langle \boldsymbol{\gamma}_{1,1}, \boldsymbol{\gamma}_{1,1} \right\rangle = \left\| \int_0^T C e^{(T-s)A} B_1 \boldsymbol{\epsilon}^1(s) ds \right\|^2 > 0. \tag{26}$$

However this is generally not the case. Consider, for example, the case where

$$\|C e^{TA} \boldsymbol{y}_0 + C \int_0^T e^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds\|^2 \leq \|C e^{TA} \boldsymbol{y}_0\|^2,$$

for any $\boldsymbol{\epsilon} \in E_{ad}$, meaning that

$$\boldsymbol{\epsilon}^1 = 0 \in \underset{\boldsymbol{\epsilon} \in E_{ad}}{\arg\max} \|C e^{TA} \boldsymbol{y}_0 + C \int_0^T e^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds\|^2$$

and therefore

$$\left\| \int_0^T C e^{(T-s)A} B_1 \boldsymbol{\epsilon}^1(s) ds \right\|^2 = 0.$$

E.g. for $N = M = P = 2$:

Example 1
Let

$$A = C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{y}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and E_{ad} *some bounded set of* $L^2(0, T, \mathbb{R}^2)$ *with*

$$\forall \boldsymbol{\epsilon} \in E_{ad} : -\frac{1}{T} \leq \boldsymbol{\epsilon}(s)_1 \leq 0, \ (0 \leq s \leq T).$$

Then we get

$$\begin{aligned} C\boldsymbol{y}(B_1, \boldsymbol{\epsilon}) &= C(e^{TA} \boldsymbol{y}_0 + \int_0^T e^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds) \\ &= \begin{pmatrix} e^T \\ e^T \end{pmatrix} + \int_0^T \begin{pmatrix} e^{T-s} & 0 \\ 0 & e^{T-s} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon}(s)_1 \\ \boldsymbol{\epsilon}(s)_2 \end{pmatrix} ds \\ &= \begin{pmatrix} e^T \\ e^T \end{pmatrix} + \int_0^T \begin{pmatrix} e^{T-s} \boldsymbol{\epsilon}(s)_1 \\ 0 \end{pmatrix} ds. \end{aligned}$$

Since $-\frac{1}{T} \leq \boldsymbol{\epsilon}(s)_1 \leq 0$ *for all* $0 < s < T$*, we have:*

$$0 \geq \int_0^T e^{T-s} \boldsymbol{\epsilon}(s)_1 ds \geq \int_0^T -e^T \frac{1}{T} ds = -e^T,$$

hence

$$|e^T + \int_0^T e^{T-s} \boldsymbol{\epsilon}(s)_1 ds| \leq e^T$$

and therefore

$$\|C\boldsymbol{y}_T(B_1, \boldsymbol{\epsilon})\|^2 = \left\| \begin{pmatrix} e^T \\ e^T \end{pmatrix} + \int_0^T \begin{pmatrix} e^{T-s} \boldsymbol{\epsilon}(s)_1 \\ 0 \end{pmatrix} ds \right\|^2 \leq \left\| \begin{pmatrix} e^T \\ e^T \end{pmatrix} \right\|^2 = \left\| C e^{TA} \boldsymbol{y}_0 \right\|^2.$$

So any control $\boldsymbol{\epsilon} \in E_{ad}$ *with* $\int_0^T C e^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds = 0$ *solves the initialization problem* (5).

One solution to this issue is to modify the initialization problem. We will discuss this further in Section 1.3.1.1.

However, even if another initialization problem is used, there is still the possibility that we get $[W_1]_{1,1} = 0$ for all $\boldsymbol{\epsilon} \in E_{ad}$, depending on the matrices A and C and the choice of B_1. This leads to the second issue of the algorithm related to the choice of the basis, as discussed in the following section.

1.2.3.2 Observability-rank and choice of basis

The general idea is that "one cannot recover what cannot be observed". For example, let $B_1 \in \mathbb{R}^{N\times M}$ be such that the columns of B_1 are in the kernel of the observability matrix $\mathcal{O}_N(C, A)$,

$$\mathcal{O}_N(C, A)B_1 = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{N-1} \end{pmatrix} B_1 = 0.$$

Then we have $CA^jB_1 = 0$ for all $0 \leq j \leq N-1$. Therefore we get for any $\boldsymbol{\epsilon}$

$$\begin{aligned} \int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds &= \int_0^T C\sum_{j=0}^{\infty}\frac{(T-s)^j}{j!}A^jB_1\boldsymbol{\epsilon}(s)ds \\ &\overset{C.H.}{=} \int_0^T \sum_{j=0}^{N-1}\beta_j(s)\underbrace{CA^jB_1}_{=0}\boldsymbol{\epsilon}(s)ds \\ &= 0. \end{aligned}$$

By C.H. we mean the Cayley-Hamilton theorem (see [HJ13, p.109f.]). Since we are considering a linear problem, this also means that

$$\begin{aligned} C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}) &= Ce^{TA}\boldsymbol{y}_0 + C\int_0^T e^{(T-s)A}\Big[\sum_{j=1}^{NM}\boldsymbol{\alpha}_j^\star B_j\Big]\boldsymbol{\epsilon}(s)ds \\ &= Ce^{TA}\boldsymbol{y}_0 + \boldsymbol{\alpha}_1^\star\underbrace{\int_0^T Ce^{(T-s)A}B_1\boldsymbol{\epsilon}(s)ds}_{=0} + C\int_0^T e^{(T-s)A}\Big[\sum_{j=2}^{NM}\boldsymbol{\alpha}_j^\star B_j\Big]\boldsymbol{\epsilon}(s)ds \\ &= Ce^{TA}\boldsymbol{y}_0 + C\int_0^T e^{(T-s)A}\Big[\sum_{j=2}^{NM}\boldsymbol{\alpha}_j^\star B_j\Big]\boldsymbol{\epsilon}(s)ds \end{aligned}$$

for any $\boldsymbol{\epsilon}$. Therefore it is impossible to observe the effect of $\boldsymbol{\alpha}_1^\star B_1$, meaning we will not be able to get any information about the component $\boldsymbol{\alpha}_1^\star$ of the solution. A similar problem occurs for any other B_j, $2 \leq j \leq NM$ with all columns in the kernel of $\mathcal{O}_N(C, A)$. Then the j-th column and row of $\widehat{W}$ (see (12)) are zero and therefore $\widehat{W}$ is only positive semi-definite with $\boldsymbol{e}_j \in \ker(\widehat{W})$ ($\boldsymbol{e}_j$ the j-th canonical vector in $\mathbb{R}^{NM}$). This means that there are infinite many solutions to problem (15) in the sense that any $\boldsymbol{\alpha}$ with an arbitrary $\boldsymbol{\alpha}_j \in \mathbb{R}$ and $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i^\star$ for $i \neq j$ solves problem (15). Therefore the true $\boldsymbol{\alpha}_j^\star$ cannot be fully reconstructed.

Similar problems also occur when the difference of two matrices in the basis is in the kernel of the observability matrix. In this case, their effect can not be distinguished and therefore any linear combination is a solution (compare Example 2).

We will discuss an appropriate choice for the basis $\mathcal{B}$ in Section 1.3.1.2.

1.2.4 Convergence analysis

Let us begin by assuming that all problems discussed in the previous do not occur. In fact the following assumptions are sufficient for our convergence analysis.

Assumtion 2
Let $A \in \mathbb{R}^{N\times N}$ *and* $C \in \mathbb{R}^{P\times N}$ *be such that the system is fully observable, i.e.* $\operatorname{rank} \mathcal{O}_N(C, A) = N$.

Assumtion 3
Let $\boldsymbol{\epsilon}^1$ *be such that* (26) *holds.*

In order to prove our main convergence result, we need to show that for the discriminatory-step problem (22) the function value in the optimum is always strictly positive.

Proposition 7
Let $\widehat{W}^k_{[1:k,1:k]}$ *be positive definite and* $\boldsymbol{\alpha}^k$ *the solution to the fitting-step problem* (6)*. Then, under Assumption 2, any solution* $\boldsymbol{\epsilon}^{k+1}$ *of the discriminatory-step problem* (7) *satisfies*

$$J_D(\boldsymbol{\epsilon}^{k+1}) = \left\| \int_0^T Ce^{(T-s)A}\Big(B_{k+1} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l\Big)\boldsymbol{\epsilon}^{k+1}(s)ds \right\|^2 > 0.$$

proof. We define

$$\widetilde{B} := B_{k+1} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l.$$

Since the matrices $B_1, \ldots, B_{k+1}$ are assumed to be linearly independent, $\widetilde{B}$ is non-zero. Consider for any $\delta \in (0, T)$ the control $\tilde{\boldsymbol{\epsilon}} \in E_{ad}$ defined as

$$\tilde{\boldsymbol{\epsilon}}(s) := \begin{cases} 0, & 0 \le s < \delta, \\ \boldsymbol{e}_i, & \delta \le s \le T, \end{cases}$$

where $\boldsymbol{e}_i \in \mathbb{R}^M$ is the i-th canonical vector and $i \in \{1, \ldots, M\}$ is such that the i-th column of $\widetilde{B}$, which we will denote by $\widetilde{B}_i$, is non-zero. Then we get

$$\begin{aligned} \int_0^T Ce^{(T-s)A}\widetilde{B}\tilde{\boldsymbol{\epsilon}}(s)ds &= \int_\delta^T Ce^{(T-s)A}\widetilde{B}_i ds \\ &= \int_\delta^T C\Big[\sum_{j=0}^{\infty} \frac{(T-s)^j A^j}{j!}\Big]\widetilde{B}_i ds \end{aligned}$$

$$\overset{(\star)}{=} c\Big[\sum_{j=0}^{\infty}\int_{\delta}^{T}\frac{(T-s)^j}{j!}ds\,CA^j\Big]\widetilde{B}_i$$
$$=\Big[\sum_{j=0}^{\infty}\frac{(T-\delta)^{j+1}}{(j+1)!}CA^j\Big]\widetilde{B}_i$$
$$=\sum_{j=0}^{\infty}\beta_j(\delta)CA^j\widetilde{B}_i,$$

where we used the dominated convergence theorem[1] in $(\star)$. Since the observability matrix is full rank by Assumption 2 and $\widetilde{B}_i \neq 0$, there exists an index $0 \leq j \leq N-1$ such that $CA^j\widetilde{B}_i \neq 0$. We also know that (non-constant) analytic functions are zero only in isolated points[2]. Therefore there exists a $0 < \delta < T$ such that $\sum_{j=0}^{\infty}\beta_j(\delta)CA^j\widetilde{B}_i \neq 0$.

In conclusion we get that there exists an $\tilde{\boldsymbol{\epsilon}} \in E_{ad}$ such that

$$\int_0^T Ce^{(T-s)A}B_1\tilde{\boldsymbol{\epsilon}}(s)ds \neq 0$$

and hence

$$J_D(\boldsymbol{\epsilon}^{k+1}) \geq \left\|\int_0^T Ce^{(T-s)A}\Big(B_{k+1}-\sum_{l=1}^{k}\boldsymbol{\alpha}_l^k B_l\Big)\tilde{\boldsymbol{\epsilon}}(s)ds\right\|^2 > 0.$$

□

Now we can prove our main convergence result.

Theorem 1
Let $\{\boldsymbol{\epsilon}^1,\dots,\boldsymbol{\epsilon}^{NM}\} \subset E_{ad}$ be a family of controls generated by Algorithm 1. Then, under Assumptions 2 and 3, the least-squares problem (4) is uniquely solvable with $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$.

proof. By Lemma 1 and Proposition 5 it is sufficient to show that the matrix $\widehat{W}$ corresponding to the controls $\boldsymbol{\epsilon}^1,\dots,\boldsymbol{\epsilon}^{NM}$ generated by the algorithm is positive definite. We show this by induction.

Base case: From Assumption 3 we have that $[W_1]_{1,1}$ is positive.

Inductive step: Assume that the submatrix $\widehat{W}^k_{[1:k,1:k]}$ is positive definite. If the submatrix $\widehat{W}^k_{[1:k+1,1:k+1]}$ is also positive definite, any control $\boldsymbol{\epsilon}^{k+1}$ will lead to a positive definite submatrix

$$\widehat{W}^{k+1}_{[1:k+1,1:k+1]} = \widehat{W}^k_{[1:k+1,1:k+1]} + [W_{k+1}]_{[1:k+1,1:k+1]}.$$

Now assume that the submatrix $\widehat{W}^k_{[1:k+1,1:k+1]}$ has a non-trivial kernel. By Lemma 3 and Proposition 6 the fitting-step problem is uniquely solvable with solution $\boldsymbol{\alpha}^k$. Then, by Lemma 5, the vector $\boldsymbol{v} = [\boldsymbol{\alpha}^k,\ -1]$ spans the kernel of $\widehat{W}^k_{[1:k+1,1:k+1]}$. Finally using Proposition 7 we get that the solution $\boldsymbol{\epsilon}^{k+1}$ to the discriminatory-step problem satisfies

$$0 < J_D(\boldsymbol{\epsilon}^{k+1}) = \Big\langle v, [W(\boldsymbol{\epsilon}^{k+1})]_{[1:k+1,1:k+1]}v\Big\rangle = \Big\langle v, [W_{k+1}]_{[1:k+1,1:k+1]}v\Big\rangle.$$

[1] [Rud87, Theorem 1.34]
[2] [Rud87, Theorem 10.18]

Hence the matrix $[W_{k+1}]_{[1:k+1,1:k+1]}$ is positive definite on the span of $\boldsymbol{v}$ and therefore $\widehat{W}^{k+1}_{[1:k+1,1:k+1]}$ is positive definite.

Repeating the argument inductively for all iterations we obtain that $\widehat{W}$ is positive definite. □

Remark 5

Under Assumptions 2 and 3 the original min-max problem (3) *is also uniquely solvable with* $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$. *To see this we first note that the min-max problem can be rewritten in terms of* $W(\boldsymbol{\epsilon})$*:*

$$\begin{aligned}\|C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon})\|^2 &= \left\| \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^{NM}(\boldsymbol{\alpha}_l - \boldsymbol{\alpha}_l^\star)B_l\Big)\boldsymbol{\epsilon}(s)ds \right\|^2 \\ &= \Big\langle (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star), W(\boldsymbol{\epsilon})(\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star)\Big\rangle.\end{aligned}$$

Similar to the proof of Proposition 7 and using the observability of the system, one can show that for any $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{NM}$ *with* $\hat{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}^\star$ *there exists a control* $\boldsymbol{\epsilon}(\hat{\boldsymbol{\alpha}})$ *such that*

$$0 < \Big\langle (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star), W(\boldsymbol{\epsilon}(\hat{\boldsymbol{\alpha}}))(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star)\Big\rangle.$$

Therefore the unique solution to (3) *is* $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$.

1.3 Improvements for the algorithm

1.3.1 More about the problems explained in Section 1.2.3

1.3.1.1 Initialization

To avoid the situation of Example 1 we change the initialization problem to

$$\max_{\boldsymbol{\epsilon}\in E_{ad}} \quad \left\| \int_0^T Ce^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds \right\|^2 . \tag{27}$$

This can be interpreted as comparing the state at the final time T with a control $\boldsymbol{\epsilon} \in E_{ad}$ to one with no control (control equal to zero), since

$$\begin{aligned} C\boldsymbol{y}_T(B_1, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B_1, 0) &= Ce^{TA}\boldsymbol{y}_0 + \int_0^T Ce^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds - Ce^{TA}\boldsymbol{y}_0 \\ &= \int_0^T Ce^{(T-s)A} B_1 \boldsymbol{\epsilon}(s) ds. \end{aligned}$$

We also have well-posedeness for this problem.

Proposition 8
The initialization problem (27) *is well-posed.*

The proof is analogous to the one in Proposition 2, only with $\boldsymbol{y}_0 = 0$.

1.3.1.2 Particular choice of the basis

Definition 1
We define the observability rank of a pair $(A, C) \in \mathbb{R}^{N\times N} \times \mathbb{R}^{P\times N}$, *with* $P \leq N$, *as*

$$\mathcal{R} := \mathcal{R}(N, A, C) := \operatorname{rank} \mathcal{O}_N(C, A) = \operatorname{rank} \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{N-1} \end{pmatrix} \tag{28}$$

We can show, that the method can recover exactly $\mathcal{R}M$ components of $\boldsymbol{\alpha}^\star$, if the matrices $\{B_i\}_{i=1}^K$ are chosen in the following way:

Assumtion 4
Let $\{\boldsymbol{v}_j\}_{j=1}^N \subset \mathbb{R}^N$ *be a basis of* $\mathbb{R}^N$*, such that*

$$\begin{aligned} \boldsymbol{v}_j &\notin \ker \mathcal{O}_N(C,A), \qquad j = 1,\dots,\mathcal{R}, \\ \boldsymbol{v}_j &\in \ker \mathcal{O}_N(C,A), \qquad j = \mathcal{R}+1,\dots,N, \end{aligned}$$

meaning that

$$\operatorname{span}\{\boldsymbol{v}_j\}_{j=\mathcal{R}+1}^N = \ker \mathcal{O}_N(C,A). \tag{29}$$

Let the family $\{B_k\}_{k=1}^{NM}$ *be defined as follows:*

$$\begin{aligned} B_1 &= \boldsymbol{v}_1 \boldsymbol{e}_1^T \\ B_2 &= \boldsymbol{v}_1 \boldsymbol{e}_2^T \\ &\vdots \\ B_M &= \boldsymbol{v}_1 \boldsymbol{e}_M^T \\ B_{M+1} &= \boldsymbol{v}_2 \boldsymbol{e}_1^T \\ &\vdots \\ B_{NM} &= \boldsymbol{v}_N \boldsymbol{e}_M^T \end{aligned}$$

where $\boldsymbol{e}_l \in \mathbb{R}^M$ *are the canonical vectors.*

Remark 6

1. *We have* $N = \operatorname{rank} \mathcal{O}_N(C,A) + \dim(\ker \mathcal{O}_N(C,A))$*, therefore the basis in Assumption 4 is well defined.*
2. *Since all the* $\boldsymbol{v}_j$ *are linearly independent, the set* $\{B_k\}_{k=1}^{NM}$ *form a basis of the matrix space* $\mathrm{Mat}_{N\times M}(\mathbb{R})$*.*
3. *One way to compute the vectors* $\boldsymbol{v}_j$ *is to use the singular value decomposition (SVD) of the observability matrix* $\mathcal{O}_N(C,A) = U\Sigma V^T$*. The last* $N-\mathcal{R}$ *columns of* V *span the kernel of* $\mathcal{O}_N(C,A)$*, while all columns together form a basis of* $\mathbb{R}^N$*. Therefore one possible choice would be* $\boldsymbol{v}_j = V_{[:,j]}$*,* $j = 1,\dots,N$*.*

1.3.1.3 Improved algorithm and convergence analysis

We first note that under Assumption 4 we can reduce the least-squares problem (4) to the first $\mathcal{R}M$ elements of the basis $\{B_k\}_{k=1}^{NM}$ or, respectively, the first $\mathcal{R}M$ coefficients. We define

$$B_{\mathcal{R}}(\boldsymbol{\alpha}^\star) = \sum_{j=1}^{\mathcal{R}M} \boldsymbol{\alpha}_j^\star B_j.$$

Lemma 6
Under Assumption 4 the least-squares problem (4) *is equivalent to*

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{\mathcal{R}M}} \quad \sum_{j=1}^{NM} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B_{\mathcal{R}}(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2. \tag{30}$$

proof. We start by noticing that for $1 \le l \le NM$, $0 \le s \le T$

$$\begin{aligned} Ce^{(T-s)A}B_l &= C\sum_{j=0}^{\infty} \frac{(T-s)^j}{j!} A^j B_l \\ &\overset{(\star)}{=} C\Big[\sum_{j=0}^{N-1} \tilde{\beta}_j(s) A^j\Big] B_l \\ &= \Big[\tilde{\beta}_0(s) I_N, \tilde{\beta}_1(s) I_N, \ldots, \tilde{\beta}_{N-1}(s) I_N\Big] \mathcal{O}_N(C,A) B_l, \end{aligned}$$

where $\tilde{\beta}_j(s)$ are the respective coefficients from the Cayley-Hamilton theorem, we use in $(\star)$.

If $l \in \{\mathcal{R}M+1, \ldots, NM\}$, then $B_l = \boldsymbol{v}_j \boldsymbol{e}_i^T$ with $j \ge \mathcal{R}+1$, hence $\boldsymbol{v}_j \in \ker \mathcal{O}_N(C,A)$ and therefore

$$\mathcal{O}_N(C,A)B_l = \underbrace{\mathcal{O}_N(C,A)\boldsymbol{v}_j}_{=0} \boldsymbol{e}_i^T = 0.$$

So we have for all $l \in \{\mathcal{R}M+1, \ldots, NM\}$ and $s \in [0,T]$

$$Ce^{(T-s)A}B_l = 0,$$

and therefore for any $\boldsymbol{\epsilon}$

$$\int_0^T Ce^{(T-s)A}B_l\boldsymbol{\epsilon}(s)ds = 0.$$

Now we can write the least-squares problem (4) as

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \sum_{j=1}^{NM} \left\| \sum_{l=1}^{NM} (\boldsymbol{\alpha}_l^\star - \boldsymbol{\alpha}_l) \int_0^T Ce^{(T-s)A}B_l\boldsymbol{\epsilon}^j(s)ds \right\|^2 \\ &= \sum_{j=1}^{NM} \left\| \sum_{l=1}^{\mathcal{R}M} (\boldsymbol{\alpha}_l^\star - \boldsymbol{\alpha}_l) \int_0^T Ce^{(T-s)A}B_l\boldsymbol{\epsilon}^j(s)ds \right\|^2. \end{aligned}$$

□

This means that we consider only the first $\mathcal{R}M$ elements of the basis. Recalling the convergence analysis from Section 1.2.4, it is therefore reasonable to assume that generating only $\mathcal{R}M$ controls (one for each relevant basis element) would be enough to get the existence of a unique solution for problem (30).

Assume that the basis $\mathcal{B} = (B_1, \dots, B_{NM})$ satisfies Assumption 4, we can now formulate an improved version of Algorithm 1, where we only perform fitting and discriminatory steps for the "observable" first $\mathcal{R}M$ basis elements.

Algorithm 2 Improved Greedy Reconstruction Algorithm

1: Solve the initialization problem (27) which gives the field $\boldsymbol{\epsilon}^1$ and set $k = 1$.
2: **while** $k \leq \mathcal{R}M - 1$ **do**
3: Fitting step: Find $\boldsymbol{\alpha}^k \in \mathbb{R}^k$ that solves (6).
4: Discriminatory step: Find $\boldsymbol{\epsilon}^{k+1}$ that solves (7).
5: Update $k \leftarrow k + 1$.
6: **end while**

The idea of proving convergence for this algorithm is analogous to the one in Section 1.2.4. Hence we have to show that the control generated by the initialization satisfies (26) and that the result from Proposition 7 is still true in this case. As in Section 1.2.1, we define $\widehat{W}$, W_k and $\widehat{W}^k$ and write all problems in matrix form, using the basis from Assumption 4 and replacing NM with $\mathcal{R}M$.

Proposition 9
Under Assumptions 4 any solution of the initialization problem (27) *satisfies the positivity condition* (26).

The proof is similar to Proposition 7. We only use B_1 and $\boldsymbol{v}_1$ instead of $\widetilde{B}$ and $\widetilde{B}_i$. The existence of an index $0 \leq j \leq N-1$ such that $CA^j\boldsymbol{v}_1 \neq$ is now guaranteed by $\boldsymbol{v}_1 \notin \ker \mathcal{O}_N(C, A)$.

Similarly one can prove that the result from Proposition 7 is still true.

Proposition 10
Let $\widehat{W}^k_{[1:k,1:k]}$, *and* $\boldsymbol{\alpha}^k$ *the solution to the fitting-step problem* (6). *Then, under Assumption 4, any solution* $\boldsymbol{\epsilon}^{k+1}$ *of the discriminatory-step problem* (7) *satisfies for* $k = 1, \dots \mathcal{R}M - 1$

$$J_D(\boldsymbol{\epsilon}^{k+1}) = \Big\langle \boldsymbol{v}, [W_{k+1}]_{[1:k+1,1:k+1]}\boldsymbol{v} \Big\rangle > 0,$$

where $\boldsymbol{v} := [\boldsymbol{\alpha}^k, \ -1]^\top$.

proof. We have

$$\begin{aligned} J_D(\boldsymbol{\epsilon}) &= \left\| \int_0^T Ce^{(T-s)A}B_{k+1}\boldsymbol{\epsilon}(s)ds - \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^k \boldsymbol{\alpha}_l^k B_l\Big)\boldsymbol{\epsilon}(s)ds \right\|^2 \\ &= \left\| C\Big(\int_0^T e^{(T-s)A}\Big(B_{k+1} - \sum_{l=1}^k \boldsymbol{\alpha}_l^k B_l\Big)\boldsymbol{\epsilon}(s)ds\Big) \right\|^2 \end{aligned}$$

Since $k+1 \leq \mathcal{R}M$ we have $\mathcal{O}_N(C,A)\Big(B_{k+1} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l\Big) \neq 0$ and the result follows by the proof of Proposition 7. □

Now we can prove convergence for Algorithm 2.

Theorem 2
Let $\{\boldsymbol{\epsilon}^1, \dots, \boldsymbol{\epsilon}^{\mathcal{R}M}\} \subset E_{ad}$ be a family of controls generated by Algorithm 1. Then, under Assumption 4, the least-squares problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{\mathcal{R}M}} \sum_{j=1}^{\mathcal{R}M} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B_{\mathcal{R}}(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2 \tag{31}$$

is uniquely solvable with $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i^\star$ for $i = 1, \dots, \mathcal{R}M$.

The proof is the same as for Theorem 1, only using Propositions 9 and 10 instead of Assumption 3 and Proposition 7.

Remark 7

1. *For $k \geq \mathcal{R}M + 1$ we have $\mathcal{O}_N(C,A)B_k = 0$ and therefore*

$$C\boldsymbol{y}_T(B_k, \boldsymbol{\epsilon}) = Ce^{TA}\boldsymbol{y}_0 + \underbrace{\int_0^T Ce^{(T-s)A} B_k \boldsymbol{\epsilon}(s) ds}_{=0} = Ce^{TA}\boldsymbol{y}_0$$

   for any $T > 0$ and $\boldsymbol{\epsilon}$ (compare Section 1.2.3.2). This means that for $k \geq \mathcal{R}M + 1$ the effect of B_k on the state $\boldsymbol{y}$ can not be observed and therefore it is impossible to recover the respective coefficients of $\boldsymbol{\alpha}^\star$.

2. *Let $\boldsymbol{\alpha}^{approx}$ be the solution to (31). Then we get the a-priori error estimate*

$$B^\star - B_{\mathcal{R}}(\boldsymbol{\alpha}^{approx}) = \sum_{j=\mathcal{R}M+1}^{NM} \boldsymbol{\alpha}_j^\star B_j.$$

3. *Following the argumentation from the proof of Lemma 6 we can show that the min-max problem (3) is equivalent to*

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{\mathcal{R}M}} \max_{\boldsymbol{\epsilon} \in E_{ad}} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}) \right\|^2. \tag{32}$$

   Analogously to Remark 5 we can conclude that, under Assumption 4, problem (32) is uniquely solvable with $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i^\star$ for $i = 1, \dots, \mathcal{R}M$.

The remaining questions are, whether one can always recover $\mathcal{R}M$ coefficients for any basis and whether there might exist a basis, for which one can recover more than $\mathcal{R}M$ coefficients.

The answer to the first question is that this is generally not the case, as one can see from the following example.

Example 2

Consider a two-dimensional case with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ C = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The observability matrix and its SVD are

$$\mathcal{O}_2(C,A) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} =: U\Sigma V^T.$$

Now we have

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \notin \ker \mathcal{O}_N(C,A), \quad \boldsymbol{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in \ker \mathcal{O}_N(C,A).$$

Therefore a basis satisfying Assumption 4 would be

$$B_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, B_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, B_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Define

$$B^\star = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = 1 \cdot B_1 + 1 \cdot B_2 + 1 \cdot B_3 + 1 \cdot B_4.$$

Hence we have $\boldsymbol{\alpha}^\star = [1\ 1\ 1\ 1]^T$.
By Theorem 2 with the controls generated by the algorithm we can recover exactly $\boldsymbol{\alpha}_1^\star = 1$ *and* $\boldsymbol{\alpha}_2^\star = 1$, *resulting in*

$$B^{approx} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

Consider now another basis defined by

$$\widehat{B}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \widehat{B}_2 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \widehat{B}_3 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \widehat{B}_4 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

We get

$$\mathcal{O}_2(C,A)(\widehat{B}_2 - \widehat{B}_1) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\mathcal{O}_2(C,A)(\widehat{B}_4 - \widehat{B}_3) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

This means that any $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\ \boldsymbol{\alpha}_3\ \boldsymbol{\alpha}_4] \in \mathbb{R}^4$ *with* $\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = 1$ *and* $\boldsymbol{\alpha}_3 + \boldsymbol{\alpha}_4 = 1$ *will solve the least-squares problem* (4), *since for any control* $\boldsymbol{\epsilon}$ *we have*

$$\begin{aligned} C\boldsymbol{y}_T(B^\star,\boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}),\boldsymbol{\epsilon}) &= Ce^{TA}\boldsymbol{y}_0 + C\int_0^T e^{(T-s)A}B^\star\boldsymbol{\epsilon}(s)ds - Ce^{TA}\boldsymbol{y}_0 - C\int_0^T e^{(T-s)A}B(\boldsymbol{\alpha})\boldsymbol{\epsilon}(s)ds \\ &= \int_0^T Ce^{(T-s)A}\Big(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3+\boldsymbol{\alpha}_4 \\ \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_4 \end{pmatrix}\Big)\boldsymbol{\epsilon}(s)ds \\ &= \int_0^T \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} e^{T-s} & 0 \\ 0 & e^{T-s} \end{pmatrix} \begin{pmatrix} 1-(\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2) & 1-(\boldsymbol{\alpha}_3+\boldsymbol{\alpha}_4) \\ 1-\boldsymbol{\alpha}_2 & 1-\boldsymbol{\alpha}_4 \end{pmatrix}\boldsymbol{\epsilon}(s)ds \\ &= \int_0^T \begin{pmatrix} e^{T-s} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1-\boldsymbol{\alpha}_2 & 1-\boldsymbol{\alpha}_4 \end{pmatrix}\boldsymbol{\epsilon}(s)ds \\ &= 0. \end{aligned}$$

Therefore we get

$$\widehat{B}^{approx} = \begin{pmatrix} 1 & 1 \\ \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_4 \end{pmatrix},$$

and hence

$$\|B^\star - B_{\mathcal{R}}(\boldsymbol{\alpha}^{approx})\|_F^2 = (1-\boldsymbol{\alpha}_2)^2 + (1-\boldsymbol{\alpha}_4)^2$$

for any $\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_4 \in \mathbb{R}$, where by $\|\cdot\|_F$ we denote the Frobenius norm.

This example also shows that for a random basis, one can not say anything about the quality of the coefficients or the difference between $B(\boldsymbol{\alpha})$ and $B^\star$. Even if one would by chance guess the right coefficients (in this case $\boldsymbol{\alpha}_2 = 1, \boldsymbol{\alpha}_4 = 1$) there would be no way to verify it, since their effect is not observable.

The answer to the question, whether there might exist a basis, for which one can recover more than $\mathcal{R}M$ coefficient, is given by the following proposition.

Proposition 11
There exists no basis for which one can recover more than $\mathcal{R}M$ coefficients.

proof. Consider a basis $\mathcal{B} = \{B_k\}_{k=1}^{NM} \subset \mathbb{R}^{N\times M}$ that satisfies Assumption 4 and an arbitrary basis $\widehat{\mathcal{B}} = \{\widehat{B}_k\}_{k=1}^{NM} \subset \mathbb{R}^{N\times M}$. Any element $\widehat{B} \in \widehat{\mathcal{B}}$ can be written as a linear combination of B_k in the following sense

$$\widehat{B} = \sum_{j=1}^{NM} \lambda_j B_j$$

where $\lambda_j \in \mathbb{R}$. Multiplying with the observability matrix, we get

$$\begin{aligned} \mathcal{O}_N(C,A)\widehat{B} &= \mathcal{O}_N(C,A)\Big[\sum_{j=1}^{NM} \lambda_j B_j\Big] \\ &= \sum_{j=1}^{NM} \lambda_j \mathcal{O}_N(C,A) B_j \\ &= \sum_{j=1}^{\mathcal{R}M} \lambda_j \mathcal{O}_N(C,A) B_j, \end{aligned}$$

where in the last step we used that $\mathcal{O}_N(C,A)B_j = 0$ for $j \in \{\mathcal{R}+1, \dots, N\}$. Now define the set $\mathcal{D} = \{D_k\}_{k=1}^{NM}$ as

$$D_k := \mathcal{O}_N(C,A)\widehat{B}_k, \quad k = 1, \dots, NM.$$

From the calculations above we can conclude, that at most $\mathcal{R}M$ elements of $\mathcal{D}$ are linearly independent. This means that for $NM - \mathcal{R}M$ elements of $\widehat{\mathcal{B}}$ there exists a linear combination of the other $\mathcal{R}M$ elements such that the observation at final time T is identical for any control $\boldsymbol{\epsilon}$. Therefore one can at most recover $\mathcal{R}M$ coefficients for the basis $\widehat{\mathcal{B}}$. □

Remark 8

1. *It is important to note that in the last part we only discussed about quality of coefficients in the respective basis. The difference in observation at final time T between the approximated matrix and the true one is the same for any basis, since in any basis there are exactly $\mathcal{R}M$ matrices that are linearly independent after multiplication with the observability matrix.*

2. *This improved choice of basis can be very impactful in situations, where the system is not "very observable", i.e. the observability rank is very low ($\mathcal{R} \ll N$), and "online" experiments are costly. In these cases using the improved algorithm reduces the number of controls generated and therefore also the number of experiments needed from NM to $\mathcal{R}M$.*
This is already noticeable in a three-dimensional case ($N = M = 3$) with an observability rank of $\mathcal{R} = 1$, where the improvement reduces the number of controls by $\approx 67\%$ from 9 to only 3. We will discuss numerical examples in Section 1.4.

1.3.2 Other improvements

1.3.2.1 Rank-revealing/swap strategy

Until now we have assumed that we need one control for each element in the basis $\mathcal{B}$. In other words we have assumed that the rank of $\widehat{W}^k$ is exactly k, improving by one for each control we add.
However in practice a control chosen by the algorithm will often improve the rank of $\widehat{W}^k$ by more than one (see for example Section 1.4.4). This means that one control can be used to identify multiple coefficients of the basis $\mathcal{B}$ at once. Therefore we could theoretically skip some iterations, where the coefficient is already identifiable, and further reduce the number of generated controls, while still obtaining a full rank matrix $\widehat{W}$.
One way to implement this is given in the following sub-algorithm before every fitting step.

Algorithm 3 Sub-algorithm swap strategy

1: Compute the full matrix $\widehat{W}^k$ and set $p = k$.
2: **for** $j \in \{k+1, \dots, \mathcal{R}M\}$ **do**
3: **if** $\bigcup_{i=1}^{p} \widehat{W}^k_{[:,i]}$ and $\widehat{W}^k_{[:,j]}$ are linearly independent **then**
4: Swap the j-th and $(p+1)$-th column and row of $\widehat{W}^k$ (see Figure 1) and the j-th and $(p+1)$-th element of $\mathcal{B}$.
5: Set $p = p + 1$.
6: **end if**
7: **end for**
8: Set $k = p + 1$.

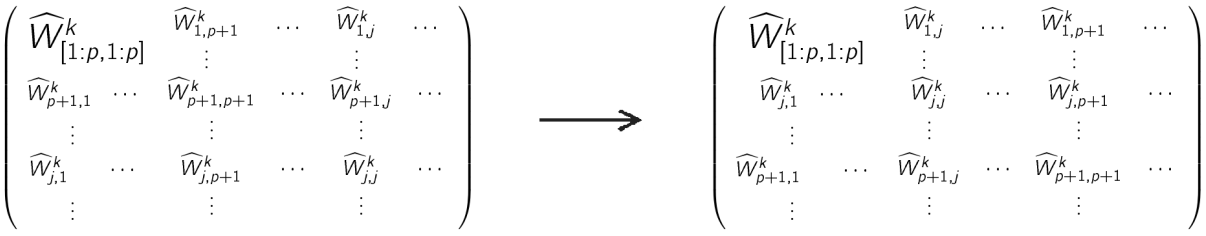


Figure 1 Swapping the j-th and $p+1$-th column and row of $\widehat{W}^k$.

Remark 9
The drawback of this approach is that one has to assemble the whole matrix $\widehat{W}^k$ in each iteration, which increases the computational time needed by the algorithm. In exchange we (potentially) generate less controls and therefore have to perform less laboratory experiments to generate the data $C\boldsymbol{y}(B^\star, \boldsymbol{\epsilon}^j)$ ($1 \leq j \leq$ #controls generated).

1.3.2.2 Extended greedy-testing strategy

Until now we always considered a full basis of the matrix space $Mat_{N\times M}(\mathbb{R})$. What if we were given a fixed basis with dimension $K < NM$ or if there were restrictions that our computed approximation of $B^\star$ had to fulfil? In these cases it might not be possible to choose a basis that has similar properties as in Assumption 4. However we still know that we can recover at most $\mathcal{R}M$ elements for this basis and that $\mathcal{R}M$ controls are sufficient.

In this context we introduce the following adjustment to the algorithm that aims at selecting $\mathcal{R}M$ recoverable elements from any basis of higher dimension.

Algorithm 4 Extended Greedy Reconstruction Algorithm

Require: A basis $\mathcal{B} = (B_1, \dots, B_K)$.

1: Solve the initialization problem

$$\max_{i\in\{1,\dots,K\}} \max_{\boldsymbol{\epsilon}\in L^2(0,T)} \|C\boldsymbol{y}_T(B_i, \boldsymbol{\epsilon})\|^2$$

which gives the field $\boldsymbol{\epsilon}^1$ and the index i_1.

2: Switch B_1 and B_{i_1} in $\mathcal{B}$ and set $k = 1$.

3: **while** $k \leq \min\{\mathcal{R}M, K\} - 1$ **do**

4: **for** $i = k+1, \dots, K$ **do**

5: Fitting step: Find $(\boldsymbol{\alpha}^i_j)_{j=1,\dots,k}$ that solve the problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^k} \sum_{j=1}^{k} \left\| C\boldsymbol{y}_T(B_i, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2. \tag{33}$$

6: **end for**

7: Extended discriminatory step: Find $\boldsymbol{\epsilon}^{k+1}$ and i_{k+1} that solve the problem:

$$\max_{i\in\{k+1,\dots,K\}} \max_{\boldsymbol{\epsilon}\in E_{ad}} \left\| C\boldsymbol{y}_T(B_i, \boldsymbol{\epsilon}) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}^i), \boldsymbol{\epsilon}) \right\|^2. \tag{34}$$

8: Switch B_{k+1} and $B_{i_{k+1}}$ in $\mathcal{B}$ and update $k \leftarrow k+1$.

9: **end while**

We basically extend the greedy character of the algorithm to the choice of the next basis element, meaning that in each iteration we consider all remaining basis elements as the potential next one and we select the one which yields the largest function value in the respective maximization. In other words we want to find the basis element for which we can split the observation the most from all previous basis elements.

Remark 10

1. *Unlike the swap strategy from the last section, this improvement does not necessarily increase the computation time of the algorithm, since one can parallelize the fitting and discriminatory steps for the remaining basis elements in each iteration k. Then the only additional step needed is a comparison of $\mathcal{R}M - k$ function values, to find the maximal one, which requires a negligible computational time.*
2. *This adjustment would even allow us to enlarge the set $\{B_k\}_{k=1}^K$ with some elements that are linearly dependent on $B_1, \dots, B_K$ and let the extended greedy algorithm pick the $\mathcal{R}M$ "best" ones. Note that a selected element B_{k+1} will unlikely be linearly dependent on previous elements (before or after multiplication with the observability matrix). Otherwise there would exist an $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that $B_{k+1} - \sum_{j=1}^k \boldsymbol{\alpha}_j B_j = 0$ (or respectively*

$\mathcal{O}_N(C, A)(B_{k+1} - \sum_{j=1}^{k} \boldsymbol{\alpha}_j B_j) = 0$*). This* $\boldsymbol{\alpha}$ *would therefore be a solution of the fitting step problem with cost function value equal to zero. However the corresponding cost functional of the discriminatory step problem would also be zero for any control* $\boldsymbol{\epsilon}$*.*

3. *This improvement is also applicable in other setting, for example the bilinear Schrödinger-type equation that we will consider in Chapter 2 (compare Section 2.5.4).*

1.4 Discretization and numerical results

We want to validate some of our theoretical results in numerical experiments and point out possible numerical instabilities. In order to numerically solve all sub-step problems of the algorithm we first need to introduce first-order optimality conditions. Throughout this section we will consider the improved initialization problem (27).

1.4.1 First order optimality conditions

We already covered this part for the fitting step (6) in Proposition 6. A similar result can be obtained for the main identification problem (4). We rewrite the cost functional as

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^{NM} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}), \boldsymbol{\epsilon}^j) \right\|^2 \tag{35}$$

$$= \sum_{j=1}^{NM} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - \int_0^T Ce^{(T-s)A}\Big(\sum_{l=1}^{NM} \boldsymbol{\alpha}_l B_l\Big)\boldsymbol{\epsilon}^j(s)ds \right\|^2 = \sum_{j=1}^{NM} \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - \Gamma^j\boldsymbol{\alpha} \right\|^2, \tag{36}$$

with Γ^j defined in (23). Using normal equations we get the following result.

Proposition 12
Let $\widehat{W}$ be positive definite. Then problem (4) is uniquely solvable with sufficient first-order optimality condition

$$\widehat{W}\boldsymbol{\alpha} = \sum_{j=1}^{NM} (\Gamma^j)^T (C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j)). \tag{37}$$

Remark 11
Recall that, in true applications, the term $C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j)$ is fully obtained by experimental results and can therefore not be rewritten as we have done before (expanding $B^\star$) for our theoretical convergence results.

For the initialization problem (27) and the discriminatory step problem (7), which are both optimal control problems, we use a "Optimize-Before-Discretize" approach, meaning that we discretize the gradient of the continuous problems. As we discussed before, the two problems can both be written as

$$\max_{\boldsymbol{\epsilon}\in E_{ad}} \left\| \int_0^T Ce^{(T-s)A}\widetilde{B}\boldsymbol{\epsilon}^{k+1}(s)ds \right\|^2, \tag{38}$$

with $\widetilde{B} = B_1$ for the initialization and $\widetilde{B} = B_{k+1} - \sum_{l=1}^{k} \boldsymbol{\alpha}_l^k B_l$ for the discriminatory step. Therefore it is sufficient to compute the first-order optimality condition for problem (38). From now on we consider $\boldsymbol{y}(t) = \boldsymbol{y}(\widetilde{B}, \boldsymbol{\epsilon})$. We denote by J the cost functional of (38)

$$J(\boldsymbol{y}, \boldsymbol{\epsilon}) := \|C\boldsymbol{y}(T)\|^2 = \left\| \int_0^T Ce^{(T-s)A}\widetilde{B}\boldsymbol{\epsilon}^{k+1}(s)ds \right\|^2$$

and construct the reduced cost functional

$$\widehat{J}(\boldsymbol{\epsilon}) := J(\boldsymbol{y}(\boldsymbol{\epsilon}), \boldsymbol{\epsilon}).$$

We introduce the adjoint equation

$$\begin{cases} -\dot{\boldsymbol{\lambda}}(t) = A^T \boldsymbol{\lambda}(t) \\ \boldsymbol{\lambda}(T) = 2C^T C \boldsymbol{y}(T), \end{cases} \tag{39}$$

and the linearised state equation

$$\begin{cases} \dot{\delta \boldsymbol{y}}(t) = A \delta \boldsymbol{y}(t) + \widetilde{B} \delta \boldsymbol{\epsilon} \\ \delta \boldsymbol{y}(0) = 0, \end{cases}$$

for any control variation $\delta\boldsymbol{\epsilon}$. Now the L^2 gradient of $\widehat{J}$ can be written as

$$\begin{aligned} \left\langle \nabla \widehat{J}(\boldsymbol{\epsilon}), \delta \boldsymbol{\epsilon} \right\rangle_{L^2(0,T;\mathbb{R}^M)} &= \left\langle \frac{\partial J(\boldsymbol{y}, \boldsymbol{\epsilon})}{\partial \boldsymbol{y}}, \delta \boldsymbol{y} \right\rangle_{H^{-1}(0,T,\mathbb{R}^N), H^1(0,T;\mathbb{R}^N)} + \underbrace{\left\langle \frac{\partial J(\boldsymbol{y}, \boldsymbol{\epsilon})}{\partial \boldsymbol{\epsilon}}, \delta \boldsymbol{\epsilon} \right\rangle_{L^2(0,T;\mathbb{R}^M)}}_{=0} \\ &= \left\langle 2C^T C \boldsymbol{y}(T), \delta \boldsymbol{y}(T) \right\rangle \\ &= \left\langle \boldsymbol{\lambda}(T), \delta \boldsymbol{y}(T) \right\rangle, \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{H^{-1}(0,T,\mathbb{R}^N), H^1(0,T;\mathbb{R}^N)}$ is the dual pairing. Integrating by parts we obtain

$$\begin{aligned} \left\langle \boldsymbol{\lambda}(T), \delta \boldsymbol{y}(T) \right\rangle &= \left\langle \lambda(0), \underbrace{\delta \boldsymbol{y}(0)}_{=0} \right\rangle + \left\langle \dot{\boldsymbol{\lambda}}, \delta \boldsymbol{y} \right\rangle_{L^2(0,T;\mathbb{R}^N)} + \left\langle \boldsymbol{\lambda}, \dot{\delta \boldsymbol{y}} \right\rangle_{L^2(0,T;\mathbb{R}^N)} \\ &= \left\langle -A^T \boldsymbol{\lambda}, \delta \boldsymbol{y} \right\rangle_{L^2(0,T;\mathbb{R}^N)} + \left\langle \boldsymbol{\lambda}, A \delta \boldsymbol{y} + \widetilde{B} \delta \boldsymbol{\epsilon} \right\rangle_{L^2(0,T;\mathbb{R}^N)} \\ &= \left\langle \widetilde{B}^T \boldsymbol{\lambda}, \delta \boldsymbol{\epsilon} \right\rangle_{L^2(0,T;\mathbb{R}^M)} \end{aligned}$$

and therefore

$$\left\langle \nabla \widehat{J}(\boldsymbol{\epsilon}), \delta \boldsymbol{\epsilon} \right\rangle_{L^2(0,T;\mathbb{R}^M)} = \left\langle \widetilde{B}^T \boldsymbol{\lambda}, \delta \boldsymbol{\epsilon} \right\rangle_{L^2(0,T;\mathbb{R}^M)}.$$

Since we restrict our control space to the convex and closed set E_{ad}, we need to introduce a variational inequality. If $\hat{\boldsymbol{\epsilon}}$ is a maximizer for (38) and $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{\lambda}})$ solve the corresponding state and adjoint equation, then

$$0 \geq \left\langle \nabla \widehat{J}(\hat{\boldsymbol{\epsilon}}), \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}} \right\rangle_{L^2(0,T;\mathbb{R}^M)} = \left\langle \widetilde{B}^T \hat{\boldsymbol{\lambda}}, \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}} \right\rangle_{L^2(0,T;\mathbb{R}^M)}.$$

for any $\boldsymbol{\epsilon} \in E_{ad}$ (compare [Cia20, Section 3.1]). We obtain the first order optimality system

$$\dot{\hat{\boldsymbol{y}}}(t) = A\hat{\boldsymbol{y}}(t) + \widetilde{B}\hat{\boldsymbol{\epsilon}}(t) \quad \text{in } (0,T), \quad \hat{\boldsymbol{y}}(0) = \boldsymbol{y}_0, \tag{40a}$$

$$-\dot{\hat{\boldsymbol{\lambda}}}(t) = A^T \hat{\boldsymbol{\lambda}}(t) \quad \text{in } (0,T), \quad \hat{\boldsymbol{\lambda}}(T) = 2C^T C \hat{\boldsymbol{y}}(T), \tag{40b}$$

$$0 \geq \left\langle \widetilde{B}^T \hat{\boldsymbol{\lambda}}, \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}} \right\rangle_{L^2(0,T;\mathbb{R}^M)} \quad \forall \boldsymbol{\epsilon} \in E_{ad}. \tag{40c}$$

1.4.2 Discretization and numerical solvers

For the discretization we consider a uniform time grid with step size $h = T/N_t$ and introduce the discrete L^2-product

$$\left\langle \boldsymbol{v}, \boldsymbol{w} \right\rangle_{L^2_h} := h \sum_{j=0}^{N_t} \boldsymbol{v}_j^T \boldsymbol{w}_j.$$

For the discrete set E_{ad}^h we use simple box constraints:

$$E_{ad}^h = \{\boldsymbol{\epsilon} \in L_h^2 \mid |(\boldsymbol{\epsilon}_j)_i| \leq b,\ 1 \leq j \leq N_t,\ 1 \leq i \leq M\}, \tag{41}$$

for some $b > 0$.
In order to solve the discrete version of our system (1), we use a trapezoidal method that corresponds to the Crank-Nicolson scheme. This means that we compute the approximated solution $\boldsymbol{y}_j = \boldsymbol{y}(jh)$, $0 < j \leq N_t$ as the solution to the following equation.

$$\boldsymbol{y}_j = \boldsymbol{y}_{j-1} + \frac{h}{2}\Big((A\boldsymbol{y}_{j-1} + B\boldsymbol{\epsilon}_{j-1}) + (A\boldsymbol{y}_j + B\boldsymbol{\epsilon}_j)\Big).$$

The solutions to the discrete versions of the main identification and the fitting step can be calculated by solving the linear systems (37) and (24), respectively.

For the initialization and the discriminatory step problem we also discretize the adjoint equation (39) with a Crank-Nicolson scheme and obtain the following discrete optimality system.

$$\hat{\boldsymbol{y}}_j = \hat{\boldsymbol{y}}_{j-1} + \frac{h}{2}\Big((A\hat{\boldsymbol{y}}_{j-1} + \widetilde{B}\hat{\boldsymbol{\epsilon}}_{j-1}) + (A\hat{\boldsymbol{y}}_j + \widetilde{B}\hat{\boldsymbol{\epsilon}}_j)\Big), \qquad j = 1, \dots, N_t, \tag{42a}$$

$$\hat{\boldsymbol{y}}_0 = \boldsymbol{y}_0, \tag{42b}$$

$$\hat{\boldsymbol{\lambda}}_j = \hat{\boldsymbol{\lambda}}_{j+1} + \frac{h}{2}\Big((A^T\hat{\boldsymbol{\lambda}}_{j+1}) + (A^T\hat{\boldsymbol{\lambda}}_j)\Big), \qquad j = N_t - 1, \dots, 0, \tag{42c}$$

$$\hat{\boldsymbol{\lambda}}_{N_t} = 2C^TC\hat{\boldsymbol{y}}_{N_t}, \tag{42d}$$

$$0 \geq \left\langle \widetilde{B}^T\hat{\boldsymbol{\lambda}}, \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}} \right\rangle_{L_h^2} \qquad \forall \boldsymbol{\epsilon} \in E_{ad}^h, \tag{42e}$$

where we again use $\widetilde{B} = B_1$ for the initialization and $\widetilde{B} = B_{k+1} - \sum_{l=1}^k \boldsymbol{\alpha}_l^k B_l$ for the discriminatory step. The discretized reduced cost functional can be written as

$$\widehat{J}_h(\boldsymbol{\epsilon}) = \left\| h \sum_{j=0}^{N_t} Ce^{(T-jh)A}\widetilde{B}\boldsymbol{\epsilon}_j \right\|^2. \tag{43}$$

To solve these optimal control problems we use a projected gradient method (compare [Cia20, Section 5.1], [Kel99, Section 5.4]), meaning that given a current iterate $\boldsymbol{\epsilon}^c$ we choose the new iterate $\boldsymbol{\epsilon}^{new}$ as

$$\boldsymbol{\epsilon}^{new} = \mathbb{P}_{E_{ad}^h}[\boldsymbol{\epsilon}^c + \sigma\nabla\widehat{J}_h(\boldsymbol{\epsilon}^c)]\Big(= \mathbb{P}_{E_{ad}^h}[\boldsymbol{\epsilon}^c + \sigma\widetilde{B}^T\hat{\boldsymbol{\lambda}}]\Big),$$

where $\mathbb{P}$ is the projection onto the set E_{ad}^h (in our case pointwise: $\mathbb{P}_{E_{ad}^h}[(\boldsymbol{\epsilon}_j)_i] = \min\{\max\{(\boldsymbol{\epsilon}_j)_i, b\}, b\}$). The parameter σ is a (positive) step length obtained by a projected line-search and satisfying the sufficient increase condition

$$\widehat{J}_h(\boldsymbol{\epsilon}(\sigma)) - \widehat{J}_h(\boldsymbol{\epsilon}) \geq \frac{\eta}{\sigma}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}(\sigma)\right\|,$$

where η is a given parameter (we will consider $\eta = 10^{-2}$ in all of our examples) and

$$\boldsymbol{\epsilon}(\sigma) = \mathbb{P}_{E_{ad}^h}[\boldsymbol{\epsilon} + \sigma\nabla\widehat{J}_h(\boldsymbol{\epsilon})].$$

Remark 12
Note that the initial guess for the projected gradient method has to be non-zero, since $\boldsymbol{\epsilon}^0 \equiv 0$ is a minimizer and therefore a stationary point of $\widehat{J}_h$ with $\nabla\widehat{J}_h(0) = 0$, meaning that the method would always return $\boldsymbol{\epsilon} = 0$ as solution. In all of our following examples in this chapter we will consider $\boldsymbol{\epsilon}^0 \equiv 1$.

1.4.3 Algorithm with improved basis

Throughout this and all following sections in this chapter we consider

$$N_t = 1000, \quad \boldsymbol{y}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad b = 100.$$

The simplest example to demonstrate the different effects of the improved basis from Assumption 4 is a three-dimensional setting, where we consider A equal to the identity and different observers C with varying ranks/number of rows.

Example 3
Consider $T = 10$, $N = M = 3$, $A = \mathcal{I}_3$ and the following observers with randomly chosen entries

$$C_1 = \begin{pmatrix} 0.815 & 0.913 & 0.279 \\ 0.906 & 0.632 & 0.547 \\ 0.127 & 0.098 & 0.958 \end{pmatrix}, \; C_2 = \begin{pmatrix} 0.046 & 0.824 & 0.317 \\ 0.097 & 0.695 & 0.950 \end{pmatrix}, \; C_3 = \begin{pmatrix} 0.709 & 0.755 & 0.276 \end{pmatrix},$$

meaning that

$$\mathcal{R}_1 := \operatorname{rank} \mathcal{O}_3(C_1, A) = 3 = N, \quad \mathcal{R}_2 := \operatorname{rank} \mathcal{O}_3(C_2, A) = 2, \quad \mathcal{R}_3 := \operatorname{rank} \mathcal{O}_3(C_3, A) = 1.$$

For the basis $\mathcal{B}$ we will compare three different settings, the unit basis, a random basis and an improved basis satisfying Assumption 4. The improved basis is chosen using the SVD of the observability matrix (compare Remark 6 point three). Note that the improved basis is different for each observer.
The true control matrix $B^\star$ is chosen with random entries but is the same for all cases:

$$B^\star = \begin{pmatrix} -12.546 & -10.318 & -5.964 \\ 1.938 & 1.978 & 7.340 \\ -5.351 & -8.298 & -3.770 \end{pmatrix}.$$

We begin by comparing the results of the final identification (4) *and* (31) *respectively.*

$\boldsymbol{\alpha}^\star$	$\boldsymbol{\alpha}^{sol}(C_1)$	$\boldsymbol{\alpha}^{sol}(C_2)$	$\boldsymbol{\alpha}^{sol}(C_3)$
−12.55	−12.55	−0.51	−5.50
−10.32	−10.32	−0.74	−5.01
−5.96	−5.96	−0.20	0.17
1.94	1.94	1.66	−5.86
1.98	1.98	1.76	−5.33
7.34	7.34	7.21	0.18
−5.35	−5.35	−6.38	−2.14
−8.20	−8.20	−9.11	−1.95
−3.77	−3.77	−4.26	0.06

Table 1 Solutions of the main identification (4) for different observers using the unit basis.

$\boldsymbol{\alpha}^\star$	$\boldsymbol{\alpha}^{sol}(C_1)$	$\boldsymbol{\alpha}^{sol}(C_2)$	$\boldsymbol{\alpha}^{sol}(C_3)$
−5.21	−5.21	6.85	1.74
9.86	9.86	11.80	4.21
−36.63	−36.63	−5.47	0.69
8.57	8.57	10.19	−0.31
−4.32	−4.32	8.92	−2.35
−14.21	−14.21	0.90	−1.12
−33.77	−33.77	0.64	1.22
31.74	31.74	−0.30	1.52
21.66	21.66	9.34	−0.25

Table 2 Solutions of the main identification (4) for different observers using the random basis.

$\boldsymbol{\alpha}^\star(C_1)$	$\boldsymbol{\alpha}^{sol}(C_1)$	$\boldsymbol{\alpha}^\star(C_2)$	$\boldsymbol{\alpha}^{sol}(C_2)$	$\boldsymbol{\alpha}^\star(C_3)$	$\boldsymbol{\alpha}^{sol}(C_3)$
$\begin{pmatrix} 9.65 \\ 1.58 \\ 9.71 \\ 9.57 \\ 4.85 \\ 8.00 \\ 1.42 \\ 4.22 \\ 9.16 \end{pmatrix}$	$\begin{pmatrix} 9.65 \\ 1.58 \\ 9.71 \\ 9.57 \\ 4.85 \\ 8.00 \\ 1.42 \\ 4.22 \\ 9.16 \end{pmatrix}$	$\begin{pmatrix} 3.12 \\ -5.82 \\ 4.91 \\ -7.91 \\ -2.40 \\ -8.02 \\ 12.09 \\ 9.62 \\ 5.79 \end{pmatrix}$	$\begin{pmatrix} 3.12 \\ -5.82 \\ 4.91 \\ -7.91 \\ -2.40 \\ -8.02 \end{pmatrix}$	$\begin{pmatrix} 8.32 \\ 7.57 \\ -0.25 \\ -5.69 \\ -9.39 \\ -8.47 \\ -7.08 \\ -6.06 \\ -8.18 \end{pmatrix}$	$\begin{pmatrix} 8.32 \\ 7.57 \\ -0.25 \end{pmatrix}$

Table 3 Solutions of the improved identification (31) for different observers using the corresponding improved basis.

As we can see, for the unit and random basis the algorithm can only recover the actual coefficients if the system is observable (second column of Table 1 and 2). Otherwise some coefficients can lie somewhat near to the actual ones (see the lower six entries in the third column of Table 1) but in most cases they are very different. This is due to linear dependencies of the basis elements after multiplication with the observability matrix (compare Example 2 and the proof of Proposition 11). However in case of the improved basis, the (improved) algorithm recovers exactly $\mathcal{R}M = \mathcal{R} \cdot 3$ of the true coefficients, as expected.

Next we want to look at the relative error of the computed solution $B^{sol} = B(\boldsymbol{\alpha}^{sol})$ and the true $B^\star$.

	unit basis	random basis	improved basis
C_1	0	0	0
C_2	0.7925	1.030	0.7925
C_3	0.8847	0.828	0.8847

Table 4 Relative norm difference $\frac{\|B^\star - B^{sol}\|_2}{\|B^\star\|_2}$ for the three different bases and observers.

The error for the improved basis is of course exactly the norm of the coefficients that are not recoverable multiplied with their corresponding basis elements, therefore it is increasing with a decreasing rank of the observability matrix. The error for the unit basis seems to be the same as for the improved basis and indeed the computed solution B^{sol} is the same for the unit and improved basis in this example. However this is generally not the case and very much depending on the method and especially the starting value used to solve the main identification problem (4)*, since it has multiple solutions if the observability matrix does not have full rank. Finally for the random basis we get an error that is smaller than for the other bases in case of $C = C_2$ and a larger one in case of $C = C_3$. Also here the error is decreasing with a decreasing rank of observability, meaning that for $C = C_3$ we get a computed solution that is closer to the true $B^\star$ as for $C = C_2$ even though we can observe less.*
In summary this confirms that for a basis which is not chosen in a way as in Assumption 4, we can not say anything about the quality of the computed solution B^{sol} a-priori.

1.4.4 Rank-revealing improvement

When we look at the rank of $\widehat{W}^k$ in Example 3 for the improved basis and $C = C_1$ (rank $\mathcal{O}_3(C_1, A) = 3 = N \rightsquigarrow \widehat{W}^k \in \mathbb{R}^{9\times 9}$) we get the following results.

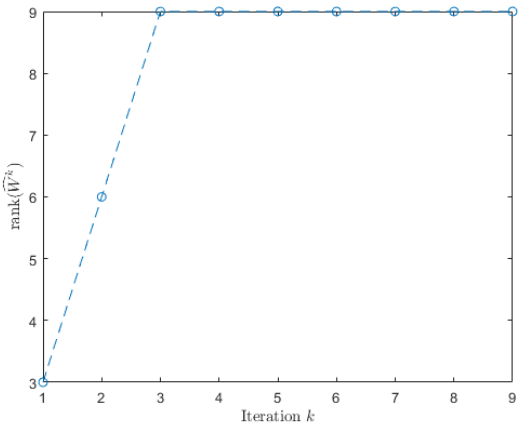


Figure 2 Rank progress of $\widehat{W}^k$ for $C = C_1$.

As we can see the rank of $\widehat{W}^k$ only increases in the first three iterations, each time by exactly 3. This means that $\widehat{W}^3$ is already full rank and therefore the first three controls would be enough to make our main identification problem 15 uniquely solvable. Hence the swap strategy from Section 1.3.2.1 would stop after three iterations, hence generating 6 controls less and therefore reducing the number of experiments to be made by $\approx 67\%$.

In this example this can also be seen very clearly when looking at the corresponding controls.

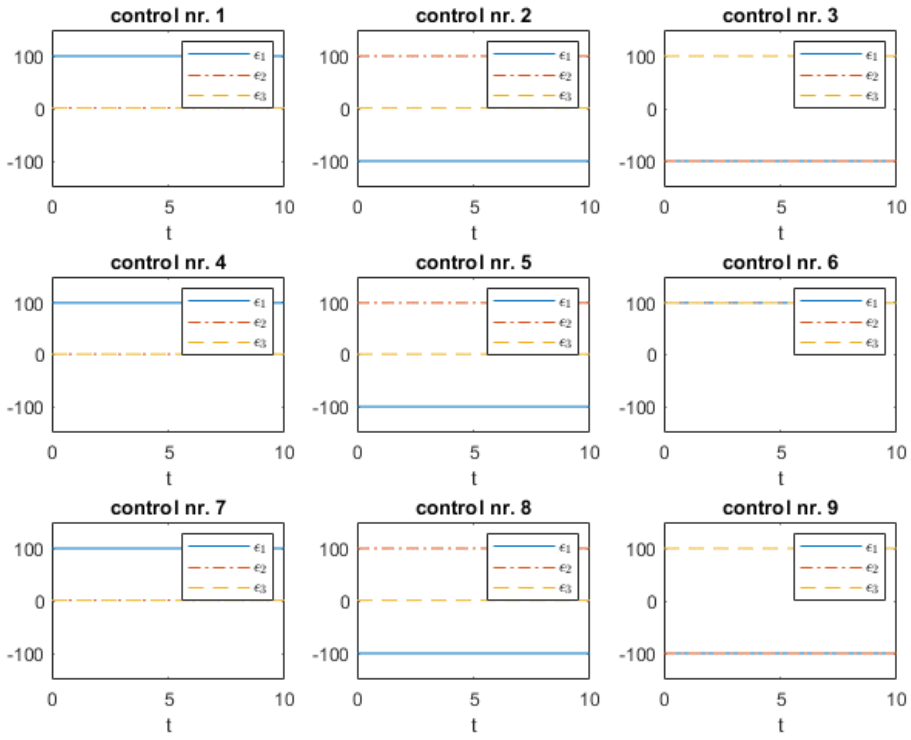


Figure 3 All nine controls for $C = C_1$ and the improved basis.

Here the controls nr. 4 and 7 are identical to the first one, controls nr. 5 and 8 are identical to the second one and control nr. 9 is identical to the third one. So even without the swap strategy we would only end up with 4 distinct controls. However this is due to the matrices A and C we chose in this example. Generally it can happen that $\widehat{W}^k$ is full rank for $k < \mathcal{R}M$ but all $\mathcal{R}M$

controls generated by the algorithm are distinct from one another.

Remark 13
Notice that in Figure 3 is that all controls are constant, either hitting the bound (± 100) or remaining at the starting value ($\equiv 1$). This is very common for the linear case, since all controls are the solution to a maximization problem ((27) or (7)) which are only bounded because the feasible set of controls E_{ad} is bounded (in this case by $b = 100$, compare (41)). Also, by construction, each basis element has only one column that is non-zero, meaning that only the corresponding entry of the control can influence the state $\boldsymbol{y}$. This explains why some entries of the controls remain at the starting value.

For cases where the observability matrix is not full rank we can also look at Example 3 for the improved basis with $C = C_2$ ($\operatorname{rank} \mathcal{O}_3(C_2, A) = 2 \rightsquigarrow \widehat{W}^k \in \mathbb{R}^{6\times 6}$) and $C = C_3$ ($\operatorname{rank} \mathcal{O}_3(C_3, A) = 1 \rightsquigarrow \widehat{W}^k \in \mathbb{R}^{3\times 3}$).

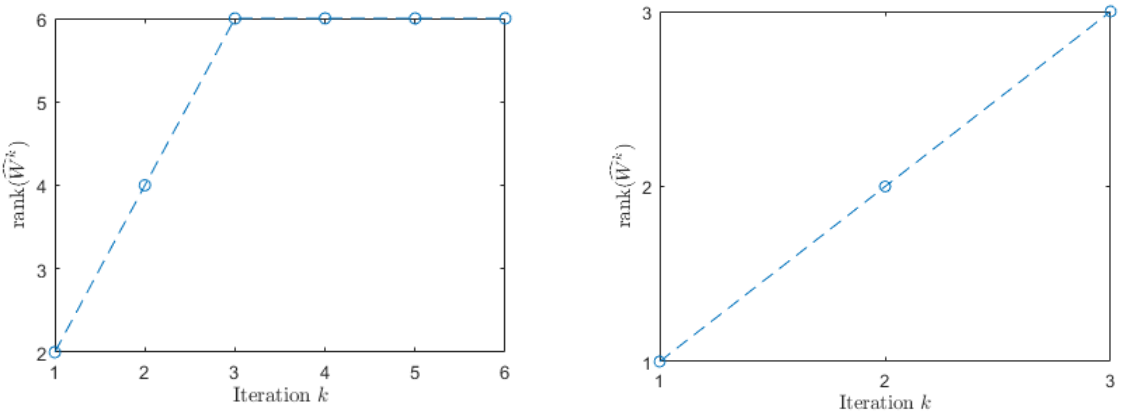


Figure 4 Rank progress of $\widehat{W}^k$ for $C = C_2$ (left) and $C = C_3$ (right).

It seems that the rank of $\widehat{W}^k$ can be increased by up to the rank of the observability matrix, meaning that M can be enough to get a full rank matrix. This can be especially effective when the observability matrix has full rank N, because in that case the algorithm with the improved basis 2 generates the same amount of controls as the standard algorithm 1 with an arbitrary basis.

From the examples above one might conclude that it would be enough to check the rank of $\widehat{W}^k$ in each iteration and that there is not need to swap any columns or rows in $\widehat{W}^k$. However for a different basis (that also satisfies Assumption 4) or even just a different order of our basis, the rank can stagnate also in early iterations. This can be observed in Figure 6, where we compare the algorithm with (red line) and without (blue line) the swap strategy for square matrices A and C with random entries ($N = M = P = 10$, observability matrix with full rank $\mathcal{R} = N = 10$) and a randomly ordered (improved) basis. The algorithm with the swap strategy generates exactly $M = 10$ controls, improving the rank of $\widehat{W}^k$ by $\mathcal{R} = 10$ in each iteration. The algorithm without the swap strategy generates, by construction, $NM = 100$ controls, but the corresponding matrix $\widehat{W}^k$ also only reaches full rank after 23 iterations. This means that, in each iteration, the swap strategy is able to identify and skip elements in the basis that are already identifiable, further reducing the number of controls generated.

In conclusion we can remark, that this improvement can significantly reduce the number of generated controls, especially in systems with large dimensions ($N, M \gg 1$, see Figure 5).

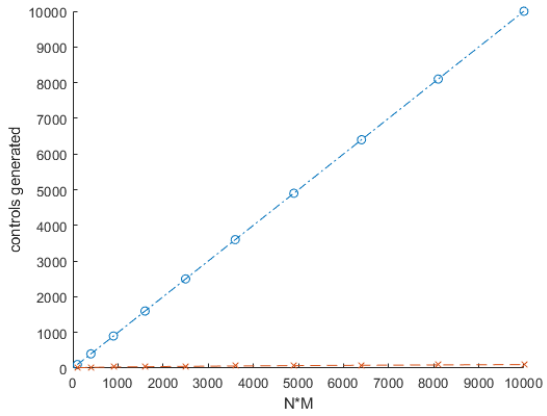


Figure 5 Number of controls generated by the algorithm with (red line) and without (blue line) the swap strategy for square matrices A and C with random entries.

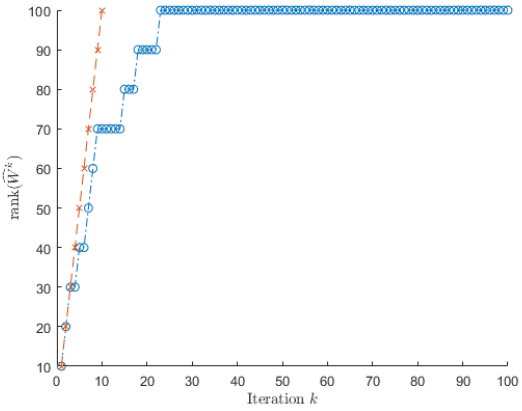


Figure 6 Rank of the matrix $\widehat{W}^k$ in each iteration for random 10×10 matrices A and C for the algorithm with (red line) and without (blue line) the swap strategy.

Remark 14

- *Note that our examples only show that results like these are possible, not that they are generally occurring. Especially in cases, where $P < N$ and $A \in \mathbb{R}^{N\times N}$ and $C \in \mathbb{R}^{P\times N}$ are matrices with random entries, the maximal increase in rank per iteration is usually smaller than M. Sometimes the matrix $\widehat{W}^k$ never even reaches full rank, not even after $\mathcal{R}M$ iterations, although it is guaranteed by our theory (compare Theorem 2). This is due to numerical instabilities/inaccuracies. We will discuss one source of error in the Section 1.4.6.*

1.4.5 Extended greedy-testing strategy

We consider a three-dimensional example, where $B^\star$ is assumed to be symmetric and our basis is the canonical basis for the space of symmetric matrices in $\mathbb{R}^{3\times 3}$.

Example 4
Let $T = 1$, $N = M = 3$, $P = 1$ and

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix},$$

meaning that

$$\mathcal{O}_3(C,A) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{3}-3}{6} & \frac{\sqrt{3}+3}{6} \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{3}+3}{6} & \frac{\sqrt{3}-3}{6} \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{pmatrix} =: U\Sigma V^T$$

and therefore

$$\mathcal{R}M = \operatorname{rank} \mathcal{O}_3(C,A)M = 3. \tag{44}$$

Consider the canonical basis $\mathcal{B}_{sym}$ for the space of symmetric matrices in $\mathbb{R}^{3\times 3}$, with

$$B_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$B_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad B_5 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad B_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We will compare three different settings. First the standard algorithm 1 with the full basis $\mathcal{B}_{sym}$, therefore generating six controls. Secondly the standard algorithm with the first three elements B_1, B_2, B_3, generating three controls. And thirdly the extended greedy-testing strategy 4, also generating $\min\{\mathcal{R}M, 6\} = 3$ controls. Since some of the six basis elements could be (and in this case indeed are) linearly dependent after multiplication with the observability matrix, we can no longer expect to obtain the true coefficients. However we can still measure the difference in observation at final time T (the cost function value of our main identification problem 4). We get the following results.

	full basis	B_1, B_2, B_3	greedy improvement
$\sum_j \left\| C\boldsymbol{y}_T(B^\star, \boldsymbol{\epsilon}^j) - C\boldsymbol{y}_T(B(\boldsymbol{\alpha}^{sol}), \boldsymbol{\epsilon}^j) \right\|$	$1.30 \cdot 10^{-5}$	9.81	$5.74 \cdot 10^{-7}$
Eigenvalues of $\widehat{W}$	$2.61 \cdot 10^5$ $1.38 \cdot 10^5$ $1.55 \cdot 10^4$ $1.52 \cdot 10^{-11}$ $-1.71 \cdot 10^{-12}$ $-1.24 \cdot 10^{-11}$	$8.86 \cdot 10^4$ $3.01 \cdot 10^4$ $-2.07 \cdot 10^{-13}$	$2.38 \cdot 10^5$ $1.18 \cdot 10^5$ $5.87 \cdot 10^4$

Table 5 Cost function value of main identification and eigenvalues of $\widehat{W}$ for the three different settings.

Note that, in order to solve the discretized main identification problem, we used the fminunc-function of MATLAB (which is also the way we will do it in the bilinear case) since the matrix $\widehat{W}$ does not have full rank for all three settings.

As we can see the observations of standard algorithm for the full basis are very close to the true observations. We also have that $\widehat{W} \in \mathbb{R}^{6\times 6}$ has three eigenvalues that are strictly positive and three that are basically zero. This means that $\operatorname{rank} \widehat{W} = 3 = \mathcal{R}M$, which is the highest rank we can expect. However if we just consider the first three basis elements B_1, B_2, B_3, the corresponding $\widehat{W} \in \mathbb{R}^{3\times 3}$ has only two eigenvalues that are strictly positive and still one that is basically zero. The difference in observation at time T between true and computed matrix B is also much bigger than for the full basis. Finally, looking at the result from the greedy improvement, we observe that the observations at time T are almost identical. Also all three eigenvalues of the corresponding $\widehat{W} \in \mathbb{R}^{3\times 3}$ are strictly positive, meaning it has have full rank (the condition number is $\kappa(\widehat{W}) = 4.05$). The three elements that were selected are B_4, B_5 and B_6.

We can conclude that, as expected, three basis elements are sufficient to get optimal results and, most importantly, our greedy improvement is able to select exactly those three. It will in fact choose these three elements, no matter what order of B_1-B_6 we start with. Three controls are also sufficient, indeed three controls generated by the standard algorithm with the full basis (and one generated by the standard algorithm with B_1, B_2, B_3) are equal to our initial guess for the projected gradient method ($\boldsymbol{\epsilon}^0 \equiv 1$).

1.4.6 Numerical instabilities

We consider a simple two-dimensional example where A is the identity and C is diagonal.

Example 5
Let $T = 10$, $N = M = P = 2$ *and*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 10^8 & 0 \\ 0 & 1 \end{pmatrix},$$

meaning that

$$\mathcal{O}_2(C,A) = \begin{pmatrix} 10^8 & 0 \\ 0 & 1 \\ 10^8 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} & 0 \\ 0 & \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} \sqrt{2}\cdot 10^8 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} =: U\Sigma V^T$$

and therefore

$$\mathcal{R}M = \text{rank } \mathcal{O}_2(C,A)M = 4 = NM. \tag{45}$$

For this case Theorem 1 (theoretically) guarantees that, using the controls generated by the algorithm, we can recover all 4 true coefficients of $\boldsymbol{\alpha}^\star$*, for any basis* $\mathcal{B}$ *that spans* $\mathbb{R}^{3\times 3}$*. However if we numerically try to solve the main identification* (31) *with the canonical basis (which in this case also satisfies Assumption 4) we can only recover 2 coefficients for any* $\boldsymbol{\alpha}^\star$*. For example setting*

$$B^\star = \begin{pmatrix} 6.557 & 0.357 \\ 8.491 & 9.340 \end{pmatrix} \tag{46}$$

we get the following result.

$\boldsymbol{\alpha}^\star$	$\boldsymbol{\alpha}^{sol}$
$\begin{pmatrix} 6.557 \\ 0.357 \\ 8.491 \\ 9.340 \end{pmatrix}$	$\begin{pmatrix} 6.557 \\ 0.357 \\ 0 \\ 0 \end{pmatrix}$

Table 6 Solutions of the main identification 31 for the canonical basis.

The numerical cause for this discrepancy between theoretical and numerical results is the ratio of the largest and smallest eigenvalues of $\widehat{W}$*, which in this case has the eigenvalues*

$$\lambda_1 = 6.84\cdot 10^{12}, \quad \lambda_2 = 1.74\cdot 10^{13}, \quad \lambda_3 = 6.84\cdot 10^{28}, \quad \lambda_4 = 1.74\cdot 10^{29}.$$

Since solving the main identification problem is equivalent to solving the linear system (37)*, the stability the solution depends on the condition number of* $\widehat{W}$*, which in this case is*

$$\kappa(\widehat{W}) = \frac{\lambda_{max}(\widehat{W})}{\lambda_{min}(\widehat{W})} = \frac{\lambda_4}{\lambda_1} = 2.55\cdot 10^{16}.$$

Hence the system is highly ill-conditioned and it is therefore not surprising that we are not able to numerically recover all coefficients.

We can generalize these findings by giving a lower bound for the condition number of $\widehat{W}$ depending on the maximal and minimal singular value of the observability matrix.
Consider arbitrary $N, M, P \in \mathbb{R}$, $T > 0$, $A \in \mathbb{R}^{N\times N}$, $C \in \mathbb{R}^{P\times N}$ and a basis $\mathcal{B} \subset \mathbb{R}^{N\times M}$ defined using the SVD of the observability matrix

$$\mathcal{O}_N(C,A) = U\Sigma V^T, \tag{47}$$

as mentioned in Remark 6. Let $\sigma_1 \geq \ldots \geq \sigma_{\mathcal{R}} > 0$ be the positive singular values of the observability matrix. Then we have

$$\kappa(\widehat{W}) = \frac{\lambda_{max}(\widehat{W})}{\lambda_{min}(\widehat{W})},$$

with

$$\lambda_{min}(\widehat{W}) = \min_{\|x\|=1} \left\langle x, \widehat{W}x \right\rangle \leq \left\langle \boldsymbol{e}_{\mathcal{R}M}, \widehat{W}\boldsymbol{e}_{\mathcal{R}M} \right\rangle \leq [\widehat{W}]_{[\mathcal{R}M,\mathcal{R}M]},$$

$$\lambda_{max}(\widehat{W}) = \max_{\|x\|=1} \left\langle x, \widehat{W}x \right\rangle \geq \left\langle \boldsymbol{e}_1, \widehat{W}\boldsymbol{e}_1 \right\rangle \geq [\widehat{W}]_{[1,1]}.$$

We also have

$$\begin{aligned}
[\widehat{W}]_{[1,1]} &= \sum_{j=1}^{\mathcal{R}M} \left\langle \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}^j), \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}^j) \right\rangle = \sum_{j=1}^{\mathcal{R}M} \left\| \boldsymbol{\gamma}_1(\boldsymbol{\epsilon}^j) \right\|^2 \\
&= \sum_{j=1}^{\mathcal{R}M} \left\| \int_0^T Ce^{(T-s)A} B_1 \boldsymbol{\epsilon}^j(s) ds \right\|^2 \\
&\overset{C.H.}{=} \sum_{j=1}^{\mathcal{R}M} \left\| \int_0^T \sum_{i=0}^{N-1} \beta_i(s) CA^i B_1 \boldsymbol{\epsilon}^j(s) ds \right\|^2 \\
&= \sum_{j=1}^{\mathcal{R}M} \left\| \int_0^T \sum_{i=0}^{N-1} \beta_i(s) CA^i \boldsymbol{v}_1 \boldsymbol{\epsilon}_1^j(s) ds \right\|^2 \\
&\overset{(47)}{=} \sum_{j=1}^{\mathcal{R}M} \left\| \sum_{i=0}^{N-1} \int_0^T \beta_i(s) \sigma_1 [u_1]_{[(Ni+1):(Ni+N)]} \boldsymbol{\epsilon}_1^j(s) ds \right\|^2 \\
&= \sigma_1^2 \underbrace{\sum_{j=1}^{\mathcal{R}M} \left\| \sum_{i=0}^{N-1} \Big(\int_0^T \beta_i(s) \boldsymbol{\epsilon}_1^j(s) ds \Big) [u_1]_{[(Ni+1):(Ni+N)]} \right\|^2}_{=:c_1<\infty, \text{ since } \boldsymbol{\epsilon}_1^j \text{ are bounded}} \\
&= c_1 \sigma_1^2.
\end{aligned}$$

Note that c_1 is independent of $\sigma_{\mathcal{R}}$. Analogous we can show that

$$[\widehat{W}]_{[\mathcal{R}M,\mathcal{R}M]} = c_{\mathcal{R}} \sigma_{\mathcal{R}}^2,$$

where $c_{\mathcal{R}}$ is independent of σ_1. In conclusion we get

$$\kappa(\widehat{W}) \geq \frac{c_1 \sigma_1^2}{c_{\mathcal{R}} \sigma_{\mathcal{R}}^2}, \tag{48}$$

meaning that the ratio of the largest and smallest singular value of the observability matrix quadratically affects the condition number of $\widehat{W}$.

In the setting of Example 5 we have $\sigma_1 = \sqrt{2} \cdot 10^8, \sigma_{\mathcal{R}} = \sigma_2 = \sqrt{2}$ and therefore

$$\frac{\sigma_1^2}{\sigma_2^2} = 10^{16} < 2.55 \cdot 10^{16} = \kappa(\widehat{W}).$$

Remark 15
This bound is, in general, not tight, meaning that a small ratio $\frac{\sigma_1^2}{\sigma_{\mathcal{R}}^2}$ does not generally imply that system (37) *is stable.*

CHAPTER 2

Bilinear system

Now we move to the original setting for which the algorithm was introduced in [MS09]: Schrödinger-type equations.

2.1 Notation

Throughout this chapter we will adopt some of the notation used in [MS09]. We assume N to be a natural number. We denote by $\|\cdot\|_2$ the Euclidean norm for complex vectors in $\mathbb{C}^N$ and by $\langle\cdot,\cdot\rangle$ the corresponding standard inner product. Finally $\langle\cdot,\cdot\rangle_{L^2}$ denotes the usual inner product of $L^2 := L^2(0,T;\mathbb{C}^N)$.

2.2 The identification problem and a greedy identification algorithm

2.2.1 Problem formulation

Consider the Schrödinger equation for the state of a quantum system $\boldsymbol{\psi}$

$$\begin{cases} i\dot{\boldsymbol{\psi}}(t) = [H + \epsilon(t)\mu]\boldsymbol{\psi}(t), \quad t \in (0,T] \\ \boldsymbol{\psi}(0) = \boldsymbol{\psi}_0, \end{cases} \tag{49}$$

where the internal Hamiltonian H is assumed to be known and the goal is to identify the (unknown) dipole moment operator μ that couples the system to a time-dependent external laser field $\epsilon \in L^2(0,T)$, which represents a control in this context. We assume the system to be finite dimensional, meaning that H and μ are Hermitian matrices in $\mathbb{C}^{N\times N}$ and $\boldsymbol{\psi}(t) \in \mathbb{C}^N$. The initial condition $\boldsymbol{\psi}_0 \in \mathbb{C}^N$ is subject to the constraint $\|\boldsymbol{\psi}_0\|_2 = 1$, where $\|\cdot\|_2$ denotes the Euclidean norm for complex vectors.

The true dipole operator $\mu = \mu^\star$ is assumed to lie in a space spanned by a basis $\mathcal{B}_\mu = \{\mu_1,\dots,\mu_K\}$, $K \in \mathbb{N}$. To identify $\mu^\star$ we can experimentally measure the value $\varphi(\mu,\epsilon) := \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu,\boldsymbol{\epsilon}) \right\rangle$, where $\boldsymbol{\psi}_T(\mu,\boldsymbol{\epsilon})$ is the solution of (49) at a given time $T > 0$, for a laser field ϵ and a matrix μ. The value $\boldsymbol{\psi}_1 \in \mathbb{C}^N$ is a fixed state with $\|\boldsymbol{\psi}_1\|_2 = 1$ and acts as our "observer". As in the linear setting, we assume that the measurements are not affected by any type of noise. The goal is to find a good approximation $\mu \in \mathcal{B}_\mu$ of $\mu^\star$ for which the norm difference at time T between observed experimental data $\varphi(\mu^\star,\epsilon)$ and numerical data $\varphi(\mu,\epsilon)$ is the smallest for any laser field $\epsilon \in L^2(0,T)$. In other words we want to find a matrix μ that solves

$$\min_{\mu\in\mathcal{B}_\mu} \max_{\epsilon\in L^2(0,T)} |\varphi(\mu^\star,\epsilon) - \varphi(\mu,\epsilon)|^2. \tag{50}$$

Once more we can write $\mu^\star = \sum_{l=1}^K \boldsymbol{\alpha}_l^\star \mu_l =: \mu(\boldsymbol{\alpha}^\star)$ and hence rewrite (50) as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \max_{\epsilon\in L^2(0,T)} |\varphi(\mu^\star,\epsilon) - \varphi(\mu(\boldsymbol{\alpha}),\epsilon)|^2. \tag{51}$$

Again the idea of the algorithm is to generate a set of laser fields $(\epsilon^1, \dots, \epsilon^K) \subset L^2(0,T)$ that attempt to discern numerical data for any two $\mu(\tilde{\boldsymbol{\alpha}}), \mu(\hat{\boldsymbol{\alpha}}) \in \mathcal{B}_\mu$, $\mu(\tilde{\boldsymbol{\alpha}}) \neq \mu(\hat{\boldsymbol{\alpha}})$, without doing any experiments. With this set one then solves the least-squares problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \sum_{j=1}^{K} |\varphi(\mu^\star, \epsilon^j) - \varphi(\mu(\boldsymbol{\alpha}), \epsilon^j)|^2, \tag{52}$$

where one has to perform exactly K laboratory experiments. Clearly $\boldsymbol{\alpha}^\star$ also solves problem (52).

2.2.2 Algorithm in the bilinear setting

We adapt the improved initialization from the linear setting (compare Section 1.3.1.1) to the algorithm stated in [MS09], meaning that we want to find the control that maximizes the difference between the observation corresponding to the controlled state and the observation $\varphi(0,0)$ corresponding to the uncontrolled state. Denoting $\mu^k(\boldsymbol{\alpha}) := \sum_{l=1}^{k} \boldsymbol{\alpha}_l \mu_l$, the algorithm can then be written as follows.

Algorithm 5 Greedy Reconstruction Algorithm for Selective Laser Fields Computation

Require: A basis $\mathcal{B}_\mu = (\mu_1, \dots, \mu_K)$.

1: Solve the initialization problem

$$\max_{\boldsymbol{\epsilon}\in L^2(0,T)} |\varphi(\mu_1, \epsilon) - \varphi(0,0)|^2. \tag{53}$$

which gives the field ϵ^1 and set $k = 1$.

2: **while** $k \leq K - 1$ **do**

3: Fitting step: Find $(\boldsymbol{\alpha}_j^k)_{j=1,\dots,k}$ that solve the problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^k} \sum_{j=1}^{k} |\varphi(\mu_{k+1}, \epsilon^j) - \varphi(\mu^k(\boldsymbol{\alpha}), \epsilon^j)|^2. \tag{54}$$

4: Discriminatory step: Find ϵ^{k+1} that solves the problem:

$$\max_{\boldsymbol{\epsilon}\in L^2(0,T)} |\varphi(\mu_{k+1}, \epsilon) - \varphi(\mu^k(\boldsymbol{\alpha}^k), \epsilon)|^2. \tag{55}$$

5: Update $k \leftarrow k + 1$.

6: **end while**

With the improved initialization (53) we make it unlikely that $\epsilon \equiv 0$ is selected as first control ϵ^1, since $\epsilon \equiv 0$ is a minimum point for (53) and we wish to maximize this quantity. The general idea of the algorithm is the same as described in Section 1.1.2.

In practise we will add a regularization term to the problems (53), (54) and (55) in the algorithm to ensure well-posedeness. Therefore we consider the regularized initialization problem

$$\max_{\boldsymbol{\epsilon}\in L^2(0,T)} |\varphi(\mu_1, \epsilon) - \varphi(0,0)|^2 - \beta \|\epsilon\|^2_{L^2(0,T)}, \tag{56}$$

the regularized fitting step problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^k} \sum_{j=1}^{k} |\varphi(\mu_{k+1}, \epsilon^j) - \varphi(\mu^k(\boldsymbol{\alpha}), \epsilon^j)|^2 + \nu \|\boldsymbol{\alpha}\|_2^2. \tag{57}$$

and the regularized discriminatory step problem

$$\max_{\boldsymbol{\epsilon}\in L^2(0,T)} |\varphi(\mu_{k+1},\epsilon) - \varphi(\mu^k(\boldsymbol{\alpha}^k),\epsilon)|^2 - \beta \|\epsilon\|^2_{L^2(0,T)}\,, \tag{58}$$

where we denote by $\beta, \nu > 0$ some regularization parameters.
In case of the initialization and the discriminatory step problem, penalizing the $L^2(0,T)$-norm of ϵ also enables us to use efficient monotonic schemes in order to solve the optimal control problems, which we will introduce in Section 2.4. From now on we refer to the regularized problems (56)-(58), when talking about the initialization, fitting step and discriminatory step problem.

2.2.3 Well-posedeness

For the main identification (52) this is clear, since the cost functional is bounded from below by zero and the cost functional value for $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$ is equal to zero, meaning that $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$ is a minimizer to this problem.

It remains to show that the three problems in the algorithm, namely (56), (57) and (58), are well-posed.

2.2.3.1 Initialization and discriminatory step problem

In order to prove well-posedeness for the optimal control problems (56) and (58), we follow the argumentation in [BCS17, p.75ff.]. Therefore we first need some intermediate lemmas. We start by proving existence and uniqueness of a solution $\boldsymbol{\psi} \in H^1(0,T;\mathbb{C}^N)$ to equation (49) for any given control $\epsilon \in L^2(0,T)$.

Lemma 7
For each control $\epsilon \in L^2(0,T)$ there exists a unique solution $\boldsymbol{\psi} \in H^1(0,T;\mathbb{C}^N)$ to problem (49), satisfying $\|\boldsymbol{\psi}(t)\|_2 = \|\boldsymbol{\psi}_0\|_2$ for all $t \in [0,T]$.

proof. For a given $\epsilon \in L^2(0,T)$ we define $f : \mathbb{C}^N \times \mathbb{R} \to \mathbb{C}^N$ as $f(\boldsymbol{\psi}, t) := [H + \epsilon(t)\mu]\boldsymbol{\psi}$. Since $\epsilon \in L^2(0,T)$, $f(\boldsymbol{\psi}, t)$ is measurable in t. It is also linear and continuous with respect to $\boldsymbol{\psi}$. Furthermore, for $\boldsymbol{\psi}, \phi \in \mathbb{C}^N$, it holds that

$$\begin{aligned}
\|f(\boldsymbol{\psi},t) - f(\phi,t)\|_2 &= \|[H+\epsilon(t)\mu]\boldsymbol{\psi} - [H+\epsilon(t)\mu]\phi\|_2 \\
&= \|[H+\epsilon(t)\mu](\boldsymbol{\psi}-\phi)\|_2 \\
&\le \|[H+\epsilon(t)\mu]\|_{\mathcal{L}} \|\boldsymbol{\psi}-\phi\|_2 \\
&\le \underbrace{\Big(\|H\|_{\mathcal{L}} + |\epsilon(t)|\,\|\mu\|_2\Big)}_{=:\gamma(t)} \|\boldsymbol{\psi}-\phi\|_2 \\
&= \gamma(t)\,\|\boldsymbol{\psi}-\phi\|_2\,,
\end{aligned}$$

where $\|\cdot\|_{\mathcal{L}}$ denotes the spectral norm of a matrix. Hence f is Lipschitz continuous in $\boldsymbol{\psi}$. Note that, since $[0,T]$ is bounded and therefore $\epsilon \in L^1(0,T)$, we have $\gamma \in L^1(0,T)$. We also have for each fixed $\boldsymbol{\psi} \in \mathbb{C}^N$ that

$$\|f(\boldsymbol{\psi},t)\|_2 = \|[H+\epsilon(t)\mu]\boldsymbol{\psi}\|_2 \le \|H+\epsilon(t)\mu\|_{\mathcal{L}}\,\|\boldsymbol{\psi}\|_2 = \gamma(t)\,\|\boldsymbol{\psi}\|_2$$

for almost all $t \in [0,T]$. Hence we can apply Carathéodory's existence theorem (see [Son98, Theorem 54]) to get the existence of a unique solution $\boldsymbol{\psi} \in AC([0,T],\mathbb{C}^N)$ for (49) for any $T > 0$ and any initial value $\boldsymbol{\psi}_0 \in \mathbb{C}^N$.

It remains to show that $\boldsymbol{\psi} \in H^1(0,T;\mathbb{C}^N)$. We have

$$i\dot{\boldsymbol{\psi}}(t) = [H + \epsilon(t)\mu]\boldsymbol{\psi}(t) \tag{59}$$

and by taking the complex conjugate on both sides we also get

$$-i\dot{\bar{\boldsymbol{\psi}}}(t) = [H + \epsilon(t)\mu]\bar{\boldsymbol{\psi}}(t), \tag{60}$$

since H and μ are Hermitian. Hence we can write

$$\begin{aligned}
\frac{d}{dt}\left\|\boldsymbol{\psi}(t)\right\|_2^2 &= \frac{d}{dt}\bar{\boldsymbol{\psi}}(t)^T\boldsymbol{\psi}(t) \\
&= \dot{\bar{\boldsymbol{\psi}}}(t)^T\boldsymbol{\psi}(t) + \bar{\boldsymbol{\psi}}(t)^T\dot{\boldsymbol{\psi}}(t) \\
&= (i[H + \epsilon(t)\mu]\bar{\boldsymbol{\psi}}(t))^T\boldsymbol{\psi}(t) - (i\bar{\boldsymbol{\psi}})^T[H + \epsilon(t)\mu]\boldsymbol{\psi}(t) \\
&= 0
\end{aligned}$$

to obtain $\left\|\boldsymbol{\psi}(t)\right\|_2 = \left\|\boldsymbol{\psi}_0\right\|_2$ for all $t \in (0,T)$ and therefore $\left\|\boldsymbol{\psi}\right\|_{L^2} = \sqrt{T}\left\|\boldsymbol{\psi}_0\right\|_2 < \infty$. We also have

$$\begin{aligned}
\left\|\dot{\boldsymbol{\psi}}\right\|_{L^2}^2 &= \left\|[H + \epsilon\mu]\boldsymbol{\psi}\right\|_{L^2}^2 = \int_0^T \left\|[H + \epsilon(t)\mu]\boldsymbol{\psi}(t)\right\|_2^2 dt \\
&\leq \int_0^T \left\|H\boldsymbol{\psi}(t)\right\|_2^2 + 2\left\|H\boldsymbol{\psi}(t)\right\|_2\left\|\epsilon(t)\mu\boldsymbol{\psi}(t)\right\|_2 + \left\|\epsilon(t)\mu\boldsymbol{\psi}(t)\right\|_2^2 dt \\
&\overset{Young}{\leq} \int_0^T 2\left\|H\boldsymbol{\psi}(t)\right\|_2^2 + 2\left\|\epsilon(t)\mu\boldsymbol{\psi}(t)\right\|_2^2 dt \\
&\leq \int_0^T 2\left\|H\right\|_{\mathcal{L}}^2\left\|\boldsymbol{\psi}(t)\right\|_2^2 + 2|\epsilon(t)|^2\left\|\mu\right\|_{\mathcal{L}}^2\left\|\boldsymbol{\psi}(t)\right\|_2^2 dt \\
&\leq 2T\left\|H\right\|_{\mathcal{L}}^2\left\|\boldsymbol{\psi}_0\right\|_2^2 + 2\left\|\epsilon\right\|_{L^2(0,T)}^2\left\|\mu\right\|_{\mathcal{L}}^2\left\|\boldsymbol{\psi}_0\right\|_2^2 < \infty
\end{aligned}$$

and therefore $\boldsymbol{\psi} \in H^1(0,T;\mathbb{C}^N)$. □

Remark 16
By the calculations in the proof of Lemma 7 we also obtain that there exist $c_1, c_2 \geq 0$ such that

$$\left\|\boldsymbol{\psi}\right\|_{H^1(0,T;\mathbb{C}^N)} \leq c_1\left\|\boldsymbol{\psi}_0\right\|_2 + c_2\left\|\boldsymbol{\psi}_0\right\|_2\left\|\epsilon\right\|_{L^2(0,T)}. \tag{61}$$

Now we are ready to prove the main result.

Proposition 13
The initialization problem (56) *and the discriminatory step problem* (58) *are well posed.*

proof. We start by noticing that $\varphi(0,0) = \varphi(0,\epsilon)$. Therefore we can write both cost functionals of (56) and (58) as

$$J(\epsilon) := |\varphi(\tilde{\mu},\epsilon) - \varphi(\hat{\mu},\epsilon)|^2 - \beta\left\|\epsilon\right\|_{L^2(0,T)}^2,$$

where $\tilde{\mu} = \mu_1$ and $\hat{\mu} = 0$ or, respectively, $\tilde{\mu} = \mu_{k+1}$ and $\hat{\mu} = \mu^k(\boldsymbol{\alpha}^k)$. Next we obtain that $J(\epsilon)$ is bounded from above.

$$\begin{aligned} J(\epsilon) &= |\varphi(\tilde{\mu}, \epsilon) - \varphi(\hat{\mu}, \epsilon)|^2 - \beta \|\epsilon\|^2_{L^2(0,T)} \\ &\leq |\left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\tilde{\mu}, \epsilon) \right\rangle - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\hat{\mu}, \epsilon) \right\rangle|^2 \\ &\leq 2 \underbrace{\|\boldsymbol{\psi}_1\|_2^2}_{=1} \underbrace{\|\boldsymbol{\psi}_T(\tilde{\mu}, \epsilon)\|_2^2}_{=1} + 2 \underbrace{\|\boldsymbol{\psi}_1\|_2^2}_{=1} \underbrace{\|\boldsymbol{\psi}_T(\hat{\mu}, \epsilon)\|_2^2}_{=1} \\ &= 4. \end{aligned}$$

Hence there exists a maximizing sequence $\{\epsilon_n\}_{n\in\mathbb{N}} \subset L^2(0,T)$,

$$\lim_{n\to\infty} J(\epsilon_n) = \sup_{\epsilon \in L^2(0,T)} J(\epsilon) =: \widetilde{J}.$$

Since $0 \leq J(0)$, we can restrict the maximization to the set $\{\epsilon \in L^2(0,T) \mid J(\epsilon) \geq 0\}$. It is also true that $J(\epsilon) \geq 0 \Rightarrow \beta \|\epsilon\|^2_{L^2(0,T)} \leq 4$, meaning that

$$\{\epsilon \in L^2(0,T) \mid J(\epsilon) \geq 0\} \subset \{\epsilon \in L^2(0,T) \mid \beta \|\epsilon\|^2_{L^2(0,T)} \leq 4\} =: \widetilde{U}.$$

Therefore, if a maximizer exists, it must be an element of $\widetilde{U}$ and we can equivalently reformulate our maximization problem in $\widetilde{U}$, which is a closed, convex and bounded subset of $L^2(0,T)$. Hence we can assume, without loss of generality, that $\beta \|\epsilon_n\|^2_{L^2(0,T)} \leq 4$ for all $n \in \mathbb{N}$, which implies that the sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ is (uniformly) bounded in $L^2(0,T)$. Since $L^2(0,T)$ is a Hilbert space and hence reflexive, there exists a weakly in $L^2(0,T)$ convergent subsequence $\epsilon_{\tilde{n}} \rightharpoonup \tilde{\epsilon}$. Using Lemma 7 and (61), we obtain that the corresponding sequences of unique solutions to (49), denoted by $\tilde{\boldsymbol{\psi}}_{\tilde{n}} := \boldsymbol{\psi}(\tilde{\mu}, \epsilon_{\tilde{n}})$ and $\hat{\boldsymbol{\psi}}_{\tilde{n}} := \boldsymbol{\psi}(\hat{\mu}, \epsilon_{\tilde{n}})$, are (uniformly) bounded in $H^1(0,T;\mathbb{C}^N)$. By reflexivity of $H^1(0,T;\mathbb{C}^N)$, there exist two weakly in $H^1(0,T;\mathbb{C}^N)$ convergent subsequences $\tilde{\boldsymbol{\psi}}_{\tilde{n}_k} \rightharpoonup \tilde{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}_{\tilde{n}_k} \rightharpoonup \hat{\boldsymbol{\psi}}$. By the Sobolev compact embedding $H^1(0,T;\mathbb{C}^N) \Subset C([0,T];\mathbb{C}^N)$, we also have $\tilde{\boldsymbol{\psi}}_{\tilde{n}_k} \to \tilde{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}_{\tilde{n}_k} \to \tilde{\boldsymbol{\psi}}$ in $C([0,T];\mathbb{C}^N)$.

Multiplying the system equation (49) from the right with a test function $v \in H^1(0,T;\mathbb{C}^N)$ and integrating over $(0,T)$ we get

$$\begin{aligned} \int_0^T \left\langle \dot{\tilde{\boldsymbol{\psi}}}_{\tilde{n}_k}(t) + i[H - \epsilon_{\tilde{n}_k}(t)\tilde{\mu}]\tilde{\boldsymbol{\psi}}_{\tilde{n}_k}(t), v(t) \right\rangle dt = 0 \\ \int_0^T \left\langle \dot{\hat{\boldsymbol{\psi}}}_{\tilde{n}_k}(t) + i[H - \epsilon_{\tilde{n}_k}(t)\hat{\mu}]\hat{\boldsymbol{\psi}}_{\tilde{n}_k}(t), v(t) \right\rangle dt = 0. \end{aligned}$$

From $\epsilon_{\tilde{n}} \rightharpoonup \tilde{\epsilon}$ in $L^2(0,T)$ and $\tilde{\boldsymbol{\psi}}_{\tilde{n}_k} \to \tilde{\boldsymbol{\psi}}$ in $C([0,T];\mathbb{C}^N)$ we obtain that

$$\begin{aligned} \left| \left\langle \epsilon_{\tilde{n}_k}\tilde{\mu}\tilde{\boldsymbol{\psi}}_{\tilde{n}_k}, v \right\rangle_{L^2} - \left\langle \tilde{\epsilon}\tilde{\mu}\tilde{\boldsymbol{\psi}}, v \right\rangle_{L^2} \right| &= \left| \left\langle \epsilon_{\tilde{n}_k}\tilde{\mu}\tilde{\boldsymbol{\psi}}_{\tilde{n}_k}, v \right\rangle_{L^2} - \left\langle \epsilon_{\tilde{n}_k}\tilde{\mu}\tilde{\boldsymbol{\psi}}, v \right\rangle_{L^2} + \left\langle \epsilon_{\tilde{n}_k}\tilde{\mu}\tilde{\boldsymbol{\psi}}, v \right\rangle_{L^2} - \left\langle \tilde{\epsilon}\tilde{\mu}\tilde{\boldsymbol{\psi}}, v \right\rangle_{L^2} \right| \\ &\leq \left| \left\langle \epsilon_{\tilde{n}_k}\tilde{\mu}(\tilde{\boldsymbol{\psi}}_{\tilde{n}_k} - \tilde{\boldsymbol{\psi}}), v \right\rangle_{L^2} \right| + \left| \left\langle (\epsilon_{\tilde{n}_k} - \tilde{\epsilon})\tilde{\mu}\tilde{\boldsymbol{\psi}}, v \right\rangle_{L^2} \right| \\ &\leq \underbrace{\|\epsilon_{\tilde{n}_k}\|_{L^2}}_{\leq C} \|\tilde{\mu}\| \underbrace{\|\tilde{\boldsymbol{\psi}}_{\tilde{n}_k} - \tilde{\boldsymbol{\psi}}\|_{C([0,T];\mathbb{C}^N)}}_{\to 0} \|v\|_{L^2} + \underbrace{\left| \left\langle (\epsilon_{\tilde{n}_k} - \tilde{\epsilon})\tilde{\mu}\tilde{\boldsymbol{\psi}}, v \right\rangle_{L^2} \right|}_{\to 0} \\ &\to 0, \end{aligned}$$

and similarly $\left| \left\langle \epsilon_{\tilde{n}_k}\hat{\mu}\hat{\boldsymbol{\psi}}_{\tilde{n}_k}, v \right\rangle_{L^2} - \left\langle \tilde{\epsilon}\hat{\mu}\hat{\boldsymbol{\psi}}, v \right\rangle_{L^2} \right| \to 0$ for all $v \in H^1(0,T;\mathbb{C}^N)$. Since $\dot{\tilde{\boldsymbol{\psi}}}_{\tilde{n}_k} \rightharpoonup \dot{\tilde{\boldsymbol{\psi}}}$, $\dot{\hat{\boldsymbol{\psi}}}_{\tilde{n}_k} \rightharpoonup \dot{\hat{\boldsymbol{\psi}}}$ in L^2, we get

$$\begin{aligned} \int_0^T \left\langle \dot{\tilde{\boldsymbol{\psi}}}_{\tilde{n}_k}(t) + i[H - \epsilon_{\tilde{n}_k}(t)\tilde{\mu}]\tilde{\boldsymbol{\psi}}_{\tilde{n}_k}(t), v(t) \right\rangle dt &\to \int_0^T \left\langle \dot{\tilde{\boldsymbol{\psi}}}(t) + i[H - \tilde{\epsilon}(t)\tilde{\mu}]\tilde{\boldsymbol{\psi}}(t), v(t) \right\rangle dt, \\ \int_0^T \left\langle \dot{\hat{\boldsymbol{\psi}}}_{\tilde{n}_k}(t) + i[H - \epsilon_{\tilde{n}_k}(t)\hat{\mu}]\hat{\boldsymbol{\psi}}_{\tilde{n}_k}(t), v(t) \right\rangle dt &\to \int_0^T \left\langle \dot{\hat{\boldsymbol{\psi}}}(t) + i[H - \tilde{\epsilon}(t)\hat{\mu}]\hat{\boldsymbol{\psi}}(t), v(t) \right\rangle dt, \end{aligned}$$

for all $v \in H^1(0,T;\mathbb{C}^N)$. Hence

$$i\dot{\tilde{\boldsymbol{\psi}}}(t) = [H - \tilde{\epsilon}(t)\tilde{\mu}]\tilde{\boldsymbol{\psi}}(t), \qquad i\dot{\hat{\boldsymbol{\psi}}}(t) = [H - \tilde{\epsilon}(t)\hat{\mu}]\hat{\boldsymbol{\psi}}(t)$$

a.e. in $(0,T)$ and $\tilde{\boldsymbol{\psi}}(0) = \hat{\boldsymbol{\psi}}(0) = \boldsymbol{\psi}_0$. By uniqueness of the limit and uniqueness of the solution of (49), we get $\tilde{\boldsymbol{\psi}} = \boldsymbol{\psi}(\tilde{\mu}, \tilde{\epsilon})$ and $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}(\hat{\mu}, \tilde{\epsilon})$. Recalling that norms are weakly lower semicontinuous, we get

$$\begin{aligned}
\widetilde{J} &= \lim_{k\to\infty} J(\boldsymbol{\epsilon}_{\tilde{n}_k}) = \lim_{k\to\infty} \left(\left| \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\tilde{\mu}, \epsilon_{\tilde{n}_k}) \right\rangle \right|^2 - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\hat{\mu}, \epsilon_{\tilde{n}_k}) \right\rangle \right|^2 - \beta \left\| \epsilon_{\tilde{n}_k} \right\|^2_{L^2(0,T)} \Big) \\
&= \lim_{k\to\infty} \left| \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}_{\tilde{n}_k} \right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}_{\tilde{n}_k}(T) \right\rangle \right|^2 - \liminf_{k\to\infty} \beta \left\| \epsilon_{\tilde{n}_k} \right\|^2_{L^2(0,T)} \\
&\leq \left| \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) \right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}(T) \right\rangle \right|^2 - \beta \left\| \tilde{\epsilon} \right\|^2_{L^2(0,T)} \\
&= \left| \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\tilde{\mu}, \tilde{\epsilon}) \right\rangle - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\hat{\mu}, \tilde{\epsilon}) \right\rangle \right|^2 - \beta \left\| \tilde{\epsilon} \right\|^2_{L^2(0,T)} = J(\tilde{\epsilon}) \\
&\leq \widetilde{J}.
\end{aligned}$$

This implies that $\tilde{\epsilon}$ is a maximizer for the respective maximization problem. □

2.2.3.2 Fitting step problem

Let $k \in \{1, \ldots, K-1\}$. Similar to the proof of Lemma 7 one can prove existence and uniqueness of a solution $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\alpha}) \in H^1(0,T;\mathbb{C}^N)$ to equation (49) (with $\mu = \mu^k(\boldsymbol{\alpha}) = \sum_{l=1}^k \boldsymbol{\alpha}_l \mu_l$) for any given $\boldsymbol{\alpha} \in \mathbb{R}^k$.

Lemma 8
For each $\boldsymbol{\alpha} \in \mathbb{R}^k$ there exists a unique solution $\boldsymbol{\psi} \in H^1(0,T;\mathbb{C}^N)$ to problem (49) (with $\mu = \mu^k(\boldsymbol{\alpha})$) satisfying $\|\boldsymbol{\psi}(t)\|_2 = \|\boldsymbol{\psi}_0\|_2$ for all $t \in [0,T]$.

One can also similarly obtain that there exist $c_1, c_2 \geq 0$ such that

$$\|\boldsymbol{\psi}\|_{H^1(0,T;\mathbb{C}^N)} \leq c_1 \|\boldsymbol{\psi}_0\|_2 + c_2 \|\boldsymbol{\psi}_0\|_2 \|\boldsymbol{\alpha}\|_2 . \tag{62}$$

Proposition 14
The least-squares problem (57) is well-posed.

proof. We denote by J the cost functional of (57)

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^k |\varphi(\mu_{k+1}, \epsilon^j) - \varphi(\mu^k(\boldsymbol{\alpha}), \epsilon^j)|^2 + \nu \|\boldsymbol{\alpha}\|_2^2 .$$

Consider an arbitrary $\boldsymbol{\alpha}_0 \in \mathbb{R}^k$. Since $\nu > 0$, we have for all $\boldsymbol{\alpha} \in \mathbb{R}^k$ with $\|\boldsymbol{\alpha}\|_2^2 > J(\boldsymbol{\alpha}_0)$ that

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^k |\varphi(\mu_{k+1}, \epsilon^j) - \varphi(\mu^k(\boldsymbol{\alpha}), \epsilon^j)|^2 + \nu \|\boldsymbol{\alpha}\|_2^2 \geq \nu \|\boldsymbol{\alpha}\|_2^2 > J(\boldsymbol{\alpha}_0).$$

Hence we can restrict the search for a minimum to the set

$$\widetilde{U} := \{\boldsymbol{\alpha} \in \mathbb{R} \mid \|\boldsymbol{\alpha}\|_2^2 \leq \frac{1}{\nu} J(\boldsymbol{\alpha}_0)\},$$

which is a non-empty ($\boldsymbol{\alpha}_0 \in \widetilde{U}$), closed and bounded subset of $\mathbb{R}^k$, meaning it is also compact. Since J is bounded from below by zero, we can find a minimizing sequence $\{\boldsymbol{\alpha}_n\}_{n\in\mathbb{N}} \subset \widetilde{U}$,

$$\lim_{n\to\infty} J(\boldsymbol{\alpha}_n) = \inf_{\boldsymbol{\alpha}\in\mathbb{R}^k} J(\boldsymbol{\alpha}) =: \widetilde{J}.$$

Since $\widetilde{U}$ is compact, there exists a convergent subsequence $\boldsymbol{\alpha}_{\tilde{n}} \to \tilde{\boldsymbol{\alpha}} \in \widetilde{U}$. Using (62) we obtain that the corresponding sequences of solutions $\{\boldsymbol{\psi}^j_{\tilde{n}}\}_{\tilde{n}\in\mathbb{N}}$, with $\boldsymbol{\psi}^j_{\tilde{n}} = \boldsymbol{\psi}^j(\boldsymbol{\alpha}_{\tilde{n}}) := \boldsymbol{\psi}(\mu^k(\boldsymbol{\alpha}_{\tilde{n}}), \epsilon^j)$ (for $j \in \{1, \dots, k\}$), are bounded in $H^1(0, T; \mathbb{C}^N)$. By reflexivity of $H^1(0, T; \mathbb{C}^N)$, there exist weakly in $H^1(0, T; \mathbb{C}^N)$ convergent subsequences $\boldsymbol{\psi}^j_{\tilde{n}_l} \rightharpoonup \tilde{\boldsymbol{\psi}}^j$ and by the Sobolev compact embedding $H^1(0, T; \mathbb{C}^N) \Subset C([0, T]; \mathbb{C}^N)$, these subsequences also converge uniformly to $\tilde{\boldsymbol{\psi}}^j$ in $C([0, T]; \mathbb{C}^N)$.

Since $\boldsymbol{\alpha}_{\tilde{n}} \to \tilde{\boldsymbol{\alpha}}$ in $\mathbb{R}^k$, $\boldsymbol{\psi}^j_{\tilde{n}_l} \to \tilde{\boldsymbol{\psi}}^j$ in $C([0, T]; \mathbb{C}^N)$ and $\dot{\boldsymbol{\psi}}^j_{\tilde{n}_l} \rightharpoonup \dot{\tilde{\boldsymbol{\psi}}}^j$ in L^2, we get

$$i\dot{\tilde{\boldsymbol{\psi}}}^j = i[H + \epsilon^j(t)\mu(\tilde{\boldsymbol{\alpha}})]\tilde{\boldsymbol{\psi}}^j, \qquad j = 1, \dots, k$$

similar to the proof of Proposition 13. Hence we have $\tilde{\boldsymbol{\psi}}^j = \boldsymbol{\psi}^j(\tilde{\boldsymbol{\alpha}})$ and therefore

$$\begin{aligned}
\widetilde{J} &= \lim_{l\to\infty} J(\boldsymbol{\alpha}_{\tilde{n}_l}) = \lim_{l\to\infty} \sum_{j=1}^{k} \left| \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu_{k+1}, \epsilon^j) \right\rangle - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu^k(\boldsymbol{\alpha}_{\tilde{n}_l}), \epsilon^j) \right\rangle \right|^2 + \nu \left\| \boldsymbol{\alpha}_{\tilde{n}_l} \right\|_2^2 \\
&= \lim_{l\to\infty} \sum_{j=1}^{k} \left| \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu_{k+1}, \epsilon^j) \right\rangle - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\tilde{n}_l} \right\rangle \right|^2 + \nu \left\| \boldsymbol{\alpha}_{\tilde{n}_l} \right\|_2^2 \\
&= \sum_{j=1}^{k} \left| \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu_{k+1}, \epsilon^j) \right\rangle - \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}^j \right\rangle \right|^2 + \nu \|\tilde{\boldsymbol{\alpha}}\|_2^2 \\
&= \sum_{j=1}^{k} \left| \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu_{k+1}, \epsilon^j) \right\rangle - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_T(\mu(\tilde{\boldsymbol{\alpha}}), \epsilon^j) \right\rangle \right|^2 + \nu \|\tilde{\boldsymbol{\alpha}}\|_2^2 \\
&= J(\tilde{\boldsymbol{\alpha}}) \geq \widehat{J}
\end{aligned}$$

This implies that $\tilde{\boldsymbol{\alpha}}$ is a minimizer for the fitting step problem (57). □

2.3 Discretization and first-order optimality conditions

In order to numerically solve problems (52), (56),(57) and (58) we first have to develop optimality conditions for all our problems and introduce a discretization scheme for our system (49).

2.3.1 First order optimality conditions

2.3.1.1 Main identification and fitting step problems

We can consider both inverse problems, namely the main identification (52) and the fitting step problem (57), at the same time, since the (reduced) cost functional in both cases can be written as

$$\hat{J}(\boldsymbol{\alpha}) = \sum_{j=1}^{L} |\varphi(\mu_{new}, \epsilon^j) - \varphi(\mu^L(\boldsymbol{\alpha}), \epsilon^j)|^2 + \nu \|\boldsymbol{\alpha}\|_2^2,$$

where $L = K$, $\mu_{new} = \mu^\star$ and $\nu = 0$ for the main identification and $L = k$, $\mu_{new} = \mu_{k+1}$ and $\nu > 0$ for the fitting step problem at iteration $k \in \{1, \dots, K-1\}$. Note that the $\varphi(\mu_{new}, \epsilon^j)$ do not depend on $\boldsymbol{\alpha}$. We denote by $\boldsymbol{\psi}^j_{\boldsymbol{\alpha}} := \boldsymbol{\psi}(\mu^L(\boldsymbol{\alpha}), \epsilon^j)$ the solution of (49) for $\epsilon = \epsilon^j$ and $\mu = \mu^L(\boldsymbol{\alpha})$ for $j \in \{1, \dots, L\}$. Now we can also write the cost functional in its unreduced form

$$\begin{aligned} J(\psi^1_{\boldsymbol{\alpha}}, \dots, \psi^L_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) &= \sum_{j=1}^{L} \left| \varphi(\mu_{new}, \epsilon^j) - \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \right|^2 + \nu \left\| \boldsymbol{\alpha} \right\|_2^2 \\ &= \sum_{j=1}^{L} \varphi(\mu_{new}, \epsilon^j)^2 - 2Re\Big(\varphi(\mu_{new}, \epsilon^j) \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \Big) + \left\langle \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \\ &\quad + \nu \left\| \boldsymbol{\alpha} \right\|_2^2 . \end{aligned}$$

The derivative of $\hat{J}(\boldsymbol{\alpha})$ in a direction $\delta\boldsymbol{\alpha} \in \mathbb{R}^L$ can then be computed as follows. We have

$$D\hat{J}(\boldsymbol{\alpha})(\delta\boldsymbol{\alpha}) = D_{\boldsymbol{\alpha}} J(\psi^1_{\boldsymbol{\alpha}}, \dots, \psi^L_{\boldsymbol{\alpha}}, \boldsymbol{\alpha})(\delta\boldsymbol{\alpha}) + \sum_{j=1}^{L} D_{\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}} J(\psi^1_{\boldsymbol{\alpha}}, \dots, \psi^L_{\boldsymbol{\alpha}}, \boldsymbol{\alpha})(\delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}),$$

where $\delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}$ are the solutions to the linearized state equations

$$\begin{cases} i\dot{\delta\boldsymbol{\psi}}^j_{\boldsymbol{\alpha}}(t) &= [H + \epsilon^j(t)\mu^L(\boldsymbol{\alpha})]\delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(t) + \epsilon^j(t)\mu^L(\delta\boldsymbol{\alpha})\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(t) \\ \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(0) &= 0. \end{cases} \tag{63}$$

For $j \in \{1, \dots, L\}$ we introduce the adjoint equation

$$\begin{cases} i\dot{\boldsymbol{\lambda}}^j(t) &= [H + \epsilon^j(t)\mu^L(\boldsymbol{\alpha})]\boldsymbol{\lambda}^j(t) \\ \boldsymbol{\lambda}^j(T) &= \Big(\left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle - \varphi(\mu_{new}, \epsilon^j) \Big) \boldsymbol{\psi}_1 \end{cases} \tag{64}$$

and get

$$\begin{aligned} D_{\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}} J(\psi^1_{\boldsymbol{\alpha}}, \dots, \psi^L_{\boldsymbol{\alpha}}, \boldsymbol{\alpha})(\delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}) &= -2Re\Big(\varphi(\mu_{new}, \epsilon^j) \left\langle \boldsymbol{\psi}_1, \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \Big) + \left\langle \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \\ &\quad + \left\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \\ &= -2Re\Big(\varphi(\mu_{new}, \epsilon^j) \left\langle \boldsymbol{\psi}_1, \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \Big) + 2Re\left\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle \\ &= 2Re\Big\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \underbrace{\Big(\left\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T) \right\rangle - \varphi(\mu_{new}, \epsilon^j) \Big) \boldsymbol{\psi}_1}_{=\boldsymbol{\lambda}^j(T)} \Big\rangle \\ &= 2Re\left\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \boldsymbol{\lambda}^j(T) \right\rangle. \end{aligned}$$

Integrating by parts and using that H and $\mu^L(\boldsymbol{\alpha})$ are Hermitian, we get

$$\begin{aligned}
\Big\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T), \boldsymbol{\lambda}^j(T)\Big\rangle &= \Big\langle \delta\dot{\boldsymbol{\psi}}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2} + \Big\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \dot{\boldsymbol{\lambda}}^j\Big\rangle_{L^2} + \Big\langle \underbrace{\delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(0)}_{=0}, \boldsymbol{\lambda}^j(0)\Big\rangle \\
&\overset{(63)}{\underset{(64)}{=}} \Big\langle -i[H+\epsilon^j\mu^L(\boldsymbol{\alpha})]\delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}} + \epsilon^j\mu^L(\delta\boldsymbol{\alpha})\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2} \\
&\quad + \Big\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, -i[H+\epsilon^j\mu^L(\boldsymbol{\alpha})]\boldsymbol{\lambda}^j\Big\rangle_{L^2} \\
&= \Big\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, i[H+\epsilon^j\mu^L(\boldsymbol{\alpha})]\boldsymbol{\lambda}^j\Big\rangle_{L^2} + \Big\langle \epsilon^j\mu^L(\delta\boldsymbol{\alpha})\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2} \\
&\quad - \Big\langle \delta\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, i[H+\epsilon^j\mu^L(\boldsymbol{\alpha})]\boldsymbol{\lambda}^j\Big\rangle_{L^2} \\
&= \Big\langle \epsilon^j\Big(\sum_{l=1}^{L}\delta\boldsymbol{\alpha}_l\mu_l\Big)\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2} \\
&= \sum_{l=1}^{L}\delta\boldsymbol{\alpha}_l\Big\langle \epsilon^j\mu_l\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2} \\
&= \Big\langle \Big[\Big\langle \epsilon^j\mu_1\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2}, \dots, \Big\langle \epsilon^j\mu_L\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2}\Big]^T, \delta\boldsymbol{\alpha}\Big\rangle.
\end{aligned}$$

Since $D_{\boldsymbol{\alpha}}J(\psi^1_{\boldsymbol{\alpha}}, \dots, \psi^L_{\boldsymbol{\alpha}}, \boldsymbol{\alpha})(\delta\boldsymbol{\alpha}) = 2\nu\Big\langle \boldsymbol{\alpha}, \delta\boldsymbol{\alpha}\Big\rangle$ we have

$$D\hat{J}(\boldsymbol{\alpha})(\delta\boldsymbol{\alpha}) = \underbrace{\Big\langle 2\nu\boldsymbol{\alpha} + \sum_{j=1}^{L} 2Re\Big[\Big\langle \epsilon^j\mu_1\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2}, \dots, \Big\langle \epsilon^j\mu_L\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2}\Big]^T, \delta\boldsymbol{\alpha}\Big\rangle}_{=:\nabla\hat{J}(\boldsymbol{\alpha})}.$$

Therefore we obtain the corresponding optimality system

$$i\dot{\boldsymbol{\psi}}^j_{\boldsymbol{\alpha}}(t) = [H+\epsilon^j(t)\mu^L(\boldsymbol{\alpha})]\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(t), \qquad j=1,\dots,L \tag{65}$$

$$\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(0) = \boldsymbol{\psi}_0, \qquad j=1,\dots,L \tag{66}$$

$$i\dot{\boldsymbol{\lambda}}^j(t) = [H+\epsilon^j(t)\mu^L(\boldsymbol{\alpha})]\boldsymbol{\lambda}^j(t), \qquad j=1,\dots,L \tag{67}$$

$$\boldsymbol{\lambda}^j(T) = \Big(\Big\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}^j_{\boldsymbol{\alpha}}(T)\Big\rangle - \varphi(\mu_{new}, \epsilon^j)\Big)\boldsymbol{\psi}_1, \qquad j=1,\dots,L \tag{68}$$

$$\nabla\hat{J}(\boldsymbol{\alpha}) = 2\nu\boldsymbol{\alpha} + \sum_{j=1}^{L} 2Re\Big[\Big\langle \epsilon^j\mu_1\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2}, \dots, \Big\langle \epsilon^j\mu_L\boldsymbol{\psi}^j_{\boldsymbol{\alpha}}, \boldsymbol{\lambda}^j\Big\rangle_{L^2}\Big]^T = 0, \tag{69}$$

where the state and adjoint equations hold in $(0,T)$.

2.3.1.2 Initialization and discriminatory step problems

As in the proof of the well-posedeness, we can again consider both optimal control problems (56) and (58) simultaneously by denoting the (reduced) cost functional as

$$\hat{J}(\epsilon) = |\varphi(\tilde{\mu},\epsilon) - \varphi(\hat{\mu},\epsilon)|^2 - \beta\,\|\epsilon\|^2_{L^2(0,T)}\,,$$

where $\tilde{\mu} = \mu_1$, $\hat{\mu} = 0$ in case of the initialization and $\tilde{\mu} = \mu_{k+1}$, $\hat{\mu} = \mu^k(\boldsymbol{\alpha}^k)$ in case of the discriminatory step problem. We denote by $\tilde{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}$ the solutions of (49) corresponding to $\mu = \tilde{\mu}$ and $\mu = \hat{\mu}$ respectively. Now we can write the cost functional in its unreduced form

$$\begin{aligned}
J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon) &= |\Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T)\Big\rangle - \Big\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}(T)\Big\rangle|^2 - \beta\,\|\epsilon\|^2_{L^2(0,T)} \\
&= \Big\langle \tilde{\boldsymbol{\psi}}(T), \boldsymbol{\psi}_1\Big\rangle\Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T)\Big\rangle - 2Re\Big\langle \tilde{\boldsymbol{\psi}}(T), \boldsymbol{\psi}_1\Big\rangle\Big\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}(T)\Big\rangle + \Big\langle \hat{\boldsymbol{\psi}}(T), \boldsymbol{\psi}_1\Big\rangle\Big\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}(T)\Big\rangle \\
&\quad - \beta\,\|\epsilon\|^2_{L^2(0,T)}\,.
\end{aligned}$$

The derivative of $\hat{J}(\epsilon)$ in a direction $\delta\epsilon \in L^2(0,T)$ can be computed as follows. We have

$$D\hat{J}(\epsilon)(\delta\epsilon) = D_\epsilon J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\epsilon) + D_{\tilde{\boldsymbol{\psi}}} J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\tilde{\boldsymbol{\psi}}) + D_{\hat{\boldsymbol{\psi}}} J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\hat{\boldsymbol{\psi}}),$$

where $\delta\tilde{\boldsymbol{\psi}}$ and $\delta\hat{\boldsymbol{\psi}}$ are the solutions to the linearized state equations

$$\begin{cases} i\dot{\delta\tilde{\boldsymbol{\psi}}}(t) &= [H + \epsilon(t)\tilde{\mu}]\delta\tilde{\boldsymbol{\psi}}(t) + \delta\epsilon(t)\tilde{\mu}\tilde{\boldsymbol{\psi}}(t) \\ \delta\tilde{\boldsymbol{\psi}}(0) &= 0, \end{cases} \tag{70}$$

$$\begin{cases} i\dot{\delta\hat{\boldsymbol{\psi}}}(t) &= [H + \epsilon(t)\hat{\mu}]\delta\hat{\boldsymbol{\psi}}(t) + \delta\epsilon(t)\hat{\mu}\hat{\boldsymbol{\psi}}(t) \\ \delta\hat{\boldsymbol{\psi}}(0) &= 0. \end{cases} \tag{71}$$

We introduce the adjoint equations

$$\begin{cases} i\dot{\tilde{\boldsymbol{\lambda}}}(t) &= [H + \epsilon(t)\tilde{\mu}]\tilde{\boldsymbol{\lambda}}(t) \\ \tilde{\boldsymbol{\lambda}}(T) &= \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T) \Big\rangle \boldsymbol{\psi}_1, \end{cases} \tag{72}$$

$$\begin{cases} i\dot{\hat{\boldsymbol{\lambda}}}(t) &= [H + \epsilon(t)\hat{\mu}]\hat{\boldsymbol{\lambda}}(t) \\ \hat{\boldsymbol{\lambda}}(T) &= \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T) \Big\rangle \boldsymbol{\psi}_1, \end{cases} \tag{73}$$

therefore

$$\begin{aligned} D_{\tilde{\boldsymbol{\psi}}} J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\tilde{\boldsymbol{\psi}}(T)) &= 2Re\Big\langle \delta\tilde{\boldsymbol{\psi}}(T), \boldsymbol{\psi}_1 \Big\rangle \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) \Big\rangle - 2Re\Big\langle \delta\tilde{\boldsymbol{\psi}}(T), \boldsymbol{\psi}_1 \Big\rangle \Big\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}(T) \Big\rangle \\ &= 2Re\Big\langle \delta\tilde{\boldsymbol{\psi}}(T), \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T) \Big\rangle \boldsymbol{\psi}_1 \Big\rangle \\ &= 2Re\Big\langle \delta\tilde{\boldsymbol{\psi}}(T), \tilde{\boldsymbol{\lambda}}(T) \Big\rangle \end{aligned}$$

and similarly

$$D_{\hat{\boldsymbol{\psi}}} J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\hat{\boldsymbol{\psi}}) = -2Re\Big\langle \delta\hat{\boldsymbol{\psi}}(T), \hat{\boldsymbol{\lambda}}(T) \Big\rangle.$$

Integrating by parts we get

$$\begin{aligned} \Big\langle \delta\tilde{\boldsymbol{\psi}}(T), \tilde{\boldsymbol{\lambda}}(T) \Big\rangle &= \Big\langle \dot{\delta\tilde{\boldsymbol{\psi}}}, \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} + \Big\langle \delta\tilde{\boldsymbol{\psi}}, \dot{\tilde{\boldsymbol{\lambda}}} \Big\rangle_{L^2} + \Big\langle \underbrace{\delta\tilde{\boldsymbol{\psi}}(0)}_{=0}, \tilde{\boldsymbol{\lambda}}(0) \Big\rangle \\ &\overset{(70)}{=} \Big\langle -i\Big([H + \epsilon\tilde{\mu}]\delta\tilde{\boldsymbol{\psi}} + \delta\epsilon\tilde{\mu}\tilde{\boldsymbol{\psi}}\Big), \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} + \Big\langle \delta\tilde{\boldsymbol{\psi}}, \dot{\tilde{\boldsymbol{\lambda}}} \Big\rangle_{L^2} \\ &= \Big\langle \delta\tilde{\boldsymbol{\psi}}, \underbrace{i[H + \epsilon\tilde{\mu}]\tilde{\boldsymbol{\lambda}} + \dot{\tilde{\boldsymbol{\lambda}}}}_{\overset{(72)}{=}0} \Big\rangle_{L^2} + i\Big\langle \delta\epsilon\tilde{\mu}\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} \\ &= i\Big\langle \Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\lambda}} \Big\rangle, \delta\epsilon \Big\rangle_{L^2(0,T)}, \end{aligned}$$

therefore

$$D_{\tilde{\boldsymbol{\psi}}} J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\tilde{\boldsymbol{\psi}}(T)) = \Big\langle 2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\lambda}} \Big\rangle, \delta\epsilon \Big\rangle_{L^2(0,T)}$$

and similarly

$$D_{\hat{\boldsymbol{\psi}}} J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\hat{\boldsymbol{\psi}}) = \Big\langle -2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}} \Big\rangle, \delta\epsilon \Big\rangle_{L^2(0,T)}.$$

Since $D_\epsilon J(\tilde{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}, \epsilon)(\delta\epsilon) = 2\beta\Big\langle \epsilon, \delta\epsilon \Big\rangle_{L^2(0,T)}$, we conclude that

$$D\hat{J}(\epsilon)(\delta\epsilon) = \Big\langle \underbrace{2\beta\epsilon + 2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\lambda}} \Big\rangle - 2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}} \Big\rangle}_{=:\nabla\hat{J}(\epsilon)}, \delta\epsilon \Big\rangle_{L^2(0,T)}.$$

Hence we obtain the corresponding optimality system

$$i\dot{\tilde{\boldsymbol{\psi}}}(t) = [H + \epsilon(t)\tilde{\mu}]\tilde{\boldsymbol{\psi}}(t), \qquad \tilde{\boldsymbol{\psi}}(0) = \boldsymbol{\psi}_0, \tag{74}$$

$$i\dot{\hat{\boldsymbol{\psi}}}(t) = [H + \epsilon(t)\hat{\mu}]\hat{\boldsymbol{\psi}}(t), \qquad \hat{\boldsymbol{\psi}}(0) = \boldsymbol{\psi}_0, \tag{75}$$

$$i\dot{\tilde{\boldsymbol{\lambda}}}(t) = [H + \epsilon(t)\tilde{\mu}]\tilde{\boldsymbol{\lambda}}(t), \qquad \tilde{\boldsymbol{\lambda}}(T) = \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T) \Big\rangle \boldsymbol{\psi}_1, \tag{76}$$

$$i\dot{\hat{\boldsymbol{\lambda}}}(t) = [H + \epsilon(t)\hat{\mu}]\hat{\boldsymbol{\lambda}}(t), \qquad \hat{\boldsymbol{\lambda}}(T) = \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T) \Big\rangle \boldsymbol{\psi}_1, \tag{77}$$

$$\nabla \hat{J}(\epsilon) = 2\beta\epsilon + 2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\lambda}} \Big\rangle - 2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}} \Big\rangle = 0, \quad \text{in } (0, T). \tag{78}$$

2.3.2 Discretization

We consider a uniform grid with step size $h = T/N_t$ and introduce the discrete L^2-product

$$\Big\langle \boldsymbol{v}, \boldsymbol{w} \Big\rangle_{L_h^2} := h \sum_{i=0}^{N_t} \bar{\boldsymbol{v}}_i^T \boldsymbol{w}_i.$$

In order to solve the discrete version of our system (49), we use a second-order Strang splitting (compare [Str68]), meaning that we compute the approximated solution $\boldsymbol{\psi}_i = \boldsymbol{\psi}(ih)$, $0 < i \leq N_t$ as the solution to the following equation.

$$\boldsymbol{\psi}_{i+1} = e^{-iH\frac{h}{2}} e^{-i\epsilon_i \mu h} e^{-iH\frac{h}{2}} \boldsymbol{\psi}_i. \tag{79}$$

Hence the discrete (*Optimize-Before-Discretize*) optimality system for the main identification and fitting step problems (compare (65)-(69)) is

$$\begin{aligned}
(\boldsymbol{\psi}_{\boldsymbol{\alpha}}^j)_{i+1} &= e^{-iH\frac{h}{2}} e^{-i\epsilon_i \mu h} e^{-iH\frac{h}{2}} (\boldsymbol{\psi}_{\boldsymbol{\alpha}}^j)_i, && j = 1, \dots, L, && i = 0, \dots, N_t - 1, \\
(\boldsymbol{\psi}_{\boldsymbol{\alpha}}^j)_0 &= \boldsymbol{\psi}_0, && j = 1, \dots, L, && \\
\boldsymbol{\lambda}_{i-1}^j &= e^{iH\frac{h}{2}} e^{i\epsilon_{i-1} \mu h} e^{iH\frac{h}{2}} \boldsymbol{\lambda}_i^j, && j = 1, \dots, L, && i = N_t, \dots, 1, \\
\boldsymbol{\lambda}_{N_t}^j &= \Big(\Big\langle \boldsymbol{\psi}_1, (\boldsymbol{\psi}_{\boldsymbol{\alpha}}^j)_{N_t} \Big\rangle - \varphi(\mu_{new}, \epsilon^j) \Big) \boldsymbol{\psi}_1, && j = 1, \dots, L, &&
\end{aligned}$$

$$\nabla \hat{J}_h(\boldsymbol{\alpha}) = 2\nu\boldsymbol{\alpha} + \sum_{j=1}^{L} 2Re\Big[\Big\langle \epsilon^j \mu_1 \boldsymbol{\psi}_{\boldsymbol{\alpha}}^j, \boldsymbol{\lambda}^j \Big\rangle_{L_h^2}, \dots, \Big\langle \epsilon^j \mu_L \boldsymbol{\psi}_{\boldsymbol{\alpha}}^j, \boldsymbol{\lambda}^j \Big\rangle_{L_h^2} \Big]^T = 0.$$

The discrete optimality system for the initialization and discriminatory step problems (compare (74)-(78)) is

$$\tilde{\boldsymbol{\psi}}_{i+1} = e^{-iH\frac{h}{2}} e^{-i\epsilon_i \tilde{\mu} h} e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\psi}}_i, \qquad i = 0, \dots, N_t - 1, \tag{80}$$

$$\tilde{\boldsymbol{\psi}}_0 = \boldsymbol{\psi}_0, \tag{81}$$

$$\hat{\boldsymbol{\psi}}_{i+1} = e^{-iH\frac{h}{2}} e^{-i\epsilon_i \hat{\mu} h} e^{-iH\frac{h}{2}} \hat{\boldsymbol{\psi}}_i, \qquad i = 0, \dots, N_t - 1, \tag{82}$$

$$\hat{\boldsymbol{\psi}}_0 = \boldsymbol{\psi}_0, \tag{83}$$

$$\tilde{\boldsymbol{\lambda}}_{i-1} = e^{iH\frac{h}{2}} e^{i\epsilon_{i-1} \tilde{\mu} h} e^{iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_i, \qquad i = N_t, \dots, 1, \tag{84}$$

$$\tilde{\boldsymbol{\lambda}}_{N_t} = \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}_{N_t} - \hat{\boldsymbol{\psi}}_{N_t} \Big\rangle \boldsymbol{\psi}_1, \tag{85}$$

$$\hat{\boldsymbol{\lambda}}_{i-1} = e^{iH\frac{h}{2}} e^{i\epsilon_{i-1} \hat{\mu} h} e^{iH\frac{h}{2}} \hat{\boldsymbol{\lambda}}_i, \qquad i = N_t, \dots, 1, \tag{86}$$

$$\hat{\boldsymbol{\lambda}}_{N_t} = \Big\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}_{N_t} - \hat{\boldsymbol{\psi}}_{N_t} \Big\rangle \boldsymbol{\psi}_1, \tag{87}$$

$$(\nabla \hat{J}_h(\epsilon))_i = 2\beta\epsilon_i + 2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}_i, \tilde{\boldsymbol{\lambda}}_i \Big\rangle - 2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\lambda}}_i \Big\rangle = 0, \qquad i = 0, \dots, N_t. \tag{88}$$

2.4 Monotonic schemes

In order to solve the two optimal control problems in the algorithm (the initialization (56) and discriminatory step problem (58)), we will use a so-called "monotonic scheme". Given an arbitrary control $\epsilon \in L^2(0,T)$, the idea of monotonic schemes is to find a control $\epsilon' \in L^2(0,T)$ such that $J(\epsilon') \geq J(\epsilon)$ by using certain properties of the cost functional J. A more detailed introduction to monotonic schemes for optimal control in quantum systems can be found in [MST06].

We will consider two different approaches to these monotonic schemes, one for the continuous system and one for the discretized system.

2.4.1 Monotonic scheme for the continuous system

As before we denote the two cost functionals of (56) and (58) as

$$J(\epsilon) = |\varphi(\tilde{\mu}, \epsilon) - \varphi(\hat{\mu}, \epsilon)|^2 - \beta \|\epsilon\|^2_{L^2(0,T)},$$

where $\tilde{\mu} = \mu_1$, $\hat{\mu} = 0$ in case of the initialization and $\tilde{\mu} = \mu_{k+1}$, $\hat{\mu} = \mu^k(\boldsymbol{\alpha}^k)$ in case of the discriminatory step problem. Let $\epsilon \in L^2(0,T)$ and denote by $\tilde{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}$ the solutions of (49) corresponding to $\mu = \tilde{\mu}$ and $\mu = \hat{\mu}$, respectively. Consider another control $\epsilon' \in L^2(0,T)$, again denoting by $\tilde{\boldsymbol{\psi}}'$ and $\hat{\boldsymbol{\psi}}'$ the corresponding solutions to (49) with $\mu = \tilde{\mu}$ and $\mu = \hat{\mu}$ respectively. Using the adjoint equations (72) and (73) from Section 2.3.1.2 and defining $\Delta\boldsymbol{\psi}(T) := \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T)$, $\Delta\boldsymbol{\psi}'(T) := \tilde{\boldsymbol{\psi}}'(T) - \hat{\boldsymbol{\psi}}'(T)$, we obtain

$$\begin{aligned}
J(\epsilon') - J(\epsilon) &= \left|\left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}'(T)\right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}'(T)\right\rangle\right|^2 - \beta \left\|\epsilon'\right\|^2_{L^2(0,T)} \\
&\quad - \left|\left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T)\right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}(T)\right\rangle\right|^2 + \beta \|\epsilon\|^2_{L^2(0,T)} \\
&= \left\langle \tilde{\boldsymbol{\psi}}'(T) - \hat{\boldsymbol{\psi}}'(T), \boldsymbol{\psi}_1\right\rangle\left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}'(T) - \hat{\boldsymbol{\psi}}'(T)\right\rangle - \beta \int_0^T \epsilon'^2(t)dt \\
&\quad - \left\langle \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T), \boldsymbol{\psi}_1\right\rangle\left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}(T) - \hat{\boldsymbol{\psi}}(T)\right\rangle + \beta \int_0^T \epsilon^2(t)dt \\
&= \left\langle \Delta\boldsymbol{\psi}'(T), \boldsymbol{\psi}_1\right\rangle\left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'(T)\right\rangle - \left\langle \Delta\boldsymbol{\psi}(T), \boldsymbol{\psi}_1\right\rangle\left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}(T)\right\rangle \\
&\quad - \beta \int_0^T \epsilon'^2(t) - \epsilon^2(t)dt \\
&= \left\langle \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T), \boldsymbol{\psi}_1\right\rangle\left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T)\right\rangle \\
&\quad + 2Re\left\langle \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T), \boldsymbol{\psi}_1\right\rangle\left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}(T)\right\rangle - \beta \int_0^T \epsilon'^2(t) - \epsilon^2(t)dt \\
&= \left|\left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T)\right\rangle\right|^2 \\
&\quad + 2Re\left\langle \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T), \tilde{\boldsymbol{\lambda}}(T)\right\rangle - \beta \int_0^T \epsilon'^2(t) - \epsilon^2(t)dt,
\end{aligned}$$

where

$$\left\langle \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T), \tilde{\boldsymbol{\lambda}}(T)\right\rangle = \left\langle \tilde{\boldsymbol{\psi}}'(T) - \tilde{\boldsymbol{\psi}}(T), \tilde{\boldsymbol{\lambda}}(T)\right\rangle - \left\langle \hat{\boldsymbol{\psi}}'(T) - \hat{\boldsymbol{\psi}}(T), \hat{\boldsymbol{\lambda}}(T)\right\rangle.$$

Integrating by parts we get

$$\begin{aligned}
\Big\langle \tilde{\boldsymbol{\psi}}'(T) - \tilde{\boldsymbol{\psi}}(T), \tilde{\boldsymbol{\lambda}}(T) \Big\rangle &= \Big\langle \dot{\tilde{\boldsymbol{\psi}}}' - \dot{\tilde{\boldsymbol{\psi}}}, \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} + \Big\langle \tilde{\boldsymbol{\psi}}' - \tilde{\boldsymbol{\psi}}, \dot{\tilde{\boldsymbol{\lambda}}} \Big\rangle_{L^2} + \Big\langle \underbrace{\tilde{\boldsymbol{\psi}}'(0)}_{=\boldsymbol{\psi}_0} - \underbrace{\tilde{\boldsymbol{\psi}}(0)}_{=\boldsymbol{\psi}_0}, \tilde{\boldsymbol{\lambda}}(0) \Big\rangle \\
&= \Big\langle -i[H + \epsilon'\tilde{\mu}]\tilde{\boldsymbol{\psi}}' + i[H + \epsilon\tilde{\mu}]\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} + \Big\langle \tilde{\boldsymbol{\psi}}' - \tilde{\boldsymbol{\psi}}, \dot{\tilde{\boldsymbol{\lambda}}} \Big\rangle_{L^2} \\
&= \Big\langle \Big(-i[H + \epsilon'\tilde{\mu}] + i[H + \epsilon\tilde{\mu}] \Big)\tilde{\boldsymbol{\psi}}', \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} \\
&\quad + \Big\langle -i[H + \epsilon\tilde{\mu}](\tilde{\boldsymbol{\psi}}' - \tilde{\boldsymbol{\psi}}), \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} + \Big\langle \tilde{\boldsymbol{\psi}}' - \tilde{\boldsymbol{\psi}}, \dot{\tilde{\boldsymbol{\lambda}}} \Big\rangle_{L^2} \\
&= \Big\langle -i(\epsilon' - \epsilon)\tilde{\mu}\tilde{\boldsymbol{\psi}}', \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} + \Big\langle \tilde{\boldsymbol{\psi}}' - \tilde{\boldsymbol{\psi}}, \underbrace{\dot{\tilde{\boldsymbol{\lambda}}} + i[H + \epsilon\tilde{\mu}]\tilde{\boldsymbol{\lambda}}}_{=0} \Big\rangle_{L^2} \\
&= i\Big\langle (\epsilon' - \epsilon)\tilde{\mu}\tilde{\boldsymbol{\psi}}', \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2}
\end{aligned}$$

and analogously

$$\Big\langle \hat{\boldsymbol{\psi}}'(T) - \hat{\boldsymbol{\psi}}(T), \hat{\boldsymbol{\lambda}}(T) \Big\rangle = i\Big\langle (\epsilon' - \epsilon)\hat{\mu}\hat{\boldsymbol{\psi}}', \hat{\boldsymbol{\lambda}} \Big\rangle_{L^2}.$$

Concluding we have

$$\begin{aligned}
J(\epsilon') - J(\epsilon) &= \Big| \Big\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T) \Big\rangle \Big|^2 - \beta \int_0^T \epsilon'^2(t) - \epsilon^2(t) dt \\
&\quad + 2Im\Big\langle (\epsilon' - \epsilon)\tilde{\mu}\tilde{\boldsymbol{\psi}}', \tilde{\boldsymbol{\lambda}} \Big\rangle_{L^2} - 2Im\Big\langle (\epsilon' - \epsilon)\hat{\mu}\hat{\boldsymbol{\psi}}', \hat{\boldsymbol{\lambda}} \Big\rangle_{L^2} \\
&= \Big| \Big\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T) \Big\rangle \Big|^2 + \int_0^T (\epsilon'(t) - \epsilon(t)) \Big(-\beta(\epsilon'(t) + \epsilon(t)) \\
&\quad + 2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}'(t), \tilde{\boldsymbol{\lambda}}(t) \Big\rangle - 2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}'(t), \hat{\boldsymbol{\lambda}}(t) \Big\rangle \Big) dt.
\end{aligned}$$

Hence a sufficient condition for $J(\epsilon') \geq J(\epsilon)$ is

$$(\epsilon'(t) - \epsilon(t)) \Big(-\beta(\epsilon'(t) + \epsilon(t)) + 2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}'(t), \tilde{\boldsymbol{\lambda}}(t) \Big\rangle - 2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}'(t), \hat{\boldsymbol{\lambda}}(t) \Big\rangle \Big) \geq 0, \tag{89}$$

for all $t \in [0, T]$. There are many ways to find an $\epsilon' \in L^2(0, T)$ such that (89) holds. The idea presented in [MS09, Section 5.1] is defining $\epsilon'(t)$ for each $t \in [0, T]$ as the solution of

$$\epsilon'(t) - \epsilon(t) = \frac{1}{\beta} \Big(2Im\Big\langle \tilde{\mu}\tilde{\boldsymbol{\psi}}'(t), \tilde{\boldsymbol{\lambda}}(t) \Big\rangle - 2Im\Big\langle \hat{\mu}\hat{\boldsymbol{\psi}}'(t), \hat{\boldsymbol{\lambda}}(t) \Big\rangle - \beta(\epsilon'(t) - \epsilon(t)) \Big), \tag{90}$$

leading to

$$J(\epsilon') - J(\epsilon) = \Big| \Big\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'(T) - \Delta\boldsymbol{\psi}(T) \Big\rangle \Big|^2 + \beta \int_0^T (\epsilon'(t) - \epsilon(t))^2 dt \geq 0.$$

However, the guarantee of monotonicity is not passed down to the discrete system, meaning that an $\epsilon' \in L_h^2(0, T)$ satisfying the discretized version of (90) (with $t = t_j$ for $j \in \{0, \ldots, N_t - 1\}$) would in general not guarantee that $J_h(\epsilon') \geq J_h(\epsilon)$. Therefore we will introduce a monotonic scheme that is directly based on the discrete cost functional in the next section.

Remark 17

One could also call this approach "Monotonize-Before-Discretize", since we derive a sufficient monotonicity condition before discretizing the system.

2.4.2 Monotonic scheme for the discretized system

For the unoptimized discrete system ((80)-(87)), we notice that the last entry of the control ϵ_{N_t} does not play any role. Therefore we introduce the discrete $L^2(0,T)$-product

$$\left\langle \tilde{\epsilon}, \hat{\epsilon} \right\rangle_{L_h^2(0,T)} := h \sum_{i=0}^{N_t-1} \tilde{\epsilon}_i \hat{\epsilon}_i.$$

Now we want to develop a monotonic scheme starting from the discrete cost functional

$$J_h(\epsilon) = \left| \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}_{N_t} \right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}_{N_t} \right\rangle \right|^2 - \beta \left\| \epsilon \right\|_{L_h^2(0,T)}, \tag{91}$$

for $\epsilon \in L_h^2(0,T)$ and where by $\tilde{\boldsymbol{\psi}}$, $\hat{\boldsymbol{\psi}}$ we now denote the solutions to (80)-(81),(82)-(83) respectively. Consider another control $\epsilon' \in L_h^2(0,T)$ and denote by $\tilde{\boldsymbol{\psi}}'$, $\hat{\boldsymbol{\psi}}'$ the corresponding solutions to (80)-(81),(82)-(83) with $\epsilon = \epsilon'$. Using the adjoint equations (84)-(85),(86)-(87) and defining $\Delta\boldsymbol{\psi}_{N_t} := \tilde{\boldsymbol{\psi}}_{N_t} - \hat{\boldsymbol{\psi}}_{N_t}$, $\Delta\boldsymbol{\psi}'_{N_t} := \tilde{\boldsymbol{\psi}}'_{N_t} - \hat{\boldsymbol{\psi}}'_{N_t}$, we obtain

$$\begin{aligned}
J_h(\epsilon') - J_h(\epsilon) &= \left| \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}'_{N_t} \right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}'_{N_t} \right\rangle \right|^2 - \beta \left\| \epsilon' \right\|^2_{L_h^2(0,T)} \\
&\quad - \left| \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}_{N_t} \right\rangle - \left\langle \boldsymbol{\psi}_1, \hat{\boldsymbol{\psi}}_{N_t} \right\rangle \right|^2 + \beta \left\| \epsilon \right\|^2_{L_h^2(0,T)} \\
&= \left\langle \tilde{\boldsymbol{\psi}}'_{N_t} - \hat{\boldsymbol{\psi}}'_{N_t}, \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}'_{N_t} - \hat{\boldsymbol{\psi}}'_{N_t} \right\rangle - \beta h \sum_{i=0}^{N_t-1} \epsilon_i'^2 \\
&\quad - \left\langle \tilde{\boldsymbol{\psi}}_{N_t} - \hat{\boldsymbol{\psi}}_{N_t}, \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \tilde{\boldsymbol{\psi}}_{N_t} - \hat{\boldsymbol{\psi}}_{N_t} \right\rangle + \beta h \sum_{i=0}^{N_t-1} \epsilon_i^2 \\
&= \left\langle \Delta\boldsymbol{\psi}'_{N_t}, \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'_{N_t} \right\rangle - \left\langle \Delta\boldsymbol{\psi}_{N_t}, \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}_{N_t} \right\rangle \\
&\quad - \beta h \sum_{i=0}^{N_t-1} (\epsilon_i'^2 - \epsilon_i^2) \\
&= \left\langle \delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t}, \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t} \right\rangle \\
&\quad + 2Re \left\langle \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t}, \boldsymbol{\psi}_1 \right\rangle \left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}_{N_t} \right\rangle - \beta h \sum_{i=0}^{N_t-1} (\epsilon_i'^2 - \epsilon_i^2) \\
&= \left| \left\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t} \right\rangle \right|^2 \\
&\quad + 2Re \left\langle \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t}, \tilde{\boldsymbol{\lambda}}_{N_t} \right\rangle - \beta h \sum_{i=0}^{N_t-1} (\epsilon_i'^2 - \epsilon_i^2),
\end{aligned}$$

where

$$\left\langle \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t}, \tilde{\boldsymbol{\lambda}}_{N_t} \right\rangle = \left\langle \tilde{\boldsymbol{\psi}}'_{N_t} - \tilde{\boldsymbol{\psi}}_{N_t}, \tilde{\boldsymbol{\lambda}}_{N_t} \right\rangle - \left\langle \hat{\boldsymbol{\psi}}'_{N_t} - \hat{\boldsymbol{\psi}}_{N_t}, \hat{\boldsymbol{\lambda}}_{N_t} \right\rangle.$$

Now, we study the two terms at the right-hand side of this equality. Since $\tilde{\boldsymbol{\psi}}'_0 = \tilde{\boldsymbol{\psi}}_0 = \boldsymbol{\psi}_0$ and $H, \tilde{\mu}$ are real, symmetric matrices, we get

$$\begin{aligned}
\Big\langle \tilde{\boldsymbol{\psi}}'_{N_t} - \tilde{\boldsymbol{\psi}}_{N_t}, \tilde{\boldsymbol{\lambda}}_{N_t} \Big\rangle &= \sum_{i=0}^{N_t-1} \Big\langle \tilde{\boldsymbol{\psi}}'_{i+1} - \tilde{\boldsymbol{\psi}}_{i+1}, \tilde{\boldsymbol{\lambda}}_{i+1} \Big\rangle - \Big\langle \tilde{\boldsymbol{\psi}}'_i - \tilde{\boldsymbol{\psi}}_i, \tilde{\boldsymbol{\lambda}}_i \Big\rangle \\
&= \sum_{i=0}^{N_t-1} \Big\langle e^{-iH\frac{h}{2}} e^{-i\epsilon'_i \tilde{\mu} h} e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\psi}}'_i, e^{-iH\frac{h}{2}} e^{-i\epsilon_i \tilde{\mu} h} e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_i \Big\rangle \\
&\quad - \Big\langle e^{-iH\frac{h}{2}} e^{-i\epsilon_i \tilde{\mu} h} e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\psi}}_i, \tilde{\boldsymbol{\lambda}}_{i+1} \Big\rangle - \Big\langle \tilde{\boldsymbol{\psi}}'_i, \tilde{\boldsymbol{\lambda}}_i \Big\rangle + \Big\langle \tilde{\boldsymbol{\psi}}_i, e^{iH\frac{h}{2}} e^{i\epsilon_i \tilde{\mu} h} e^{iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_{i+1} \Big\rangle \\
&= \sum_{i=0}^{N_t-1} \Big\langle \tilde{\boldsymbol{\psi}}'_i, e^{iH\frac{h}{2}} e^{i\epsilon'_i \tilde{\mu} h} e^{-i\epsilon_i \tilde{\mu} h} e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_i \Big\rangle - \Big\langle \tilde{\boldsymbol{\psi}}_i, e^{iH\frac{h}{2}} e^{i\epsilon_i \tilde{\mu} h} e^{iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_{i+1} \Big\rangle \\
&\qquad\qquad - \Big\langle \tilde{\boldsymbol{\psi}}'_i, \tilde{\boldsymbol{\lambda}}_i \Big\rangle + \Big\langle \tilde{\boldsymbol{\psi}}_i, e^{iH\frac{h}{2}} e^{i\epsilon_i \tilde{\mu} h} e^{iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_{i+1} \Big\rangle \\
&= \sum_{i=0}^{N_t-1} \Big\langle \tilde{\boldsymbol{\psi}}'_i, e^{iH\frac{h}{2}} \Big(e^{i(\epsilon'_i - \epsilon_i)\tilde{\mu} h} - \mathcal{I} \Big) e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_i \Big\rangle
\end{aligned}$$

and analogously

$$\Big\langle \hat{\boldsymbol{\psi}}'_{N_t} - \hat{\boldsymbol{\psi}}_{N_t}, \hat{\boldsymbol{\lambda}}_{N_t} \Big\rangle = \sum_{i=0}^{N_t-1} \Big\langle \hat{\boldsymbol{\psi}}'_i, e^{iH\frac{h}{2}} \Big(e^{i(\epsilon'_i - \epsilon_i)\hat{\mu} h} - \mathcal{I} \Big) e^{-iH\frac{h}{2}} \hat{\boldsymbol{\lambda}}_i \Big\rangle.$$

Concluding we have

$$\begin{aligned}
J_h(\epsilon') - J_h(\epsilon) &= \Big| \Big\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t} \Big\rangle \Big|^2 + 2Re \sum_{i=0}^{N_t-1} \Big\langle \tilde{\boldsymbol{\psi}}'_i, e^{iH\frac{h}{2}} \Big(e^{i(\epsilon'_i - \epsilon_i)\tilde{\mu} h} - \mathcal{I} \Big) e^{-iH\frac{h}{2}} \tilde{\boldsymbol{\lambda}}_i \Big\rangle \\
&\quad - 2Re \sum_{i=0}^{N_t-1} \Big\langle \hat{\boldsymbol{\psi}}'_i, e^{iH\frac{h}{2}} \Big(e^{i(\epsilon'_i - \epsilon_i)\hat{\mu} h} - \mathcal{I} \Big) e^{-iH\frac{h}{2}} \hat{\boldsymbol{\lambda}}_i \Big\rangle - \beta h \sum_{i=0}^{N_t-1} (\epsilon'^2_i - \epsilon^2_i) \\
&\geq h \sum_{i=0}^{N_t-1} (\epsilon'_i - \epsilon_i) \Big(- 2Im \Big\langle \tilde{\boldsymbol{\psi}}'_i, \tilde{\mu}_h(\epsilon'_i, \epsilon_i) \tilde{\boldsymbol{\lambda}}_i \Big\rangle - 2Im \Big\langle \hat{\boldsymbol{\psi}}'_i, \hat{\mu}_h(\epsilon'_i, \epsilon_i) \hat{\boldsymbol{\lambda}}_i \Big\rangle - \beta(\epsilon'_i + \epsilon_i) \Big),
\end{aligned}$$

where the matrices $\tilde{\mu}_h(\epsilon'_i, \epsilon_i), \hat{\mu}_h(\epsilon'_i, \epsilon_i)$ are defined as

$$\begin{aligned}
\tilde{\mu}_h(\epsilon'_i, \epsilon_i) &= e^{iH\frac{h}{2}} \frac{e^{i(\epsilon'_i - \epsilon_i)\tilde{\mu} h} - \mathcal{I}}{ih(\epsilon'_i - \epsilon_i)} e^{-iH\frac{h}{2}} \\
\hat{\mu}_h(\epsilon'_i, \epsilon_i) &= e^{iH\frac{h}{2}} \frac{e^{i(\epsilon'_i - \epsilon_i)\hat{\mu} h} - \mathcal{I}}{ih(\epsilon'_i - \epsilon_i)} e^{-iH\frac{h}{2}}.
\end{aligned}$$

Hence a sufficient condition for $J_h(\epsilon') \geq J_h(\epsilon)$ is

$$(\epsilon'_i - \epsilon_i) \Big(- 2Im \Big\langle \tilde{\boldsymbol{\psi}}'_i, \tilde{\mu}_h(\epsilon'_i, \epsilon_i) \tilde{\boldsymbol{\lambda}}_i \Big\rangle - 2Im \Big\langle \hat{\boldsymbol{\psi}}'_i, \hat{\mu}_h(\epsilon'_i, \epsilon_i) \hat{\boldsymbol{\lambda}}_i \Big\rangle - \beta(\epsilon'_i + \epsilon_i) \Big) \geq 0, \quad i = 0, \ldots, N_t - 1.$$

Similarly to 90 we can now choose ϵ' as the solution of

$$\epsilon'_i - \epsilon_i = \frac{1}{\beta} \Big(- 2Im \Big\langle \tilde{\boldsymbol{\psi}}'_i, \tilde{\mu}_h(\epsilon'_i, \epsilon_i) \tilde{\boldsymbol{\lambda}}_i \Big\rangle - 2Im \Big\langle \hat{\boldsymbol{\psi}}'_i, \hat{\mu}_h(\epsilon'_i, \epsilon_i) \hat{\boldsymbol{\lambda}}_i \Big\rangle - \beta(\epsilon'_i + \epsilon_i) \Big), \quad i = 0, \ldots, N_t - 1,$$

leading to

$$J_h(\epsilon') - J_h(\epsilon) = \Big| \Big\langle \boldsymbol{\psi}_1, \Delta\boldsymbol{\psi}'_{N_t} - \Delta\boldsymbol{\psi}_{N_t} \Big\rangle \Big|^2 + \beta h \sum_{i=0}^{N_t-1} (\epsilon'_i - \epsilon_i)^2 \geq 0. \tag{92}$$

Remark 18

1. *Since we will use this approach in our numerical experiments, we give an overview on the proof of convergence in the next section.*
2. *One could also call this approach "Discretize-Before-Monotonize", since we discretize the system, before deriving a sufficient monotonicity condition.*

2.4.3 Convergence of the monotonic scheme for the discretized system

We follow the approach presented in [Sal07]. We start by showing that the sequence $\{\epsilon^k\}_{k\in\mathbb{N}}$, iteratively defined as the solution to

$$\epsilon_i^{k+1}-\epsilon_i^k = \frac{1}{\beta}\Big(-2Im\Big\langle\tilde{\boldsymbol{\psi}}_i^{k+1},\tilde{\mu}_h(\epsilon_i^{k+1},\epsilon_i^k)\tilde{\boldsymbol{\lambda}}_i^k\Big\rangle-2Im\Big\langle\hat{\boldsymbol{\psi}}_i^{k+1},\hat{\mu}_h(\epsilon_i^{k+1},\epsilon_i^k)\hat{\boldsymbol{\lambda}}_i^k\Big\rangle-\beta(\epsilon_i^{k+1}+\epsilon_i^k)\Big), \quad (93)$$

for $i = 0,\dots,N_t-1$, is bounded.

Lemma 9
Consider an initial field ϵ^0, $\left\|\epsilon^0\right\|_\infty > \infty$, where $\|\cdot\|_\infty$ denotes the essential supremum norm. Then the sequence $\{\epsilon^k\}_{k\in\mathbb{N}}$, iteratively defined as the solution to (93), is uniformly bounded.

For a proof compare [MST06, Theorem 3] (in the notation used there, we consider $\delta = 1$,$\eta = 0$ in our approach).

Now we can prove that the convergence of the function values $J_h(\epsilon^k)$.

Lemma 10
There exists $\widetilde{J}_{\epsilon^0} \geq 0$ such that

$$\lim_{k\to\infty} J_h(\epsilon^k) = \widetilde{J}_{\epsilon^0} \quad (94)$$

proof. From (92) we know that the sequence $\{J_h(\epsilon^k)\}_{k\in\mathbb{N}}$ is monotone non-decreasing and by Lemma 9 and (79) we obtain that the sequence is also bounded from above. This proves the existence of a $\widetilde{J}_{\epsilon^0}$ such that (94) holds. □

The next step is to show that each limit point of $\{\epsilon^k\}_{k\in\mathbb{N}}$ is also a critical point of J_h.

Lemma 11
Let S_{ϵ^0} be the set containing all limit points of $\{\epsilon^k\}_{k\in\mathbb{N}}$ and S_{crit} be the set containing all critical points of J_h (with gradient equal to zero). It holds that

$$S_{\epsilon^0} \subset S_{crit}$$

and also

$$d(\epsilon^k, S_{\epsilon^0}) \to 0,$$

where $d(\epsilon^k, S_{\epsilon^0})$ denotes the (2-norm) distance between ϵ^k and S_{ϵ^0}.

proof. For the first part of the statement we refer to [Sal07, Lemma 3.3]. The second one follows because $\{\epsilon^k\}_{k\in\mathbb{N}}$ is uniformly bounded by Lemma 9 and since if the statement was false, $\{\epsilon^k\}_{k\in\mathbb{N}}$ would have a (still uniformly bounded) subsequence that had no limit points, which is a contradiction. □

From the monotonicity of $\{J_h(\epsilon^k)\}_{k\in\mathbb{N}}$ it is also clear that

$$J_h(S_{\epsilon^0}) = \widetilde{J}_{\epsilon^0}.$$

Now it remains to show that the sequence $\{\epsilon^k\}_{k\in\mathbb{N}}$ itself is convergent.

Theorem 3
For $\beta > 0$ large enough, the sequence $\{\epsilon^k\}_{k\in\mathbb{N}}$ defined by (93) converges towards a critical point of J_h.

In [Sal07] this is proven with a few technical lemmas. Instead we will give a short sketch of the proof. The idea is to use the shifted cost functional

$$J_h^s(\epsilon) := \widetilde{J}_{\epsilon^0} - J_h(\epsilon)$$

and show that for the sequence $\{(J_h^s(\epsilon^k))^\theta\}_{k\in\mathbb{N}}$ ($\theta \in (0, \frac{1}{2}]$) there exists a $c > 0$ such that

$$(J_h^s(\epsilon^k))^\theta - (J_h^s(\epsilon^{k+1}))^\theta \geq c \left\|\epsilon^{k+1} - \epsilon^k\right\|. \tag{95}$$

This is done by using a special case of the Łojasiewicz inequality, where we utilize that for β large enough, the Hessian of J_h is diagonally dominant and therefore invertible (see [Sal07, Lemma 2.1]). Since $\{J_h(\epsilon^k)\}_{k\in\mathbb{N}}$ is monotonically increasing and bounded from above, $\{(J_h^s(\epsilon^k))^\theta\}_{k\in\mathbb{N}}$ is monotonically decreasing and bounded from below. Therefore $\{(J_h^s(\epsilon^k))^\theta\}_{k\in\mathbb{N}}$ is a Cauchy sequence and by (95) we conclude that $\{\epsilon^k\}_{k\in\mathbb{N}}$ is also a Cauchy sequence.

2.5 Numerical results

Before we show concrete numerical results, we will first discuss some practical implementation techniques to improve the performance of the algorithm.

2.5.1 Diagonalization and numerical solvers

In order to solve the initialization (56) and discriminatory step problem (58), we will use the monotonic scheme for the discrete system from Section 2.4.2. Therefore we need to solve equation (93) N_t-times in each iteration, meaning that we have to compute an exponential matrix of the form $e^{\pm i\epsilon_i\mu h}$ many times, for fixed matrices μ but with changing controls ϵ. Since all matrices μ that we are considering are Hermitian they can be diagonalized, meaning there exists a unitary matrix $P \in \mathbb{C}^{N\times N}$, containing the eigenvectors of μ, and a diagonal matrix $D \in \mathbb{R}^{N\times N}$ (eigenvalues of Hermitian matrices are real) such that $\mu = PDP^{-1}$. Therefore we get

$$e^{\pm i\epsilon_i\mu h} = e^{P(\pm ih\epsilon_i D)P^{-1}} = e^P e^{\pm ih\epsilon_i D} e^{P^{-1}} = P \begin{pmatrix} e^{\pm ih\epsilon_i D_{1,1}} & & 0 \\ & \ddots & \\ 0 & & e^{\pm ih\epsilon_i D_{N,N}} \end{pmatrix} P^{-1}.$$

This way we have to invest additional computational time into the pre-computation of P and D (for example with the built-in MATLAB function `eig`) once for every μ, but reduce the computation of an exponential matrix to the computation of N exponential values in each iteration, for each discrete time step.

The initial value for the monotonic schemes is chosen to be $\epsilon \equiv 0$. For the termination condition, we consider the difference between the new and the old cost function value ($J_h(\epsilon') - J_h(\epsilon)$). If the increase in function value is too small, we stop the monotonic scheme and return the last control ϵ'. We also have to be careful about the monotonicity of the cost function. Since equation (93) is non-linear, there is no numerical guarantee to get an exact solution and therefore a control that indeed yields a larger cost function value. Also the convergence result for our monotonic scheme is based on the assumption that β is larger than some constant that is depending on the matrices $\tilde{\mu}$ and $\hat{\mu}$, which themselves depend on varying $\boldsymbol{\alpha}$ given by the fitting step. Therefore we integrate a simple fail-safe, that stops the monotonic scheme whenever a decrease in function value is detected and then starts it again with a larger value for β, making the problem more concave.

Regarding the main identification (52) and fitting step problem (57) we note that these problems are non-linear in $\boldsymbol{\alpha}$, meaning that there can be many local minima. In our implementation we therefore do multiple searches from different starting points for both inverse problems, to have a better chance at finding the global minimum. Hence it is crucial to have a fast method rather than a precise one. The built-in function `fminunc` of MATLAB (that uses) has proven to be very efficient for this problem. However, in the future, one could also investigate other methods, that might work better with this kind of problems.

2.5.2 Selective laser fields in a three-dimensional example

Some results of this and the following sections are also published in [BCS20].

We start with the same setting as in [MS09], meaning that we consider a small, three-dimensional example ($N = 3$) with

$$H = 10^{-2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \quad \boldsymbol{\psi}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\psi}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{96}$$

$N_t = 500$ and $T = 4000\pi$, which corresponds to 20 periods of the smallest frequency of the system. We assume $\mu \in \mathbb{R}^{3\times 3}$ to be symmetric and consider two different choices of basis $\mathcal{B}_\mu$ for the space of 3×3 symmetric matrices, the canonical and a random one. We obtain the fields displayed in Figures 7 and 8.

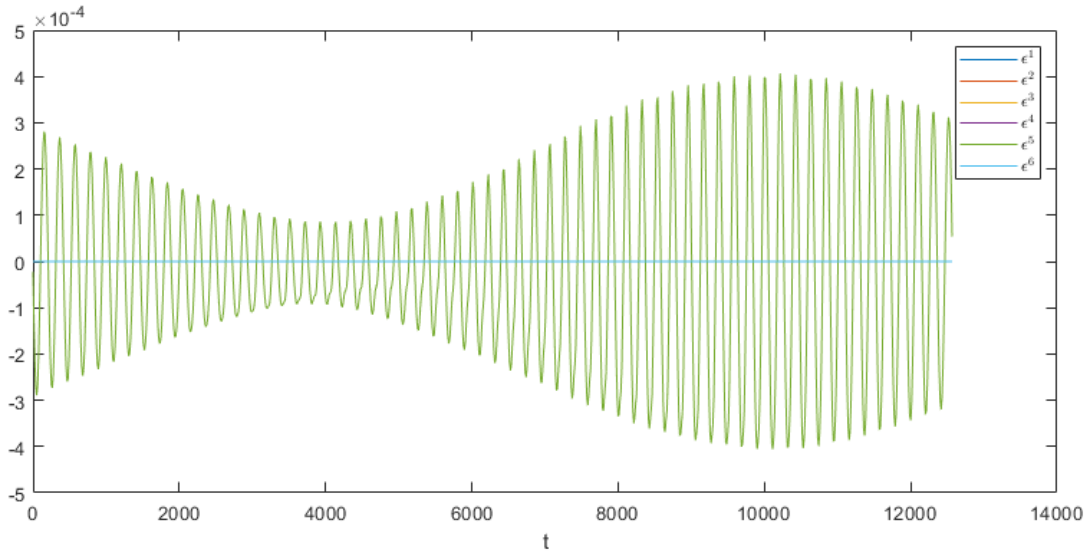


Figure 7 Laser fields for the canonical basis of 3×3 symmetric matrices.

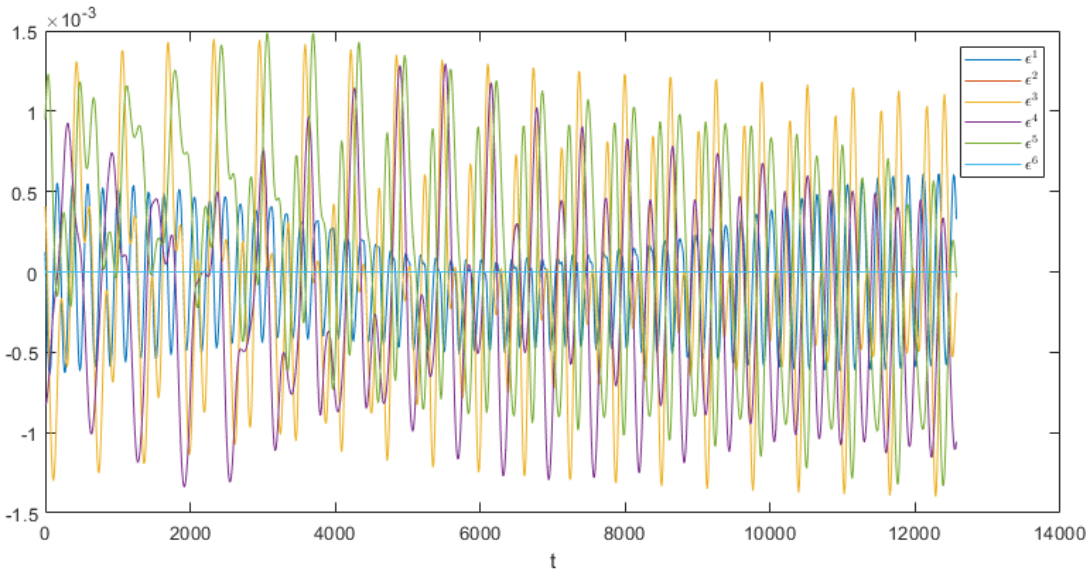


Figure 8 Laser fields for the random basis of 3×3 symmetric matrices.

We can already see from this example that the choice of the basis can be crucial also in the bilinear case. For the canonical basis, only one of the six obtained fields is non-zero, whereas for the random basis, all but two fields are non-zero. The result for the canonical basis also do not change if we use a non-zero starting value for the monotonic schemes. It is also not caused by a value of the regularization parameter β that has grown too large by our fail-safe that enlarges β if the monotonic scheme detects a decrease in function value. The values of β, for which the monotonic schemes terminated, are shown in Table 7 (the starting value is $\beta = 0.015$).

	init	disc 1	disc 2	disc 3	disc 4	disc 5
canonical basis	0.015	0.015	0.015	0.015	74.8	0.015
random basis	49.9	33.3	252.5	252.5	378.8	49.9

Table 7 Final values of the regularization parameter β for all monotonic schemes and the two different bases.

As we can see, the values of β are larger for the random bases than for the canonical basis. The controls that are zero do not even correspond to the largest value of β, meaning that the fields being equal to zero does not seem to have a correlation with a strong penalization of their norm.

However if we choose a different observer $\boldsymbol{\psi}_1 = [\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}]^T$, we get the fields for the canonical basis and the random basis shown in Figures 9 and 10 respectively.

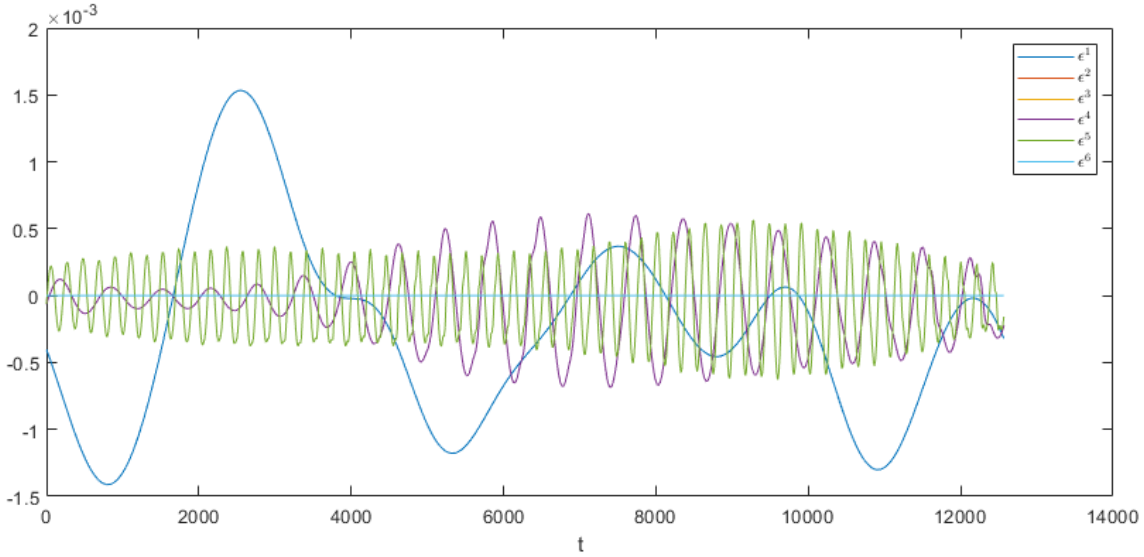


Figure 9 Laser fields for the canonical basis of 3×3 symmetric matrices.

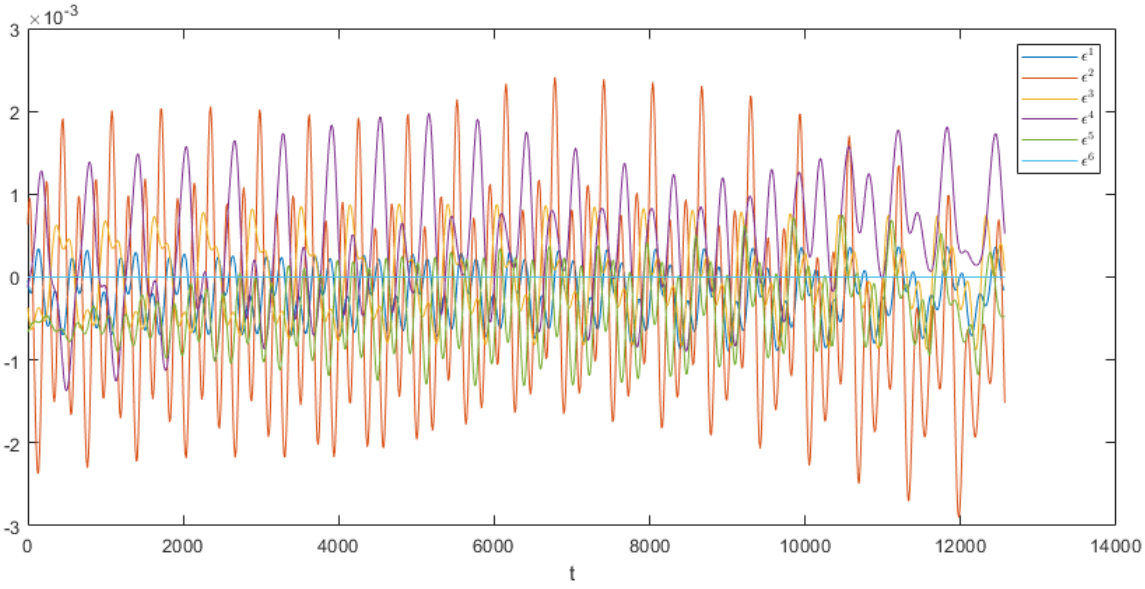


Figure 10 Laser fields for the random basis of 3×3 symmetric matrices.

We observe that now three fields are non-zero in the case of the canonical basis and five in case of the random basis.

2.5.3 Local and global minima of the inverse problems

As we mentioned before, both inverse problems, namely the main identification (52) and the fitting step problem (57), are non-linear in $\boldsymbol{\alpha}$, meaning that there can be many local minima. For the main identification the implications are clear, i.e. the final identification method does not converge to the true dipole operator $\mu^\star$. But there are also implications for the fitting step, i.e. we might not actually find a "defect of selectivity" and the resulting discriminatory step is trying to split observations, that can already be separated by the fields generated so far. To compensate for this effect, we solve all inverse problems multiple times (five times in our implementation) with different starting points. However there is still no guarantee that we are able to find the global minimum. To visualize this, we consider the following true dipole moment matrix $\mu^\star$ with random entries

$$\mu^\star = \begin{pmatrix} -0.0391 & -0.0216 & 0.0307 \\ -0.0216 & -0.0642 & 0.3091 \\ 0.0307 & 0.3091 & 0.3961 \end{pmatrix},$$

and the fields from Section 2.5.2 for the random basis, using the observable $\boldsymbol{\psi}_1 = [\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}]^T$. Now we solve the main identification (52) for 100 random initial vectors $\boldsymbol{\alpha}$ that lie in a 6-dimensional hypercube, with different radii, around the true coefficient vector $\boldsymbol{\alpha}^\star$ in the respective basis. We then count the number of times the minimization converges to $\boldsymbol{\alpha}^\star$, by which we mean that norm difference between the computed $\boldsymbol{\alpha}^{approx}$ and $\boldsymbol{\alpha}^\star$ is smaller than 10^{-2}. For the purpose of comparison we also do the same calculations for a set of trigonometric fields that were not generated by the algorithm (in this case $j \cdot 10^{-3} \sin(3jt + \frac{2\pi j}{6})$, for $j = 1, \ldots, 6$), using the random basis. We obtain the results shown in Table 8.

Radius	0.01	0.1	0.3	0.5	0.7
canonical basis	70	10	1	1	0
random basis	91	27	50	38	21
trigonometric fields	77	7	3	2	0

Table 8 Number of minimizations with random initialization that converge to the global minimum (out of 100).

We observe that there is a clear difference in the performance when it comes to the choice of the basis. Especially for large radii, the fields generated by the algorithm for the random basis (third row) converge much more often to the global minimum than for the canonical basis (second row). The small dent of 27 for the random basis and a radius of 0.1 is most likely due to statistical reasons and since we chose the tolerance, for a convergence to the global minimum, to be exactly 0.01. If we look more closely at the norm difference of computed solutions $\boldsymbol{\alpha}^{approx}$ and $\boldsymbol{\alpha}^\star$ in Table 11 we can see that in a lot of cases the norm difference is between 0.01 and 0.1. Note that the trigonometric fields are corresponding to the random basis, meaning that we can not make a meaningful comparison to the fields for the canonical basis without further computations. However compared to the fields generated by the algorithm for the random basis, the performance of the trigonometric fields is again worse, falling off significantly for larger radii. In this case we can also explain this difference in performance by looking at the graph of the cost functional for the main identification. In order to plot this graph, we use the random basis from our examples and fix four of the six coefficients of $\boldsymbol{\alpha}$ to be equal to their corresponding coefficient in $\boldsymbol{\alpha}^\star$. We then plot the graph of the cost function for the remaining two coefficients, for the trigonometric fields and the ones generated by the algorithm (Figure 10). The results are shown in Figures 11-13.

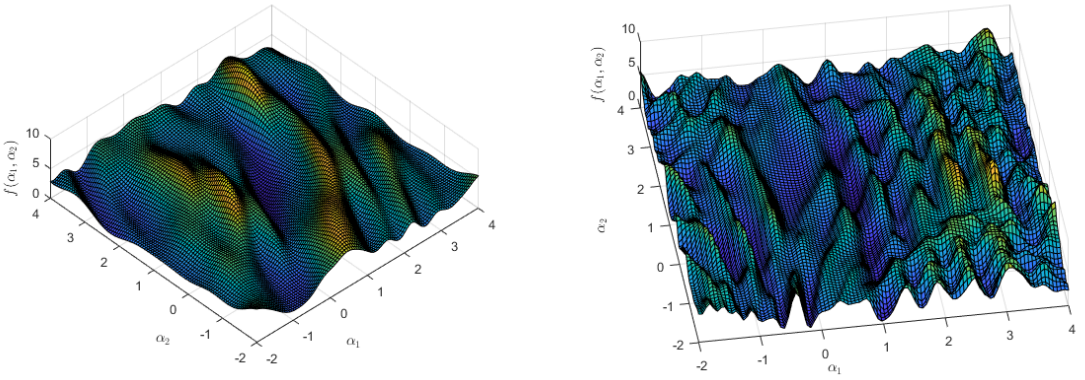


Figure 11 Graph of the cost function for the main identification, where $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^\star$ for $j \in \{3, \ldots, 6\}$, using the fields generated by the algorithm (left) and the trigonometric fields (right).

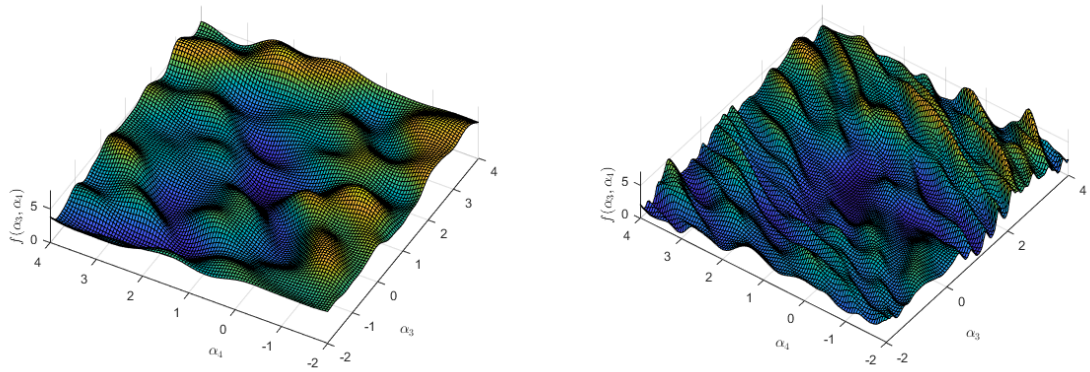


Figure 12 Graph of the cost function for the main identification, where $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^\star$ for $j \in \{1, 2, 5, 6\}$, using the fields generated by the algorithm (left) and the trigonometric fields (right).

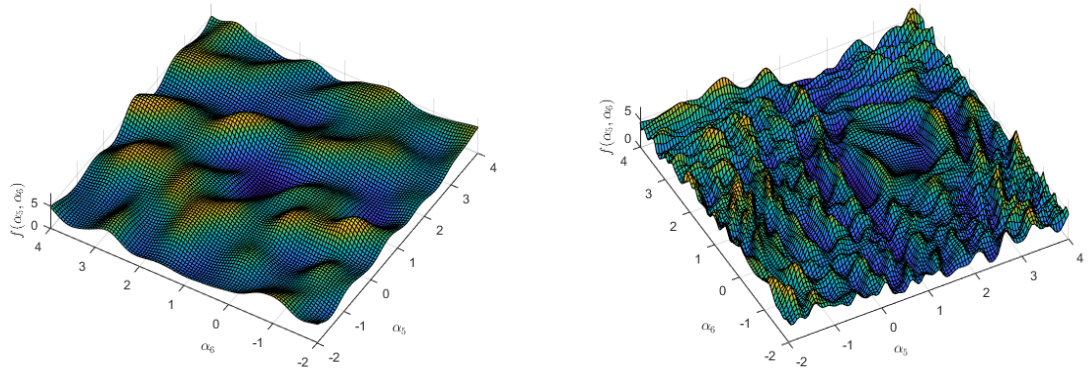


Figure 13 Graph of the cost function for the main identification, where $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^\star$ for $j \in \{1, \ldots, 4\}$, using the fields generated by the algorithm (left) and the trigonometric fields (right).

As we can see, the cost functional for the trigonometric fields has many more local minima and also the pit around the global minimum (in which one might expect the minimization method to converge to the global minimum) is much more narrow than for the fields generated by the algorithm. With these results one can justify investing computational time to generate the fields with the algorithm, but there is also the question whether there is a way of improving the algorithm to generate more robust fields.

2.5.4 Extended greedy-testing improvement

2.5.4.1 Motivation and statement of the improved algorithm

The results from last sections show that, as in the linear setting, the choice of basis and observer is crucial for the algorithm's ability to generate robust selective fields. Based on the extended greedy-testing strategy from the linear setting (compare Section 1.3.2.2) we want to introduce an improved version of the algorithm, that is able to select K "good" basis elements from a larger set.

Algorithm 6 Improved Greedy Reconstruction Algorithm for Selective Controls

Require: A set $\{\mu_1, \dots, \mu_L\}$, a tolerance $Tol > 0$.

Set $k = 1$ and solve the improved initialization problem

$$\max_{n\in\{1,\dots,L\}} \max_{\epsilon\in L^2(0,T)} |\varphi(\mu^n, \epsilon) - \varphi(0,0)|^2 - \beta \|\epsilon\|^2_{L^2(0,T)} \tag{97}$$

to get the field ϵ^1 and the index n_1.
Set $\mathcal{L} = \{n_1\}$ and $\mathcal{J} = \{1, \dots, L\} \setminus \{n_1\}$.
while $k \leq K - 1$ **do**
  Improved fitting step:
  Set $\mathcal{I} = \emptyset$, $f_{min} = 10^4$ and $i_{min} = 0$.
  for $i \in \mathcal{J}$ **do**
    if $\mu_i \in \text{span}(\{\mu_\ell\}_{\ell\in\mathcal{L}})$ **then**
      Set $\mathcal{J} = \mathcal{J} \setminus \{i\}$
    else
      Find $(\boldsymbol{\alpha}^i_j)_{j=1,\dots,k}$ that solves the problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^k} \sum_{m=1}^{k} |\varphi(\mu_i, \epsilon^m) - \varphi(\sum_{j=1}^{k} \boldsymbol{\alpha}_j \mu_j, \epsilon^m)|^2 \tag{98}$$

      with function value $f_i(\boldsymbol{\alpha}^i)$ in the minimum.
      if $f_i(\boldsymbol{\alpha}^i) < Tol$ **then**
        $\mathcal{I} = \mathcal{I} \cup \{i\}$.
      end if
      if $f_i(\boldsymbol{\alpha}^i) < f_{min}$ **then**
        Set $i_{min} = i$ and $f_{min} = f_i(\boldsymbol{\alpha}^i)$.
      end if
    end if
  end for
  if $\mathcal{I} = \emptyset$ **then**
    Set $\mathcal{I} = \{i_{min}\}$.
  end if
  Improved discriminatory step:
  Find ϵ^k and n_k that solve the problem:

$$\max_{n\in\mathcal{I}} \max_{\epsilon\in L^2(0,T)} |\varphi(\mu^n, \epsilon) - \varphi(\sum_{j=1}^{k-1} \boldsymbol{\alpha}^n_j \mu^j, \epsilon)|^2 - \beta \|\epsilon\|^2_{L^2(0,T)} .$$

  Set $\mathcal{L} = \mathcal{L} \cup \{n_k\}$, $\mathcal{J} = \{1, \dots, K\} \setminus \{n_k\}$.
end while

The main idea behind this improved algorithm is, to provide L different elements ($L > K$) from the space that $\mu^\star$ lies in, some of which will inevitably be linearly dependent. The improved algorithm then picks K linearly independent elements that yield the largest function value in the (improved) discriminatory step. In other words it picks the K most selective elements.

The algorithm starts with an improved initialization, where we maximize over all L elements to get the first field. In each of the following $K-1$ iterations, we test all remaining elements whether they are linearly dependent on the ones chosen so far. For all that are not, we do a fitting step and check whether the cost function value is small enough (whether we really detected a defect of selectivity). Remember that by "doing a fitting step", we still mean that we solve the non-linear inverse problem (98) for a few (in our case 5) random starting points to have a chance at finding the global minimum. For the elements that return a small enough function value, we then do a discriminatory step and select the new field as the one yielding the largest function value in the discriminatory step (the one that compensates the corresponding defect the most). If none of the remaining elements yield a small enough function value in the fitting step, we consider only the one (denoted by i_{min}) with the smallest function value in the (improved) discriminatory step.

Remark 19

1. *The improved algorithm does not only return the fields $\epsilon^1, \dots, \epsilon^K$ but also an improved basis $\tilde{\mathcal{B}} = \{\mu_1, \dots, \mu_K\}$.*
2. *As we mentioned in the linear setting, this improvement does not necessarily increase the computational time, since all additional steps can be parallelized.*

2.5.4.2 Numerical experiments

We consider the setting from the last sections with the observable $\boldsymbol{\psi}_1 = [\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}]^T$. We provide the improved algorithm with the set $\{\mu_1, \dots, \mu_{12}\}$, where the first six elements are the canonical basis and the last six are the random basis. The fields returned by the improved algorithm are shown in Figure 14.

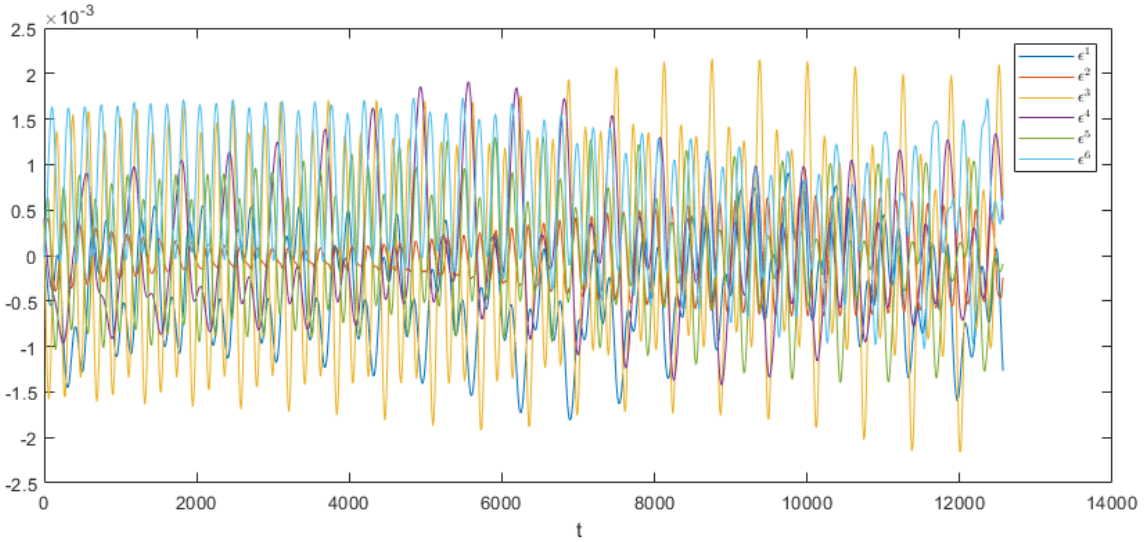


Figure 14 Laser fields generated by the improved algorithm.

We observe that all six improved fields are non-zero, meaning that this set of fields is richer than all other sets we have seen. Therefore the identification has access to more data to compare

the effects of different coefficient combinations of $\boldsymbol{\alpha}$.

To compare the robustness of these fields to the other ones we considered so far, we again do 100 random initial minimizations for the main identification in the same 6-dimensional hypercubes around the corresponding $\boldsymbol{\alpha}^\star$ as before and count how computed solutions are (in norm) closer than 0.01 to the true solution $\boldsymbol{\alpha}^\star$. We obtain the results in Table 9.

Radius	0.01	0.1	0.3	0.5	0.7
canonical basis	70	10	1	1	0
random basis	91	27	50	38	21
improved basis	100	100	100	97	65

Table 9 Number of minimizations with random initialization that converge to the global minimum (out of 100).

As we can see, the fields of the improved algorithm are much more robust, converging to the global minimum for all initial values for a radius up to 0.3. Also for larger radii the performance for the improved fields is much better than for the other fields. In this setting it is also worth taking a closer look at the norm difference between the computed solutions and the true solution $\boldsymbol{\alpha}^\star$. Therefore we compare in Table 10, 11 and 12 the number of computed solutions that is (in norm) closer to the global minimum than a given tolerance.

Tolerance	10^{-4}	10^{-3}	0.01	0.1	1
Rad. 0.01	0	0	70	100	100
Rad. 0.1	0	2	10	97	100
Rad. 0.3	0	0	1	30	89
Rad. 0.5	0	1	1	12	46
Rad. 0.7	0	0	0	3	21

Table 10 Minimizations for the canonical basis that converge close enough to the global minimum (out of 100).

Tolerance	10^{-4}	10^{-3}	0.01	0.1	1
Rad. 0.01	0	10	91	100	100
Rad. 0.1	0	7	27	97	100
Rad. 0.3	0	15	50	89	96
Rad. 0.5	0	1	38	55	61
Rad. 0.7	0	6	21	25	28

Table 11 Minimizations for the random basis that converge close enough to the global minimum (out of 100).

Tolerance	10^{-4}	10^{-3}	0.01	0.1	1
Rad. 0.01	89	100	100	100	100
Rad. 0.1	66	99	100	100	100
Rad. 0.3	71	100	100	100	100
Rad. 0.5	74	97	97	97	98
Rad. 0.7	56	65	65	65	74

Table 12 Minimizations for the improved basis that converge close enough to the global minimum (out of 100).

These tables clearly show that, with the fields generated by the improved algorithm, the main identification is able to converge much more often and also much closer to the global minimum than with the other fields. None of the random starting points converge within a tolerance of 10^{-4} for the canonical and the random basis, while, even for the largest radius of 0.7, over 50 percent of the random starting points converge within this tolerance for the improved basis. Also under 30 percent of the random starting points even converge within a tolerance of 1 for the other bases. One possible interpretation would be that the improved fields and basis make the main identification problem, in some sence, more smooth but also more convex, meaning that it is able to globally converge from more distant initial point but can also locally converge closer

to the true solution. However to verify this statement one would have to further investigate the properties of the algorithm in the bilinear setting.

Looking at the elements of the improved basis, we notice that it consists of exactly three elements from both, the canonical and the random basis. It seems that the improved algorithm was able to select the "good" basis elements from both bases and to thereby generate much more robust fields than the standard algorithm for any of the two bases alone.

Conclusion and outlook

In Chapter 1 we have outlined the way the algorithm works in the setting of a linear ordinary differential equation. We have also developed a new convergence theory and shown how the observability of the system is a limiting factor for the ability to identify the unknown operator. We also developed an improved choice for a basis of the full matrix space such that we can recover the most amount of information about the true unknown matrix. For this improved basis choice we also presented a corresponding convergence theory. In addition, we proposed two other working improvements, one of which was able to, in some sense, minimize the number of controls being generated by the algorithm, while keeping the same level of identifiability. All improvements were then supported by numerical experiments, confirming their capabilities.
One point of future works could be a further investigation of cases, where restrictions are imposed onto the basis $\mathcal{B}$. In this context one could discuss whether it is still possible to provide an improved basis similar to the one we proposed in Assumption 4. There are also still numerical cases where one can observe instabilities that can not be explained with a discrepancy in eigenvalues of the observability matrix.

In Chapter 2 we described the algorithm in its original setting of Schrödinger-type equations. We introduced monotonic schemes in order to solve quantum control problems and briefly discussed their convergence. Finally we showed that the choice of the basis and the observer is also crucial in the bilinear setting. In this context we presented an improved algorithm, based on one of the improvements developed in Chapter 1. In a numerical example the improvement showed great promise regarding the selectivity of the generated fields.
In the future one could investigate the observability of these bilinear systems and try to find criteria for a good basis choice. One possible idea would be to consider, locally, the linearized version of the system and try to use some of the results we presented in Chapter 1. The choice of the procedure for the identification could also be further investigated, there might exist methods that converge to the global minimum, even if the initial guess is further away.

Since the idea of the algorithm is independent of the underlying system structure, it could also be applied to more sophisticated operator reconstruction settings. Another interesting point could be the development of a "hybrid" method that uses the fields generated by the algorithm, but is also able to adaptively generate new fields "on the fly", i.e. during the laboratory experiments.

Bibliography

[Atk11] A. C. Atkinson. *Optimum Experimental Design*, pages 1037–1039. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[BCS17] A. Borzì, G. Ciaramella, and M. Sprengel. *Formulation and Numerical Solution of Quantum Control Problems.* Number 16 in Computational science & engineering. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, 2017.

[BCS20] S. Buchwald, G. Ciaramella, and J. Salomon. An optimized greedy reconstruction algorithm for dipole momentum operators. Technical report, 2020.

[BMR09] S. Bonnabel, M. Mirrahimi, and P. Rouchon. Observer-based Hamiltonian identification for quantum systems. *Automatica*, 45(5):1144 – 1155, 2009.

[Cia20] G. Ciaramella. *Optimal Control of Ordinary Differential Equations*, 2020.

[DR14] A. Donovan and H. Rabitz. Exploring the Hamiltonian inversion landscape. *Phys. Chem.*, 16:15615–15622, 2014.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix Analysis, Second Edition.* Cambridge University Press, USA, 2013.

[Kel99] C. T. Kelley. Iterative methods for optimization. *Siam*, 18, 1999.

[LBMRT07] C. Le Bris, M. Mirrahimi, H. Rabitz, and G. Turinici. Hamiltonian identification for quantum systems: well-posedness and numerical approaches. *ESAIM: Control, Optimisation and Calculus of Variations*, 13(2):378–395, 5 2007.

[MS09] Y. Maday and J. Salomon. A greedy algorithm for the identification of quantum systems. In *PROCEEDINGS OF THE 48TH IEEE CONFERENCE ON DECISION AND CONTROL, 2009 HELD JOINTLY WITH THE 2009 28TH CHINESE CONTROL CONFERENCE (CDC/CCC 2009)*, IEEE Conference on Decision and Control, pages 375–379. IEEE; Honeywell; Quanser; United Technologies; Googol Tech; MathWorks; Natl Instruments, 2009. Joint 48th IEEE Conference on Decision and Control (CDC) / 28th Chinese Control Conference (CCC), Shanghai, PEOPLES R CHINA, DEC 15-18, 2009.

[MST06] Y. Maday, J. Salomon, and G. Turinici. Monotonic time-discretized schemes in quantum control. *NUMERISCHE MATHEMATIK*, 103(2):323–338, APR 2006.

[Rud87] W. Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA, international edition, 1987.

[Sal07] J. Salomon. Convergence of the time-discretized monotonic schemes. *ESAIM-MATHEMATICAL MODELLING AND NUMERICAL ANALYSIS-MODELISATION MATHEMATIQUE ET ANALYSE NUMERIQUE*, 41(1):77–93, JAN-FEB 2007.

[Son98] E. Sontag. *Mathematical Control Theory: Deterministic Finite-Dimensional Systems.* 01 1998.

[Str68] G. Strang. On the Construction and Comparison of Difference Schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, September 1968.

[TKO92] D. Tannor, V. Kazakov, and V. Orlov. Control of photochemical branching: Novel

procedures for finding optimal pulses and global upper bounds. -1:347, 01 1992.

[WDQ+18] Y. Wang, D. Dong, B. Qi, J. Zhang, I. R. Petersen, and H. Yonezawa. A quantum Hamiltonian identification algorithm: Computational complexity and error analysis. *IEEE Trans. Autom. Control*, 63(5):1388–1403, 2018.

[XWLJ19] S. Xue, R. Wu, D. Li, and M. Jiang. A gradient algorithm for Hamiltonian identification of open quantum systems. *arXiv preprint - arXiv:1905.09990*, 2019.

[ZR98] W. Zhu and H. A. Rabitz. A rapid monotonically convergent iteration algorithm for quantum optimal control over the expectation value of a positive definite operator. *Journal of Chemical Physics*, 109(2):385–391, December 1998.

[ZS14] J. Zhang and M. Sarovar. Quantum Hamiltonian identification from measurement time traces. *Phys. Rev. Lett.*, 113:080401, 2014.

[ZSG+12] W.W. Zhou, S. Schirmer, E. Gong, H. Xie, and M. Zhang. Identification of markovian open system dynamics for qubit systems. *Chinese Sci. Bull.*, 57(18):2242–2246, 2012.