Efficient Data Collection Strategies for Rapid Learning in Physical Environments

by

Gal Shulkind

B.A. and B.Sc., Technion Israel Institute of Technology (2006)
M.Sc., Technion Israel Institute of Technology (2012)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author .....
Department of Electrical Engineering and Computer Science
March 9, 2018

Certified by .....
Gregory W. Wornell
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Efficient Data Collection Strategies for Rapid Learning in Physical Environments

by

Gal Shulkind

Submitted to the Department of Electrical Engineering and Computer Science
on March 9, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

With the ubiquity of intelligent systems capable of sensing, inferring and acting upon their surroundings, it becomes critical to learn rapidly about unknown systems or environments. However, obtaining empirical data is often costly and involves setting up time consuming experiments or deploying specialized sensors. We are interested in deriving scalable algorithms and system architectures that facilitate efficient data collection, maximizing inference quality under limited resource budget.

In this work, we consider efficient data collection strategies in several applications involving physical environments. We study the problem of learning dynamical systems with initial approximated models, where we prescribe methods for choosing near optimal experimental parameters to collect empirical data. We study the problem of antenna array topology design where we prescribe configurations allowing efficient scene inference under various measurement schemes and budget constraints. We introduce a novel nonlinear radar modality and discuss efficient design techniques for this setting. Finally, we introduce a novel methodology for optical imaging of non line of sight hidden scenes by utilizing occlusions and investigate how to achieve efficient illumination of the scene for fast hidden target interrogation.

Thesis Supervisor: Gregory W. Wornell
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Reflecting back on my years at MIT I am incredibly thankful for all that I have experienced and humbled by how much I have grown. I am grateful to have been given this opportunity and am eager now to continue my journey, as hopeful as ever that I will be able to continue working to make a positive impact on the world around me.

My growth at MIT would not have been possible without the help and support I have received from many who have shaped my experience and made my time here special and I wish to take this opportunity to thank them.

I would like to offer my sincerest gratitude to my thesis advisor, Prof. Gregory Wornell for his guidance and mentorship. Over the course of the last five years, over countless meetings, conversations and discussions, he has, little by little, shaped my image as a researcher. His quest for deep understanding of fundamental principles and his crystal clear presentation and communication style have inspired me and will remain a gold standard for me for the remainder of my career.

I wish to thank the other members of my thesis committee, Prof. Stefanie Jegelka and Prof. Pablo Parrilo. Prof. Jegelka's class on submodular optimization inspired me to think about the problems and ideas that lie at the core of this thesis. Her advice and help were key in the shaping of this work. Prof. Parrilo, who taught me optimization and participated in the early work on non-line-of-sight imaging influenced my work with his ideas and suggestions.

I want to thank my collaborators who taught me a great deal about research and have helped me to develop some of the ideas I explore in this thesis. Dr. Lior Horesh hosted me for a summer internship at IBM research where I also had the great pleasure of working with Prof. Haim Avron. Dr. Milutin Pajovic and Dr. Philip Orlik hosted me for a summer internship at Mitsubishi Electric Research Laboratories, and Dr. Jeffrey Paulsen and Dr. Lalitha Venkataramanan hosted me for a summer internship at Schlumberger-Doll Research Center. These experiences helped me to understand how research is applied in industry and have made me a better scientist and engineer. I also wish to thank Prof. Yuval Kochman for multiple discussions about radar and Drs. Christos Thrampoulidis and Feihu Xu with

whom I have had the great pleasure of developing and exploring ideas in non-line-of-sight imaging.

As a member of the Signals, Information and Algorithms laboratory at the Research Laboratory of Electronics at MIT I have had the pleasure of interacting with multiple friends and colleagues who have enriched my life with ideas, stories and experiences: James Krieger, Da Wang, Ying-Zong Huang, Ligong Wang, Qing He, Gauri Joshi, Atulya Yellepeddi, Hongchao Zhou, David Romero, Lisa Zhang, Ganesh Ajjanagadde, Joshua Lee, Christos Thrampoulidis, Ankit Singh Rawat, Adam Yedidia, Amichai Painsky, Toros Arikan, and Gary Lee, as well as many other guests and visitors. A special thank you also goes to Tricia O'Donnell for always being there to help, support and offer kind words and interesting stories.

Most importantly, I want to express my gratitude to my family in Israel and here in the United States. To my mother, who always believed in me and who taught me how to tell a compelling story, and to my father, a pillar of solid rock in my life whose limitless love and support gave me the confidence to leave my home country and explore. And to my wife, Marilyn, for accompanying me in this journey from close by and from far away, and for helping me to realize my dreams.

In memory of Yechiel Shwartz and Shmuel Shulkind.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In a wide variety of applications we are interested in utilizing data in the form of collected measurements to make inference about an object of interest whose properties we want to learn. Typically, upon modeling the various components of the environment, such as the system under study, the measurement process, and the noise generation mechanism we derive algorithms and formulas for estimating unknown feature values given empirical measurements. In many such learning setups it is often assumed that data is a-priori available and emphasis is placed on analyzing different modeling aspects of the problem and designing efficient computational methods for performing inference. However, obtaining data is often a major concern in real world applications, where collection of a single data point may involve setting up an experiment or placing a dedicated sensor such that it can be costly in terms of time and other resources. Deriving efficient and deliberate data collection strategies, supporting desired performance levels while minimizing experimental resources, is thus important for lowering costs and allowing economical learning. Furthermore, in settings where rapid learning is crucial, such as when the environment is evolving with time, or when decisions have to be taken quickly in response to changes in the state of the system, efficient data collection schemes minimize the acquisition time and allow timely estimation of the system state, which is crucial for enabling on-time response. Intuitively, the efficacy of such data collection strategies is tied to the accuracy and certainty of the model describing the

system under study. When nothing is a-priori known there is little we can do in the way of devising efficient a-priori learning strategies, whereas when the model describing the system at play is highly specified and accurate our hope is to be able to do much better.

The challenging problem of designing efficient data collection strategies for learning has been studied by different communities under different titles such as experimental design, sampling and active learning, and it is notoriously difficult and rich with computationally intractable formulations that are known to be out of reach for efficient solvers. It is only in the past decade that new formulations have emerged for some of these problems, facilitating approximately optimal solutions that can be efficiently computed in some situations. However, these may still be hard to compute and they are not tailored to specific applications where specialized modeling issues may call for a detailed study and adapted solutions.

In this work, we identify several physical environments and settings where obtaining data efficiently is of key importance in determining overall system performance. In these settings an efficient data collection strategy may have far reaching applications in reducing, e.g. system latency, cost, weight and size. Each setup we consider illuminates a unique challenge and poses domain specific trade-offs which we take into account. Ultimately, our goal is to tailor solutions for these application domains and offer novel formulations that enable more economical design and improved performance.

1.2 Our Research Goals and Focus

In this work we focus on several domains of interest:

- We discuss the problem of collecting samples from a Gaussian process where we extend recently proposed sampling formulations and suggest alternatives that are more computationally efficient and lead to similar performance.

- We formulate the problem of learning misspecified dynamical models, discuss some of its features and derive an approximately optimal strategy for choosing experimental parameters for rapid learning.

- We explore the problem of antenna array design for single and multiple wavelength sce-

narios, introduce several measurement collection schemes and derive efficient approximately optimal solvers for array topology design under various budget constraints.

- We introduce a novel radar modality, where the interaction between the probing field and the targets is nonlinear and explore efficient signal set and antenna array topology design in this setting.

- We discuss the problem of non-line-of-sight optical imaging where state of the art systems require collection of ultra fast sub-picosecond measurements. We introduce a novel imaging modality based on exploiting occluders in the scene and study efficient scene interrogation strategies for fast image acquisition.

1.3 Thesis Overview and Structure

In Chapter 2 we revisit the problem of sensor placement in a Gaussian Process (GP). A GP is a mathematical tool that can be used to capture statistical assumptions about a wide range of physical and mathematical phenomena, such as temperature variations in a room or sample values of a function, and perform inference over unknown quantities given observed measurements. An efficient sensor placement configuration in a GP is one enabling high quality inference over unobserved locations with a fixed number of measurements. The problem of designing near optimal sensor placement configurations in GPs has previously been addressed, but has proven to be computationally challenging, especially in high dimensional settings. We introduce a novel approach towards efficient sensor placement in such settings, and show that near optimal solutions may be efficiently obtained.

We then turn to address the problem of learning mis-specified dynamical systems. Dynamical models are ubiquitous in describing natural and man made phenomena. It is often the case where a model describing such system is mis-specified, e.g. it has been derived using specialized domain knowledge that does not completely capture the exact dynamics, for example neglecting to account for weak non-idealities that exist in every real world environment but are hard to predict theoretically. In situations such as this we can augment the approximated model and enhance its accuracy by taking advantage of empirical data col-

lected from the system. We deploy our sensor placement technique to address the question of designing efficient experiments for empirical data collection in this setup. This setting allows us to probe the interplay between domain knowledge and empirical data and study how the former can direct efficient use of the latter, as we show by deriving theoretical bounds and demonstrate via numerical experiments.

In Chapter 3 we turn our attention to designing antenna arrays for far field sensing. Traditionally, antenna arrays have been designed to meet desired beam pattern specifications such as main lobe width and sideband suppression levels which implicitly support fixed reconstruction fidelity thresholds with respect to a set of scenes of interest. We introduce a Bayesian setting where prior information on scenes, e.g. in the form of smoothness assumptions or statistics of expansion coefficients in some countable base, is known, and propose an array geometry design scheme that facilitates efficient collection of measurements under these assumptions. We show that our method can computationally efficiently achieve near optimal designs. Moreover, by adapting relevant results from the growing body of literature on submodular optimization we demonstrate how to design antenna arrays under combinatorial placement constraints, and in multiple wavelength settings where we find specialized designs for measurement fusion and robust operation.

In Chapter 4 we study the problem of optically imaging non line of sight (NLOS) hidden scenes through diffuse optical reflections. State of the art systems for NLOS imaging utilize Time of Flight (TOF) measurements that are collected with ultra fast sub-picosecond detectors, implying high cost and complicated setups. In contrast, we introduce a novel approach for NLOS imaging by opportunistically exploiting occlusions in the scene. We show that in many interesting settings our method obviates the need for collecting costly time resolved information. In this setting, to achieve low acquisition times it is crucial to design efficient scene interrogation strategies. We utilize our experimental design methodology developed in previous chapters to address this problem. Our resulting imaging system represents an instance of a rich and promising new imaging modality with important potential implications for imaging science.

In Chapter 5 we further specialize our treatment for antenna array design. Motivated by commercial imaging applications where the interaction mechanism between the scene and the

probing field is nonlinear we consider novel imaging models and study the role this interaction plays in carrying information about the scene. Focusing on the classic directions of arrival (DOA) estimation problem we derive theoretical results that demonstrate how power-law type interactions can lead to enhanced target identifiability with a fixed number of sensors, and suggest corresponding array topologies and signal sets that support this enhanced level of performance.

1.4 Bibliographical Notes

A preliminary version of Chapter 2 appears in:

- Gal Shulkind, Lior Horesh, and Haim Avron. Experimental design for non-parametric correction of misspecified dynamical models. *arXiv:1705.00956*, 2017.

A preliminary version of Chapter 3 appears in:

- Gal Shulkind, Stefanie Jegelka, and Gregory W Wornell. Sensor array design through submodular optimization. *arXiv:1705.06616*, 2017.

- Gal Shulkind, Stefanie Jegelka, and Gregory W Wornell. Multiple wavelength sensing array design. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 3424–3428. IEEE, 2017.

A preliminary version of Chapter 4 appears in:

- Feihu Xu, Gal Shulkind, Christos Thrampoulidis, Jeffrey H Shapiro, Antonio Torralba, Franco N C Wong, and Gregory W Wornell. Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *arXiv:1802.03529*, 2018. First three authors contributed equally.

- Christos Thrampoulidis, Gal Shulkind, Feihu Xu, William T Freeman, Jeffrey H Shapiro, Antonio Torralba, Franco NC Wong, and Gregory W Wornell. Exploiting occlusion in Non-Line-of-Sight active imaging. *arXiv:1711.06297*, 2018. First two authors contributed equally.

A preliminary version of Chapter 5 appears in:

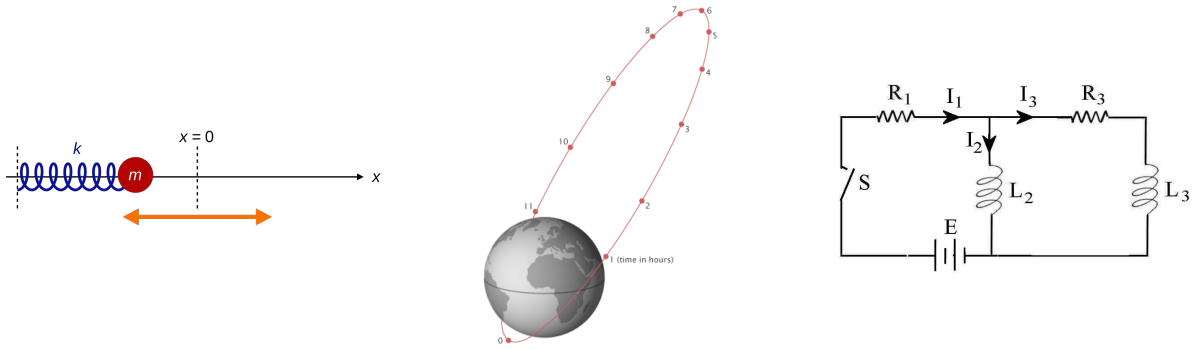
- Gal Shulkind, Gregory W Wornell, and Yuval Kochman. Direction of arrival estimation in MIMO radar systems with nonlinear reflectors. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 3016–3020. IEEE, 2016.

Chapter 2

Experimental Design for Learning Misspecified Dynamical Systems

Dynamical models are ubiquitous in describing natural and man-made phenomena, such as particle motion in a force field, liquid and gas flow, acoustic wave propagation and electrical signals in electronic circuits. An accurate dynamical model describing a system under study is useful for making predictions, for example about the future value of state variables given their initial conditions at some fixed time. Unfortunately, in many situations of interest we do not have access to an accurate dynamical model describing a system, and instead we only have available a crude approximated model. We will be interested in estimating an accurate model for the system by augmenting the misspecified model with a correction term learned from empirical data measurements, which are to be gathered by experimenting with the system in question.

Experimenting with real world systems is often time-consuming and costly such that efficiently designing an informative series of such experiments is important. We study the problem of designing experiments for rapidly and efficiently learning to correct such misspecified dynamical models using empirical data measurements. In this chapter, we formulate a scheme for representing such problems by utilizing Gaussian Processes and show, by using the theory of submodular functions, that while the problem of optimal experimental design is NP-hard, near optimal (up to a constant) designs are achievable through use of computationally efficient solvers.



Physical environments where differential equations are used to model state space evolution and approximate models can be derived from fundamental physical principles: (Left) Mass attached to a spring (Middle) Satellite in orbit around a planet (Right) Electrical circuit. [free use figures obtained online]

Figure 2-1: Illustration of several dynamical systems.

2.1 Introduction

The evolution of a wide variety of dynamical systems can be described by mathematical models which embody differential equations [34]. Such dynamical models are employed ubiquitously for description, prediction, and decision-making in a wide variety of environments and applications, e.g. as illustrated in Figure 2-1.

Acknowledging that "essentially all models are wrong" [10], we recognize that an exact description of a physical system is almost never within reach, and all models are misspecified to some extent, that is, their accuracy is limited. One such scenario is in situations where domain knowledge, available via an expert familiar with the setup, is employed to derive a crude low-complexity initial description for the dynamics of a system of interest [88, 43, 37, 117, 64, 105, 98]. However, real-world phenomena often involve additional, weak effects, that are not accounted for in typical expert derived models, rendering such models inadequate representations of reality. For example, in designing electrical circuits, ideal linear models are often assumed for circuit components such as resistors, capacitors and inductors, however, available components tend to exhibit weak but complicated non-linear characteristics not accounted for by the approximate models [18, 73, 24]. Another example is in deriving models for flow systems, where idealized models may be assumed for the medium and its boundaries, neglecting weak nonlinear phenomena and deteriorating the fidelity of

the resulting models [29, 116]. More broadly, diverse physical phenomena such as friction, nonlinearities, saturation, breakdown and many other non idealities are often not taken into account when formulating crude models.

Other common sources of model misspecification may be related to simplified representation of the domain geometry [109], isotropic modeling of anisotropic medium [107, 122, 1] and conscious or non-conscious choices made regarding the numerical solution of the underlying system: immature truncation of infinite expansions, round-off errors, approximate solutions of linear or non-linear terms, etc. [45, 48, 123, 111].

Such prominent modeling errors may result in discrepancies between predicted and observed system behavior and creep into simulation-based insights, ramifications of which could be inaccurate state descriptions, unstable model inferences, or erroneous control output, designs or decisions.

In lieu of deriving an approximate model which may confer an inadequate representation of the system's dynamic, a common alternative is to take a completely agnostic, data-driven approach and apply either parametric or non-parametric techniques to learn the dynamics purely based on empirical data collected from the system [40, 74, 88, 74]. However, such an agnostic approach fails to utilize the existing approximate information about the system model, often lead to models of limited interpretability, and usually relies on the availability of a large set of training examples to derive complex models of sufficient fidelity [47, 65].

In this study we explore a third approach of combining these two information sources effectively: on the one hand a crude misspecified system model as derived based on domain knowledge, and on the other hand empirical measurements and data to complement the misspecified model. Our goal is to learn a generalized representation for the system dynamics, based on the approximate model and the empirical data. We focus on understanding how this learning process can be performed efficiently, with only a limited budget for experiments to probe the system and collect empirical data points [38, 46, 106]. Specifically, we explore the role of the initial approximate model in guiding the design of experiments for collection of empirical data that best informs the model correction objective.

We are interested in exploring efficient methods for modeling such mis-specified dynamical systems, and analyzing how empirical measurements can enhance our models. We focus on a

non-parametric approach for modeling the unknown components of the system dynamics [49] that requires weak, implicit assumptions regarding the desired correction and its smoothness. These formulations offer great versatility in defining non-linear functional representations [4, 76, 7] and are therefore applicable to a broad class of problems. In this study, we take the approach proposed by Kennedy and O'Hagen [49] and articulate the misspecified correction term as a Gaussian Processes (GP).

With the GP formulation set we have a natural framework for concisely recording a correction term for the system dynamics and a matching Bayesian setting for making inference from empirical measurements from the system output and the correction term. Namely, we assume that the system can be prepared at one of multiple initial conditions, upon which it is allowed to evolve freely and we have access to noisy output measurements. The discrepancy between the actual system output and the output predicted based on the misspecified model drives the inference process and updates estimates for the correction term.

As setting up experiments with physical systems is often costly in term of both time and other resources, one of our main goals will be to devise efficient strategies for data collection in this setup, that facilitate rapid learning. Our end-goal is to accelerate the learning curve, and infer the dynamic correction term with a minimal number of observations. In this study, we consider a Bayesian *D-optimal* experimental design [25, 90] where we maximize the information gain through collection of informative data. While this problem is computationally difficult, following the work of Krause and Golovin [54] we prove that an approximation to the mutual information in our setting is a monotonic submodular set function. Based upon this observation, we gain access to the wealth of optimization machinery available for optimization of submodular set functions [78, 83, 15, 86], and thereby provide solid performance guarantees that allows us to probe the trade-off between the availability of domain knowledge and empirical data, as we show by deriving theoretical bounds and demonstrate via numerical experiments. Specifically, we demonstrate the use of a computationally efficient solver that guarantees approximately optimal (up to a constant factor) solution to the optimal experimental design problem.

2.2 Problem Formulation

The behavior of a broad variety of dynamical models can be described by Ordinary Differential Equations (ODE) [34]. Consider a misspecified system of such first order ODEs[1]:

$$\frac{d}{dt}\mathbf{y}(t) = \mathbf{G}(\mathbf{y}(t)) + \mathbf{F}(\mathbf{y}(t)) \tag{2.1}$$

with t time, $\mathbf{y}(t) = [y_1(t), \ldots, y_d(t)]^\top$ a state vector of interest, and

$$\mathbf{F}(\mathbf{y}(t)) \equiv [F_1(\mathbf{y}(t)), \ldots, F_d(\mathbf{y}(t))]^\top, \mathbf{G}(\mathbf{y}(t)) \equiv [G_1(\mathbf{y}(t)), \ldots, G_d(\mathbf{y}(t))]^\top$$

vector valued functions, $\mathbf{F}(\cdot), \mathbf{G}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ governing the system dynamics.

We are interested in settings where the temporal evolution of $\mathbf{y}(t)$ is dominated by the driving term $\mathbf{G}(\cdot)$, whereas the correction term $\mathbf{F}(\cdot)$ is assumed to have only a small effect over short time spans. Concretely, define the auxiliary system

$$\frac{d}{dt}\mathbf{y}_G(t) = \mathbf{G}(\mathbf{y}_G(t)) \tag{2.2}$$

then our interest is in the regime where, initialized in the same state $\mathbf{y}(0) = \mathbf{y}_G(0)$ the two systems track each other closely over some prescribed time span t_f. Specifically, we assume that for all $t \in [0, t_f]$ we have $\mathbf{y}(t)$ is close to $\mathbf{y}_G(t)$ in a sense that will be quantitatively defined in Section 2.5. One way of ensuring this is requiring $\|\mathbf{F}(\cdot)\| \ll \|\mathbf{G}(\cdot)\|$ over some domain $\mathcal{D} \subseteq \mathbb{R}^d$, and short enough time spans t_f. The model (2.1) is misspecified in the sense that $\mathbf{G}(\cdot)$ is assumed known, whereas the small additive correction function $\mathbf{F}(\cdot)$ is not available. Situations like this may arise, e.g. when we have at our disposal some approximation $\mathbf{G}(\cdot)$ to a system of interest, perhaps derived via expert domain knowledge, which does not fully capture the true dynamics driving the system.

Our goal is to utilize system evolution paths $\mathbf{y}(t)$ as observed in experiments to learn a representation for the correction term $\mathbf{F}(\cdot)$. The resulting 'corrected' model allows making accurate predictions about the system evolution. We specifically focus on designing efficient

[1]Higher order ODE systems may be converted into first order ODE form by defining new state variables, e.g. see Section 2.6.

experiments that facilitate rapid learning of the correction term under a limited experimental budget.

2.2.1 Initial Conditions

We consider applications where we are at liberty to perform a limited number of at most K experiments to facilitate learning the correction term $\mathbf{F}(\cdot)$. The kth experiment entails preparing the system at some fixed initial conditions at time zero $\mathbf{y}^{(k)}(0) \in \mathcal{Y}$ and observing its subsequent evolution $\mathbf{y}^{(k)}(t)$ for $t > 0$ as determined[2] by the (not fully known) model (2.1). We take the set $\mathcal{Y}$ to be a finite collection of possible experimental conditions that we may choose to start the system from. It may be a finely discretized grid over a continuous region of accessible initial conditions, e.g. expressing power constraints $\mathcal{Y} \subset \{\mathbf{y} : \|\mathbf{y}\|_2^2 \leq P\}$, or otherwise meeting an application specific set of restrictions.

Let $\mathcal{Y}_0 \subseteq \mathcal{Y}, |\mathcal{Y}_0| \leq K$ be the set of selected initial conditions that seed the K experiments. Informative prescription of $\mathcal{Y}_0$ is a primary concern in this study, as in many practical scenarios experiments are costly and it is important to design them carefully in order to extract as much information as possible from the limited set of measurements.

2.2.2 An Observation Model

The empirical evolution data $\mathbf{y}^{(k)}(t)$, $k = 1, \ldots, K$ allows us to probe the system dynamics and learn a representation for the correction term. To set the framework we specify a discrete and noisy observation model.

In this work we assume readings are collected on a discrete time grid. As the kth experiment unfolds the system evolves from the initial state $\mathbf{y}^{(k)}(0) \in \mathcal{Y}_0$ according to $\mathbf{y}^{(k)}(t)$ and we gain access to T temporal observations on a discrete time grid $t \in \mathcal{T} = \{t_1, \ldots, t_T\}$. Let $\mathcal{Y}_m \equiv \left\{\mathbf{y} | \exists k, i \text{ s.t. } \mathbf{y} = \mathbf{y}^{(k)}(t_i)\right\}$ be the set of size $\tilde{K} \equiv |\mathcal{Y}_m| = KT$ of system states recorded during the K experiments seeded by states in $\mathcal{Y}_0$.

For a given trajectory $\mathbf{y}(t)$, it is apparent from (2.1) that the correction term $\mathbf{F}(\cdot)$ can

[2]Fixing initial conditions at time $t = t_0$ the output of the first order ODE system (2.1) is determined for all $t > t_0$ [34].

be evaluated at points along the path via

$$\mathbf{F}(\mathbf{y}(t)) = \frac{d}{dt}\mathbf{y}(t) - \mathbf{G}(\mathbf{y}(t)) \tag{2.3}$$

With oracle access to the derivative $\frac{d}{dt}\mathbf{y}(t)$ we could attain point samples of $\mathbf{F}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}_m$ through (2.3) as $\mathbf{G}(\mathbf{y})$ is assumed known. However, with only discrete samples on the trajectory we do not have access to $\frac{d}{dt}\mathbf{y}(t)$. Instead, we assume access to noisy derivative estimates[3] $\frac{d}{dt}\tilde{\mathbf{y}}^{(k)}(t_i)$ given according to:

$$\frac{d}{dt}\tilde{\mathbf{y}}^{(k)}(t_i) = \frac{d}{dt}\mathbf{y}^{(k)}(t_i) + \boldsymbol{\epsilon}^{k,i} \qquad k = 1, \dots, K, \quad i = 1, \dots, T \tag{2.4}$$

with $\boldsymbol{\epsilon}^{k,i} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$ i.i.d. Gaussian noise. We form noisy estimates for the correction term $\tilde{\mathbf{F}}(\mathbf{y}^{(k)}(t_i))$ by substituting:

$$\begin{aligned}\tilde{\mathbf{F}}(\mathbf{y}^{(k)}(t_i)) &\equiv \frac{d}{dt}\tilde{\mathbf{y}}^{(k)}(t_i) - G(\mathbf{y}^{(k)}(t_i)) = \frac{d}{dt}\mathbf{y}^{(k)}(t_i) - G(\mathbf{y}^{(k)}(t_i)) + \boldsymbol{\epsilon}^{k,i} \\ &= \mathbf{F}(\mathbf{y}^{(k)}(t_i)) + \boldsymbol{\epsilon}^{k,i}\end{aligned} \tag{2.5}$$

For the sequel, we sometimes ease notations by writing $\mathbf{f}^j \equiv \mathbf{F}(\mathbf{y}^j)$, and $\tilde{\mathbf{f}}^j \equiv \tilde{\mathbf{F}}(\mathbf{y}^j)$ for the noisy readings $\mathbf{y}^j \in \mathcal{Y}_m, \quad j = 1, \dots, \tilde{K}$. In these symbols the noisy measurement model (2.5) reads:

$$\tilde{\mathbf{f}}^j = \mathbf{F}(\mathbf{y}^j) + \boldsymbol{\epsilon}^j \qquad j = 1, \dots, \tilde{K} \tag{2.6}$$

and $\boldsymbol{\epsilon}^j \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$ are Gaussian i.i.d..

2.3 Correction Estimation

The experimental framework detailed in the last section resulted in a set of $\tilde{K}$ noisy point estimates for the correction term $\tilde{\mathbf{F}}(\mathcal{Y}_m) = \left\{\tilde{\mathbf{F}}(\mathbf{y}) | \mathbf{y} \in \mathcal{Y}_m\right\}$ which form our training set. Our interest lies in estimating $\mathbf{F}(\cdot)$ over some domain $\mathcal{D} \subseteq \mathbb{R}^d$, however even in the noiseless

[3] E.g. by simple numerical differences, or more signal tailored techniques performing smoothing over the trajectory [20, 111].

setting and in the limit where the sampling interval approaches zero, we generally cannot achieve a dense cover over $\mathcal{D}$ with a finite number of trajectories $\mathbf{y}^{(k)}(t)$. Thus, some structure or prior information must be assumed for the correction term, such as degree of smoothness or adherence to a specific functional form, to allow for its estimation from the collected data.

In this section we take a Bayesian approach, setting a Gaussian Process (GP) formulation for the problem [50], allowing to express prior knowledge over the correction term $\mathbf{F}(\cdot)$ and enabling inference from the finite number of collected noisy samples to the underlying values over the entire domain $\mathcal{D}$. The estimated correction term may subsequently be used to make evolution predictions for arbitrary initial conditions. Please find a brief review of the GP framework in Appendix B.

2.3.1 Gaussian Processes

To correct the ODE model we assume a probabilistic setting in which $\mathbf{F}(\mathbf{y})$ is a vector-valued GP $\mathbf{F}(\mathbf{y}) \sim \mathcal{GP}(\mathbf{m}(\mathbf{y}), \mathbf{k}(\mathbf{y}, \mathbf{y}'))$ defined over some bounded region $\mathcal{D} \subseteq \mathbb{R}^d$ with $\mathbf{m}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ the mean function and $\mathbf{k}(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ the covariance function [92]. Every finite collection of sample points $\{\mathbf{F}(\mathbf{y}^1), \mathbf{F}(\mathbf{y}^2), \ldots\}$ is then distributed as multivariate Gaussian. The mean vector is retrieved by stacking $\mathbf{m}(\mathbf{y}^1), \mathbf{m}(\mathbf{y}^2), \ldots$ and the second order statistics are given according to $\mathbb{E}[[\mathbf{F}(\mathbf{y}^i)]_m [\mathbf{F}(\mathbf{y}^j)]_n] = [\mathbf{k}(\mathbf{y}^i, \mathbf{y}^j)]_{m,n}$ [3]. For the sequel we make the simplifying assumptions $\mathbf{m}(\mathbf{y}) \equiv 0$ and $\mathbf{k}(\mathbf{y}, \mathbf{y}') = k(\mathbf{y}, \mathbf{y}')\mathbf{I}_d$, i.e. the vector components are zero mean, independent and share a common scalar kernel function, as in the usual scalar-valued GP setting. Our techniques and methods can be generalized to the biased and correlated-components setting, but we restrict our model here for brevity.

Let $\tilde{\mathbf{F}}(\mathcal{A})$ be a set of noisy measurements collected at some set of sampling points $\mathcal{A}$: $\tilde{\mathbf{F}}(\mathcal{A}) = \{\mathbf{F}(\mathbf{y}) + \boldsymbol{\epsilon} | \mathbf{y} \in \mathcal{A}\}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$ is i.i.d. additive noise. We are interested in predicting the value of the process in unobserved locations. The posterior for $\mathbf{F}(\mathcal{B}) = \{\mathbf{F}(\mathbf{y}) | \mathbf{y} \in \mathcal{B}\}$ where $\mathcal{B}$ is some arbitrary set of sampling points is given according

to $\mathbf{F}(\mathcal{B})|\tilde{\mathbf{F}}(\mathcal{A}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{B}|\mathcal{A}}, \boldsymbol{\Sigma}_{\mathcal{B}|\mathcal{A}})$ with [3]:

$$\boldsymbol{\mu}_{\mathcal{B}|\mathcal{A}}=\mathbf{k}(\mathcal{B},\mathcal{A})[\mathbf{k}(\mathcal{A},\mathcal{A})+\boldsymbol{\Sigma}]^{-1}\tilde{\mathbf{F}}(\mathcal{A}) \tag{2.7}$$

$$\boldsymbol{\Sigma}_{\mathcal{B}|\mathcal{A}}=\mathbf{k}(\mathcal{B},\mathcal{B})-\mathbf{k}(\mathcal{B},\mathcal{A})[\mathbf{k}(\mathcal{A},\mathcal{A})+\boldsymbol{\Sigma}]^{-1}\mathbf{k}(\mathcal{A},\mathcal{B}) \tag{2.8}$$

and $\mathbf{k}(\mathcal{S}_1,\mathcal{S}_2) \in \mathbb{R}^{|\mathcal{S}_1|d\times|\mathcal{S}_2|d}$ has block structure with elements $[\mathbf{k}(\mathbf{y}^i,\mathbf{y}^j)]_{mn}$ for all $\mathbf{y}^i \in \mathcal{S}_1, \mathbf{y}^j \in \mathcal{S}_2$ and $m,n=1,\ldots,d$ and $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_{|\mathcal{A}|}$.

The GP formalism facilitates expression of prior knowledge over unknown functions $\mathbf{F}(\cdot)$, as determined by the choice of kernel, capturing notions of similarity between values at different positions. Popular choices for the kernel function include the Gaussian RBF $k(\mathbf{y},\mathbf{y}')=\exp(-\frac{1}{2\sigma_k^2}\|\mathbf{y}-\mathbf{y}'\|^2)$ with σ_k^2 the kernel bandwidth and the polynomial kernel $k(\mathbf{y},\mathbf{y}')=(1+\langle\mathbf{y},\mathbf{y}'\rangle)^m$ with $m\in\mathbb{N}^+$ the order. The Gaussian RBF kernel is of particular interest as it is universal in the sense that with a large enough training set, estimation according to (2.7) can approximate any continuous bounded function on a compact domain [76]. With the GP model set, the value of $\mathbf{F}(\mathbf{y})$ at any $\mathbf{y}\in\mathcal{D}$ may be estimated according to (2.7) based on the noisy measurements $\tilde{\mathbf{F}}(\mathcal{Y}_m)$.

2.3.2 Feature Space Representation

With the assumptions of the last subsection, the d-dimensional vector-valued GP $\mathbf{F}(\mathbf{y})$ is comprised of d independent GPs $F_i(\mathbf{y}) \sim \mathcal{GP}(0,k(\mathbf{y},\mathbf{y}'))$, $i=1,\ldots,d$. We follow [124, 92, 21] and review the correspondence between these GPs and equivalent linear regression models in the feature space.

Mercer's theorem guarantees the existence of a sequence of eigenfunctions $\{\phi_j(\mathbf{y})\}, j=1,2,\ldots$ such that $k(\mathbf{y},\mathbf{y}')=\sum_j \phi_j(\mathbf{y})\phi_j(\mathbf{y}')=\langle\boldsymbol{\phi}(\mathbf{y}),\boldsymbol{\phi}(\mathbf{y}')\rangle$ where $\boldsymbol{\phi}(\mathbf{y})=[\phi_1(\mathbf{y}),\phi_2(\mathbf{y}),\ldots]^\top$ is the feature transformation from the input space to the feature space and $\langle\cdot;\cdot\rangle$ is an inner product.

Let $\theta_{ij}\sim\mathcal{N}(0,1)$, $i=1,\ldots,d$, $j=1,2,\ldots$ be a sequence of i.i.d. standard Gaussian variables. For notational convenience we define $\boldsymbol{\theta}_i \equiv [\theta_{i1},\theta_{i2},\ldots]^\top$, $i=1,\ldots,d$ and $\boldsymbol{\Theta}=$

$[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d]^\top$. We will see that the following identity holds in distribution:

$$F_i(\mathbf{y}) = \sum\nolimits_j \theta_{ij}\phi_j(\mathbf{y}) \equiv \langle \boldsymbol{\theta}_i, \boldsymbol{\phi}(\mathbf{y}) \rangle \qquad i = 1, \ldots, d \tag{2.9}$$

i.e. the GP inference of Section 2.3.1 is equivalent to a Bayesian linear regression model in the feature space.

To see that (2.9) holds notice that both sides of the equality are zeros mean GPs over $\mathbf{y}$. The covariance function of the left hand term is $k(\mathbf{y}, \mathbf{y}')$ by definition. The covariance function of the right hand term is

$$\mathbb{E}\left[\sum\nolimits_j \theta_{ij}\phi_j(\mathbf{y}) \sum\nolimits_{j'} \theta_{ij'}\phi_{j'}(\mathbf{y}')\right] = \sum\nolimits_{jj'} \mathbb{E}\left[\theta_{ij}\theta_{ij'}\right] \phi_j(\mathbf{y})\phi_{j'}(\mathbf{y}') = \sum\nolimits_j \phi_j(\mathbf{y})\phi_j(\mathbf{y}') = k(\mathbf{y}, \mathbf{y}') \tag{2.10}$$

Given noisy data $\tilde{\mathbf{F}}(\mathcal{Y}_m) = \mathbf{F}(\mathcal{Y}_m) + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$ i.i.d. noise, inference in the GP can be equivalently performed by estimating the regression coefficients $\boldsymbol{\Theta}$ and making predictions for $\mathbf{F}(\mathbf{y})$ as per (2.9).

2.4 Informative Sampling in a GP

In this section we study the problem of sampling in a GP. Our main interest lies in understanding how to choose an efficient sampling strategy under constraints, one that facilitates high fidelity inference over the GP from the collected measurements. This will become useful in subsequent sections for designing efficient experimental schemes for rapid learning of misspecified systems.

2.4.1 Sampling Setup

Consider the problem of estimating $\mathbf{F}(\mathbf{y})$ for any $\mathbf{y} \in \mathcal{D}$ by collecting noisy samples $\tilde{\mathbf{F}}(\mathcal{S})$ over some set $\mathcal{S} \subseteq \mathcal{Y}$, with $\mathcal{Y} \subset \mathcal{D}$ a set of allowed sampling positions, and applying the inference methodology of Section 2.3. The quality of inference strongly depends on the sampling set $\mathcal{S}$, which we also refer to as a sensor placement configuration. For example, if the set $\mathcal{S}$ is highly

localized in some region in $\mathcal{D}$ it is reasonable to expect that inference of $\mathbf{F}(\cdot)$ becomes more accurate in that region on the expense of farther locals in $\mathcal{D}$. We are generally interested in estimating $\mathbf{F}(\cdot)$ over the whole of $\mathcal{D}$ and so we are interested in developing a mechanism that allows this.

An efficient sampling configuration is one supporting high quality inference under a fixed budget constraint on the number of sensors used, i.e. imposing $\mathcal{S} : \mathcal{S} \subseteq \mathcal{Y}$, $|\mathcal{S}| \leq K$. The problem of designing sensor placement configurations for GPs has been extensively studied in the past, but only recently new methodologies guaranteeing near optimal designs have emerged [58]. While these are computationally tractable in low dimensional settings the computational complexity does not scale favorably with the dimension, showing an exponential dependence, as we discuss next.

2.4.2 Mutual Information Criterion

Krause et al. [58] have proposed the following strategy for choosing $\mathcal{S}$. Let $\mathcal{U} \subset \mathcal{D}$ be a set of test points where estimation quality will be assessed, and $\mathcal{V} = \mathcal{Y} \cup \mathcal{U}$. Krause suggests choosing $\mathcal{S}$ according to:

$$G(\mathcal{S}) \equiv I(\tilde{\mathbf{F}}(\mathcal{S}); \mathbf{F}(\mathcal{V} \backslash \mathcal{S})) \tag{2.11}$$

$$\mathcal{S}^{\star} = \text{argmax}_{\mathcal{S}:\mathcal{S}\subseteq\mathcal{Y},|\mathcal{S}|\leq K} G(\mathcal{S}) \tag{2.12}$$

Intuitively, the idea is to find a set $\mathcal{S}$ which is informative over positions in $\mathcal{V} \backslash \mathcal{S} = \mathcal{U} \cup (\mathcal{Y} \backslash \mathcal{S})$ in the sense of maximally decreasing the entropy of the process value in these positions.

At this point, please review Appendix A for a brief survey of submodular functions and optimization problems involving their maximization. It is shown in [58] that $G(\mathcal{S})$ is a submodular function. They further show that when $|\mathcal{U}|$ is large enough compared to K then $G(\mathcal{S})$ is additionally approximately monotone. Using results in Appendix A we then have that (2.11) is amenable under some restrictions for efficient near optimal solution via conventional submodular optimization techniques.

This result is important in that it allows for a guaranteed near optimal solution for the sensor placement problem under the specified criterion, however it heavily relies on the choice

of the test set $\mathcal{U}$, and the computational complexity may be shown to scale as $O(K|\mathcal{V}|^4)$. Specifically, in high dimensions if we were to choose the set $\mathcal{U}$ to be an ϵ-net covering of some set $\mathcal{D}$ of fixed per-axis length, both $|\mathcal{U}|$ and $|\mathcal{V}|$ grow exponentially and the computational complexity becomes prohibitive.

2.4.3 Feature Space Information Criterion

We introduce a feature space based sensor placement criterion, and show that near optimal solutions may be efficiently obtained, in some cases even in high dimensional spaces. Our key insight is that in some situations invoking the feature space representation for the GP may result in a succinct representation that is amenable to efficient manipulation. The computational complexity of our approach can be orders of magnitude lower than that of [58]. Specifically, we show that it scales nicely with the dimension, in lieu of the aforementioned exponential trend. We demonstrate the efficacy of our approach via numerical experiments in Section 2.6.

Invoking the feature space representation of Section 2.3.2 we see that performing inference in the GP based on a ground set of noisy measurements $\tilde{\mathbf{F}}(\mathcal{S})$ may be viewed as first estimating $\boldsymbol{\Theta}$ and then applying (2.9) to retrieve estimates for the rest of $\mathcal{D}$. From this viewpoint, the estimation error in $\mathbf{F}(\mathbf{y})$ originates from the error in $\boldsymbol{\Theta}$ and so our goal is to decrease these as much as possible by maximizing the quality of inference from $\tilde{\mathbf{F}}(\mathcal{S})$ to $\boldsymbol{\Theta}$. Various statistical criteria have been developed for quantifying the quality of inference between observations and underlying random variables [17, 11, 25, 90]. Here we follow D-Bayes optimality [8].

In this framework, the uncertainty associated with $\boldsymbol{\Theta}$ is quantified through the Shannon entropy $H(\cdot)$. Before the experiment we have initial uncertainty $H(\boldsymbol{\Theta})$ which is revised to $H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S}))$ following data collection. A D-Bayes optimal design minimizes the posterior uncertainty $H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S}))$, or equivalently maximizes the mutual information:

$$G(\mathcal{S}) \equiv I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{S})) = H(\boldsymbol{\Theta}) - H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S})) \tag{2.13}$$

and an optimal experimental design under the budget constraint $\mathcal{S} \subseteq \mathcal{Y}$, $|\mathcal{S}| \leq K$ is

$$\mathcal{S}^\star = \underset{\mathcal{S}:|\mathcal{S}|\leq K, \mathcal{S}\subseteq\mathcal{Y}}{\operatorname{argmax}} \; G(\mathcal{S}) \tag{2.14}$$

We show that $G(\cdot)$ as defined above holds favorable set function properties (as defined in appendix A):

Theorem 2.1. *The utility function $G(\mathcal{S})$ is submodular and monotone.*

Proof. First we prove submodularity. Let $\mathcal{S} \subset \mathcal{Y}$ and $\mathbf{y} \in \mathcal{Y}\backslash\mathcal{S}$. Expanding the mutual information according to $I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{S})) = H(\tilde{\mathbf{F}}(\mathcal{S})) - H(\tilde{\mathbf{F}}(\mathcal{S})|\boldsymbol{\Theta})$ we have:

$$\begin{aligned} G(S \cup \{\mathbf{y}\}) - G(\mathcal{S}) &= H(\tilde{\mathbf{F}}(\mathcal{S}) \cup \tilde{\mathbf{F}}(\mathbf{y})) - H(\tilde{\mathbf{F}}(\mathcal{S})) - [H(\tilde{\mathbf{F}}(\mathcal{S}) \cup \tilde{\mathbf{F}}(\mathbf{y})|\boldsymbol{\Theta}) - H(\tilde{\mathbf{F}}(\mathcal{S})|\boldsymbol{\Theta})] \\ &= H(\tilde{\mathbf{F}}(\mathbf{y})|\tilde{\mathbf{F}}(\mathcal{S})) - H(\tilde{\mathbf{F}}(\mathbf{y})|\boldsymbol{\Theta}) \end{aligned} \tag{2.15}$$

where we used the conditional independence of the elements of $\tilde{\mathbf{F}}(\mathcal{S}) \cup \tilde{\mathbf{F}}(\mathbf{y})$ given $\boldsymbol{\Theta}$, so $H(\tilde{\mathbf{F}}(\mathcal{S}) \cup \tilde{\mathbf{F}}(\mathbf{y})|\boldsymbol{\Theta}) = H(\tilde{\mathbf{F}}(\mathcal{S})|\boldsymbol{\Theta}) + H(\tilde{\mathbf{F}}(\mathbf{y})|\boldsymbol{\Theta})$.

Now apply the results of (2.15) for two specific choices for $\mathcal{S}$, namely $\mathcal{S} \leftarrow \mathcal{S}^1$ and $\mathcal{S} \leftarrow \mathcal{S}^2$ such that $\mathcal{S}^1 \subseteq \mathcal{S}^2$:

$$[G(\mathcal{S}^1 \cup \{\mathbf{y}\}) - G(\mathcal{S}^1)] - [G(\mathcal{S}^2 \cup \{\mathbf{y}\}) - G(\mathcal{S}^2)] = H(\tilde{\mathbf{F}}(\mathbf{y})|\tilde{\mathbf{F}}(\mathcal{S}^1))) - H(\tilde{\mathbf{F}}(\mathbf{y})|\tilde{\mathbf{F}}(\mathcal{S}^2))) \tag{2.16}$$

Conditioning on a larger set cannot increase entropy and we have $H(\tilde{\mathbf{F}}(\mathbf{y})|\tilde{\mathbf{F}}(\mathcal{S}^1))) \geq H(\tilde{\mathbf{F}}(\mathbf{y})|\tilde{\mathbf{F}}(\mathcal{S}^2)))$ such that $G(\mathcal{S}^1 \cup \{\mathbf{y}\}) - G(\mathcal{S}^1) \geq G(\mathcal{S}^2 \cup \{\mathbf{y}\}) - G(\mathcal{S}^2)$ and G is submodular.

To prove monotonicity it is enough to show $G(\mathcal{S} \cup \{\mathbf{y}\}) - G(\mathcal{S}) \geq 0$. This time expand the mutual information according to $I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{S})) = H(\boldsymbol{\Theta}) - H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S}))$:

$$G(\mathcal{S} \cup \{\mathbf{y}\}) - G(\mathcal{S}) = H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S})) - H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S}) \cup \tilde{\mathbf{F}}(\mathbf{y}))\,. \tag{2.17}$$

Conditioning can never increase entropy so $H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S})) \geq H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{S}) \cup \tilde{\mathbf{F}}(\mathbf{y}))$ and the result follows. □

With the last theorem, and applying results from the theory of submodular optimization, we can show that it is possible to efficiently compute near optimal solutions for (2.14) using, e.g. variations on greedy selection algorithms, as we further detail in Section 2.5.3.

Computational complexity To evaluate the computational complexity of our approach when applying a greedy selection algorithm, notice that we have K steps, each culminating with the inclusion of an additional element $x \in \mathcal{Y}$ to the budding $\mathcal{S}'$. For each candidate element x and the resulting $\mathcal{S}'$ we compute $G(\mathcal{S}') = I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{S}')) = H(\tilde{\mathbf{F}}(\mathcal{S}')) - H(\tilde{\mathbf{F}}(\mathcal{S}')|\boldsymbol{\Theta}) = H(\tilde{\mathbf{F}}(\mathcal{S}')) - H(\boldsymbol{\epsilon})$.

Evaluating $H(\boldsymbol{\epsilon})$ may be done in time $O(1)$ using the Gaussian distribution of $\boldsymbol{\epsilon}$ and the analytic formula for the entropy of a Gaussian random variable[4]. The computational complexity of evaluating $G(\mathcal{S}')$ is thus equivalent to the complexity of evaluating the term $H(\tilde{\mathbf{F}}(\mathcal{S}'))$ which entails calculating a determinant for the corresponding $|\mathcal{S}'| \times |\mathcal{S}'|$ covariance matrix. The computational complexity thus scales as $O(K|\mathcal{Y}||\mathcal{S}'|^3) = O(K|\mathcal{Y}|K^3) = O(K^4|\mathcal{Y}|)$. Typically we have $|\mathcal{Y}| \ll |\mathcal{V}|$ and $K \ll |\mathcal{V}|$ such that the computational complexity can be orders of magnitude smaller than the one associated with the approach of [58]. In particular notice that the complexity does not scale with the ambient dimension.

2.5 Experimental Design for Dynamical Systems

In Section 2.3 we reviewed inference in a GP setting, and suggested applying this formulation for estimating the correction term $\mathbf{F}(\cdot)$ based on the set of noisy measurements $\tilde{\mathbf{F}}(\mathcal{Y}_m)$. The observation set $\mathcal{Y}_m$, as determined by the initial conditions set $\mathcal{Y}_0$ was assumed given and fixed.

In this section, following the introduction of Section 2.4 we study efficient experimental design in the misspecified dynamical system context. That is, our goal would be to select an informative set of experiments, parametrized through the initial conditions $\mathcal{Y}_0$, such as to facilitate rapid learning of the correction term $\mathbf{F}(\cdot)$ under a limited experimental budget constraint. We quantify the expected utility associated with choosing sets of initial conditions

[4]The entropy of a Gaussian multivariate random variable is given according to: $\mathbf{x} \in \mathbb{R}^k, \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow H(\mathbf{x}) = \log((\pi e)^k \det \boldsymbol{\Sigma})$.

and suggest an efficient near-optimal (up to a constant factor) algorithm for choosing the best such experimental setup.

2.5.1 Utility Function for Misspecified Dynamical Models

In our setting we select a set of initial conditions $\mathcal{Y}_0$ and observe the corresponding outputs. This chain of dependencies is made explicit as $\mathcal{Y}_0 \to \mathcal{Y}_m(\mathcal{Y}_0) \to \tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))$. The quality of inference, viewed as a function of the initial conditions $\mathcal{Y}_0$ is given by

$$G(\mathcal{Y}_0) \equiv I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))) \tag{2.18}$$

and an optimal experimental design under the budget constraint $|\mathcal{Y}_0| \leq K$, $\mathcal{Y}_0 \subseteq \mathcal{Y}$ is

$$\mathcal{Y}_0^\star = \underset{\mathcal{Y}_0 : |\mathcal{Y}_0| \leq K, \mathcal{Y}_0 \subseteq \mathcal{Y}}{\operatorname{argmax}} G(\mathcal{Y}_0) \tag{2.19}$$

2.5.2 Output Trajectory Proxy

The design problem (2.19) entails choosing a set $\mathcal{Y}_0$ of K initial conditions, and observing $\tilde{K}$ noisy measurements $\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))$, which are utilized for estimating $\mathbf{F}(\cdot)$ over $\mathcal{D}$.

As we are concerned with misspecified systems such that the complete system model (2.1) is unknown, we are unable to predict system trajectories based on initial conditions at time zero. In particular, we do not have a-priori access to the mapping between the sets $\mathcal{Y}_0$ and $\mathcal{Y}_m$, such that evaluation of the cost function (2.18) and thus solution of the design problem (2.19) are not possible. However, at this point our assumption that the system is only slightly misspecified in short time spans, i.e. that the correction term $\mathbf{F}(\cdot)$ introduces a small effect on the trajectory, turns out to be useful in retrieving approximate solutions.

For any given set of initial conditions $\mathcal{Y}_0$ we invoke the approximate system model (2.2) to obtain a proxy $\mathcal{Y}_g$ for the true set of future states $\mathcal{Y}_m$. Let $\mathbf{y}^{(k)}(0) \in \mathcal{Y}_0$ be the initial conditions seeding the k^{th} experiment, and designate the approximate ensuing trajectory $\mathbf{y}_G^{(k)}(t)$. Collect the approximate trajectories in $\mathcal{Y}_g \equiv \left\{\mathbf{y} | \exists k, i \text{ s.t. } \mathbf{y} = \mathbf{y}_G^{(k)}(t_i)\right\}$, and note that the set $\mathcal{Y}_g$ may be evaluated in advance given $\mathcal{Y}_0$. For example, for a linear misspecified system $\frac{d}{dt}\mathbf{y}_G(t) = \mathbf{A}\mathbf{y}_G(t)$ for some fixed $\mathbf{A} \in \mathbb{R}^{d \times d}$, the trajectories comprising $\mathcal{Y}_g$ may be

determined according to $\mathbf{y}_G^{(k)}(t_i) = e^{\mathbf{A}t_i}\mathbf{y}_G^{(k)}(0)$ where $e^{(\cdot)}$ is the matrix exponential according to the usual definition.

In what follows we propose a proxy for the cost function (2.18) where $\mathcal{Y}_g$ is used in lieu of the unknown $\mathcal{Y}_m$, and derive approximation bounds for the discrepancy between the two. We show that that these bounds scale with the deviation between the actual and approximate system outputs $\mathbf{y}(\cdot)$ and $\mathbf{y}_G(\cdot)$. Specifically, we have:

Theorem 2.2. *Let $\tilde{G}(\mathcal{Y}_0) \equiv I(\boldsymbol{\Theta}; \tilde{\boldsymbol{F}}(\mathcal{Y}_g(\mathcal{Y}_0)))$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$, and let δ be a positive constant such that the maximum covariance discrepancy between the true and approximate models is bounded according to*

$$\forall k_1, k_2, i_1, i_2 : \left| k(\boldsymbol{y}^{(k_1)}(t_{i_1}), \boldsymbol{y}^{(k_2)}(t_{i_2})) - k(\boldsymbol{y}_G^{(k_1)}(t_{i_1}), \boldsymbol{y}_G^{(k_2)}(t_{i_2})) \right| \leq \delta \,.$$

We have

$$\left| \tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0) \right| \leq -d\tilde{K} \log \left(1 - \frac{\delta (d\tilde{K})^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)} \right) \tag{2.20}$$

with $\sigma_{\min}(\cdot)$ the minimal singular value, and $\tilde{K} = KT$ the number of measurements.

Proof. See proof in Section C.1. □

Notice that the bound of Theorem 2.2 becomes looser as the noise decreases. However, notice that the value of $G(\mathcal{Y}_0)$ increases in this case in about the same proportion so the relative error remains similar. As an illustration, consider the case $\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2 \mathbf{I}$. Using notation used in the proof of Theorem 2.2 and $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_{\tilde{K}}$ we have:

$$\begin{aligned} G(\mathcal{Y}_0) &= H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))|\boldsymbol{\Theta}) = \log(\det(\mathbf{k}(\mathcal{Y}_m, \mathcal{Y}_m) + \boldsymbol{\Sigma})) - \log(\det(\boldsymbol{\Sigma})) \\ &= \log(\det(\mathbf{I} + \boldsymbol{\Sigma}^{-1}\mathbf{k}(\mathcal{Y}_m, \mathcal{Y}_m))) \end{aligned}$$

Now observe

$$\lambda_i(\mathbf{I} + \boldsymbol{\Sigma}^{-1}\mathbf{k}(\mathcal{Y}_m, \mathcal{Y}_m)) = 1 + \lambda_i(\boldsymbol{\Sigma}^{-1}\mathbf{k}(\mathcal{Y}_m, \mathcal{Y}_m)) \geq 1 + \frac{\sigma_{\min}(\mathbf{k}(\mathcal{Y}_m, \mathcal{Y}_m))}{\sigma_\epsilon^2}$$

so

$$\log(\det(\mathbf{I}+\boldsymbol{\Sigma}^{-1}\mathbf{k}(\mathcal{Y}_m,\mathcal{Y}_m))) \geq \tilde{K}\log\left(1+\frac{\sigma_{\min}(\mathbf{k}(\mathcal{Y}_m,\mathcal{Y}_m))}{\sigma_\epsilon^2}\right)$$

and we have

$$G(\mathcal{Y}_0) \geq d\tilde{K}\log\left(1+\frac{\sigma_{\min}(\mathbf{k}(\mathcal{Y}_m(\mathcal{Y}_0),\mathcal{Y}_m(\mathcal{Y}_0)))}{\sigma_\epsilon^2}\right).$$

Corollary 2.1. *Let $k(\boldsymbol{y},\boldsymbol{y}') = k(\|\boldsymbol{y}-\boldsymbol{y}'\|)$ be a shift-invariant kernel with $k(\cdot)$ Lipschitz continuous with constant L over $\mathcal{D}' \equiv \{\boldsymbol{y}^1-\boldsymbol{y}^2 | \boldsymbol{y}^1,\boldsymbol{y}^2 \in \mathcal{D}\}$, and assume $\forall k,i : \|\boldsymbol{y}^{(k)}(t_i) - \boldsymbol{y}_G^{(k)}(t_i))\| \leq \Delta$. We have*

$$\left|\tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0)\right| \leq -d\tilde{K}\log\left(1-\frac{2L\Delta(d\tilde{K})^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)}\right) \tag{2.21}$$

Proof. For any k_1,k_2,i_1,i_2 we have

$$\begin{aligned}
&\left|k(\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}^{(k_2)}(t_{i_2}))-k(\mathbf{y}_G^{(k_1)}(t_{i_1}),\mathbf{y}_G^{(k_2)}(t_{i_2}))\right| = \left|k(\|\mathbf{y}^{(k_1)}(t_{i_1})-\mathbf{y}^{(k_2)}(t_{i_2})\|)-k(\|\mathbf{y}_G^{(k_1)}(t_{i_1})-\mathbf{y}_G^{(k_2)}(t_{i_2})\|)\right| \\
&\leq L\left|\|\mathbf{y}^{(k_1)}(t_{i_1})-\mathbf{y}^{(k_2)}(t_{i_2})\|-\|\mathbf{y}_G^{(k_1)}(t_{i_1})-\mathbf{y}_G^{(k_2)}(t_{i_2})\|\right| \leq L\|(\mathbf{y}^{(k_1)}(t_{i_1})-\mathbf{y}_G^{(k_1)}(t_{i_1}))-(\mathbf{y}^{(k_2)}(t_{i_2})-\mathbf{y}_G^{(k_2)}(t_{i_2}))\| \\
&\leq L\left(\|\mathbf{y}^{(k_1)}(t_{i_1})-\mathbf{y}_G^{(k_1)}(t_{i_1})\|+\|\mathbf{y}^{(k_2)}(t_{i_2})-\mathbf{y}_G^{(k_2)}(t_{i_2})\|\right) \leq 2L\Delta
\end{aligned}$$

and the result follows by substitution in (2.20). □

Corollary 2.2. *Let $k(\boldsymbol{y},\boldsymbol{y}') = (1+\langle\boldsymbol{y},\boldsymbol{y}'\rangle)^m$ be the polynomial kernel, $B \equiv \sup_{\boldsymbol{y}\in\mathcal{D}}\|\boldsymbol{y}\|$ and assume $\forall k,i : \|\boldsymbol{y}^{(k)}(t_i) - \boldsymbol{y}_G^{(k)}(t_i))\| \leq \Delta$, then*

$$\left|\tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0)\right| \leq -d\tilde{K}\log\left(1-\frac{m\Delta(2B+\Delta)(1+B^2)^{m-1}(d\tilde{K})^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)}\right) \tag{2.22}$$

Proof. Consider the following chain of inequalities

$$
\begin{aligned}
&\left|k(\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}^{(k_2)}(t_{i_2}))-k(\mathbf{y}_G^{(k_1)}(t_{i_1}),\mathbf{y}_G^{(k_2)}(t_{i_2}))\right| \\
&=\left|\left(1+\left\langle\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}^{(k_2)}(t_{i_2})\right\rangle\right)^m-\left(1+\left\langle\mathbf{y}_G^{(k_1)}(t_{i_1}),\mathbf{y}_G^{(k_2)}(t_{i_2})\right\rangle\right)^m\right| \\
&\overset{(a)}{\leq} m(1+B^2)^{m-1}\left|\left\langle\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}^{(k_2)}(t_{i_2})\right\rangle-\left\langle\mathbf{y}_G^{(k_1)}(t_{i_1}),\mathbf{y}_G^{(k_2)}(t_{i_2})\right\rangle\right| \\
&=m(1+B^2)^{m-1}|\left\langle\mathbf{y}_G^{(k_1)}(t_{i_1})-\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}^{(k_2)}(t_{i_2})\right\rangle+\left\langle\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}_G^{(k_2)}(t_{i_2})-\mathbf{y}^{(k_2)}(t_{i_2})\right\rangle \\
&+\left\langle\mathbf{y}_G^{(k_1)}(t_{i_1})-\mathbf{y}^{(k_1)}(t_{i_1}),\mathbf{y}_G^{(k_2)}(t_{i_2})-\mathbf{y}^{(k_2)}(t_{i_2}))\right\rangle| \\
&\leq m(1+B^2)^{m-1}(\Delta B+B\Delta+\Delta^2)=m\Delta(2B+\Delta)(1+B^2)^{m-1}
\end{aligned}
$$

where (a) is due to the Lipschitz constant of the function $f(x)=(1+x)^m$ being smaller than $m(1+\sup_{x\in\mathcal{D}}|x|)^{m-1}$. The result follows by substitution in (2.20). □

Theorem 2.2 and Corollaries 2.1 and 2.2 bound the discrepancy between $G(\mathcal{Y}_0)$ and its proxy $\tilde{G}(\mathcal{Y}_0)$. As the trajectory uncertainty becomes smaller the two become more tightly aligned as quantified by our results in this subsection.

2.5.3 Near Optimal Solution

Based on the results of Theorem 2.2 and the ensuing corollaries, in lieu of problem (2.19) we pose a relaxed proxy that circumvents around the uncertainty associated with the system output. Namely, we are interested in the solution of

$$\tilde{\mathcal{Y}}_0^\star=\underset{\mathcal{Y}_0:|\mathcal{Y}_0|\leq K,\mathcal{Y}_0\subseteq\mathcal{Y}}{\operatorname{argmax}}\ \tilde{G}(\mathcal{Y}_0) \tag{2.23}$$

Generic combinatorial optimization problems such as (2.23) exhibit prohibitive computational complexity, as the solution generally involves enumeration over all possible subset combinations satisfying the constraints, which is exponential in the size of the set $|\mathcal{Y}_0|$. We prove that $\tilde{G}(\mathcal{Y}_0)$ holds favorable properties, rendering the optimization problem (2.23) amenable to approximate solution by means of computationally efficient algorithms with provable guarantees. Reviewing Appendix A our next step is to show that the set function

$\tilde{G}(\mathcal{Y}_0)$ is submodular and monotonic (similar to [56]), a fact that allows us to make use of the rich literature on submodular optimization.

Theorem 2.3. *Let $\tilde{G} : 2^{\mathcal{Y}} \to \mathbb{R}$ be the set function defined in Theorem 2.2. Then $\tilde{G}$ is monotonic (increasing) and submodular.*

Proof. First we prove submodularity. Let $\mathcal{Y}_0 \subset \mathcal{Y}$ and $\mathbf{y} \in \mathcal{Y} \backslash \mathcal{Y}_0$, such that the system output proxy for $\mathbf{y}$ is given as $\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))$. Expanding the mutual information according to $I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) = H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))|\boldsymbol{\Theta})$ we have:

$$\begin{aligned} \tilde{G}(\mathcal{Y}_0 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0) &= H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) \cup \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) - \\ &\qquad [H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) \cup \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y})|\boldsymbol{\Theta}) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)|\boldsymbol{\Theta})] \qquad (2.24) \\ &= H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\boldsymbol{\Theta}) \end{aligned}$$

where we used the conditional independence of the elements of $\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) \cup \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))$ given $\boldsymbol{\Theta}$, so $H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) \cup \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\boldsymbol{\Theta}) = H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))|\boldsymbol{\Theta}) + H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\boldsymbol{\Theta})$.

Now apply the results of (2.24) twice, for two specific choices for $\mathcal{Y}_0$, namely $\mathcal{Y}_0 \leftarrow \mathcal{Y}_0^1$ and $\mathcal{Y}_0 \leftarrow \mathcal{Y}_0^2$ such that $\mathcal{Y}_0^1 \subseteq \mathcal{Y}_0^2$:

$$\begin{aligned} &[\tilde{G}(\mathcal{Y}_0^1 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0^1)] - [\tilde{G}(\mathcal{Y}_0^2 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0^2)] \\ = \; & H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^1))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^2))) \end{aligned}$$

Conditioning on a larger set cannot increase entropy and we have $H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^1))) \geq H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^2)))$ such that $\tilde{G}(\mathcal{Y}_0^1 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0^1) \geq \tilde{G}(\mathcal{Y}_0^2 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0^2)$ and $\tilde{G}$ is submodular.

To prove monotonicity it is enough to show $\tilde{G}(\mathcal{Y}_0 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0) \geq 0$. This time expand the mutual information according to $I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) = H(\boldsymbol{\Theta}) - H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)))$:

$$\tilde{G}(\mathcal{Y}_0 \cup \{\mathbf{y}\}) - \tilde{G}(\mathcal{Y}_0) = H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) - H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) \cup \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y}))) \,. \qquad (2.25)$$

Conditioning can never increase entropy so

$$H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) \geq H(\boldsymbol{\Theta}|\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0)) \cup \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathbf{y})))$$

and the result follows. □

The class of submoudlar combinatorial optimization problems has been extensively studied in the past, as we survey in Appendix A. The computationally efficient greedy solver delineated in algorithm 3 is guaranteed to achieve a good approximation (up to a constant factor) to the optimal solution [78, 83], as stated in lemma A.1.

Proposed Method Applied to our setting, the Algorithm 3 performs successive evaluations of the proxy function $\tilde{G}(\cdot)$ for candidate sets $\mathcal{Y}_0^C \equiv \mathcal{Y}_0 \cup \{\mathbf{y}\}$ where $\mathbf{y} \in \mathcal{Y} \setminus \mathcal{Y}_0$. During the k^{th} iteration the candidate sets $\mathcal{Y}_0^C$ are of size k. We utilize the following identity to facilitate the flow of the algorithm:

$$\begin{aligned}\tilde{G}(\mathcal{Y}_0^C) =& H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^C))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^C))|\boldsymbol{\Theta}) \\ =& \log((\pi e)^{kT} \det \boldsymbol{\Sigma}_g) - \log((\pi e)^{kT} \det \boldsymbol{\Sigma}_{g|\boldsymbol{\Theta}}) = \log(\det \boldsymbol{\Sigma}_g) - \log(\det \boldsymbol{\Sigma}_{g|\boldsymbol{\Theta}}) \end{aligned} \tag{2.26}$$

In the equation above, $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Sigma}_{g|\boldsymbol{\Theta}}$ are the covariance matrices for the ensemble of kT samples $\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^C))$, taken without and with conditioning on the feature space coefficients $\boldsymbol{\Theta}$, respectively. Notice that conditioned on $\boldsymbol{\Theta}$ the measurements covariance matrix $\boldsymbol{\Sigma}_{g|\boldsymbol{\Theta}}$ is block-diagonal with block submatrices being the noise covariance matrix, and the no-conditioning covariance matrix $\boldsymbol{\Sigma}_g$ can be retrieved by adding the aforementioned noise matrix to the corresponding kernel covariance matrix $k(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^C)), \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0^C)))$.

Denoting the result of running the greedy maximization algorithm 3 on the proxy function $\tilde{G}(\mathcal{Y}_0)$ with $\tilde{\mathcal{Y}}_0^{\mathbf{gr}}$ we have our final result:

Theorem 2.4. *Let the maximum covariance discrepancy between the true and approximate models be bounded according to*

$$\forall k_1, k_2, i_1, i_2 : \left| k(\boldsymbol{y}^{(k_1)}(t_{i_1}), \boldsymbol{y}^{(k_2)}(t_{i_2})) - k(\boldsymbol{y}_G^{(k_1)}(t_{i_1}), \boldsymbol{y}_G^{(k_2)}(t_{i_2})) \right| \leq \delta$$

then we have

$$G(\tilde{\mathcal{Y}}_0^{\boldsymbol{gr}}) \geq (1 - e^{-1})(G(\mathcal{Y}_0^{\star}) + O(\log(1 - const \cdot \delta)))$$

Proof. Immediate from Lemma A.1 and Theorem 2.2. □

The last theorem demonstrates that applying the greedy maximization algorithm on the proxy function $\tilde{G}(\cdot)$ retrieves a solution $\tilde{\mathcal{Y}}_0^{\mathbf{gr}}$ which is near optimal for the original function $G(\cdot)$, which is what we want.

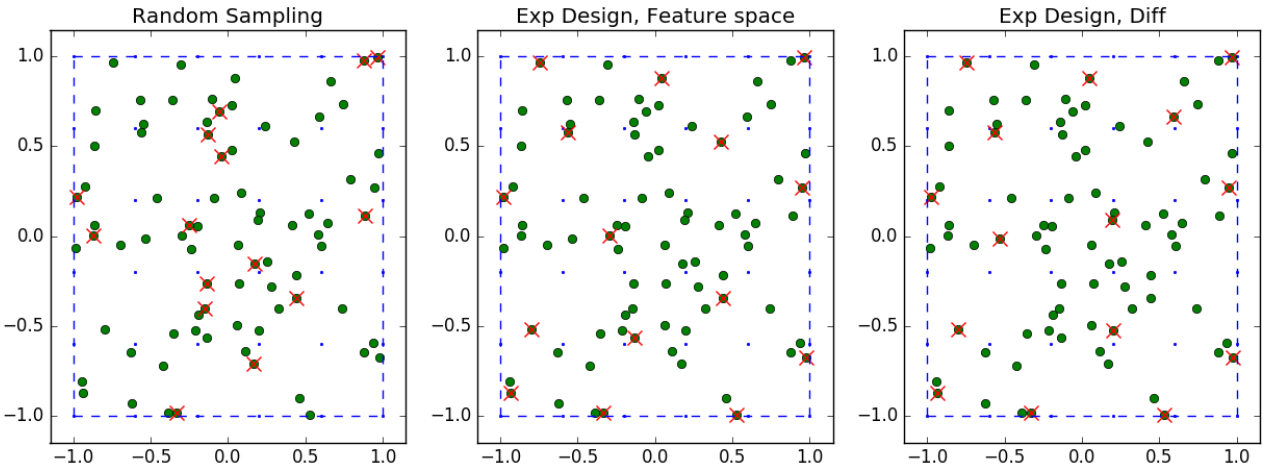
2.5.4 Leveraging Submodular Optimization Techniques

In this section we briefly survey additional results of interest from the literature on submodular maximization. We identified our approximated experimental design problem (2.23) as one of maximizing a submodular function under a cardinality constraint on a subset of $\mathcal{Y}$. With the argument identified as submodular we can define variants of the cardinality constrained problem that may be of interest in applications and retain the efficient approximation property of (2.23).

We briefly mention submodular maximization with matroid constraints [54],[15], where in lieu of (2.23) we solve:

$$Y_0^{\star} = \underset{\mathcal{Y}_0 : \mathcal{Y}_0 \in \mathcal{I}}{\operatorname{argmax}} \tilde{G}(\mathcal{Y}_0) \tag{2.27}$$

and $\mathcal{I}$ is a matroid combinatorial structure [86]. Matroids can concisely capture complicated constraints on $\mathcal{Y}_0$, for example let $\{\mathcal{Y}^i\}$ be a partition of $\mathcal{Y}$, i.e. $\bigcup_i \mathcal{Y}^i = \mathcal{Y}$, $\forall i \neq j$: $\mathcal{Y}^i \bigcap \mathcal{Y}^j = \emptyset$. With the partition in place, a constraint on $\mathcal{Y}_0$ of the form $\mathcal{Y}_0 \bigcap \mathcal{Y}^i \leq K_i$ can be shown to be a matroid constraint of the form $\mathcal{Y}_0 \in \mathcal{I}$. A constraint like this is useful for designing experiments to learn misspecified models where we cannot choose more than a limited number K_i of initial conditions to lie in any specific region $\mathcal{Y}^i$, e.g. due to some physical impediment for repeating experiments with similar conditions. It may be shown that an efficient greedy algorithm can approximate the optimal solution of problems such as those mentioned despite the exact problem being generally NP-hard.



(Left) Random configuration. (Middle) Near-optimal configuration for the feature space design. (Right) Near-optimal configuration for the Krause et al. utility function. The noise levels was set to $\sigma_n^2 = 10^{-3}$ in all figures.

Figure 2-2: Sensor placement configurations.

2.6 Numerical Experiments

In this section we discuss results of numerical experiments validating and demonstrating our techniques.

2.6.1 Comparing Sensor Placement Algorithms

We illustrate the efficacy of our sensor placement approach presented in Section 2.4 by considering a numerical experiment. The setup is a two dimensional GP with a Gaussian RBF kernel $k(x, x') = \exp(-\frac{1}{2}\|x - x'\|^2)$ and measurement noise $\sigma_n^2 = 10^{-3}$. In Figure 2-2 we compare a random sensor placement (left) to the configurations designed by the Krause approach (right) and our feature space approach (middle). Using the notation of Section 2.4.1, $\mathcal{D}$ is outlined with dotted lines, the test set $\mathcal{U}$ is marked with blue dots, the candidate set $\mathcal{Y}$ with green circles and $\mathcal{S}$ with red cross marks. Notice that for this experiment the candidate set $\mathcal{Y}$ was randomly drawn (uniformly over $\mathcal{D}$) to represent a scenario where there is some arbitrary fixed set of possible sampling positions. Evidently, our configuration closely matches the one suggested by the Krause criterion.

Figure 2-3 compares empirical prediction Mean Square Error (MSE) as measured over 500 random draws for GP realizations. For each random draw we collect empirical samples as prescribed by the different methods, and infer process values over a dense set of test points, comparing to the true underlying values. The performance of our approach is virtually identical to that of the Krause method, with reduced computational complexity.

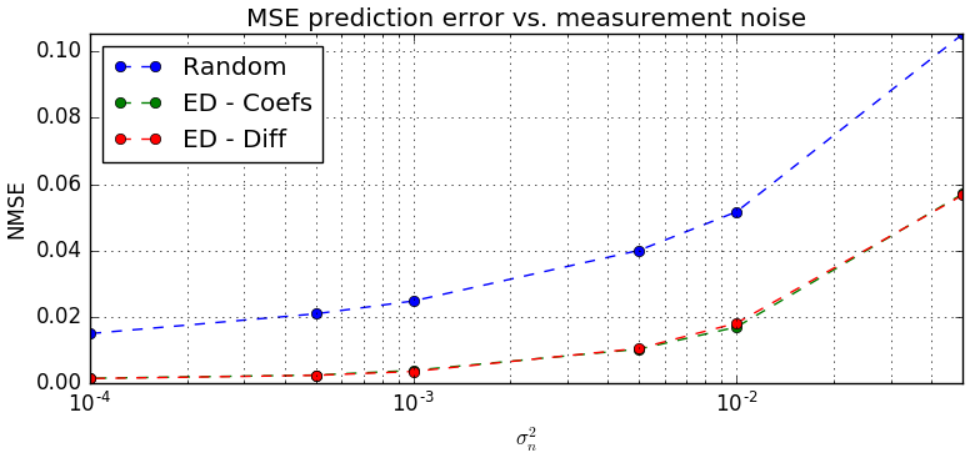


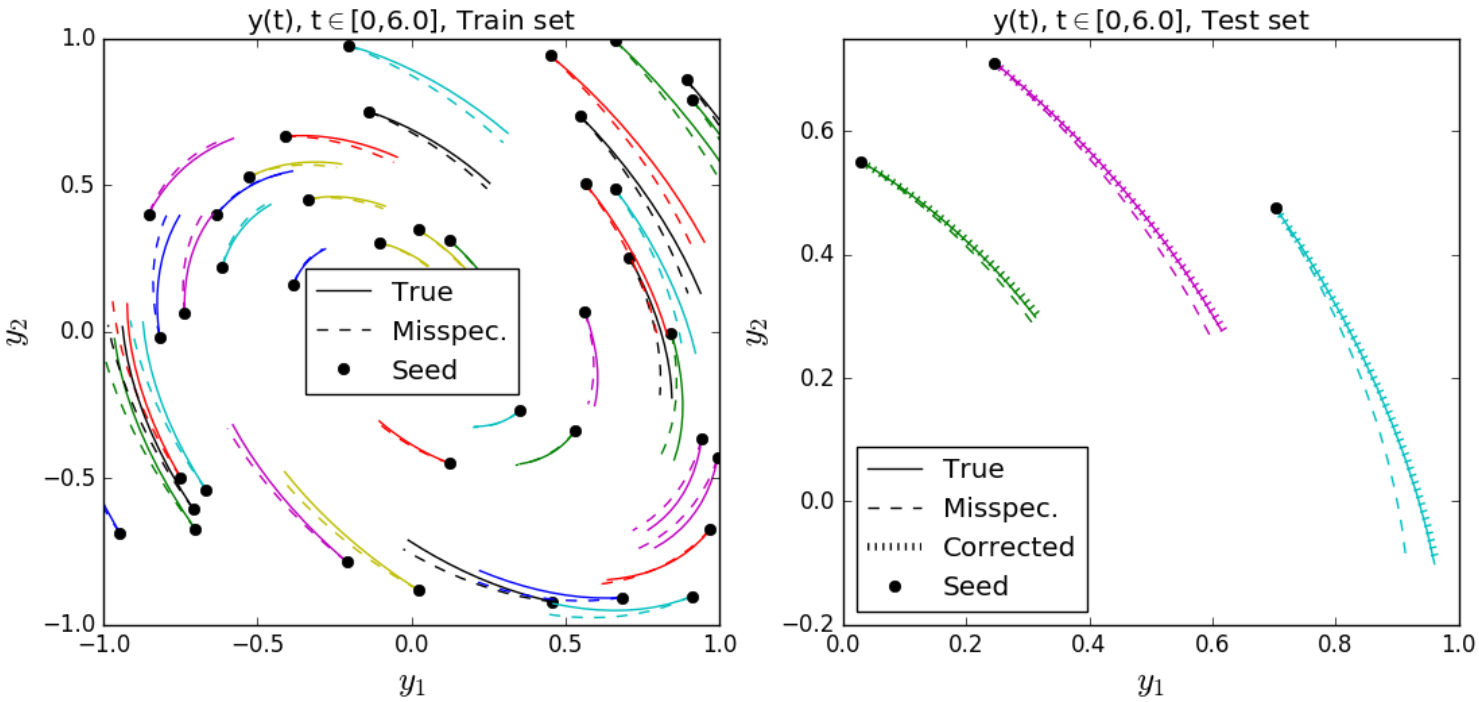
Figure 2-3: Prediction error versus noise level for several sensor configurations.

2.6.2 Correction Term Fitting via GP Regression

For the first experiment we consider a misspecified system in $d = 2$ dimensions where the known component is a fixed linear (matrix) operator, $\mathbf{G}(\mathbf{y}(t)) = \mathbf{A}\mathbf{y}(t)$ with

$$\mathbf{A} = \begin{bmatrix} +0.02 & +0.10 \\ -0.10 & -0.06 \end{bmatrix},$$

and the misspecified component is set according to $\mathbf{F}([y_1, y_2]^\top) = [0.01y_1^2, 0.01y_2^2]^\top$. We observe the system evolution over the time span $t \in [0, 6]$, collecting $T = 11$ equally-spaced time samples per experiment. The sampled time evolution sequences $\mathbf{y}^{(k)}(t)$ were computed exactly, and we have measured noisy samples $\tilde{\mathbf{F}}(\cdot)$ along the evolution path as per the observation model (2.6), where the measurement noise was taken as $\mathbf{\Sigma}_\epsilon = \sigma_\epsilon^2 \mathbf{I}$ with $\sigma_\epsilon^2 = 10^{-4}$.



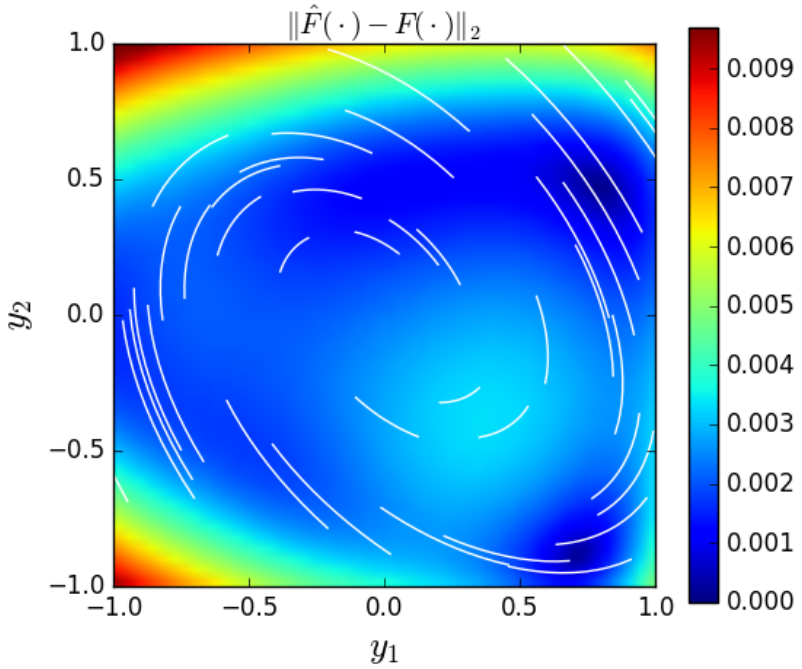
(Left) Training set, with actual evolution in solid lines and misspecified predictions in dashed lines. (Right) Test set, with corrected predictions overlaid.

Figure 2-4: Evolution trajectories for dynamical systems.

Figure 2-4 (left) depicts $K = 40$ trajectories $\mathbf{y}(t)$ (solid lines) induced by a set $\mathcal{Y}_0$ of initial conditions (black dots). Elements $\mathbf{y} \in \mathcal{Y}_0$ were drawn from a uniform distribution over the square $\mathcal{D} = [-1, +1] \times [-1, +1]$. For comparison, we overlay the corresponding trajectories of the misspecified model $\mathbf{y}_G(t)$ taking into account solely the linear driving term $\mathbf{G}(\cdot)$ (dashed lines).

For the GP regression we use a Gaussian kernel with $\sigma_w^2 = 1.0$ scaled for local variance $\frac{1}{|\mathcal{D}|}\iint_{\mathcal{D}} |F_1(\mathbf{y})|^2 d\mathbf{y} = \frac{1}{|\mathcal{D}|}\iint_{\mathcal{D}} |F_2(\mathbf{y})|^2 d\mathbf{y} = 4 \cdot 10^{-5}$. Figure 2-5 depicts the estimation error $\|\hat{\mathbf{F}}(\mathbf{y}) - \mathbf{F}(\mathbf{y})\|_2$ for $\mathbf{y} \in \mathcal{D}$, overlaid with the training sequences. As is evident from these plots the estimation fidelity is high in the regions where training data is readily available.

Finally, in Figure 2-4 (right) the estimated correction term $\hat{\mathbf{F}}(\cdot)$ was used to test prediction performance over some arbitrary set of initial conditions, and compare to the misspecified predicted evolution. The corrected curves (striped lines) are evidently closer to the true paths (solid lines) compared to the misspecified predictions (dashed lines).



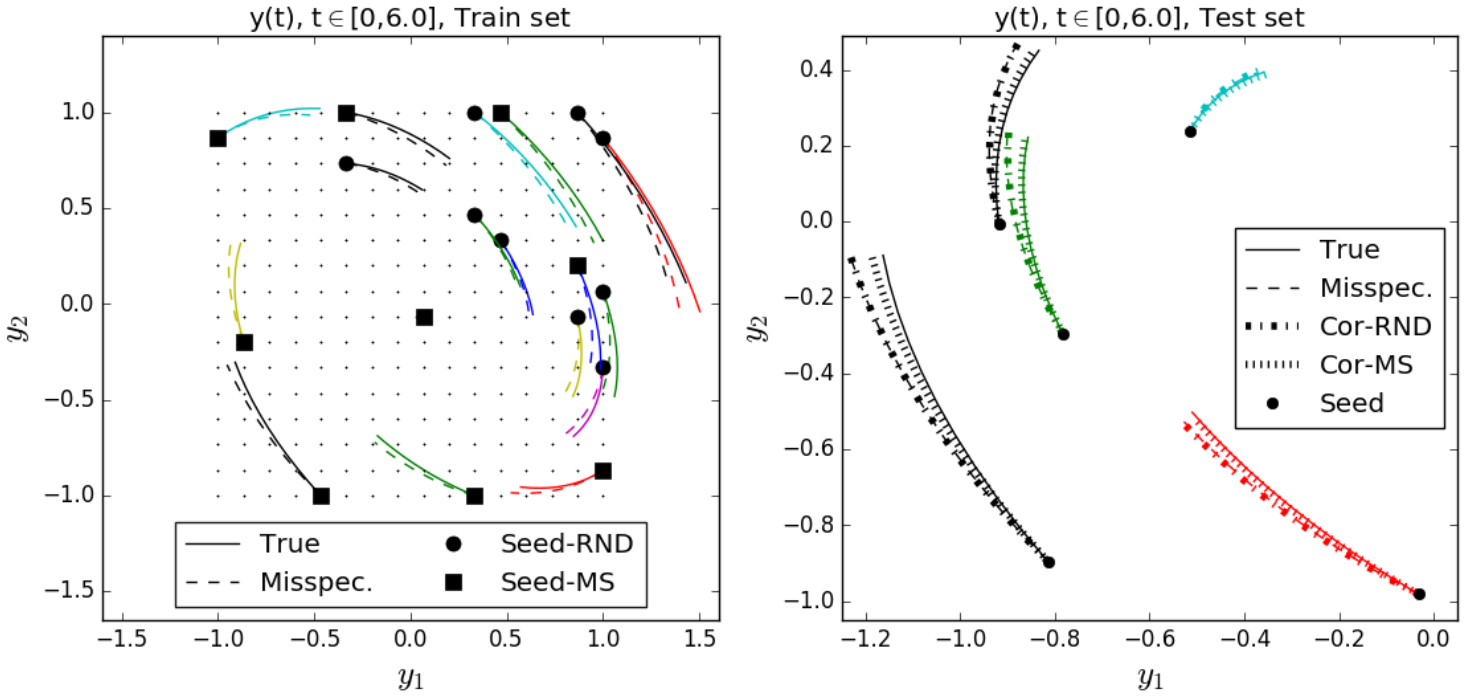
White traces depict the training set time evolution trajectories.

Figure 2-5: Misspecified driving term estimation error map.

2.6.3 Experimental Design for a Dynamical System

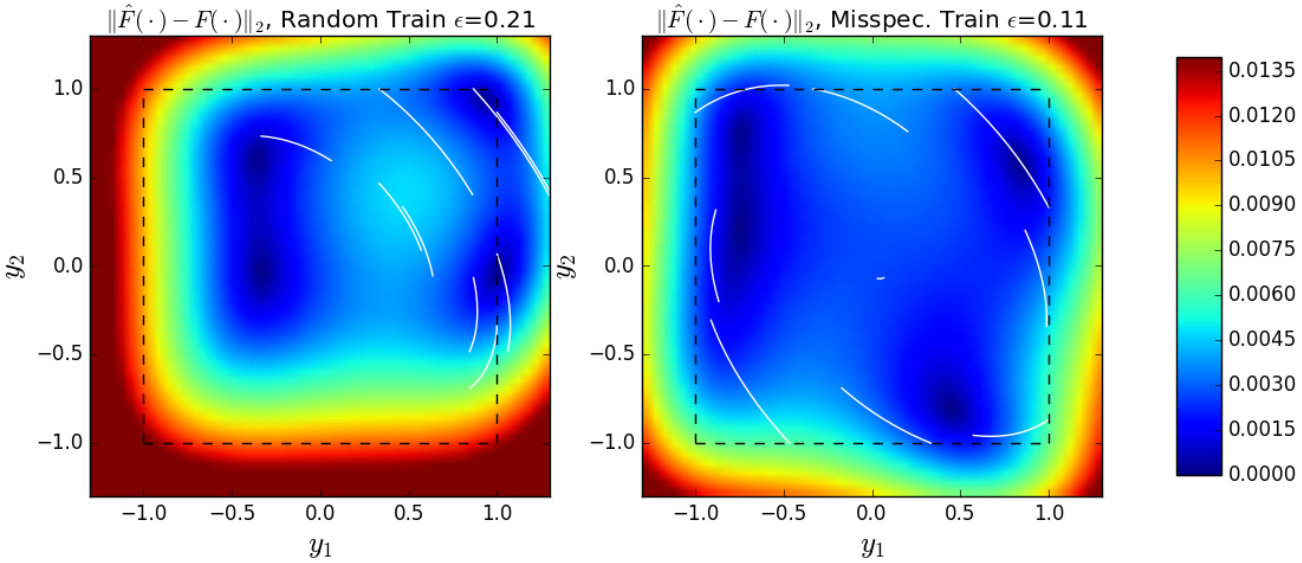
In this subsection we experiment with and implement the experimental design procedures detailed in Section 2.5. We are interested in designing a succession of $K = 9$ experiments. The experimental design entails selecting an optimal set $\mathcal{Y}_0 \subseteq \mathcal{Y}$ of initial conditions from which to start the system off. With the misspecified system as defined in the previous subsection, we take the possible selection set $\mathcal{Y}$ to be a uniformly spaced two dimensional 13×13 grid in $\mathcal{D} = [-1, +1]\times[-1, +1]$ as depicted in Figure 2-6 (left). We implement the lazy greedy algorithm and design an approximately optimal selection set $\mathcal{Y}_0$, marked with black squares in Figure 2-6 (left). Performance is compared to a seed of equal size chosen randomly over $\mathcal{Y}$ marked in black circles. Prediction performance over some arbitrary test set of initial conditions is presented in Figure 2-6 (right) and a heat map for the estimation error in $\hat{\mathbf{F}}(\cdot)$ is plotted in Figure 2-7.

Our next experiment involved changing the training set size, keeping track of estimation performance as measured according to $\int\int_{\mathcal{D}} \|\hat{\mathbf{F}}(\mathbf{y}) - \mathbf{F}(\mathbf{y})\|_2 d\mathbf{y}$ (estimated via numerical



(Left) Training data collected in two setups, first random and second based on designing experiments to match the misspecified dynamics. (Right) Example of prediction test on some arbitrary initial conditions.

Figure 2-6: Experimental design simulations for learning a dynamical system.



(Left) Random initial conditions (Right) Experimental design.

Figure 2-7: Misspecified driving term estimation error map for an experimental design simulation.

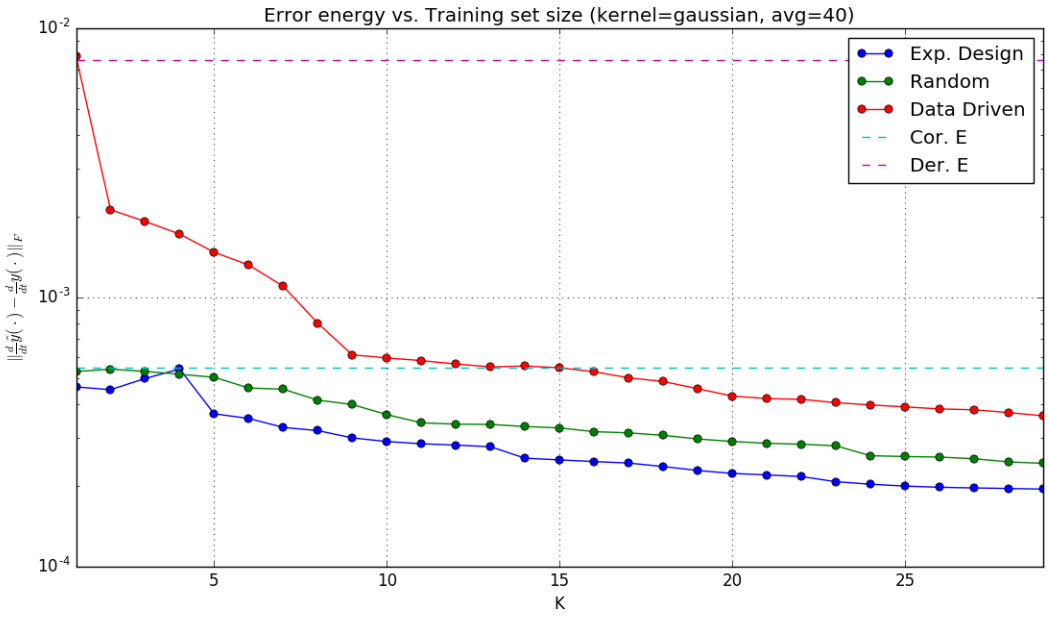


Figure 2-8: Misspecified driving term estimation error versus training set size.

integration). Our dynamical system is as previously described, and we compare several correction strategies as summarized in Figure 2-5.

The first comparison is against a fully data driven estimator, which has no knowledge (not even approximate) of the system dynamics. We use training sequences as determined by our misspecified experimental design procedure but learn the full dynamics by applying GP regression with a Gaussian RBF kernel of scaled power 10^{-2} (due to the higher energy of the unknown function when the entire driving term is to be learned) and estimate the full system dynamics. The two other estimators are the ones previously described, namely estimating just the correction component using the knowledge about the approximate (misspecified) system dynamics, done once with a random seed training set and again with a training set seeded by a choice of initial conditions determined according to our misspecified experimental design procedure.

The results are averaged over 10 realizations of this setup. Also for comparison we show the energy of the correction term $\int\int_{\mathcal{D}} \mathbf{F}(\mathbf{y}) d\mathbf{y}$ and the energy of the entire dynamics term $\int\int_{\mathcal{D}} [\mathbf{F}(\mathbf{y}) + \mathbf{G}(\mathbf{y})] \, d\mathbf{y}$ which quantify the effective error associated with the misspecified and the completely unknowable models.

Evidently the fully data driven approach is always the worst as it ignores the data embedded in the approximated model. However, with increasing number of experiments the difference between this approach and the ones taking into account the approximate dynamics tends to diminish, as the data becomes abundant and no prior assumptions about the model are needed. The approach taking into account the known component in designing the experimental setup is superior as it utilizes all available knowledge. The random training ignoring the known dynamics component incurs a cost in terms of estimation performance compared to the experimental design approach.

2.6.4 A Misspecified Gravitational Field

Experimental design is crucial when the cost of experiments is high. One plausible such scenario is when a misspecified gravitational field is estimated by running controlled experiments of placing an object and observing its free fall (such experiments are likely to be costly). Accurate models of gravitational fields can be useful in planning satellite trajectories around a planet. We use an artificial simplified simulation of the above in which we explore a problem of motion in a two-dimensional gravitational field. If the gravitational field around the planet is fully characterized then this motion can be easily simulated through the laws of mechanics. However, in our setting we assume that the gravitational field is not fully known, in reality this could happen due to e.g. nonuniform mass distribution for the planet or gravitational influence from other nearby heavy masses [80, 93, 94].

Concretely, the two dimensional space is populated with a set of fixed objects, e.g. stars, with the ith object having mass m_i and position $\boldsymbol{x}^i$ and we are interested in solving for the motion of some free-moving unit mass, i.e. a satellite, in the corresponding gravitational field. Let $\boldsymbol{x}(t) = [x_1(t), x_2(t)]^T$ be the coordinate vector of the free-moving unit mass. The equations of motion governing the time evolution of $\boldsymbol{x}(t)$ are prescribed by classical mechanics and given according to [36]:

$$\frac{d^2}{dt^2}\boldsymbol{x}(t) = -\sum_i m_i \frac{\boldsymbol{x}(t) - \boldsymbol{x}^i}{\|\boldsymbol{x}(t) - \boldsymbol{x}^i\|^3} \tag{2.28}$$

This is a second order ODE expressing Newton's second law of motion and the gravitational

field force. Namely, the acceleration experienced by the satellite is equal to the sum of forces acting on it. The force exerted on the satellite by the ith mass is aligned with the vector connecting the two and is directly proportional to m_i and inversely proportional to the squared distance between them.

The second order ODE may be converted into first order form by introducing new variables and defining the transformation

$$[y_1(t), y_2(t), y_3(t), y_4(t)]^\top \equiv [x_1(t), x_2(t), \frac{d}{dt}x_1(t), \frac{d}{dt}x_2(t)] \tag{2.29}$$

In the new variables the equations of motion read:

$$\frac{d}{dt}y_1(t) = y_3(t) \tag{2.30}$$

$$\frac{d}{dt}y_2(t) = y_4(t) \tag{2.31}$$

$$\frac{d}{dt}y_3(t) = -\sum_i m_i \frac{y_1(t) - x_1^i}{\|[y_1(t), y_2(t)]^\top - \boldsymbol{x}^i\|^3} \tag{2.32}$$

$$\frac{d}{dt}y_4(t) = -\sum_i m_i \frac{y_2(t) - x_2^i}{\|[y_1(t), y_2(t)]^\top - \boldsymbol{x}^i\|^3} \tag{2.33}$$

which is a first order system of ODEs as in (2.1).

We consider a known but misspecified model that takes into account a single fixed mass in the origin with $m_1 = 0.2$ and $\boldsymbol{x}^1 = [0, 0]^\top$. The true model however includes two additional masses $m_2 = 0.1, m_3 = 0.4$ and $\boldsymbol{x}^2 = [0, 4]^\top, \boldsymbol{x}^3 = [0.5, 3.8]^\top$. With these symbols, we have

$$\mathbf{G}(\mathbf{y}(t)) = \left[y_3(t), y_4(t), \frac{-m_1(y_1(t) - x_1^1)}{\|[y_1(t), y_2(t)]^\top - \boldsymbol{x}^1\|^3}, \frac{-m_1(y_2(t) - x_2^1)}{\|[y_1(t), y_2(t)]^\top - \boldsymbol{x}^1\|^3}\right]^\top \tag{2.34}$$

$$\mathbf{F}(\mathbf{y}(t)) = \left[0, 0, \sum_{i=2,3} \frac{-m_i(y_1(t) - x_1^i)}{\|[y_1(t), y_2(t)]^\top - \boldsymbol{x}^i\|^3}, \sum_{i=2,3} \frac{-m_i(y_2(t) - x_2^i)}{\|[y_1(t), y_2(t)]^\top - \boldsymbol{x}^i\|^3}\right]^\top \tag{2.35}$$

For this experiment the signal $\mathbf{y}(t)$ is 4 dimensional such that at any moment it captures the location as well as vector velocity of the satellite. Similarly, initial conditions are specified in this four dimensional space.

We limit our attention to correction functions of the functional form $\mathbf{F}([y_1, y_2, y_3, y_4]) =$

$[0, 0, \mathbf{F}_{3,4}([y_1, y_2])]^\top$, i.e. the gravitational field correction is strictly a function of the spatial coordinates (y_1, y_2), and has only two unknown components. We thus consider the problem of estimating $\mathbf{F}_{3,4} : \mathbb{R}^2 \to \mathbb{R}^2$, and our results and techniques naturally carry over to this scenario.

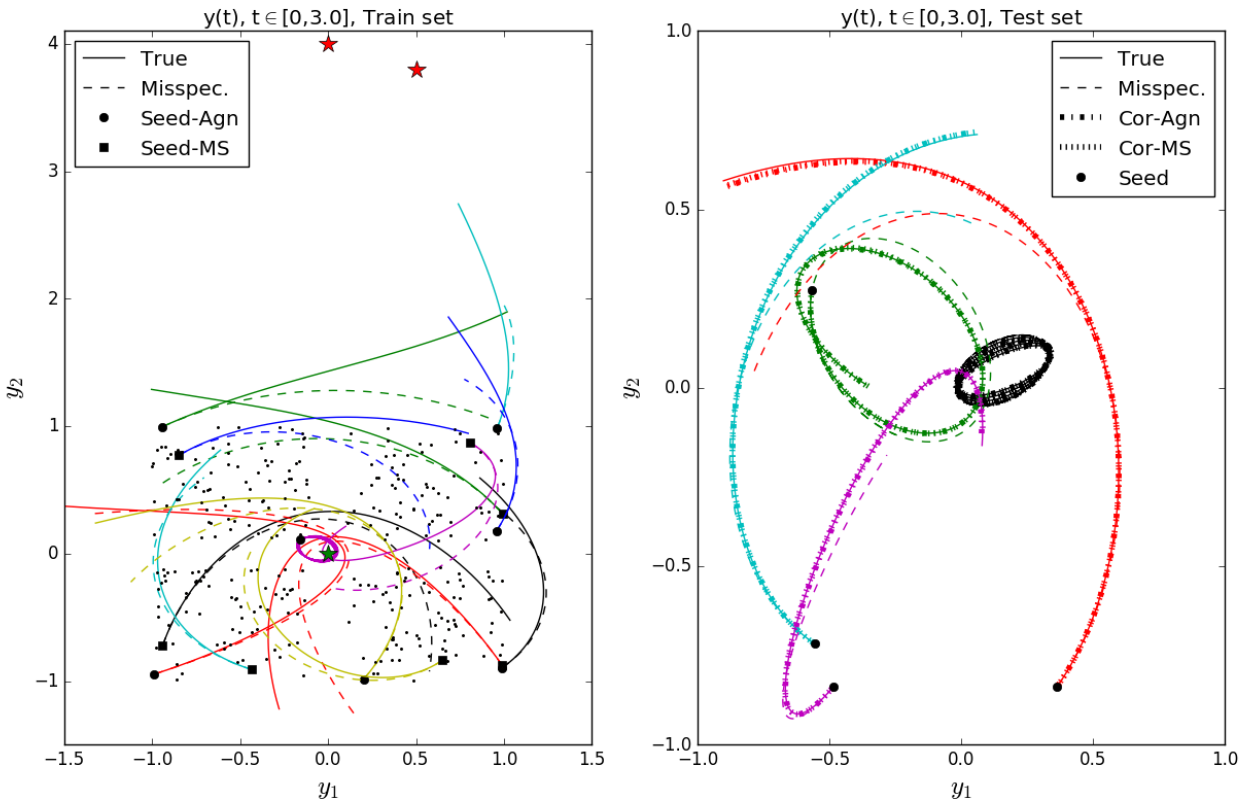
For the kernel we use a Gaussian RBF with $\sigma_k^2 = 1.0$ scaled for local variance 10^{-3} and the measurement noise is $\mathbf{\Sigma}_\epsilon = 10^{-4}\mathbf{I}$. Experiments run in the time frame $t \in [0, 3.0]$ and $T = 20$ data samples are collected per experiment. The selection set $\mathcal{Y}$ is a set of size $|\mathcal{Y}| = 300$ of initial conditions, whose spatial coordinates (y_1, y_2) are depicted in Figure 2-9 (left) in addition to the mass configuration in space. Also shown are training sets of size $K = 7$ as selected via an agnostic experimental design procedure and a misspecified aided one. In Figure 2-9 (right) we showcase prediction performance on a random test set. Both the agnostic and the misspecified designs perform well here compared to the misspecified predictions.

Figure 2-10 plots the estimation error of $\hat{\mathbf{F}}_{3,4}(\cdot)$ for the setup above for the agnostic design (left) and the misspecified guided design (right) which performs slightly better when compared according to the mean squared error over the domain of interest $\mathcal{D}$ delineated inside the dashed line.

Finally in Figure 2-11 we compare the mean square error for the two methods as a function of K, as determined empirically by averaging the results of 400 noise realizations. For reference, the dashed red line depicts the mean energy in the unknown term $\mathbf{F}_{3,4}(\cdot)$.

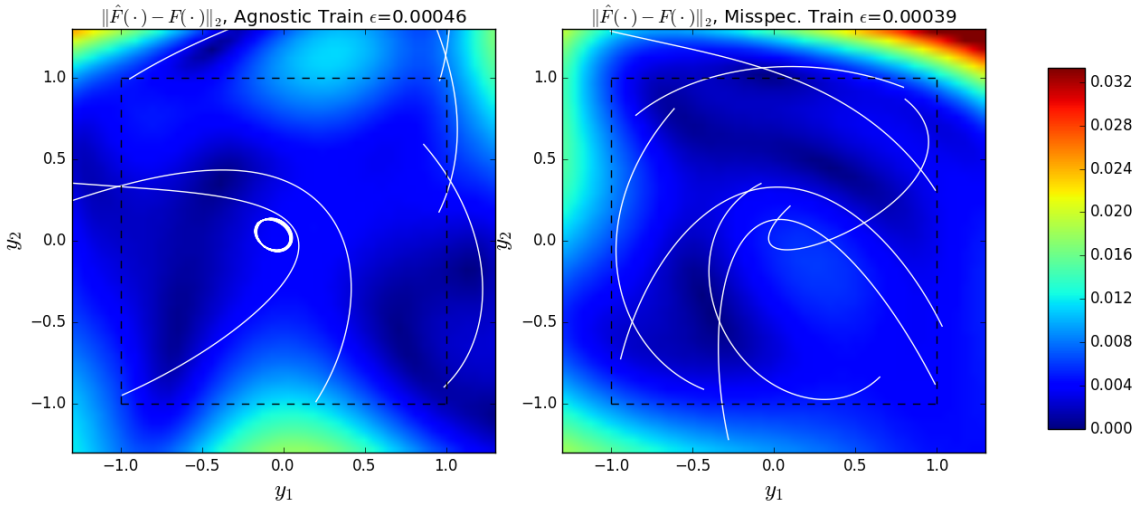
2.7 Discussion

We have introduced a flexible Gaussian Process based formalism for expressing misspecified models for dynamical systems, and a corresponding technique for making inference and learning the misspecified dynamics based on empirical data collected from system evolution sequences. We formulated a corresponding optimal experimental design problem as one of choosing informative initial conditions that facilitate rapid learning of the system, and suggested an efficient algorithm with guarantees to find approximate such designs under an experimental budget constraint.



(Left) Training sets as determined via an agnostic approach and a misspecified aided approach. (Right) Predictions over a random test set.

Figure 2-9: Experimental design simulations in a misspecified gravitational field.



(Left) Agnostic choice of training set (Right) Misspecified aided design.

Figure 2-10: Misspecified driving term estimation error map for an experimental design simulation.

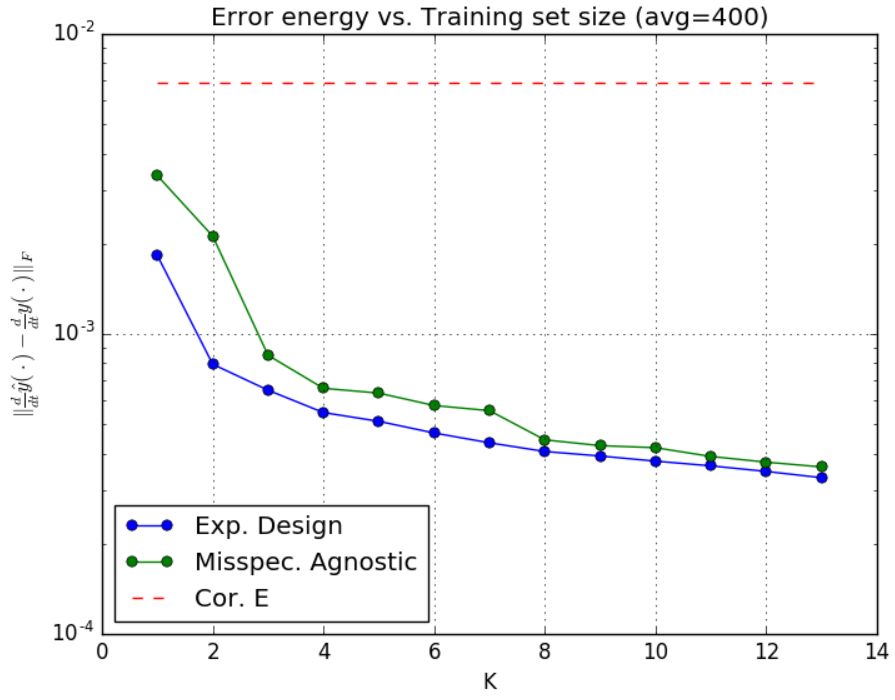


Figure 2-11: Misspecified driving term estimation error versus training set size.

Several aspects of our work may be extended. We leave the following ideas and directions for future research. In this study, we have assumed that empirical data is collected only after experimental design has been performed. However, in various configurations, it is possible to consider an online adaptive experimental design formulation, where sequential predictions are made based on past observations. While one can consider a setting in which the aforementioned design process is being re-executed following each observation (with updated knowledge), such approach may be sub-optimal. Recent studies have been considering approaches such as dynamic programming in the context of Bayesian optimization, to devise experimental design in a less myopic fashion [89, 66]. On another matter, in the current study, the design space involved a discrete lattice of prospective seed coordinates (initial conditions starting points). Alternative, spatially continuous parametrization of the seeding points, may be more appropriate in other circumstances, and may enable harnessing scalable, continuous optimization strategies for determination of the initial states. While we attempted to generalize the functional form of the correction model by the utilization of a Gaussian Process as a generic form of model correction, the overall relationship of the correction term to the misspecified model is still in the form of an additive term. This popular choice may be appropriate for a broad range of applications, but obviously, for others, more sophisticated forms could be considered.

Chapter 3

Antenna Array Design

In this chapter we consider the problem of far-field sensing by means of an antenna array. Traditional array geometry design techniques are agnostic to prior information about the far-field scene. However, in many applications such priors are available and may be utilized to design more efficient array topologies.

We formulate the problem of array geometry design with scene prior as one of finding a sampling configuration that enables efficient inference in the D-Bayes optimality sense, which can be cast in the form of a combinatorial optimization problem. While generic combinatorial optimization problems are NP-hard and resist efficient solvers, we show how for array design problems the theory of submodular optimization may be utilized to obtain efficient algorithms that are guaranteed to achieve solutions within a constant approximation factor from the optimum.

We leverage the connection between array design problems and submodular optimization and apply several results of interest. We demonstrate efficient methods for designing arrays with constraints on the sensing aperture, as well as arrays respecting combinatorial placement constraints. We further show that the problem of designing far field arrays operating at multiple wavelengths can naturally be expressed using our formulations, and we consider two relevant design paradigms. The first design paradigm is optimized for collection of measurements at multiple wavelengths, fusing these together for joint inference over an underlying scene. The second design paradigm is robust, in a sense that it is guaranteed to allow good inference over the scene at any one single wavelength at a time. We showcase

designs of arrays under both paradigms utilizing simple greedy selection algorithms, and state-of-the-art robust submodular maximization algorithms.

The novel connection between array design and submodularity opens the door for utilizing other insights and techniques from the growing body of literature on submodular optimization in the field of array design.

3.1 Introduction

Sensor arrays for spatial sensing are deployed in a wide range of applications including radar, sonar, medical imaging and radio astronomy and there is a vast literature on the topics of array design and array processing from the last century [113, 61].

A major goal in designing arrays is efficiently meeting required specifications with a limited budget of sensing elements, which are often a main determinant of system cost, size, weight and complexity. However, even in the single wavelength case, the design of array geometries is a notoriously hard task, and many applications simply utilize a uniform truncated half-wavelength design, or restrictions thereof. Indeed, the problem of designing non-uniform arrays hints at combinatorial optimization and is computationally hard as we discuss later.

Various studies tackle the problem of non-unifom array design directly. In beamforming arrays attaining a desired resolution level is often a primary concern, achieved by means of manipulating beam pattern parameters such as lobe widths and positions, and the problem of designing efficient array geometries that facilitate desired beam patterns has been widely studied in the past. Techniques involving array thinning start with a dense uniform geometry, removing elements while maintaining performance within specified bounds [41]. Other approaches consider methods such as swarm optimization [51], dynamic programming [104], genetic algorithms [42], inversion [63] and Bayesian compressive sampling [84].

In applications of estimating direction of arrival, other specialized techniques have emerged for finding efficient array designs such as optimizing the corresponding Cramer-Rao error bounds [33], or designing according to the nested array methodology for increasing the available number of degrees of freedom [112].

In virtually all the design methodologies some assumptions on the scene of interest are made. These represent beliefs, constraints or knowledge that hold over the unobserved scene. A favorable design is one meeting requirements taking into account these assumptions. For example, in direction of arrival applications we may assume some limit on the number of point targets present [33]. In beamforming we assume some separation level between objects of interest [113] that necessitates a certain resolution level, or some scene sparsity structure [59, 60].

In this chapter we study the problem of inference on a scene of interest through measurements collected at a sensor array. The approach we take for designing array geometries is novel in that we consider settings where some Bayesian prior on the scene is available at the time of design. Frequently, the same device is used to sense multiple similar scenes, where past examples are indicative of future ones. A medical imaging device, for instance, is typically used to image the same organ across different patients. In other cases, we may have prior knowledge in the form of scene properties such as smoothness or adherence to spatial constraints. We incorporate such knowledge as a prior in a Bayesian model and propose exploiting this knowledge and adapting the geometry accordingly to achieve efficient inference with a limited budget for sensors. In this Bayesian setting, sensing the scene is just performing inference in the model, and the problem of array geometry design asks to select a geometry that optimizes the quality of inference.

We show how quantifying inference quality through the D-Bayes optimality criterion [17, 8] results in a cost function for the array geometry design problem that holds the property of submodularity [83]. A submodular set function is one that exhibits diminishing marginal gains, i.e. adding additional elements results in diminishing benefit. Recently, there has been significant progress on the theory of optimizing submodular functions [31, 13, 118, 79]. In particular, these results state that, while NP-hard, submodular maximization admits variants of greedy algorithms that are guaranteed to achieve near-optimal solutions, i.e. within a constant factor. Submodularity has been used in connection with sensor placement problems, for example in [58, 99, 55], however these works are not tailored to the far field scenes and models that we focus on here.

Importantly, we address the topic of designing sensor arrays for multiple wavelength

sensing applications. Arrays operating at multiple wavelengths have been studied in various contexts such as wideband direction of arrival estimation problems [119, 120], multi-frequency synthesis in Astronomy [96], and designing wideband array patterns [97]. We consider two design paradigms for the multiple wavelength setting. The first design paradigm is optimized for collection of measurements at multiple wavelengths, fusing these together for joint inference over an underlying scene. The second design paradigm is robust, in a sense that it is guaranteed to allow good inference over the scene at any one single wavelength at a time.

Our connections and formulations open avenues for leveraging those results for efficient array design with strong guarantees. We demonstrate this by showcasing the design of array geometries under both paradigms in settings with arbitrary apertures. Together with our new formulations, the exploitation of prior knowledge leads to higher quality inference at lower cost in terms of the number of sensing elements.

3.2 Classic Antenna Array Design

In this section we briefly review classic antenna array design theory and introduce the main concepts and notation that will be useful for the rest of our discussion.

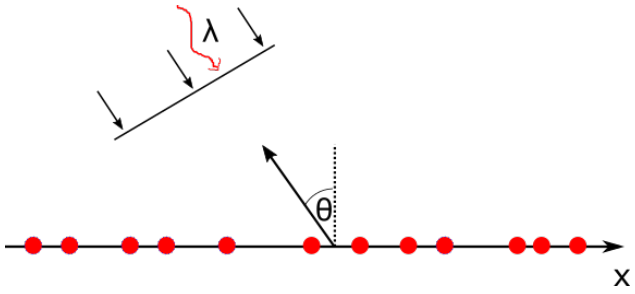
3.2.1 Antenna Array Setup

Antenna arrays are used for forming images and estimating properties of distant radiation-emitting scenes. For simplicity we consider throughout a simplified one dimensional setting although our techniques and results transfer to higher dimensional settings in a straightforward way. We assume coherent narrowband emission around a central wavelength λ and frequency $f = 2\pi\omega$[1] impinging on an observation axis x, where antenna sensing elements are purposefully placed as illustrated in Figure 3-1. The radiation-emitting scene is distributed along the θ axis. We will focus on the *far-field* setting where the scene is so distant from the observation plane x such that by the time the impinging spherical waves arrive at the observation plane they appear as planar waves as illustrated in the Figure. Let D be the spatial extent of the array (i.e. the distance between the two extreme antennas), then the

[1] We always have $f = \frac{c}{\lambda}$ with c the wave velocity, which is the speed of light for electromagnetic radiation.

Radiation frequency	Far-field cutoff distance
1Mhz	0.003m
1Ghz	3.33m
10Ghz	33.3m

Table 3.1: Far field threshold distance versus radiation frequency.



Antenna sensing elements are depicted in red.

Figure 3-1: Far-field sensing.

far-field cut-off distance is $\frac{D^2}{\lambda}$, beyond which we can safely assume far field conditions [113]. We tabulate a few example far-field cutoff distances for a $D = 1$m array vs. the EM radiation frequency in Table 3.1.

3.2.2 Radiation Propagation Model

The far field scene can be characterized by a single illumination function $\tilde{\beta}(\theta)$ tracking the radiation amplitude impinging as a function of θ. Figure 3-2 depicts a plane wave arriving from azimuth θ with respect to the array axis. The plane wave assumption implies that the waveform everywhere on the x axis is identical in amplitude but may be differing in phase depending on the varying paths the plane wave propagates through before impinging on each position. As illustrated in the figure, the additional path corresponding to location x is $x\sin(\theta)$. When a coherent wave propagates a distance equal to its wavelength λ it accrues a 2π phase. Over the additional distance $x\sin\theta$ the phase accrual is thus $2\pi \cdot \frac{x\sin\theta}{\lambda}$.

integrating over $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$ and accounting for the different amplitude impinging from

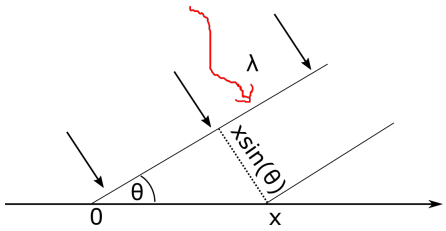


Figure 3-2: Phase accrual by a plane wave impinging on the observation axes.

each azimuth we retrieve the total measurement recorded at location x:

$$r(x) = \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \tilde{\beta}(\theta) e^{j\frac{2\pi}{\lambda} x \sin\theta} \cos\theta d\theta = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \beta(\psi) e^{j\frac{4\pi}{\lambda} x\psi} d\psi \tag{3.1}$$

where we have defined the normalized azimuth parameter $\psi \equiv \frac{1}{2}\sin\theta$ such that $-\frac{1}{2} \leq \psi \leq \frac{1}{2}$ and $\beta(\psi) \equiv \tilde{\beta}(\sin^{-1}(2\psi))$.

3.2.3 Array Topologies and Performance

Our ultimate goal is to retrieve the illumination function $\beta(\psi)$ from measurements $r(x)$. However, we usually do not have access to the function $r(x)$ for all x but rather just at those x values where an antenna is placed and is able to collect measurements which we take as $\{x_n\}$, i.e. we have access to the corresponding set of samples $\{r(x_n)\}$.

Equation (3.1) bears close resemblance to the definition of the Fourier transform. Namely, defining $f(t)$ to be the Fourier transform of $\beta(\psi)$

$$f(t) \equiv \int \beta(\psi) e^{j2\pi t\psi} d\psi \tag{3.2}$$

we immediately identify, comparing (3.1) and (3.2):

$$r(x) = f(\frac{2}{\lambda}x) \tag{3.3}$$

Using (3.3) we see that the set of samples $\{r(x_n)\}$ is equivalent to the set of samples $\{f(\frac{2}{\lambda}x_n)\}$. Notice that since the support of $\beta(\psi)$ is restricted to $\psi \in [-\frac{1}{2}, +\frac{1}{2}]$ we im-

mediately have through the Fourier transform definition (3.2) that $f(t)$ is band-limited with bandwidth 1. Thus, as a consequence of the Whittaker-Kotelnikov-Shannon sampling theorem we have that perfect reconstruction of $f(t)$, and subsequently the scene $\beta(\psi)$, is possible from an infinite set of samples taken at $t \in \{\ldots, -1, 0, +1, \ldots\}$, which corresponds to placing infinite antenna elements at $x_n = n\frac{\lambda}{2}$ which is the celebrated $\frac{\lambda}{2}$ array facilitating perfect scene reconstruction [113], see illustration in Figure 3-3.

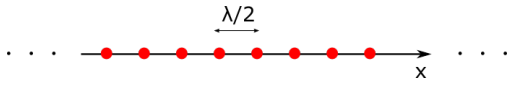


Figure 3-3: The uniform $\frac{\lambda}{2}$ array, enabling perfect scene reconstruction.

For other array configurations, perfect scene reconstruction is not always possible[2]. Traditionally, arrays are characterized through their beam pattern which for our setting may be interpreted as how reconstruction of a delta function (i.e. point target) scene would appear. Reconstruction of arbitrary scenes then results as a superposition of such shifted beam patterns. For an array with antennas positioned at $\{x_n\}$ and each antenna output is scaled with gain w_n the resulting beampattern is given according to

$$B(\psi) = \sum_i w_i e^{j\frac{4\pi}{\lambda}x_n\psi} \tag{3.4}$$

For a uniform gain finite $\frac{\lambda}{2}$ array with N elements this reads:

$$B(\psi) = \frac{1}{N}\frac{\sin(N\pi\psi)}{\sin(\pi\psi)} \tag{3.5}$$

This pattern is illustrated for finite arrays of $N = 10, 100$ elements in Figure 3-4. As N grows this approaches a delta function, but for finite N this curve is usually characterized through such parameters as the main lobe width, its distance to the secondary lobe and the attenuation of these secondary lobes with respect to the main lobe.

[2]It may be shown that any array configuration with an infinite number of antennas of average spacing $\frac{\lambda}{2}$ allows perfect reconstruction, although this may be very unstable. Alternatively, with less dense arrays perfect reconstruction is not possible [127].

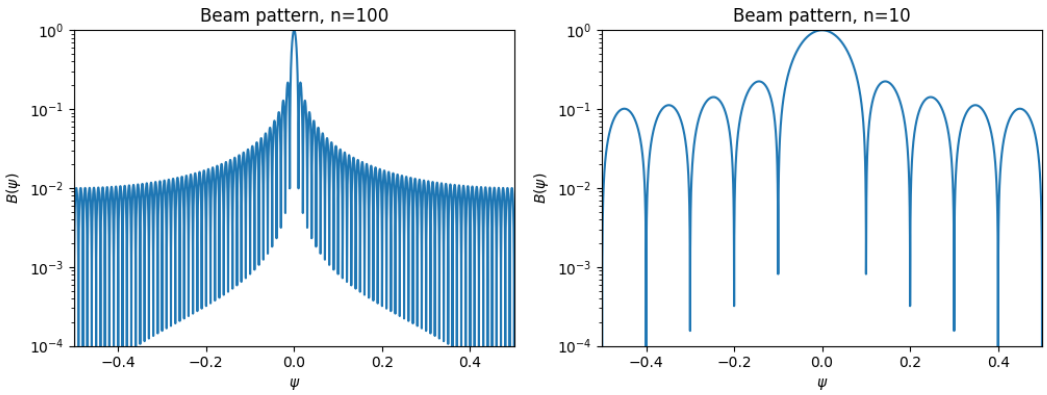


Figure 3-4: Finite $\frac{\lambda}{2}$ arrays beam patterns.

3.3 Problem Formulation

In this section we formulate the array design problem. We focus on far-field sensing applications, although, as will become apparent, the same techniques could also be generalized to other settings where the measurement process is linear. For simplicity we consider a one dimensional setting (the extension to multiple dimensions is straightforward). We begin with a review of the far-field sensing model to establish notation, and then pose the sensing problem as one of Bayesian inference.

3.3.1 Far-field Sensing

The far field sensing setup is depicted in Fig. 3-1 and Eq. (3.1). A scene of interest is located far from an observation axis x and is characterized through $\beta(\psi)$.

We are to place Q sensors along the observation axis at positions $\mathcal{S} \equiv \{x_0, \ldots, x_{Q-1}\}$, and collect measurements at wavelengths $\Lambda \equiv \{\lambda_0, \ldots, \lambda_{L-1}\}$. We aim to choose an optimal set S satisfying some constraints. We take A to be a finite selection set of possible positions (e.g. a finite grid on some section of the real line), and pose the constraint $S \subseteq A$.
The noiseless reading $r(x_q; \lambda_l)$ taken at position $x = x_q$ and wavelength $\lambda = \lambda_l$ is given according to (3.1):

$$r(x_q; \lambda_l) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \beta(\psi) e^{j\frac{4\pi}{\lambda_l} x_q \psi} d\psi \tag{3.6}$$

where we assumed that the illumination function is wavelength-invariant such that the scene appears identical when probed at different wavelengths $\lambda \in \Lambda$.

We take into account the effect of noise by introducing $\tilde{\boldsymbol{f}} = [\tilde{f}_0, \ldots, \tilde{f}_{N-1}]^\top$, a vector of $N \equiv QL$ noisy measurements modeled according to:

$$\tilde{f}_{q+Ql} \equiv r(x_q; \lambda_l) + w_{q+Ql} \qquad \begin{matrix} q = 0, \ldots, Q-1 \\ l = 0, \ldots, L-1 \end{matrix} \tag{3.7}$$

where w_n is additive noise. Stacked in $\boldsymbol{w} = [w_0, \ldots, w_{N-1}]^T$, we assume throughout that the noise is complex, circular, Gaussian $\boldsymbol{w} \sim \mathcal{CN}(0, \boldsymbol{\Sigma}_{ww})$ [81], i.i.d. across measurements, i.e., $\boldsymbol{\Sigma}_{ww} = \sigma_w^2 \boldsymbol{I}_N$, where $\boldsymbol{I}_N$ is an $N \times N$ identity matrix such that the noise is i.i.d. across different sensors.

3.3.2 Setting a Prior

The sensing problem we consider here entails the estimation of the illumination function $\beta(\psi)$ from the set of noisy measurements $\tilde{\boldsymbol{f}}$. Even in the noiseless setting this problem is gravely ill-posed as infinitely many wildly varying scenes map to any given finite set of observations[3].

To cope with this ill-posedness, some prior belief or knowledge pertaining to $\beta(\psi)$ (or equivalently its scaled Fourier transform $f(t)$) must hence be incorporated into the model, and this could be achieved in several ways. Wingham [125] proposed selecting one specific $f(t)$ of the multiple such functions consistent with the samples, namely the minimum norm solution. Alternatively, some constraints or other preferences may be imposed on the solution by penalizing the inversion. For example one may require the solution to satisfy some constraints (e.g. lie in some pre-specified sub-space of the function space), or impose regularization (e.g. on smoothness, or total variation) [23].

In what follows, we take a Bayesian approach and impose a prior on the scene $\beta(\psi)$. Subsequently, sensing is equivalent to performing inference in this model. The prior may be assigned based on past observations over the distribution of scenes or based on a-priori

[3]The mapping between $\beta(\psi)$ and a finite set of its Fourier transform samples $r(x_q; \lambda_l)$ is not bijective [127].

knowledge of scene properties as we discuss next. We consider two approaches for assigning a prior on the continuous function $\beta(\psi)$. The first involves frequency space representation and the second utilizes the Gaussian Process (GP) formulation. The choice for which to use (or to use a different prior) depends on the application and specific knowledge we have of the function in the case at hand.

Frequency Space Representation

Assigning a prior on $\beta(\psi)$ can be simplified if $\beta(\psi)$ may be expanded in a countable basis of functions such that the prior is imposed in the discrete domain of expansion coefficients. With $\beta(\psi)$ having constrained support in $|\psi| \leq \frac{1}{2}$, we can expand it by means of Fourier basis functions $\left\{e^{j2\pi m\psi} | m \in \mathbb{Z}\right\}$ in that domain [85]:

$$\beta(\psi) = \sum_m \beta_m e^{j2\pi m\psi}, \qquad \beta_m \equiv \int_{-\frac{1}{2}}^{+\frac{1}{2}} \beta(\psi) e^{-j2\pi m\psi} d\psi \tag{3.8}$$

where $\{\beta_m\}$ are the Fourier expansion coefficients, and the usual Parseval relation holds:

$$\int |\beta(\psi)|^2 d\psi = \sum_m |\beta_m|^2 \tag{3.9}$$

In lieu of the prior on $\beta(\psi)$ we impose a prior over $\{\beta_m\}$. This description is especially suited for applications involving smooth functions $\beta(\psi)$ as suggested by the following property of the Fourier series expansion [30]:

Lemma 3.1. *Let $\beta(\psi) \in C^r$ where C^r is the space of r-times continuously differentiable functions over some domain. Then $|\beta_m| \leq \frac{\alpha}{|m|^r}$ with $\alpha = \sup_\psi |\frac{\partial^r}{\partial \psi^r} \beta(\psi)|$. More generally, from the Riemann-Lebesgue lemma, Let $\beta(\psi)$ be any integrable function, then $|\beta_m| \overset{|m|\to\infty}{\longrightarrow} 0$.*

Thus, for any nicely behaved $\beta(\psi)$ the high frequency Fourier expansion coefficients diminish polynomially to zero, with an asymptotically polynomial rate determined by the level of smoothness, allowing good approximate representation through a finite subset of low frequency coefficients, which in the sequel we designate via the finite vector $\boldsymbol{\beta}$.

In the sequel we use Gaussian priors on the coefficients $\{\beta_m\}$:

$$\beta_m \sim \mathcal{CN}(0, \sigma_m^2) \tag{3.10}$$

where β_m are independent, complex, circular and Gaussian, and σ_m^2 are the corresponding variances. Using (3.9) we define the expected scene power:

$$P \equiv \mathbb{E} \int |\beta(\psi)|^2 d\psi = \sum_m \sigma_m^2 \tag{3.11}$$

The $\{\sigma_m^2\}$ can be set following some initial measurements of sample functions $\beta(\psi)$ or taking into account prior knowledge. For example if we have a-priori knowledge that $\beta(\psi) \in C^r$ we may use

$$\sigma_m^2 = \begin{cases} 1 & m = 0 \\ \frac{1}{m^{2r}} & m \neq 0 \end{cases} \tag{3.12}$$

which is a very simple distribution respecting the polynomial variance decay. For the rest of the chapter we adopt the prior in (3.12) and take $r = 1$ to promote continuously differentiable functions.

Observation Model With the prior stated in the discrete $\boldsymbol{\beta}$ domain as described above, our next goal is to circumvent $\beta(\psi)$, directly stating the problem in terms of the measurements $\tilde{f}_n$ and the coefficients β_m, replacing the continuous representation with a discrete counterpart. Substituting (3.8) into (3.6) we have (where $n = q + Ql$):

$$r(x_q; \lambda_l) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \sum_m \beta_m e^{j2\pi m\psi} e^{j\frac{4\pi}{\lambda_l} x_q \psi} d\psi = \sum_m K_{nm} \beta_m \tag{3.13}$$

where

$$K_{nm} \equiv \int_{-\frac{1}{2}}^{+\frac{1}{2}} e^{j2\pi(\frac{2}{\lambda_l} x_q + m)\psi} d\psi = \operatorname{sinc}(m + \frac{2}{\lambda_l} x_q) \tag{3.14}$$

and $\text{sinc}(x) \equiv \frac{\sin(\pi x)}{\pi x}$. Plugging this into (3.7) we retrieve the observation model

$$\tilde{f}_n = \sum\nolimits_m K_{nm}\beta_m + w_n \qquad\qquad n = 0, \ldots, N-1 \tag{3.15}$$

and the sensing problem amounts to estimating the coefficients $\boldsymbol{\beta}$ given the observation vector $\tilde{\boldsymbol{f}}$. As we have assumed Gaussian distributions throughout, the posterior $\mathbb{P}(\boldsymbol{\beta}|\tilde{\boldsymbol{f}})$ is Gaussian with a convenient analytic form. We detail this in Section 3.4.3.

Gaussian Process Prior[4]

An alternative description for the prior of the scene utilizes Gaussian Process (GP) statistics, according to $\psi \in [-\frac{1}{2}, +\frac{1}{2}]$: $\beta(\psi) \sim \mathcal{GP}(m(\psi), k(\psi, \psi'))$ with $m(\cdot) : \mathbb{R} \to \mathbb{R}$ the mean function, which we will take without loss of generality to be identically zero, and $k(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ the kernel function (Appendix B). A natural choice for the Kernel function in this setting, capturing the smoothness of the scene, is the Gaussian RBF $k(\psi, \psi') = \exp(-\frac{1}{2\sigma_f^2}\|\psi - \psi'\|^2)$ with σ_f^2 controlling the level of smoothness. The associated observation model is expressed in terms of the statistics as we detail next.

Observation Model In the GP setting, the joint statistics of any finite collection of scene points $\{\beta(\psi_1), \ldots, \beta(\psi_{M-1})\}$ and samples $\left\{\tilde{f}_0, \ldots, \tilde{f}_{N-1}\right\}$ is Gaussian, by virtue of the GP Gaussian statistics, and (3.6), (3.7) expressing the measurements $\tilde{f}_i$ as linear projections of $\beta(\psi)$ and the noise statistics as Gaussian. Thus, given a finite set of observations we can estimate any finite set of scene values by performing linear Gaussian inference (Appendix B). Concretely, to specify the Gaussian statistics of joint collections of scene point values and samples, notice that by virtue of our GP definition the mean of all entries is always zero.

[4] In the sequel we assume a frequency space prior representation, although our treatment can be adapted in a straightforward way to a GP prior.

As for the covariance, we have the following three categories:

$$\mathbb{E}\beta(\psi_i)\beta(\psi_j) = k(\psi_i, \psi_j) \tag{3.16}$$

$$\mathbb{E}\beta(\psi_i)\tilde{f}_j = \int_{-\frac{1}{2}}^{+\frac{1}{2}} k(\psi_i, \psi) e^{j\frac{4\pi}{\lambda_l} x_q \psi} d\psi \tag{3.17}$$

$$\mathbb{E}\tilde{f}_i\tilde{f}_j = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \int_{-\frac{1}{2}}^{+\frac{1}{2}} k(\psi, \psi') e^{4\pi j\left(\frac{x_q}{\lambda_l}\psi - \frac{x_{q'}}{\lambda_{l'}}\psi'\right)} d\psi + \sigma_w^2 \delta_{i,j} \tag{3.18}$$

where the first equation follows from the definition of the GP and the next two equations follow from (3.6), (3.7).

3.4 Single Wavelength Array Design

In the previous section we formulated the problem of far-field sensing in a Bayesian setting with a prior on the distribution of the underlying scene. In this section we design corresponding array geometries to facilitate efficient sensing, exploiting the model and prior. We initially consider a single wavelength setting and a limited budget for the number of sensors. Specifically, we take $L = 1$ such that up to $N = Q$ sensors are free to be placed over some aperture $\mathcal{A}$ on the real line, e.g. for simplicity we can consider $\mathcal{A} \equiv \{x | -a \leq x \leq a\}, a \in \mathbb{R}^+$. In Sections 3.5, 3.6 we consider more sophisticated combinatorial placement constraints and show that the same formulations we develop here may be adapted to those more challenging use cases.

3.4.1 Setting a Cost Function

In order to make the problem of optimal array design well-posed in our Bayesian setting we need to specify a cost function to compare different designs and choose the best one. Revisiting the observation model (3.15) we notice that the design of the array determines the coefficients K_{mn} through the set of sensor positions $\mathcal{S}$. With $\mathcal{S}$ fixed, the sensing problem amounts to performing inference leading to the posterior $\mathbb{P}(\{\beta_m\} | \tilde{\boldsymbol{f}})$.

A natural cost function in this setting quantifies the quality of inference, i.e., the information gained by performing the sensing experiments which results in updating our beliefs

about the coefficients $\{\beta_m\}$ from the prior $\mathbb{P}(\{\beta_m\})$ to the posterior $\mathbb{P}(\{\beta_m\}\,|\tilde{\boldsymbol{f}})$. This problem has been extensively studied in the context of statistical inference and experimental design [17],[8]. Here we adopt the Bayes D-optimality criterion whereby the quality of inference between measurements and hidden random variables is given by the mutual information between the two. In our setting this amounts to:

$$G(\mathcal{S}) \equiv I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \{\beta_m\}) = H(\{\beta_m\}) - H(\{\beta_m\}\,|\tilde{\boldsymbol{f}}_{\mathcal{S}}) \tag{3.19}$$

Where $I(\cdot;\cdot)$ is the mutual information and $H(\cdot)$ the Shannon entropy. The subscript $\mathcal{S}$ explicitly emphasizes the dependence of the measurements on the set of sensor positions $\mathcal{S}$.

Notice that maximizing $G(\mathcal{S})$ as a function of $\mathcal{S}$ can be equivalently viewed as minimizing the uncertainty (entropy) in $\{\beta_m\}$ given $\tilde{\boldsymbol{f}}_S$, i.e. the larger $G(\mathcal{S})$ the more we trust the values of the coefficients $\{\beta_m\}\,|\tilde{\boldsymbol{f}}_{\mathcal{S}}$.

With the cost function in place the array design problem becomes

$$\mathcal{S}^\star = \underset{\mathcal{S}\subseteq\mathcal{A},|\mathcal{S}|\leq N}{\operatorname{argmax}} I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \{\beta_m\}). \tag{3.20}$$

which is in general an NP-hard combinatorial optimization problem. However, we will show later that as opposed to other problems, we can obtain a constant-factor approximation for (3.20) using efficient computational techniques.

3.4.2 Approximate Problem

The optimization problem (3.20) assumes a continuous set $\mathcal{A}$ as well as an infinite dimensional representation for the expansion coefficients $\{\beta_m\}$. Here we approximate these with finite proxies that can be input to generic discrete solvers and measure the corresponding approximation errors.

Finite Dimensional Representation

Solving (3.20) under our model (3.15) involves manipulations of the infinite sequence $\{\beta_m\}$ which may not be amenable to computer representation. To make our formulation tractable

we approximate the infinite sequence $\{\beta_m\}$ with a finite truncated set of coefficients $\{\beta_m | m \in \mathcal{M}\}$, where $\mathcal{M}$ is some finite set. Stacked in vector form $\boldsymbol{\beta}$, we consider the simplified finite dimensional approximation of (3.15):

$$\hat{\boldsymbol{f}}_{\mathcal{S}} = \boldsymbol{K}_{\mathcal{S}}\boldsymbol{\beta} + \boldsymbol{w} \tag{3.21}$$

where $\boldsymbol{K}_{\mathcal{S}}$ is an $N \times |\mathcal{M}|$ matrix formed by restricting K_{nm} on $m \in \mathcal{M}$ and the dependency on the sampling set $\mathcal{S}$ again made explicit. Notice that hat notation is replacing the previous tilde.

We show that the approximate finite-dimensional model (3.21) is a good proxy for the original infinite-dimensional model (3.15) for a suitably selected $\mathcal{M}$. Indeed, if $\mathcal{M}$ is chosen to only exclude those β_m coefficients that in expectation contribute a marginally small part of the energy of the full infinite sequence $\{\beta_m\}$ then the mutual information derived from the approximate model (3.21) will closely track that of the infinite dimensional model in (3.15). More precisely we have the following result:

Lemma 3.2. *Let the prior on $\{\beta_m\}$ be i.i.d. according to $\beta_m \sim \mathcal{CN}(0, \sigma_m^2)$ fixed, and $\epsilon \equiv \sum_{m \notin \mathcal{M}} \sigma_m^2$ satisfies $\epsilon < \sigma_w^2 N^{-\frac{3}{2}}$. We have:*

$$-N\log(1+\frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2}) \leq I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \{\beta_m\}) - I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta}) \leq -N\log(1-\frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2})$$

Proof. See Appendix C.2. □

By virtue of the last lemma and $N\log(1 \pm \frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2}) \overset{\epsilon \to 0}{\longrightarrow} 0$ we have that $I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta})$ is an arbitrarily accurate proxy for $I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \{\beta_m\})$ for ϵ small enough, such that in lieu of problem (3.20) we now continue with the simplified finite dimensional approximation:

$$\mathcal{S}^\star = \underset{\mathcal{S} \subseteq \mathcal{A}, |\mathcal{S}| \leq N}{\operatorname{argmax}} \; I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta}) \tag{3.22}$$

and the results will be accurate to within the approximation bounds from Lemma 3.2.

Grid Discretization

Next we turn to discretizing the aperture $\mathcal{A}$ to cast the array design problem in form of a generic finite selection problem. We thus restrict the choice of sampling positions to the finite set

$$\mathcal{V} \equiv \left\{x^1, \ldots, x^{|\mathcal{V}|}\right\} \subset \mathcal{A} \tag{3.23}$$

For the sequel we take $\mathcal{V}$ to be a uniform δ-spaced grid of positions in $\mathcal{A}$ (δ-net). We adapt the array design problem (3.22) accordingly as:

$$\mathcal{S}_d^\star = \underset{\mathcal{S}\subseteq\mathcal{V},|\mathcal{S}|\leq N}{\operatorname{argmax}} I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta}) \tag{3.24}$$

with the subscript d implying discretization. The next result can be used to quantify the level of discretization δ necessary to guarantee performance close to optimum within some specified error bound:

Lemma 3.3. *With $\mathcal{V}$ a uniform grid of sampling positions with distance δ between adjacent positions we have:*

$$I(\hat{\boldsymbol{f}}_{\mathcal{S}_d^\star}; \boldsymbol{\beta}) \leq I(\hat{\boldsymbol{f}}_{\mathcal{S}^\star}; \boldsymbol{\beta}) \leq I(\hat{\boldsymbol{f}}_{\mathcal{S}_d^\star}; \boldsymbol{\beta}) + N \log(1+\frac{4\delta P(1+\delta)N^{\frac{3}{2}}}{\lambda\sigma_w^2})$$

Proof. See Appendix C.3. □

With this last lemma in place the array design problem (3.22) may be further approximated in the more convenient finite combinatorial problem form of (3.24) with guarantees on the accuracy of the resulting designs. In the sequel we assume that δ is chosen such as to meet desired accuracy levels, as prescribed in Lemma 3.3, and work with the simplified formulation (3.24).

Further notice that evaluation of the target function in (3.24) for the model (3.15) is straightforward, as all relevant distributions are Gaussian such that evaluation of the mutual informations may be accomplished using the formula for the entropy of Gaussian random vectors as referenced in the previous chapter, for $X \in \mathbb{R}^k, X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ we have $H(X) =$

$\log((\pi e)^k \det\boldsymbol{\Sigma})$.

3.4.3 Scene Inference

With the observation model of (3.21), coupled with a Gaussian distribution for the noise vector $\boldsymbol{w}$ and a Gaussian prior for the coefficients vector $\boldsymbol{\beta}$, calculation of the posterior distribution $\boldsymbol{\beta}|\hat{\boldsymbol{f}}_{\mathcal{S}}$ is particularly simple and can be performed analytically. Concretely, we have as a result of all random variables being Gaussian $\boldsymbol{\beta}|\hat{\boldsymbol{f}}_{\mathcal{S}} \sim \mathcal{CN}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and the parameters are given according to the conventional Gaussian conditional parameters:

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \boldsymbol{\Sigma}_{\beta\hat{f}} \boldsymbol{\Sigma}_{\hat{f}\hat{f}}^{-1} \hat{\boldsymbol{f}}_{\mathcal{S}} \\ \hat{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{\beta\beta} - \boldsymbol{\Sigma}_{\beta\hat{f}} \boldsymbol{\Sigma}_{\hat{f}\hat{f}}^{-1} \boldsymbol{\Sigma}_{\beta\hat{f}}^{\dagger} \end{aligned} \tag{3.25}$$

where $\boldsymbol{\Sigma}_{\beta\hat{f}} = \boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{K}_S^{\dagger}$, $\boldsymbol{\Sigma}_{\hat{f}\hat{f}} = \boldsymbol{K}_{\mathcal{S}}\boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{K}_{\mathcal{S}}^{\dagger} + \boldsymbol{\Sigma}_{ww}$ and $\boldsymbol{\Sigma}_{\beta\beta} = \mathrm{diag}[\sigma_1^2, \ldots, \sigma_{\mathcal{M}}^2]$.

3.4.4 Optimization

The next step is to prescribe an efficient algorithm for the solution of (3.24). As we will show shortly (3.24) is an instance of a known NP-hard problem such that it is widely believed that no efficient algorithm for its solution exists. However, due to the structure of the cost function an efficient approximation algorithm is known to exist with strong theoretical guarantees. In this subsection we survey the relevant results and adapt them to our needs.

Submodularity

We begin by invoking the submodularity property of set functions (Appendix A). As it turns out, our cost function is monotonic and submodular as the next result shows (this is similar to corollary 4 in [56]. We reproduce it here with adaptations to our setting):

Lemma 3.4. *Let $\mathcal{V}$ be defined as before, and define the set function $G : 2^{\mathcal{V}} \to \mathbb{R}$ according to $G(\mathcal{S}) = I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta})$. Then G is submodular and monotonic (increasing).*

Proof. Expanding the mutual information according to $I(x;y) = H(x) - H(x|y)$ we have:

$$\begin{aligned} G(\mathcal{S} \cup \{x\}) - G(\mathcal{S}) &= H(\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}}) - H(\hat{\boldsymbol{f}}_{\mathcal{S}}) \\ - [H(\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}}|\boldsymbol{\beta}) - H(\hat{\boldsymbol{f}}_{\mathcal{S}}|\boldsymbol{\beta})] &= H(\hat{\boldsymbol{f}}_{\{x\}}|\hat{\boldsymbol{f}}_{\mathcal{S}}) - H(\hat{\boldsymbol{f}}_{\{x\}}|\boldsymbol{\beta}) \end{aligned} \tag{3.26}$$

where in the last equality we used the conditional independence of the components of $\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}}$ given $\boldsymbol{\beta}$. Substituting $\mathcal{T}$ for $\mathcal{S}$ we immediately get:

$$\begin{aligned} &[G(\mathcal{S} \cup \{x\}) - G(\mathcal{S})] - [G(\mathcal{T} \cup \{x\}) - G(\mathcal{T})] \\ &= H(\hat{\boldsymbol{f}}_{\{x\}}|\hat{\boldsymbol{f}}_{\mathcal{S}}) - H(\hat{\boldsymbol{f}}_{\{x\}}|\hat{\boldsymbol{f}}_{\mathcal{T}}) \end{aligned} \tag{3.27}$$

Using $\mathcal{S} \subseteq \mathcal{T}$ we have $H(\hat{\boldsymbol{f}}_{\{x\}}|\hat{\boldsymbol{f}}_{\mathcal{S}}) \geq H(\hat{\boldsymbol{f}}_{\{x\}}|\hat{\boldsymbol{f}}_{\mathcal{T}})$ such that $G(\mathcal{S} \cup \{x\}) - G(\mathcal{S}) \geq G(T \cup \{x\}) - G(\mathcal{T})$ and G is submodular.

To prove monotonicity it is enough to show $G(\mathcal{S} \cup \{x\}) - G(\mathcal{S}) \geq 0$. This time expand the mutual information according to $I(x;y) = H(y) - H(y|x)$:

$$G(\mathcal{S} \cup \{x\}) - G(\mathcal{S}) = H(\boldsymbol{\beta}|\hat{\boldsymbol{f}}_{\mathcal{S}}) - H(\boldsymbol{\beta}|\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}}) \tag{3.28}$$

Conditioning can never increase entropy so $H(\boldsymbol{\beta}|\hat{\boldsymbol{f}}_{\mathcal{S}}) \geq H(\boldsymbol{\beta}|\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}})$ and the result follows. □

Efficient Solvers

With Lemma 3.4 we have that our optimization problem (3.24) is the maximization of a monotonic submodular function. The greedy Algorithm 3 solves this problem to within the best possible approximation factor, as stated in Lemma A.1.

Combining the guarantees of Lemma A.1, Lemma 3.2 and 3.3 we derive an approximation bound on the original problem (3.20):

Corollary 3.1.

$$(1 - \frac{1}{e}) \left[I(\tilde{\boldsymbol{f}}_{\mathcal{S}^\star}; \{\beta_m\}) - N \log \frac{\lambda\sigma_w^2 + 4\delta P(1+\delta)N^{\frac{3}{2}}}{\lambda\sigma_w^2 - \epsilon\lambda N^{\frac{3}{2}}} \right] \leq I(\hat{\boldsymbol{f}}_{\mathcal{S}^{gr}}; \boldsymbol{\beta}) \tag{3.29}$$

We hence apply Algorithm 3 to solve our optimization problem (3.24). The algorithm runs in time $O(|\mathcal{V}|N)$, linear in the size of the set $\mathcal{V}$ and the number of selected elements N [78] such that it is easily implementable for problems of large size. More efficient variants of the algorithm have been introduced and studied, in particular the 'lazy greedy' Algorithm 4 was studied in [78] and was shown to offer substantial running-time improvements in practice (with an unlikely worst-case theoretical performance upper bounded by that of the conventional greedy algorithm). Our numerical experiments described in Section 3.7 implement this more efficient variant to reduce running time.

While Lemma A.1 guarantees an approximation bound of $(1-\frac{1}{e}) \approx 63\%$ for the efficient greedy algorithm this guarantee is not tight. The data dependent bound of Lemma A.2 takes $O(|\mathcal{V}| \log |\mathcal{V}|)$ evaluations of $G(\mathcal{S})$ to compute and sort and is often tighter in practice. We use Equation (A.2) to improve the distance from optimality bound in some of our numerical solutions in Section 3.7.

3.4.5 Design Example: A Simple Ideal Setting

In the previous subsections we formulated the array design problem in a setting with constraints on the aperture $\mathcal{A}$ and the number of sensors N and showed how a greedy algorithm (Algorithm 3) is guaranteed to efficiently find an approximate solution.

Here, we study a particular instance of this problem, where the Signal to Noise Ratio (SNR) is high, and the aperture is effectively unconstrained (the N sensors may be placed anywhere on the real line). Under these conditions a truncated $\frac{\lambda}{2}$-spaced array is traditionally considered the design of choice in the conventional non-Bayesian setting (this is a truncated version of the infinite $\frac{\lambda}{2}$-spaced design mentioned in Section 3.2). We show next that the truncated $\frac{\lambda}{2}$-spaced design naturally emerges as the approximately optimal solution as retrieved by our schemes in Bayesian settings where the a-priori $\boldsymbol{\beta}$ distribution satisfies some conditions. Specifically, we have the following result:

Theorem 3.1. *Consider the high SNR regime* $\frac{P}{\sigma_w^2} \to \infty$ *and assume the prior from* (3.10) *takes a symmetric, monotonically decreasing form, i.e.* $\sigma_m^2 = \sigma_{-m}^2$ *and* $\sigma_{m_1}^2 \geq \sigma_{m_2}^2$ *whenever* $0 \leq m_1 < m_2$. *In addition, take* $\mathcal{V}$ *as an arbitrarily dense set of sampling points on* $\mathbb{R}$, *and*

$\mathcal{M} = -M, \ldots, M$ *with* $M \to \infty$.

We then have that a greedy solver on (3.24) *will return a length* N, $\frac{\lambda}{2}$*-spaced truncated uniform array centered around* $x = 0$.

Proof. See Appendix C.4. □

The last theorem studies one class of simple idealized problems where the greedy solution is reminiscent of generic non-Bayesian array designs. However, notice that our formalism is also useful in more challenging design problems such as when the aperture $\mathcal{A}$ takes on arbitrary forms, and the effects of noise and application-tailored priors are considered.

3.5 Array Design with Combinatorial Constraints

In Section 3.4 we formalized the array design problem in a setting where we imposed constraints on the aperture and the number of sensors. Specifically, the constraints were $\mathcal{S} \subseteq \mathcal{V}$, $|\mathcal{S}| \leq N$. In many practical scenarios these may be too simplistic to accurately represent real world design constraints. For example in applications where sensors are heavy and mounted on support beams we may want to restrict the number of sensors in specific sections of the aperture.

In this section we briefly review key elements from matroid theory which is a branch in combinatorics [86] and survey results from submodular optimization with matroid constraints guaranteeing the existence of efficient approximate solvers for this class of problems. We continue to show that these mathematical structures may be utilized to impose constraints of interest in array design enriching the set of problems our Bayesian formulation can describe and solve.

3.5.1 Optimization with Matroid Constraints

We begin by defining matroids and their corresponding independent sets [83]:

Definition 3.1. *A finite* ***matroid*** M *is a pair* $(\mathcal{V}, \mathcal{I})$ *where* $\mathcal{V}$ *is a ground set and* $\mathcal{I}$ *is a collection of subsets of* $\mathcal{V}$ *(the* ***independent sets****) that satisfies the following properties:*

1. *The empty set is independent:* $\emptyset \in \mathcal{I}$

2. *A subset of an independent set is independent:* $\mathcal{X} \subset \mathcal{Y}, \mathcal{Y} \in \mathcal{I} \quad \Rightarrow \quad \mathcal{X} \in \mathcal{I}$

3. *If* $\mathcal{X}$ *is an independent set and* $\mathcal{Y}$ *is a larger indepedent set,* $\mathcal{X}$ *can be augmented to a larger independent set by adding an element from* $\mathcal{Y} \setminus \mathcal{X}$*:*
   $\mathcal{X}, \mathcal{Y} \in \mathcal{I}, |\mathcal{X}| < |\mathcal{Y}| \quad \Rightarrow \quad \exists e \in \mathcal{Y} \setminus \mathcal{X} \;\; s.t. \;\; \mathcal{X} \cup \{e\} \in \mathcal{I}$

A matroid structure is useful for classifying subsets of a ground set $\mathcal{V}$ into permissible subsets which belong to $\mathcal{I}$ and non permissible subsets which do no belong to $\mathcal{I}$. In the next subsection we show that using this formalism we can express interesting array design constraints.

From the theory of submodular optimization we have the following results for submodular optimization with matroid constraints [54],[16]. Let $M = (\mathcal{V}, \mathcal{I})$ be a matroid and $G(\mathcal{S})$ a monotonic, submodular set function. There exists an efficient approximate solver for the problem $\text{argmax}_{\mathcal{S} \in \mathcal{I}} G(\mathcal{S})$. Specifically, a greedy solver (maximizing the immediate marginal benefit at each step) taking the form

$$\mathcal{S}^{\text{gr}} \leftarrow \mathcal{S}^{\text{gr}} \cup \left\{ \underset{e : e \notin \mathcal{S}^{\text{gr}}, \mathcal{S}^{\text{gr}} \cup \{e\} \in \mathcal{I}}{\text{argmax}} \left[G(\mathcal{S}^{\text{gr}} \cup \{e\}) - G(\mathcal{S}^{\text{gr}}) \right] \right\} \tag{3.30}$$

and stopping when no more elements e can be added is guaranteed to achieve a one-half approximation bound:

$$G(\mathcal{S}^{\text{gr}}) \geq \frac{1}{2} \max_{\mathcal{S} \in \mathcal{I}} G(\mathcal{S}). \tag{3.31}$$

The constant factor may be tightened to $\left(1 - \frac{1}{e}\right)$ by utilizing specialized randomized algorithms [16].

3.5.2 Matroid Constraints in Array Design

Here we invoke a well known matroid structure and demonstrate its application in expressing useful array design constraints. Let $\mathcal{V}$ be a ground set of grid points where sensors are allowed

to be placed as before. Let $\mathcal{V}_1, \ldots, \mathcal{V}_K$ be a partition of the set $\mathcal{V}$, i.e. $\bigcup_k \mathcal{V}_k = \mathcal{V}, \quad \mathcal{V}_i \bigcap \mathcal{V}_j = \emptyset$, $\forall i \neq j$, and let $N, n_1, \ldots, n_K$ be a set of integers.

We define the (cardinality constrained) partition matroid [16] $M = (\mathcal{V}, \mathcal{I})$ with the following definition for the collection of independent sets $\mathcal{I}$: a subset $\mathcal{S} \subseteq \mathcal{V}$ is an independent subset $\mathcal{S} \in \mathcal{I}$ if it holds $|\mathcal{S} \cap \mathcal{V}| \leq N$, $|\mathcal{S} \cap \mathcal{V}_j| \leq n_j, \ \forall j$.

In the context of array design the partition matroid may be useful in expressing practical constraints over sensor placement configurations. For example if the subsets $\mathcal{V}_i$ represent closed line sections, e.g. a physical partitioning of the aperture into zones, and $\mathcal{I}$ represents the collection of all permissible designs then the structure of the matroid limits the number of sensors that may be placed in the *ith* zone to n_i which may be an important engineering constraint coupled with some specific application. We solve:

$$\mathcal{S}^\star = \text{argmax}_{\mathcal{S} \in \mathcal{I}} \, I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta}) \tag{3.32}$$

Applying the results from the previous subsection we immediately have an efficient approximate solver for the array design problem coupled with a partition matroid constraint. In Section 3.7 we detail such a design for a numerical example.

3.6 Multiple Wavelength Array Design Paradigms

Here we define two multiple wavelength array design paradigms, concisely formulated as combinatorial optimization problems: The **fusion** problem entails designing an array $\mathcal{S}$ for collection of measurements at a set of fixed wavelengths Λ. The full set of measurements (taken at each location in all wavelengths in Λ) is jointly used to infer $\boldsymbol{\beta}$. The set $\mathcal{S}$ is constrained to be in $\mathcal{A}$ as before, and to be of size no more than Q (remember that as in Section 3.3, we have $|\Lambda| \equiv L, \ N \equiv QL$):

$$S^\star = \text{argmax}_{\mathcal{S}: \mathcal{S} \subseteq \mathcal{A}, |\mathcal{S}| \leq Q} \, I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta}). \tag{3.33}$$

The **robust** problem entails designing an array $\mathcal{S}$ for collection of measurements at a single wavelength $\lambda \in \Lambda$, which is fixed but unknown. Thus we are interested in guaranteeing good

quality inference for any $\lambda \in \Lambda$ and solve the robust optimization problem of maximizing the worst-case performance achieved when operating the array at any single wavelength. Concretely, let $\tilde{\boldsymbol{f}}_{\mathcal{S}}^{\lambda}$ be the set of samples collected at a single wavelength λ at the set of positions $\mathcal{S}$, which satisfies the same constraints as before. The design criterion is:

$$\mathcal{S}^{\star} = \text{argmax}_{\mathcal{S}:\mathcal{S}\subseteq\mathcal{A},|\mathcal{S}|\leq Q} \min_{\lambda\in\Lambda} I(\tilde{\boldsymbol{f}}_{\mathcal{S}}^{\lambda}; \boldsymbol{\beta}) \tag{3.34}$$

3.6.1 Optimization

In this section we prescribe efficient algorithms for the solution of (3.33) and (3.34). As before, we have that due to the structure of the cost functions, efficient approximation algorithms are known to exist with strong theoretical guarantees. Define $G(\mathcal{S}) \equiv I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta})$, $G_{\lambda}(\mathcal{S}) \equiv I(\tilde{\boldsymbol{f}}_{\mathcal{S}}^{\lambda}; \boldsymbol{\beta})$, than we have similar to Lemma 3.4 that both $G(\mathcal{S})$ and $G_{\lambda}(\mathcal{S})$ are submodular and (increasing) monotonic for every λ.

Problem (3.33) is immediately recognized as an instance of a submodular optimization problem under a cardinality constraint, and retrieving an approximately optimal solution follows as before.

As for the robust problem (3.34), we briefly review some theory pertaining to robust submodular maximization. Let $\{G_i(\mathcal{S})\}$ be a set of monotone, submodular set functions and consider the robust optimization problem

$$\mathcal{S}^{\star} = \text{argmax}_{\mathcal{S}\subseteq\mathcal{A},|\mathcal{S}|\leq Q} \min_i G_i(\mathcal{S}) \tag{3.35}$$

which we immediately recognize as a generalization of (3.35). It is known that no polynomial time algorithm approximating the solution of (3.35) exists [57]. However, the following lemma suggests that for integer-valued $G_i(\mathcal{S})$, Algorithm 2 is guaranteed to achieve approximately optimal solution:

Lemma 3.5. *(Krause [57]). For any integer Q, SAT (delineated in Algorithm 2) finds a*

solution $\mathcal{S}^{sat}$ *such that*

$$\min_i G_i(\mathcal{S}^{sat}) \geq \min_i G_i(\mathcal{S}^\star) \quad \textit{and} \quad |\mathcal{S}^{sat}| \leq \alpha Q \tag{3.36}$$

$$\textit{where} \quad \alpha \equiv 1 + \log\left(\max_{s\in\mathcal{A}} \sum\nolimits_i G_i(s)\right) \tag{3.37}$$

Algorithm 1 Greedy submodular partial cover

function S=GPC($\bar{G}_c(\mathcal{S}), c$)
  $\mathcal{S} \leftarrow \emptyset$
  while $\bar{G}_c(\mathcal{S}) < c$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \text{argmax}_{s\in\mathcal{A}} \left\{\bar{G}_c(\mathcal{S}\cup\{s\}) - \bar{G}_c(\mathcal{S})\right\}$
  end while
end function

Algorithm 2 Submodular saturation algorithm

function $\mathcal{S}_{\text{BEST}}$=SAT($G_1, \ldots, G_m, A, k, \alpha$)
  $c_{\min} \leftarrow 0$; $c_{\max} \leftarrow \min_i G_i(\mathcal{A})$; $S_{\text{best}} \leftarrow \emptyset$
  while $(c_{\max} - c_{\min}) \geq \frac{1}{m}$ **do**
    $c \leftarrow (c_{\min} + c_{\max})/2$
    $\hat{\mathcal{S}} \leftarrow \text{GPC}(\frac{1}{m}\sum_i \min\{G_i(\mathcal{S}), c\}, c)$
    if $|\hat{\mathcal{S}}| > \alpha k$ **then** $c_{\max} \leftarrow c$
    else $c_{\min} \leftarrow c$; $\mathcal{S}_{\text{best}} \leftarrow \hat{\mathcal{S}}$
    end if
  end while
end function

Lemma 3.5 guarantees that SAT will find an approximate optimal set $\mathcal{S}^{\text{sat}}$ achieving performance at least as good as the true optimal set $\mathcal{S}^\star$, at a cost of using as many as αQ elements of the set $\mathcal{A}$ in lieu of the Q elements included in $\mathcal{S}^\star$.

The extension of Lemma 3.5 to non integer-valued functions is discussed in [57] (Section 7). One approach is porting these problems into integer-valued ones by scaling and rounding $G_i(\mathcal{S})$. However, this requires careful manipulations of the guarantees in Lemma 3.5. Instead, [57] follows an empirical approach making the ad-hoc choice $\alpha = 1$ for the implementation of Algorithm 2, and keeping the non integer-valued $G_i(\mathcal{S})$ unchanged. Based on extensive numerical experiments it is empirically shown that under these conditions SAT performs favorably. We follow this approach in our numerical experiments described in Section 3.7,

and empirically verify the usefulness of the above choice when applied to our non integer-valued problem.

3.7 Numerical Experiments

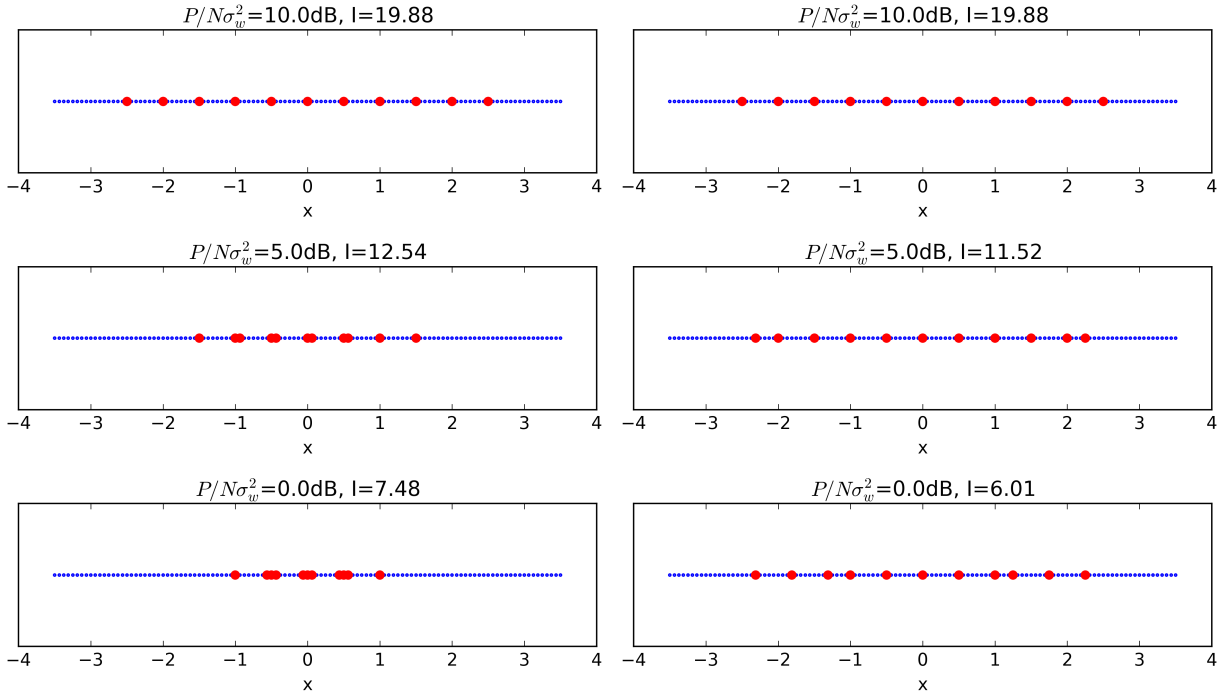
In this section we perform numerical experiments validating our theoretical results and exemplifying them. We showcase an array design with cardinality and aperture constraints as prescribed in Section 3.4, design arrays adhering to matroid constraints as prescribed in Section 3.5, and multiple wavelength arrays according to the paradigms in Section 3.6

3.7.1 Single Wavelength Array Design

Our initial setting is as follow. We fix $\lambda = 1$ throughout as in this setting the wavelength only serves to scale the x axis. The aperture is set as $\mathcal{A} = \{x| -3.5 \leq x \leq 3.5\}$ and the selection set $\mathcal{V}$ is chosen as a uniform grid of 113 positions from $\mathcal{A}$ spaced $\delta = 0.0625$ apart. We set out to design an array consisting of $N = 11$ sensor locations. The prior for $\{\beta_m\}$ is set as per (3.12) with $r = 1$ and normalized to sum to $P = 1$. For the simulations we consider the truncated vector $\boldsymbol{\beta}$ formed when restricting the set of m coefficients to 901 consecutive elements centered around the origin, i.e., we set $\mathcal{M} = \{-450, \ldots, +450\}$. For the preliminary design we implement the lazy greedy algorithm and plot the results in the left column of Fig. 3-5 as a function of the SNR which we define here as $\frac{P}{N\sigma_w^2}$. Blue markers denote the full selection set $\mathcal{V}$ and red markers delineate the active $\mathcal{S}$ selected by the algorithm.

In the high SNR regime (SNR=10dB or higher values) the resulting design is a truncated $\frac{\lambda}{2}$ uniform array as predicted according to Theorem 3.1. As the SNR decreases the reliability of the measurements deteriorates and the algorithm prefers locating antennas right next to each other on expense of widening the array as this serves to average out the noise. The corresponding antenna array beam patterns (according to Equation (3.4) with unit gain coefficients) are shown in Figure 3-6.

The performance in terms of mutual information $I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta})$ for the selected locations $\mathcal{S}$ appears in the title of the plots (in natural units). Notice for example that for the 5dB SNR design the achieved mutual information is 12.54. Using Lemma A.1 we have that the optimal



(Left) Array designs with an aperture constraint for various SNR levels. (Right) Array designs with combinatorial placement constraints for various SNR levels.

Figure 3-5: Near-optimal single wavelength antenna array designs.

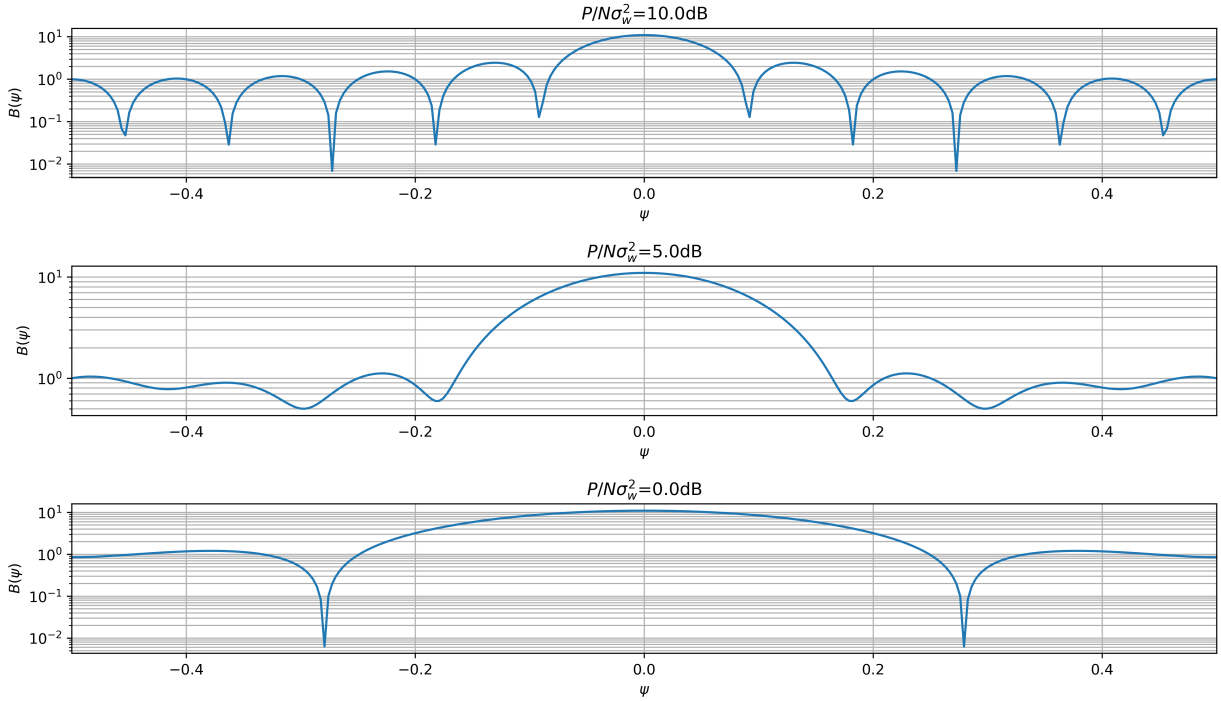


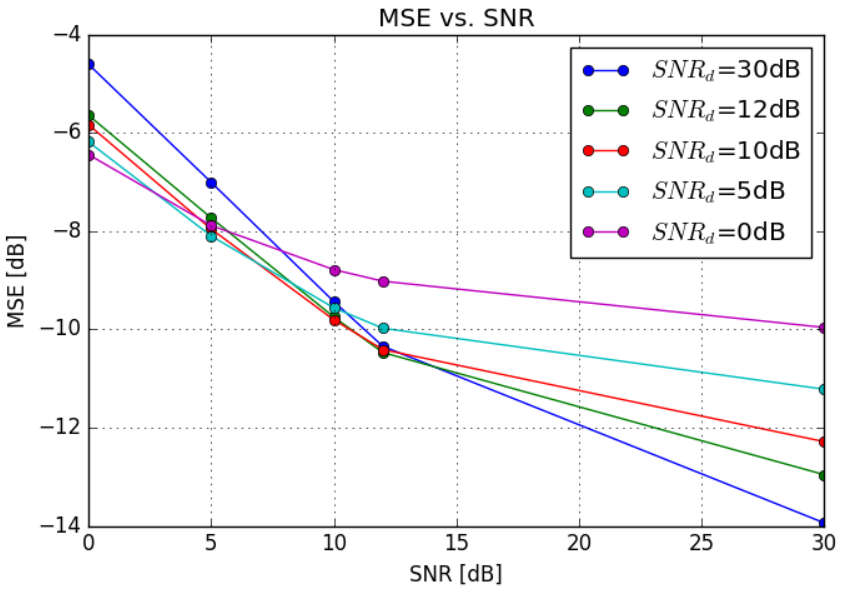
Figure 3-6: Beam patterns corresponding to near-optimal array designs.

design cannot achieve mutual information better than $\frac{1}{1-\frac{1}{e}}12.54 = 19.83$. This bound can be improved using the improved online bounding method briefly described in Appendix A to show that the optimal performance is not greater than 17.45.

The truncation level dictated by our choice of $\mathcal{M}$ translates to $\epsilon \cong 1e{-}4$ and the truncation bounds from Lemma 3.2 read (for the lower, extreme SNR case): $-0.45 \leq I(\tilde{\boldsymbol{f}}_{\mathcal{S}}; \{\beta_m\}) - I(\hat{\boldsymbol{f}}_{\mathcal{S}}; \boldsymbol{\beta}) \leq 0.47$. We empirically find that these bounds are extremely loose and $\mathcal{M}$ can be shrunk considerably without substantially compromising accuracy. To achieve a similar upper bound on $I(\hat{\boldsymbol{f}}_{\mathcal{S}^\star}; \boldsymbol{\beta}) - I(\hat{\boldsymbol{f}}_{\mathcal{S}_d^\star}; \boldsymbol{\beta})$ as per Lemma 3.3 a discretization level of $\delta \cong 2e{-}4$ is needed. However, we empirically find that our choice of $\delta = 0.0625$ is accurate enough as refining the grid further does not significantly change the design. Our lemmas prove to be pessimistic as is expected given that the proofs take into account worst-case scenarios.

The array geometries above, derived according to the formulations of Section 3.4, are designed to optimize the quality of inference between the measurements and the scene expansion coefficients $\boldsymbol{\beta}$. Many sensing applications of interest specifically involve imaging the scene, that is reconstructing $\beta(\psi)$ from the measurements. Our next experiment was designed to empirically evaluate the Mean Square Error (MSE) performance in scene reconstruction from measurements collected using the prescribed designs. First, we designed five array geometries as described above, optimized for several target SNR levels $\{30\text{dB}, 12\text{dB}, 10\text{dB}, 5\text{dB}, 0\text{dB}\}$. We set up a Monte-Carlo experiment where 1000 scenes were randomly drawn from the distribution of Section 3.3.2. For each scene, noisy measurements were collected by each of the five optimized arrays. The measurements were repeated with five different synthetic noise levels corresponding to the five target SNR levels.

We repeatedly performed maximum likelihood estimation [2] of the expansion coefficients $\boldsymbol{\beta}$, and synthesized an estimated scene $\hat{\beta}(\psi)$ according to (3.8). The MSE discrepancy between $\hat{\beta}(\psi)$ and the true scene is depicted in Fig. 3-7. It is evident that the quality of inference criterion is indicative of MSE performance, as each of the five geometries yielded the best MSE performance at its specified target SNR level.



Arrays are designed for specific target SNR levels, while actual tested SNR is swept.

Figure 3-7: Scene reconstruction performance for near-optimal arrays.

3.7.2 Array Design with Matroid Constraints

Next we solve a corresponding set of design problems with matroid constraints installed to limit the number of sensors in given aperture segments. Specifically, we use the (cardinality constrained) partition matroid from Section 3.5.2 with $N = 11, n_i = 1, \forall i$ and $\mathcal{V}_i$ spanning consecutive line segments of length 0.5: $\mathcal{V}_i = [-0.25 + 0.5{\cdot}i, +0.25 + 0.5{\cdot}i)$. The matroid constraints limit the proximity between sensor elements, which may be a useful requirement in practical applications. We plot the results for the matroid constrained designs in the right column of Fig. 3-5. Notice that while the theoretical guarantees pertaining to the greedy matroid optimization scheme of Section 3.5 is $\frac{1}{2}$ compared to $1 - \frac{1}{e}$ with the cardinality constraints of Section 3.4, the actual performance achieved in the constrained design instances is not far from those achieved with the simple cardinality constraints.

3.7.3 Multiple Wavelength Array Design

Fusion Problem

First, we design arrays for the fusion setting as per (3.33) using the lazy greedy submodular optimization algorithm. We take $\Lambda = \{1, 1.1, 1.2\}$, the sensor position selection set $\mathcal{A}$ is a uniformly spaced grid of 161 positions in $|x| \leq 10$, and we design an array consisting of $Q = 7$ elements. $\boldsymbol{\beta}$ is formed by approximating $\{\beta_m\}$ via the 101 lowest frequency coefficients[5], and the prior for β_m is set as per (3.12) with $r = 1$. We normalize $\{\sigma_m^2\}$ for unit average scene power P using Parseval:

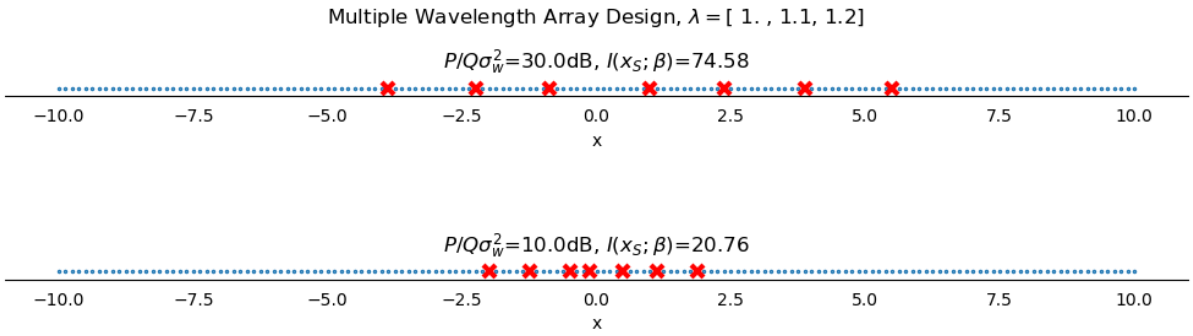
$$P \equiv \mathbb{E} \int |\beta(\psi)|^2 d\psi = \mathbb{E} \sum\nolimits_m |\beta_m|^2 = \sum\nolimits_m \sigma_m^2 = 1 \tag{3.38}$$

The results are summarized in Fig. 3-8, where blue markers denote the selection set $\mathcal{A}$ and red markers delineate the chosen set $\mathcal{S}$. The upper subplot depicts a design for high Signal to Noise Ratio (SNR) (defined as $\frac{P}{Q\sigma_w^2}$) of 30dB. Notice how the design differs from a uniform sampler due to the multiple wavelengths involved. Adjacent pairs of antennas display varying distances to effectively tune into the several wavelengths at play, utilizing as much of the information impinging on the array from the scene as possible.

The lower subplot repeats the experiment at a lower SNR of 10dB. In comparison, this design tends to limit the spread of the sampling positions as samples become less reliable and there is value in limiting sampling diversity for the sake of concentrating more samples in valuable regions.

The performance in terms of mutual information $I(\tilde{\boldsymbol{f}}_S; \boldsymbol{\beta})$ for the selected locations $\mathcal{S}$ appears in the plot title. Notice for example that for the 10dB SNR design, the achieved mutual information is 20.76. Using Lemma A.1 we have that the optimal design cannot achieve mutual information better than $(1-\frac{1}{e})^{-1}20.76 = 32.84$.

[5]We empirically find that refining the sampling grid, or including more Fourier coefficients does not significantly change the design.



The grid of possible placement locations appears in blue and red markers delineate selected locations.

Figure 3-8: Near-Optimal multiple wavelength antenna array designs.

Robust Problem

Next, we design arrays for the robust setting as per (3.34) using the SAT algorithm of Section 3.6. We take $\mathcal{A}$ as before, $\Lambda = \{1, 2, 3, 4\}$, the number of elements is $Q = 9$ and we assume a SNR of 10dB. In Fig. 3-9 (left) we plot several designs. The top configuration is the robust design generated via Algorithm 2, fixing $\alpha = 1$. The bottom four configurations depict arrays each optimized for a single wavelength. These were generated by applying the greedy design scheme of Appendix A with a single measurement wavelength from the set Λ. The figure shows that for a single observation wavelength, we obtain configurations that are generic truncated uniform $\frac{\lambda}{2}$ arrays [113]. However, when considering observations across multiple possible wavelengths, as is done for the robust design, the resulting configuration is no longer uniform, but consists of a mixture of large and small inter-element spacings, to cater to all possibilities.

Fig. 3-9 (middle) plots the performance of these arrays in terms of the corresponding mutual information $I(\tilde{\boldsymbol{f}}_S; \boldsymbol{\beta})$ when the actual wavelength at which measurements are collected is swept in $0.9 \leq \lambda \leq 4.4$. Each of the four single-wavelength arrays (dashed lines) maximizes the mutual information when operated at the wavelength for which it was designed, as expected. However, at mismatched wavelengths performance deteriorates. In contrast, the robust array (solid line) does not perform as well as the specialized single wavelength arrays at their target wavelengths. But, while those specialized designs are very sensitive to

misspecified wavelengths, the robust design flexibly performs well across the entire range of wavelengths.

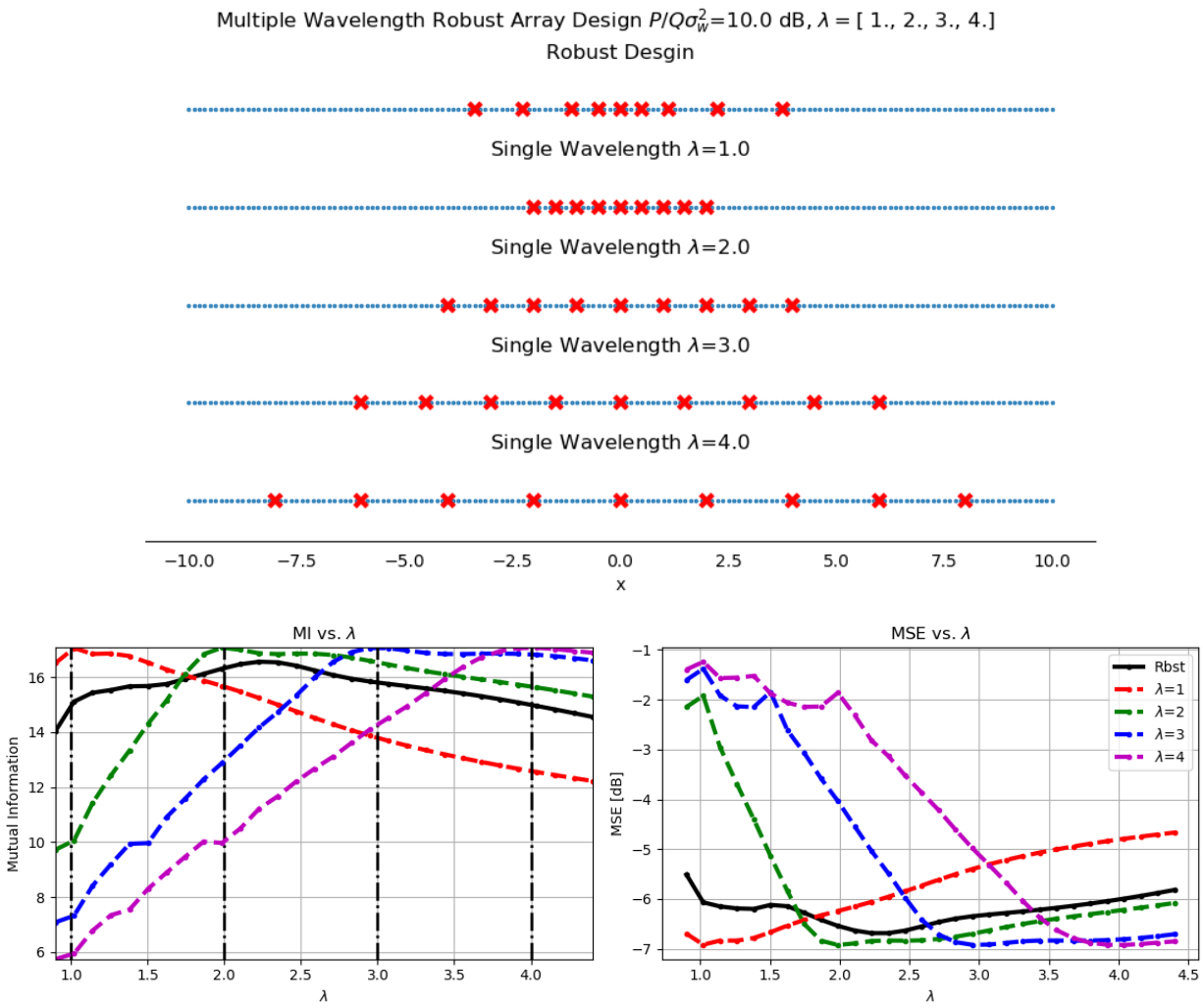
Fig. 3-9 (right) summarizes a Monte-Carlo experiment set up to empirically evaluate Mean Square Error (MSE) performance in scene reconstruction using our robust design. We have drawn 2000 scenes distributed as prescribed in Section 3.3.2, and collected corresponding noisy measurements at various wavelengths using the robust and the four single wavelength optimized arrays. We repeatedly performed maximum likelihood estimation of the expansion coefficients $\boldsymbol{\beta}$, and synthesized an estimated scene $\hat{\beta}(\psi)$ according to (3.8). The MSE discrepancy between $\hat{\beta}(\psi)$ and the true scene was recorded. Evidently, the mutual information performance of Fig. 3-9 (middle) is indicative of MSE performance, with the robust design exhibiting best worst-case results.

3.8 Discussion

Our work in this chapter revolved around the design of sensing topologies for estimating scenes in a Bayesian setting, adapting priors and performing efficient inference. Designing efficient configurations for such environments is a computationally hard task, and finding tractable solutions, even if just approximate, is a desirable goal.

We introduced a novel framework for designing sensor arrays and formulated various sensing paradigms as optimization problems, focusing on antenna arrays for single or multiple wavelength sensing with robust single-wavelength estimation or joint inference over the full spectrum. We showed that optimal solutions to these problems can be efficiently approximated by porting results and efficient solvers from the theory of submodular set function optimization.

In future work, we will be interested in tackling other related problems, such as devising adaptive designs that evolve as the scene is learned, and taking into account other measurement models (e.g. near field imaging), and physical phenomena associated with antenna arrays. We are also interested in extending our treatment to additional signal processing paradigms, for example where the unknown signal of interest has an efficient representation in some countable signal base, and the signal is drawn from some structured distribution,



(Top) Several antenna array configurations. The top most configuration is a robust design, while the next four configurations are tuned for specific wavelengths. (Bottom Left) Mutual Information vs. observation wavelength. (Bottom Right) Reconstruction MSE for Monte-Carlo experiments vs. observation wavelength.

Figure 3-9: Near-optimal multiple wavelength antenna array designs and performance.

e.g. wavelet with tree coefficients. Finally, we are interested in exploring various constraint structures that still allow for such efficient approximate solutions to be found.

Chapter 4

NLOS Optical Imaging[†]

Non-line-of-sight (NLOS) optical imaging, where we pursue techniques to recover scenes hidden around obstacles by processing optical reflections from intervening surfaces, could offer great benefits in a wide variety of applications. However, the diffuse nature of the reflections from many typical surfaces lead to the mixing of spatial information carried by reflected light, preventing scene recovery, and rendering the problem of inverting optical measurements to reconstruct hidden scenes challenging.

The NLOS optical imaging setup we study in this chapter is a fertile ground for thinking about and applying efficient data collection strategies to learn physical environments. Our main focus in this work will be to explore and understand what makes optical measurements informative about hidden scenes, allowing efficient learning, and understand how such measurements should be collected and processed to efficiently infer representations for hidden scenes under investigation.

Recently proposed NLOS imaging modalities rely on exploitation of time-resolved (TR) optical measurements, i.e. measuring the time it takes a narrow light pulse to traverse across a scene, to undue the spatial mixing introduced by reflections. However, this often requires costly, and highly specialized, carefully calibrated laboratory equipment that is capable of collecting optical measurements with extremely high temporal resolution, limiting the potential for widespread use of such systems.

As an alternative to the TR measurements inversion approach, we develop a computa-

[†]Based on joint work with Christos Thrampoulidis and Feihu Xu [126, 108].

tional imaging technique that, perhaps counter intuitively, opportunistically exploits structure in optical measurements that is introduced by occluding objects obstructing light propagation in the hidden scene to undo the spatial information mixing and allow robust, high fidelity hidden scene recovery. We demonstrate both analytically and experimentally that in some cases the presence of such occluders in the hidden scene can completely obviate the need for collecting TR measurements.

More generally, we identify opportunities in designing more accurate, robust and cost-effective NLOS imaging systems that trade-off between the use of high temporal resolution TR optical measurements and available side information about occluders present in the scene. We develop a study framework that involves a mathematical formulation for light propagation in such environments, as well as comprehensive numerical illustrations, and additionally demonstrate our results in a meter-scale experimental setup.

Our experimental demonstrations justify and motivate our utilization of Gaussian Process (GP) models to describe scenes in other chapters of this thesis, and provide a testing ground for some of the fast acquisition methods discussed in previous chapters.

4.1 Introduction

The problem of imaging non-line-of-sight (NLOS) scenes that are hidden behind obstacles has gained much attention in recent years. What makes the ability to glimpse into spaces that are not directly visible to the observer so appealing, is its numerous promising applications in a wide variety of application domains such as medical and industrial inspection, vehicle safety, scientific imagery, security and basic science, to name a few.

The problem of NLOS imaging introduces new and exciting challenges to the field of computational imaging. In contrast to classical photography where the scene of interest is in the direct line of sight of an observer, optical NLOS imaging systems only have indirect access to the scene through reflections from intervening surfaces. In many practical settings these surfaces, e.g. walls, dividers, floors, exhibit matte properties, i.e. they diffusely reflect light across the entire optical spectrum, essentially mixing the spatial information carried by the light by erasing beam orientation, thus rendering the problem of image reconstruction

challenging.

In order to undo the effect of diffuse reflections, initial demonstrations of NLOS imaging behind obstacles used ultrafast transient imaging modalities [52, 115]. In particular, they used a fast laser source to transmit optical pulses of sub-picosecond duration, and a streak camera exhibiting temporal resolution in the picosecond range. A computational imaging algorithm used the fine time-resolved light intensity measurements obtained by the streak camera to form a three dimensional reconstruction of the hidden scene. The requirements posed by this system for transmission of very narrow, high power optical pulses on the laser side, and for very high temporal resolution on the detector side, inevitably implies high cost. Thus, much of the follow-up work has focused on developing reduced cost implementations. For example, the authors of [14], used a single-pixel single photon avalanche diode (SPAD) detector for reduced power consumption and cost. A SPAD camera was also used in [32] to demonstrate tracking of hidden moving objects. In a different line of work, with the aid of modulated illumination, the authors of [44] used widespread CMOS time-of-flight sensors such as photonic mixer devices, substantially reducing the overall system cost although at the expense of reduced spatial resolution.

Motivated by the unfavorable cost-performance trade-off curve offered by existing NLOS optical imaging methods we introduce a novel imaging modality, utilizing a previously untapped resource that has traditionally been considered a nuisance and an impediment for imaging, namely occlusions in the scene. Our work explores the beneficial role played by occlusions in NLOS imaging, and develops the idea that these can facilitate more robust image reconstruction. We demonstrate, both analytically and experimentally that in some circumstances the presence of occluders in the hidden scene can completely obviate the need for collecting time-resolved (TR) measurements, enabling imaging systems of significantly reduced cost. This further allows us to use single-pixel detectors (e.g. SPADs) that, unlike in all previous works, have a wide field of view, enabling more photons to be collected per measurement, reducing the overall acquisition time.

We introduce these new concepts and ideas in the context of imaging a hidden wall of unknown reflectivity. For this problem, we develop a study framework that involves a mathematical formulation, as well as comprehensive numerical and experimental illustra-

tions. More generally, we identify opportunities in designing more accurate, robust and cost-effective NLOS imaging systems that relax the stringent temporal resolution requirements for optical measurements in the presence of occluders.

Related Work

To the best of our knowledge, our work is the first to introduce the concept of exploiting occluders for the problem of NLOS imaging, and initiates the study of their beneficial role in this setting. However, there is a variety of related work in computational imaging exploring the use and exploiting the presence of physical structure in the space between the scene of interest and the measurement system.

The most relevant instance of this idea is that of coded-aperture imaging where the role of occluder in the optical path is played by a carefully designed mask that modulates the light transferred from the scene of interest to a detector array. This is essentially a generalization of the pinhole camera [27] or its inverse the anti-pinhole camera [19]. The main motivation for using these techniques is in applications where lens fabrication is difficult, such as in x-ray or gamma-ray imaging [12].

The idea of a mask appropriately combined with a lens has been used in computational photography for motion deblurring [91], depth estimation [68], digital refocusing and recovery of 4D light-fields [114]. More recently, there has been an increased interest in using masks with appropriate computational techniques, instead of traditional lens-based cameras, to build cameras that have fewer pixels, need not be focused [22], or meet physical constraints [5]. All these methods are passive, and only very recently the authors of [95] proposed the addition of an active illumination source and time resolved sensing to speed up acquisition time in lensless imaging systems. Notably, while related, none of these studies targets or applies to the more challenging problem of NLOS image reconstruction.

In another related work, Torralba and Freeman [110] studied the potential of accidental pinhole and anti-pinhole camera images for revealing information about a scene that is outside the field of view. Their work can be seen as the most direct predecessor of this work, but there are significant differences. Their setup requires a video sequence, and assumes availability of a reference frame captured without occluder presence. Not only do we eliminate

these requirements, but also, using active illumination and more sophisticated computational techniques, we obtain significantly enhanced image reconstructions.

Very recently, Klein et. al. [53] demonstrated the ability to track moving NLOS objects from non-time resolved intensity images recorded on a visible wall. In contrast, this study focuses on imaging a static scene, thus their work is not directly applicable in our setting. In fact, we will show that without further exploiting the presence of occluders, imaging static scenes with non-TR measurements is a severely ill-posed problem.

4.2 Problem Formulation

In this section we introduce the occluded NLOS imaging setup and derive a forward model for light propagation in this environment. In particular, we formulate the problem of imaging a hidden object as a linear inverse problem, exploiting occlusions in the scene in order to obtain more accurate and robust solutions.

4.2.1 Imaging Setup

The goal of NLOS imaging systems is to perform joint estimation of both the geometry and reflectivity properties of a hidden three-dimensional scene by processing reflected light-intensity measurements, as illustrated in Figure 4-1. An observer is equipped with a laser source and a camera, and is interested in forming an image of the hidden object which is not directly visible from its vantage point, due to the wall blocking the direct line of sight path between the two. A focused laser beam is steered towards a visible illumination surface and reflects back towards the hidden object. Upon hitting the object light is reflected back towards the illumination surface and is measured by a focused camera. This forms a three-bounce problem in which light beams follow paths of the form

$$\Lambda(\text{Laser}) \to \boldsymbol{\ell} \to \mathbf{x} \to \mathbf{c} \to \Omega(\text{Camera}),$$

where $\boldsymbol{\ell}, \mathbf{c}$ lie on the illumination surface and $\mathbf{x}$ lies on the hidden object surface. We let $\mathcal{S}$ be a parametrization of the hidden object surface, and $f(\mathbf{x}),\ \mathbf{x} \in \mathcal{S}$ denote its spatially

varying reflectivity function (or, albedo).

Upon hitting the illumination and hidden surfaces the laser light reflects as dictated by the surfaces respective Bidirectional Reflectance Distribution Functions (BRDF). In what follows we will assume a Lambertian[1] reflection function for the intervening surfaces, although our models could easily accommodate other responses. Thus, for fixed $\boldsymbol{\ell}, \mathbf{c}$ the laser light follows many such three-bounce trajectories on its path to the camera by reflecting from all points $\mathbf{x}$ on the hidden object surface that have a direct unobstructed line of sight to both $\boldsymbol{\ell}$ and $\mathbf{c}$. By raster scanning the laser position $\boldsymbol{\ell}$ and changing the focal point of the camera $\mathbf{c}$, we retrieve multiple measurements corresponding to a set of K parameters $\mathcal{P} = \{(\boldsymbol{\ell}_i, \mathbf{c}_i) | i = 1, \ldots, K\}$.

The focus of our work will be an occluded NLOS imaging setup where the space between the illumination surface and the hidden object is occupied with occluders, whose effect on the imaging process is to obstruct and block light beams that propagate in their direction, as illustrated in Figure 4-1. We usually assume that the occluders are known objects, i.e. we have information about their exact position and optical parameters. This is reasonable in settings where the occluders, while being away from the observer may be visible to it, with an unobstructed line of sight, as opposed to the hidden object which is hidden from view.

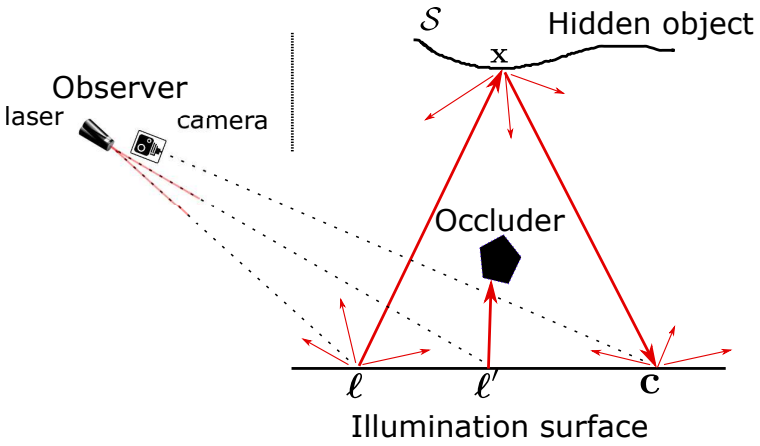
4.2.2 Forward Model

In this section we introduce a forward propagation model that determines the irradiance waveform $y_{\boldsymbol{\ell},\mathbf{c}}(t)$ measured at point $\mathbf{c}$ on the illumination surface in response to a single optical laser pulse $p(t)$ fired towards position $\boldsymbol{\ell}$.

In order to account for the presence of occluders in the scene (as illustrated in Figure 4-1), we introduce a binary *visibility function* $V(\mathbf{x}, \mathbf{z})$ which determines whether point $\mathbf{x}$ on the hidden object surface $\mathcal{S}$ and point $\mathbf{z}$ on the illumination surface are visible to each other:

$$V(\mathbf{x}, \mathbf{z}) = \begin{cases} 1, & \text{clear line of sight between } \mathbf{x} \text{ and } \mathbf{z}, \\ 0, & \text{no line of sight between } \mathbf{x} \text{ and } \mathbf{z}. \end{cases} \tag{4.1}$$

[1] As we further discuss next, a Lambertian surface reflects light diffusely, i.e an incoming focused narrow beam of light reflects isotropically in all directions, resulting in a matt surface appearance. This stands in contrast to specular reflection where an incoming light beam remains focused and narrow following reflection, resulting in a mirror-like surface appearance. This is illustrated in Figure 4-1 by the red arrow heads.



A side view illustration of the hidden-scene reconstruction setup. Red lines trace beam paths reflecting from the virtual laser points $\boldsymbol{\ell}, \boldsymbol{\ell}'$, where a laser beam hits the illumination surface towards point $\mathbf{x}$ on the hidden object. The illustrated beam emanating from $\boldsymbol{\ell}'$ is blocked by the occluder. Upon hitting the point $\mathbf{x}$ light reflects back towards a virtual camera position $\mathbf{c}$, where a focused camera is steered.

Figure 4-1: NLOS imaging setup.

With these definitions installed, the forward model is given as follows[2]:

$$y_{\boldsymbol{\ell},\mathbf{c}}(t) = \int_{\mathcal{S}} f(\mathbf{x}) \frac{V(\mathbf{x},\boldsymbol{\ell})V(\mathbf{x},\mathbf{c})}{\|\mathbf{x}-\boldsymbol{\ell}\|^2\|\mathbf{x}-\mathbf{c}\|^2} G(\mathbf{x},\boldsymbol{\ell},\mathbf{c}) p\left(t - \frac{\|\mathbf{x}-\boldsymbol{\ell}\| + \|\mathbf{x}-\mathbf{c}\|}{c}\right) \mathrm{d}\mathbf{x}. \tag{4.2}$$

Here, G is the Lambertian Bidirectional Reflectance Distribution Function (BRDF):

$$G(\mathbf{x},\boldsymbol{\ell},\mathbf{c}) \equiv \cos(\mathbf{x}-\boldsymbol{\ell},\mathbf{n}_{\boldsymbol{\ell}}) \cos(\mathbf{x}-\boldsymbol{\ell},\mathbf{n}_{\mathbf{x}}) \cos(\mathbf{x}-\mathbf{c},\mathbf{n}_{\mathbf{x}}) \cos(\mathbf{x}-\mathbf{c},\mathbf{n}_{\mathbf{c}}) \tag{4.3}$$

$\mathbf{n}_{\mathbf{x}}, \mathbf{n}_{\mathbf{c}}, \mathbf{n}_{\boldsymbol{\ell}}$ are the surface normals at $\mathbf{x}, \mathbf{c}, \boldsymbol{\ell}$, respectively and c is the speed of light. The model can easily be generalized to account for non-Lambertian BRDFs for the illumination surface and the hidden object by appropriately adjusting G.

We provide a detailed account for the forward model, by tracking light as it propagates from the laser, positioned at Λ, until it reaches the detector, positioned at Ω, accounting for the three bounces experienced along its path. In formulating the model (4.2) we ignore fixed known terms that are not functions of the spatial variable $\mathbf{x}$ as they can be pre-compensated for by the computational algorithm.

[2]A similar forward model is used in [44], and is based on well-known principles, namely quadratically decaying power attenuation for optical beams, and Lambert's cosine law for diffuse reflection. Eqn. (4.2) further accounts for possible occlusions in the scene through the visibility function.

Consider the three-bounce trajectory $\Lambda \to \boldsymbol{\ell} \to \mathbf{x} \to \mathbf{c} \to \Omega$ illustrated in Figure 4-1. The light travel time along this path is given by $\frac{1}{c}\left(\|\boldsymbol{\ell}-\Lambda\|+\|\mathbf{x}-\boldsymbol{\ell}\|+\|\mathbf{c}-\mathbf{x}\|+\|\Omega-\mathbf{c}\|\right)$ resulting in the detection of a delayed optical pulse $p(t)$ at the camera. By shifting the time axis by the fixed known delay $\frac{1}{c}\left(\|\boldsymbol{\ell}-\Lambda\|+\|\Omega-\mathbf{c}\|\right)$ we arrive at the term $p\left(t-\frac{\|\mathbf{x}-\boldsymbol{\ell}\|+\|\mathbf{x}-\mathbf{c}\|}{c}\right)$ appearing in (4.2).

As light travels from Λ to Ω it experiences quadratic power decay during the free-space propagation occurring in the sections $\boldsymbol{\ell} \to \mathbf{x}$, $\mathbf{x} \to \mathbf{c}$, $\mathbf{c} \to \boldsymbol{\Omega}$, i.e. it is multiplied by $\|\mathbf{x}-\boldsymbol{\ell}\|^{-2}\|\mathbf{c}-\mathbf{x}\|^{-2}\|\Omega-\mathbf{c}\|^{-2}$. We can compensate for the known power decay occurring in the final segment between $\mathbf{c}$ and Ω by multiplying the received signal by the fixed term $\|\Omega-\mathbf{c}\|^2$, resulting in the two remaining quadratic terms appearing in (4.2).

On the path from Λ to Ω the light experiences three reflections, at $\boldsymbol{\ell}, \mathbf{x}$ and $\mathbf{c}$. Each reflection results in scaling the light intensity by a reflectivity coefficient and the appropriate Lambertian BRDF response. The reflectivity coefficients at $\boldsymbol{\ell}$ and $\mathbf{c}$ are assumed known and can be compensated for. The reflectivity coefficient at $\mathbf{x}$, $f(\mathbf{x})$ is the objective of our reconstruction and appears as a scaling factor in (4.2). As for the BRDF terms, a reflection occuring at $\mathbf{z}$ with incoming radiation arriving from $\vec{in}$ and measured at $\vec{out}$ directions results in response $\cos(\vec{in}, \mathbf{n_z}) \cdot \cos(\vec{out}, \mathbf{n_z})$ where $\mathbf{n_z}$ is the normal at $\mathbf{z}$. Accounting for the three reflections we have six cosine factors appearing in the forward model, of which, the first and last, describing light from Λ hitting at $\boldsymbol{\ell}$ and light from $\mathbf{c}$ emerging towards Ω are fixed, known and can be compensated for, and can thus be excluded from the model. The remaining four cosine terms appear in (4.2) in the term $G(\mathbf{x}, \boldsymbol{\ell}, \mathbf{c})$.

The factors $V(\mathbf{x}, \boldsymbol{\ell})V(\mathbf{x}, \mathbf{c})$ appearing in the forward model (4.2) account for the presence of the occluder. Namely, light traveling the trajectory $\Lambda \to \boldsymbol{\ell} \to \mathbf{x} \to \mathbf{c} \to \Omega$ can make it through space unobstructed only if the direct line of sight paths $\boldsymbol{\ell} \to \mathbf{x}$ and $\mathbf{x} \to \mathbf{c}$ are unobstructed by the occluder, i.e. $V(\mathbf{x}, \boldsymbol{\ell}) = 1$ and $V(\mathbf{x}, \mathbf{c}) = 1$. If this does not hold, the path does not contribute to the signal measurement taken at the camera.

Finally, the measured intensity $y_{\boldsymbol{\ell},\mathbf{c}}(t)$ results from a superposition of all possible three bounce paths of the form $\Lambda \to \boldsymbol{\ell} \to \mathbf{x} \to \mathbf{c} \to \Omega$, which entails integration over all $\mathbf{x} \in \mathcal{S}$ as accounted for in (4.2).

4.2.3 Comments on the Forward Model

Here we comment on several details and modeling aspects related to the the forward model (4.2) and the occluded NLOS imaging setup:

Occluders

We conceptually think about the occluder (or multiple occluders) as an accidental object that happens to partially block the field of view between the illumination wall and the hidden object, and while some applications may call for purposefully designing and placing such an occluder to aid in NLOS imaging, we mostly consider its characteristics as given.

Throughout most of this work we assume that the occluder is observable from the point of view of the observer (refer to Figure 4-1) such that its parameters are fully known for the purpose of image reconstruction. The occluder may have any arbitrary shape, however its effect on the measurements in (4.2) is summarized solely through the binary visibility function $V(\mathbf{x}, \mathbf{z})$, implying that the occluder is either not blocking or fully blocking light propagating between the illumination surface and the hidden object. However, real occluders may additionally exhibit reflectance of their own which may superpose on top of the third-bounce signal from the hidden object in $y_{\boldsymbol{\ell},\mathbf{c}}(t)$. When the reflectivity pattern of the occluder is known this can be incorporated into our forward model without changing the reconstruction process.

For simplicity, in what follows we assume a fully absorbing occluder. In Section 4.10 we hint at the blind deconvolution problem where the occluder is not fully known. This can happen when the occluder itself is distant from the illumination wall and thus hidden from the observer. We argue that scene reconstruction is possible under some conditions even in that difficult setting.

While occlusions are traditionally considered a hindrance for imaging problems, we identify scenarios in which occlusions can be used in favor of better reconstruction. Introducing the visibility function in (4.2) is important for this purpose as will become apparent in Section 4.5.

Third Bounces

Our model (4.2) only accounts for the contributions in the measurements resulting from three bounces ($\Lambda \to \boldsymbol{\ell} \to \mathbf{x} \to \mathbf{c} \to \Omega$) which are informative about the hidden object. In most experimental setups (e.g. the one we report in Section 4.9) higher order bounces experience high attenuation and can be neglected in modeling the measurements.

As we detailed before, the contribution of the reflectance $f(\mathbf{x})$ at target spatial location $\mathbf{x}$ to the measurements is determined by the attenuation and temporal delays accrued by light propagating from $\boldsymbol{\ell}$ to $\mathbf{c}$ through $\mathbf{x}$, whereas the attenuation, delays and reflections associated with the paths from the laser to $\boldsymbol{\ell}$ and from $\mathbf{c}$ to the detector are known, can be compensated for, and thus are not incorporated in our model. In general, it is useful to think of $\boldsymbol{\ell}$ and $\mathbf{c}$ as 'virtual' unfocused illumination and detection points, ignoring the paths leading to and from these points, respectively.

Far-field Approximation

In the sequel we occasionally resort to using approximations to (4.2) when convenient for analysis and intuition, or useful for a clearer presentation. In particular, we will sometimes consider the *far-field* scenario in which the scene is far from the illumination surface such that $\|\mathbf{x}-\boldsymbol{\ell}\|^2\|\mathbf{x}-\mathbf{c}\|^2 \approx \text{const}$ and $G(\mathbf{x},\boldsymbol{\ell},\mathbf{c}) \approx \text{const}$.

Low Power Regime

We presented (4.2) as a light propagation model for pulsed laser illumination. Third bounce reflections are very weak in practical applications and a single pulse may result in such low reflected light intensity that just a few photons will be recorded at the detector at seemingly random times. However, in these settings repeating the pulsed illumination numerous times and averaging the recordings over many such cycles yields the temporal waveform predicted in (4.2). Thus, taking a single measurement with fixed $\boldsymbol{\ell}, \mathbf{c}$ can be a lengthy process and we are usually limited by a budget for the total number of such measurements we can take. We explore this further in Section 4.6.

Temporal Resolution of the Camera

The camera averages the incident irradiance at $\mathbf{c}$ with a finite temporal resolution Δt resulting in measurements $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau}, \tau = 1, 2, \ldots, T$,

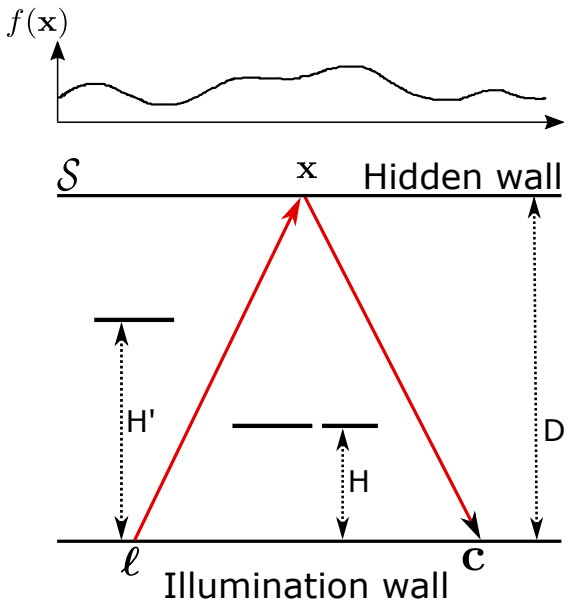
$$\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau} = \int_{(\tau-1)\Delta t}^{\tau\Delta t} \mathbf{y}_{\boldsymbol{\ell},\mathbf{c}}(t)\mathrm{d}t. \tag{4.4}$$

Since only third-bounce reflections involving the hidden object are of interest to us, with some abuse of notation we shift the time axis such that time $t = 0$ is the first instant when third bounce reflections reach the camera and $T\Delta t$ is chosen such that all relevant third-bounce reflections from the hidden object are included in the interval $[0, T\Delta t]$.

Notice that as the temporal resolution degrades, i.e. Δt grows larger, the number of samples retrieved from a single measurement configuration decreases, until finally, for Δt large enough we collect just a single sample for each measurement configuration. We refer to this limit as the non-TR limit. While it might appear that non-TR measurements are not informative about the hidden object, we will show in Section 4.5 that exploiting the presence of occluders in the scene makes NLOS imaging possible even with poor or non time-resolved measurements. A major advantage of utilizing non-TR measurements is that such measurements can be collected using simple experimental setups. Most significantly, a conventional cheap CCD detector may be used in lieu of the sensitive time-resolved detector. A CCD camera is cheaper, easier to setup and operate, and has better pixel resolution compared to a time-resolved camera. While the sensitivity of a CCD camera is worse than that of a SPAD detector, suggesting that a longer integration time might be necessary, notice that in the CCD setup the laser can be operated in a non-pulsed 'always on' mode, such that it can transmit a higher average power, resulting in faster imaging. Eq. (4.2) applies in this the non-pulsed, eq. non time-resolved, setting by taking $p(t) \equiv 1$.

4.3 Study Framework

In the previous section we introduced a general NLOS occluded imaging setup that is capable of capturing a wide variety of interesting scenarios, but is difficult for analytical study. In



A reference imaging setup in which the objective is to reconstruct the reflectivity $f(\mathbf{x})$ of a flat hidden wall that is parallel to the illumination wall at known distance D. The position and size of the fully absorbing flat occluders are known.

Figure 4-2: A simplified reference NLOS imaging setup.

this section, we specialize the general configuration and introduce a simplified reference setup that will allow a detailed study in subsequent sections.

4.3.1 Reference Imaging Setup

Our reference setup is illustrated in Figure 4-2. It is a specialized version of the general setup from Figure 4-1 and includes a planar hidden object and a parallel planar illumination surface, which we refer to as the *hidden wall* and the *illumination wall*, respectively. These two surfaces of known geometry are placed distance D apart.

In between the illumination and the hidden walls lie flat occluders, whose effect on the imaging process is captured through the visibility function defined in (4.1). As the geometry and location of the occluders is known, the visibility function can be trivially determined. The NLOS imaging objective under this setting is then to reconstruct the unknown reflectivity function $f(\mathbf{x})$ of the hidden wall from the measurements.

From (4.4) and (4.2) we immediately have that the measurements $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau}$ are linear in the unknown reflectivity function $f(\mathbf{x})$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a discretization of the hidden

wall, then, according to (4.2), each measurement $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau}$ corresponds to a measurement vector $\boldsymbol{a}_{\boldsymbol{\ell},\mathbf{c},\tau} \in \mathbb{R}^N$ such that $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau} = \boldsymbol{a}_{\boldsymbol{\ell},\mathbf{c},\tau}^\top \mathbf{f}$, where $\mathbf{f} \equiv [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]^\top$.

We collect measurements by raster scanning the laser and camera positions over a total of K configurations $(\boldsymbol{\ell}, \mathbf{c})$, obtaining T time samples per each pair. Collecting theses measurements in a vector $\mathbf{y}$ of dimension $M = K \cdot T$, this gives rise to the linear system of equations $\mathbf{y} = \mathbf{A}\mathbf{f}$ where $\mathbf{A}$ is an $M \times N$ measurement matrix whose rows are vectors $\boldsymbol{a}_{\boldsymbol{\ell},\mathbf{c},\tau}^\top$ that correspond to the chosen $(\boldsymbol{\ell}, \mathbf{c})$ pairs and temporal resolution Δt. In this study we consider measurements that are contaminated by additive noise $\boldsymbol{\epsilon}$:

$$\mathbf{y} = \mathbf{A}\mathbf{f} + \boldsymbol{\epsilon}. \tag{4.5}$$

The noise term can be thought of as a simple means to capture system modeling errors, camera quantization errors, background noise, etc..

4.3.2 Bayesian Priors

The idea of imposing Bayesian priors is well-established in image processing [9, 35]. Past studies have considered various forms of Gaussian prior distributions on the unknown target scene, including variations promoting sparse derivatives [68], and natural image statistics [77]. Such priors offer enough flexibility and at the same time are amenable to analysis and intuitive interpretation. In this work, we impose the following Gaussian prior on the reflectivity vector $\mathbf{f}$[3]:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{f}}), \tag{4.6}$$

For the covariance, in line with the modeling choices we applied in previous chapters (and also Appendix B), we impose a smoothness-promoting kernel function such that the entries of the covariance matrix are $[\Sigma_{\mathbf{f}}]_{ij} = \exp(-\frac{1}{2\pi\sigma_{\mathbf{f}}^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ and the spatial variance $\sigma_{\mathbf{f}}^2$ controls the extent of smoothness. Additionally, we consider an i.i.d. Gaussian distribution for the

[3]The zero mean assumption is somewhat simplified, but not particularly restrictive. In order to respect the nonnegative nature of the reflectivity function, a positive additive mean should be added in all models considered here, but this addition has no effect on the qualitative conclusions drawn from our results.

measurement noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ such that the Signal to Noise Ratio (SNR) in our problem is given by $\text{SNR} = \text{Tr}\big(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{A}^\top\big)/(M\sigma^2)$, where M denotes the total number of measurements. For the reconstruction, we consider the minimum mean-squared error (MMSE) estimator, which under the Gaussian framework is explicitly computable as

$$\hat{\mathbf{f}} = \boldsymbol{\Sigma}_\mathbf{f}\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{A}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{y}. \tag{4.7}$$

We measure and compare reconstruction performance in different settings using the normalized mean squared error $\text{NMSE} = \mathbb{E}\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2/\mathbb{E}\|\mathbf{f}\|_2^2$, which equals the (normalized) trace of the posterior covariance matrix

$$\text{NMSE} = \frac{1}{M}\,\text{Tr}(\boldsymbol{\Sigma}_\mathbf{f} - \boldsymbol{\Sigma}_\mathbf{f}\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{A}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{A}\boldsymbol{\Sigma}_\mathbf{f}). \tag{4.8}$$

Note that the NMSE can be evaluated before collecting measurements $\mathbf{y}$. Also, the reconstruction in (4.7) remains the optimal linear estimator under given first and second order statistics for $\mathbf{f}$, even beyond Gaussian priors.

4.4 Unoccluded Time-Resolved NLOS Imaging

In this section we study the limits of traditional NLOS imaging that is based on collecting TR optical measurements, and set up a reference against which we compare the newly proposed imaging modality that uses occlusions and no TR measurements, which we formally introduce in Section 4.5.

4.4.1 Collecting Time-Resolved Measurements

Here, focusing on the setup described in Section 4.3 (Figure 4-2) we review the main principles of TR NLOS imaging[4].

The hidden wall is indirectly illuminated with a short laser pulse and the reflected light is measured using a camera with temporal resolution Δt, as defined in (4.4). For each $(\boldsymbol{\ell}, \mathbf{c})$

[4]In this work we consider the task of imaging a hidden object of known geometry $\mathcal{S}$. TR measurements can additionally be used to estimate the geometry of the hidden object if it is unknown, e.g. [115].

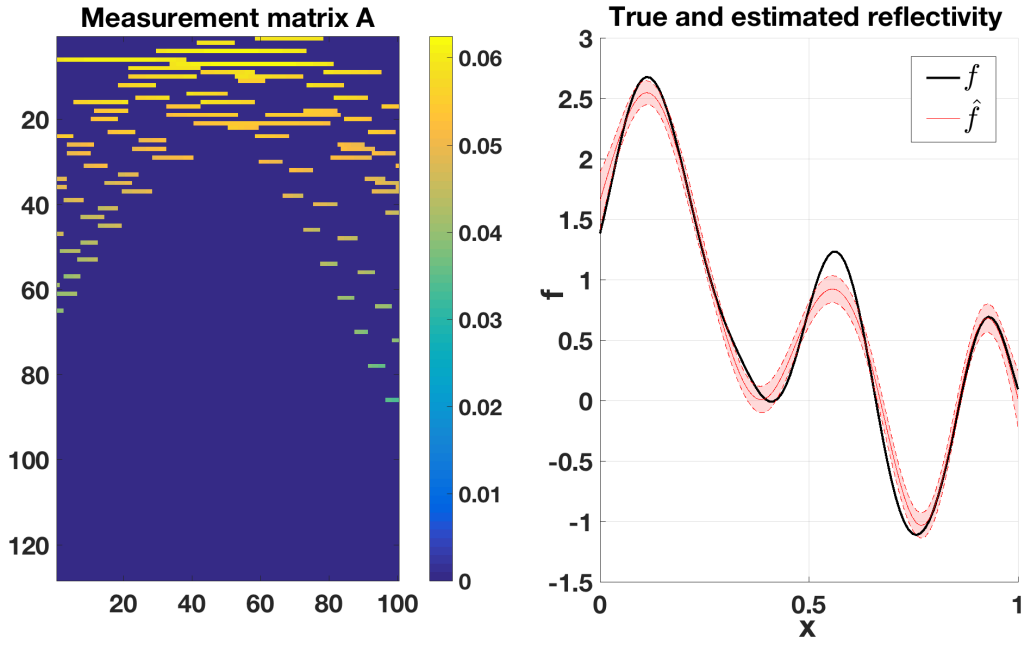
configuration pair we collect a set of T measurements $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau}$ for $\tau = 1, \ldots, T$ as discussed in Section 4.3.1. Assuming an ideal pulse $p(t) = \delta(t)$, and considering the propagation of optical pulses at the speed of light c, the measurement $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c},\tau}$ taken at time step τ forms a linear combination of the reflectivity values of only those scene patches $\mathbf{x}_i$ whose sum distance to $\boldsymbol{\ell}$ and $\mathbf{c}$ corresponds to a propagation time around $\tau\Delta t$. These patches fall within the elliptical annulus with focal points $\boldsymbol{\ell}$ and $\mathbf{c}$ described by the following equation:

$$(\tau - 1) \cdot c\Delta t \leq \|\mathbf{x}_i - \boldsymbol{\ell}\| + \|\mathbf{x}_i - \mathbf{c}\| \leq \tau \cdot c\Delta t \tag{4.9}$$

The thinner the annulus (eqv. the lower Δt), the more informative the measurements are about the reflectivity values of these patches. Furthermore, scanning the laser and camera positions $(\boldsymbol{\ell}, \mathbf{c})$, different sets of light paths are probed, each generating a different set of elliptical annuli. For a total of K $(\boldsymbol{\ell}, \mathbf{c})$-pairs, this forms the linear system of equations (4.5), with a total of $M = K \cdot T$ measurements.

Most reported experimental work utilizing TR measurements have used filtered back-projection as a heuristic for reconstructing the hidden object from the measurements (see [115]), i.e., for each potential hidden patch, sum the measurements that could result due to reflections originating from the patch (according to (4.9)). The resulting reconstruction is often blurry, but it can be computationally sharpened by applying post-processing heuristics [115]. Alternatively, others have suggested solving the linear system via some form of regularized least-squares accounting for prior scene knowledge, e.g. [44]. Here, operating in the Bayesian setting of Section 4.3, we obtain the optimal MMSE estimate for $\mathbf{f}$.

We performed a numerical simulation to demonstrate scene reconstruction performance in a TR setup. For the purposes of illustration the simulations presented here and in the sequel consider two dimensional layouts. This allows for easy visualization of important concepts such as the visibility function and the forward measurement operator, and it enables useful insights, but is otherwise non-restrictive. The room size was set such that the width of the walls is 1m, the distance between the walls is $D = 2$m and the temporal resolution was set at $\Delta t = 100$ps. $K = 8$ $(\boldsymbol{\ell}, \mathbf{c})$ pairs were randomly chosen, $\mathbf{f}$ was drawn according to the Gaussian prior with $\sigma_f^2 = 0.1$, and we set SNR = 13.7dB. The results are summarized in



(Left) Measurement matrix, where each row corresponds to a specific choice for the $(\boldsymbol{\ell}, \mathbf{c})$ pair and time index τ. The columns correspond to a discretization of the hidden wall to $N = 100$ patches. (Right) True reflectivity function versus the MMSE estimate $\hat{\mathbf{f}}$.

Figure 4-3: Scene reflectivity reconstruction from TR measurements.

Figure 4-3, where we plot the measurement matrix $\mathbf{A}$, the true reflectance $\mathbf{f}$ and the estimated $\hat{\mathbf{f}}$ with the corresponding reconstruction uncertainty depicted in shaded color around the MMSE estimator. The reconstruction uncertainty for our purposes is the square-root of the diagonal entries in the posterior covariance matrix corresponding to the standard deviation of $\hat{\mathbf{f}}_i - \mathbf{f}_i$ for the individual patches i on the wall. For this setup and resolution we collect $T = 16$ temporal samples per $(\boldsymbol{\ell}, \mathbf{c})$ pair such that the total number of measurements is $M = 8 \cdot 16 = 128$. These are the rows of $\mathbf{A}$ depicted in the figure, where each block of 8 consecutive rows corresponds to the measurements collected at a single time instant and for all $(\boldsymbol{\ell}, \mathbf{c})$ pairs. Notice that the last few blocks are zero as at those times no patch on the hidden wall contributes to the measurements.

4.4.2 Reconstruction Performance vs. Temporal-Resolution

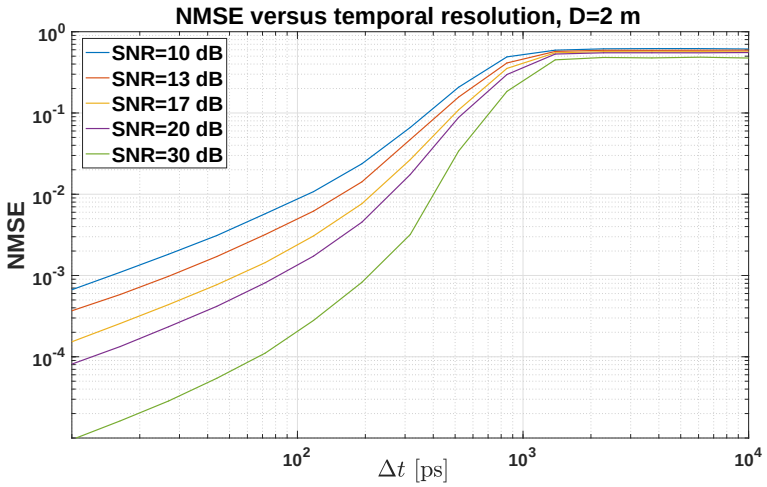
Here we explore the dependence of scene reconstruction performance on the temporal resolution of the available optical measurements. While reconstruction fidelity deteriorates when the temporal resolution becomes coarser, low-resolution equipment is less expensive and easier to set up, naturally defining a cost-performance tradeoff curve. For example, The simulation results reported in Figure 4-3 demonstrate high-fidelity reflectivity reconstruction when the available temporal resolution is fine ($\Delta t = 100$ps). However, practical technological and budget considerations limit the availability of such high resolution measurements resulting in significant deterioration of the reconstruction fidelity with less sensitive detectors, as we show next.

Let us first consider an extreme situation where the temporal resolution is so low such that the distance that light travels during a single resolution window of the detector is larger than the entire spatial extent of the hidden object[5]. As an example, for the setup in Figure 4-3 this happens when $\Delta t \gtrsim 1.5$ ns. In this extreme, which is essentially equivalent to collecting non-time-resolved measurements, each $(\boldsymbol{\ell}, \mathbf{c})$-pair effectively generates just a single scalar measurement which we denote $\mathbf{y}_{\boldsymbol{\ell},\mathbf{c}}$ and which is a linear combination of all the entries of $\mathbf{f}$. The combination coefficients are determined by the decay and cosine factors in (4.2). Focusing on the distance factors $\|\mathbf{x} - \boldsymbol{\ell}\|^{-2}\|\mathbf{x} - \mathbf{c}\|^{-2}$ for intuition, the range of values that these can take is clearly determined by the geometry of the problem, and can be very limited, e.g. when the two walls are far apart. This weak variation can result in poor conditioning of the measurement matrix $\mathbf{A}$ and subsequently poor reconstruction fidelity.

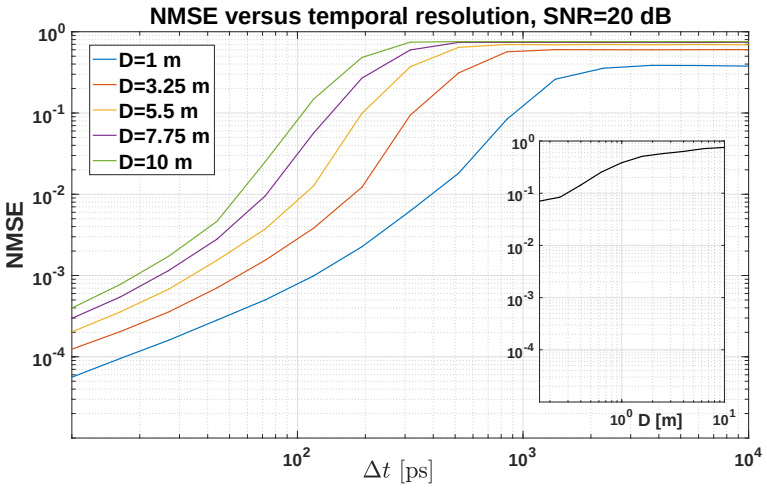
This is illustrated in Figure 4-4a where we plot the NMSE for the same setup as in Figure 4-3, with $K = 30$ measurements vs. the temporal resolution parametrized against the SNR (for each data point $(\Delta t, \text{SNR})$ we average over 10 random drawings for $(\boldsymbol{\ell}, \mathbf{c})$). We see that as temporal resolution deteriorates, reconstruction fidelity decreases. Introducing finite SNR for the purpose of this evaluation is key as reconstruction in an ideal noise-free experiment could result in high fidelity reconstruction even if $\mathbf{A}$ is ill conditioned. However, this is far from being true accounting for realistic noise levels[6] (e.g. SNR < 30dB).

[5] This happens when all hidden wall patches lie within the same elliptical annulus of (4.9)

[6] Each of the plots in Figure 4-4a correspond to a different SNR level. In practice, when comparing setups



(a) Normalized mean-squared reconstruction error versus temporal resolution (Δt), parameterized by the SNR level.



(b) Normalized mean-squared reconstruction error versus temporal resolution (Δt), parametrized by the room size.

Figure 4-4: Scene reconstruction error versus the available detector temporal resolution in TR NLOS imaging.

When imaging more distant walls, the poor conditioning of $\mathbf{A}$ further deteriorates as the distance decay factors become less varied and approach $\|\mathbf{x} - \boldsymbol{\ell}\| \approx \|\mathbf{x} - \mathbf{c}\| \approx D$, as illustrated in Figure 4-4b where reconstruction performance is parametrized against D for a fixed SNR in a setup with otherwise identical parameters as those of the first subfigure. In particular notice in this plot the limit of non-time-resolved measurements $\Delta t > 1.5$ns where the reconstruction squared error is always poor but is especially bad for larger D. This limit is separately summarized in the inset, which captures the fact that unless the room size is particularly small (i.e. just a few cm) it is hopeless to attempt high fidelity reconstruction.

Summarizing, we see that unless very high resolution measurement are available, NLOS scene reconstruction becomes ill-posed and reconstruction is not robust. In the next section we discuss the role of occluders in facilitating high fidelity reconstruction in this non-time-resolved and practical room size setting.

4.5 Imaging with Occluders

Existing NLOS imaging systems for static scenes rely on obtaining TR measurements. In this setting, occlusions are traditionally perceived as interfering with the imaging process by obstructing the optical paths propagating from the laser to the scene and back to the camera. In this section we study the role of occluders in NLOS imaging and demonstrate that in some situations not only is their presence not impeding the imaging process, but in fact it can enable high fidelity hidden scene reconstruction without the need for time resolved measurements, which would not otherwise be possible as was shown in Section 4.4.2. In subsequent sections we will mostly be interested in this regime of collecting non-TR measurements in the occluded NLOS setting.

---

of different temporal resolving capabilities the equipment involved will be technologically different such that a fair comparison does not necessarily entail assuming a fixed SNR common to all setups. Notice however the general trend of worsening reconstruction performance with diminishing temporal resolutions which holds for all SNR levels.

Informative Measurements Through Occlusions

We showed in Section 4.4 that image reconstruction performance drops significantly as the resolution of the temporally resolved optical measurement deteriorates. The inversion problem in the poor temporal resolution limit is inherently difficult as rows of $\mathbf{A}$, the linear forward operator, are smooth functions over the spatial target coordinate $\mathbf{x}$, resulting in bad-conditioning of the operator.

The situation changes drastically when the line of sight between $\boldsymbol{\ell}$ (and $\mathbf{c}$) and the hidden wall is partially obstructed by an occluder. For each measurement pair, certain segments of the hidden wall (that are different for different measurement pairs) are occluded from $\boldsymbol{\ell}$ and from $\mathbf{c}$. This is encoded in the linear forward operator $\mathbf{A}$ via zero entries on the corresponding spatial target coordinates $\mathbf{x}$, such that its rows are choppy and varied. Consequently, the inverse problem (4.5) becomes significantly better conditioned. We make this idea concrete immediately next.

Recall from Section 4.2 that in the absence of temporal resolution, when Δt effectively goes to infinity, measurements $y_{\boldsymbol{\ell},\mathbf{c}}$ correspond to integrating (4.2) over all t according to (4.4), i.e.,

$$y_{\boldsymbol{\ell},\mathbf{c}} = \int_{\mathcal{S}} f(\mathbf{x}) \frac{V(\mathbf{x},\boldsymbol{\ell})V(\mathbf{x},\mathbf{c})}{\|\mathbf{x}-\boldsymbol{\ell}\|^2\|\mathbf{x}-\mathbf{c}\|^2} G(\mathbf{x},\boldsymbol{\ell},\mathbf{c})\mathrm{d}\mathbf{x}. \tag{4.10}$$

let L be the number of distinct occluders $\mathcal{O}_i$, $i = 1,\ldots,L$ that are present in the scene. We associate a distinct (binary) visibility function $V_i(\mathbf{x},\mathbf{z})$ to each one of them. Observe then that the overall visibility function $V(\mathbf{x},\mathbf{z})$ is given as the product of the individual visibility functions, i.e. $V(\mathbf{x},\mathbf{z}) = \prod_i V_i(\mathbf{x},\mathbf{z})$. In terms of the forward operator $\mathbf{A}$, it holds that

$$\mathbf{A} = \mathbf{A}_0 \circ (\mathbf{V}_1 \circ \cdots \circ \mathbf{V}_L), \tag{4.11}$$

where $\mathbf{A}_0$ is the operator corresponding to a scene with no occluders, $\mathbf{V}_i$ is the (binary) visibility matrix with entries

$$(\mathbf{V}_i)_{(\boldsymbol{\ell},\mathbf{c}),\mathbf{x}} = V_i(\mathbf{x},\boldsymbol{\ell})V_i(\mathbf{x},\mathbf{c}), \tag{4.12}$$

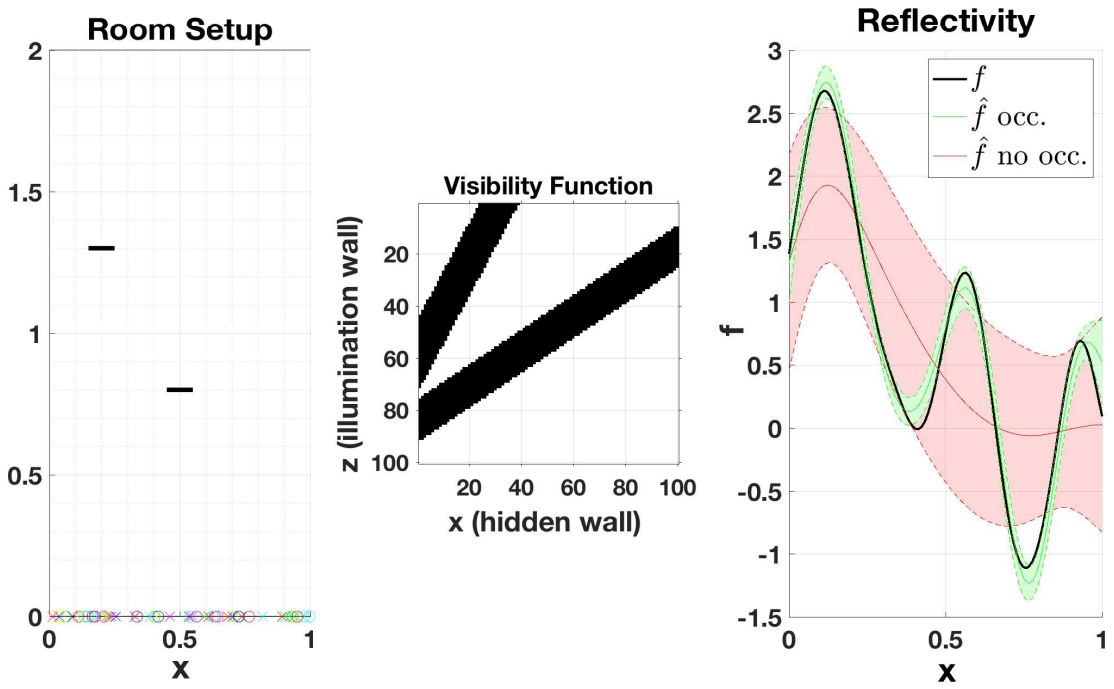
and $\circ$ denotes the Hadamard entrywise product of matrices.

As discussed in Section 4.4.2, the operator $\mathbf{A}_0$ is generally ill-conditioned as successive entries of any of its rows exhibit small and smooth variations due only to the quadratic distance attenuation and the BRDF factors G in (4.10). On the other hand, the Hadamard multiplication with nontrivial binary visibility matrices results in a well-conditioned operator. This is demonstrated through an example in Figure 4-5, which compares reconstruction performance in the presence and absence of occluders.

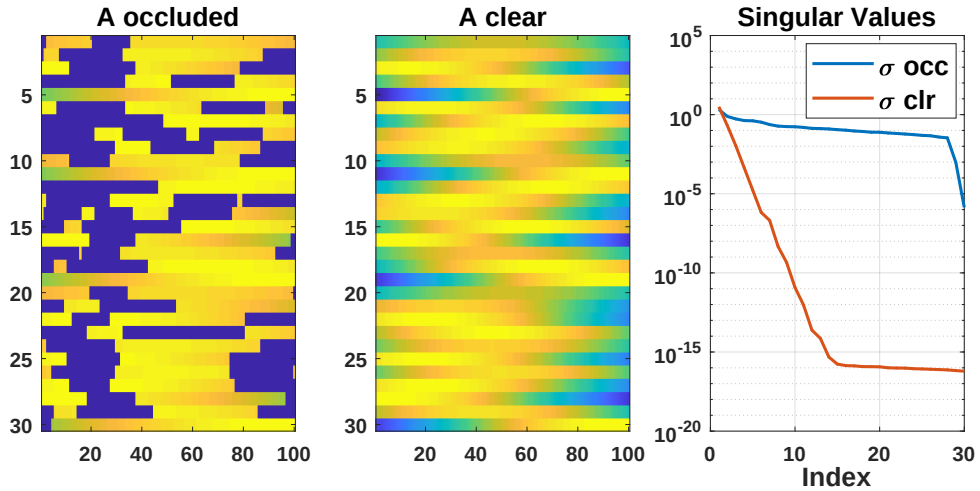
The setup, illustrated in Figure 4-5a, is as reported in previous simulations, with the addition of occluders as depicted. We collect $K = 30$ measurements with randomly drawn $\boldsymbol{\ell}, \mathbf{c}$ parameters and noise variance SNR = 25dB. The occluded measurement matrix $\mathbf{A}$ and the "un-occluded" matrix $\mathbf{A}_0$ are depicted in Figure 4-5b alongside their corresponding singular values. Observe that the singular values of $\mathbf{A}_0$ decay substantially faster than those of $\mathbf{A}$, which exhibits a much flatter spectrum. As expected, this better conditioning translates to better image reconstruction, as illustrated in the rightmost top plot: in solid red is the poor reconstruction without the occluder (NMSE = 54%), and in solid green is the successful reconstruction with the occluder (NMSE = 2.4%). The dashed lines indicate the standard deviation of the error $\hat{\mathbf{f}}_i - \mathbf{f}_i$ for each spatial coordinate $\mathbf{x}_i$, which corresponds to the square-root of the diagonal entries of the posterior covariance matrix.

4.6 Data Collection Strategies

In previous sections we have considered a setting in which a focused laser source and a focused detector generate measurements corresponding to $(\boldsymbol{\ell}, \mathbf{c})$ pairs on the illumination wall. In this section we discuss several extensions to this generic setup. First, we consider the problem of choosing an optimal set of $(\boldsymbol{\ell}, \mathbf{c})$ combinations under a budget constraint. We then extend our generic measurement setup with one where a wide field-of-view detector is employed in lieu of the focused detector. Finally, we briefly discuss a restricted setup where the focused laser and detector are constrained to align with one another.



(a) (Left) Room setup. (Middle) Binary visibility matrix, with 0 (1) depicted in black (white). (Right) Scene reflectivity reconstruction.



(b) (Left) Measurement matrix for the occluded setup and for the (Middle) unoccluded setup, in jet colormaps. (Right) The corresponding singular values.

Figure 4-5: Numerical study of NLOS imaging in the presence and absence of occluders.

4.6.1 Optimal Experimental Design

Our generic imaging setup assumed measurements were obtained for a set of $(\boldsymbol{\ell}, \mathbf{c})$, laser-camera configurations. For each $(\boldsymbol{\ell}, \mathbf{c})$ configuration the laser and detector need to mechanically be steered towards their respective orientations[7]. Once locked in position, the detector starts recording signal, however, due to the high attenuation experienced by light as it bounces three times, the signal measured at the detector may be weak and a long dwell time may be required to attain sufficient SNR levels[8]. Thus, it is evident that the size of the set $\mathcal{P}$ of $(\boldsymbol{\ell}, \mathbf{c})$ pairs is a main determinant of the image acquisition time.

Reducing the system acquisition time required to attain a minimum performance level is of key significance in designing imaging systems. In particular, in situations where the scene might be evolving over time, e.g. due to motion of objects or the imaging equipment, it is important to be able to image the hidden scene in time periods much shorter than those characterizing typical scene evolution. A key step towards achieving this goal is devising efficient schemes for choosing the set $\mathcal{P}$ under a budget constraint, as we pursue in this section.

To make things concrete, let $\mathcal{D}$ be a (uniform) discretization of the illumination wall, such that $(\boldsymbol{\ell}, \mathbf{c})$ pairs are restricted to be chosen on the product set $\mathcal{D} \times \mathcal{D}$, and suppose we are allowed to collect at most K measurements. We are then interested in choosing optimal subsets $\mathcal{P} \subset \mathcal{D} \times \mathcal{D}$ of $(\boldsymbol{\ell}, \mathbf{c})$ pairs such that $|\mathcal{P}| \leq K$. Furthermore, we want to understand, under this optimal choice, how imaging performance improves as K increases and more measurements are allowed. We explore here an efficient strategy that provides answers to these questions, consistent with the formulations we have discussed in previous chapters.

As before, our goal is to choose the set $\mathcal{P}$ such that the corresponding measurement vector $\mathbf{y}_{\mathcal{P}} := \{\mathbf{y}_{\boldsymbol{\ell},\mathbf{c}} \mid (\boldsymbol{\ell}, \mathbf{c}) \in \mathcal{P}\}$ is the most informative about the unknown reflectance $\mathbf{f}$. Denoting $I(\cdot;\cdot)$ the mutual information between two (vector) random variables, this amounts

[7] In the experiment reported in Section 4.9 laser steering is done with a small mirror mounted on a motor controlled pivot.

[8] In the experiment reported in Section 4.9 the laser illuminates the scene with a sequence of thousands of pulses to attain sufficient signal levels.

to solving

$$\mathcal{P}^\star = \underset{\mathcal{P}:\mathcal{P}\subseteq\mathcal{D}\times\mathcal{D},|\mathcal{P}|\leq K}{\operatorname{argmax}} \quad G(\mathcal{P}) \tag{4.13}$$

$$G(\mathcal{P}) \equiv I(\mathbf{y}_\mathcal{P}; \mathbf{f}). \tag{4.14}$$

The optimization problem in (4.13) is NP-hard in general. However, under the framework of Section 4.3 the objective function $G(\mathcal{P})$ is monotonic and submodular, which can be easily derived as we did in previous chapters. With this result, the theory of submodular optimization suggests an efficient greedy solver that obtains near optimal solutions $\mathcal{P}^{\text{gr}}$ satisfying: $G(\mathcal{P}^{\text{gr}}) \geq (1 - \frac{1}{e})G(\mathcal{P}^\star)$ (Appendix A).

During each of its iterations, the greedy algorithm augments the set $\mathcal{P}$ with an additional configuration pair $(\boldsymbol{\ell}, \mathbf{c})$, for a total sequence of K iterations. The solution has the property $\mathcal{P}^{\text{gr}}_K \subset \mathcal{P}^{\text{gr}}_{K+1}$, where we have used subscript notation for the budget constraint on the size of $\mathcal{P}$. The algorithm picks the next element myopically given the solution set built so far, i.e as the element that maximizes the marginal information gain.

We illustrate the efficacy of this approach via numerical simulations. For the purpose of clearly illustrating the solution in a simple setting our setup is similar to that used to generate Figure 4-5, except we only consider one of the two occluders, the one centered around $x = 0.5$m. The noise variance is kept constant at $\sigma^2 = 0.1$, and we seek an optimal set $\mathcal{P}$ of measurement configurations under a budget constraint $|\mathcal{P}| \leq K$. Figure 4-6a shows the output of the greedy algorithm for the most informative $(\boldsymbol{\ell}, \mathbf{c})$ pairs for values of K up to 30. The selected parameters, marked with red crosses are accompanied by a number indicating the iteration cycle at which they were retrieved. Notice how the first two selections are positioned to the left and right of the occluder.

Figure 4-6b validates the optimality features of the output $\mathcal{P}^{\text{gr}}$ of the greedy algorithm by comparing it to an equal size subset of measurements chosen uniformly at random. For a fixed desired NMSE the number of measurements required when randomly picking configurations can be as large as double the number required with approximately optimal selection. On the other hand, observe that under both schemes the NMSE drops significantly for the first few added measurements and the marginal benefit degrades as more measurements are added.

It is important to note that the numerical experiment presented here serves as a very basic demonstration of our optimal measurement scheme. In extremely occluded environments where access to the hidden scene through reflections is limited our approach could hopefully result in an even more dramatic improvement over randomly aiming the laser and detector.

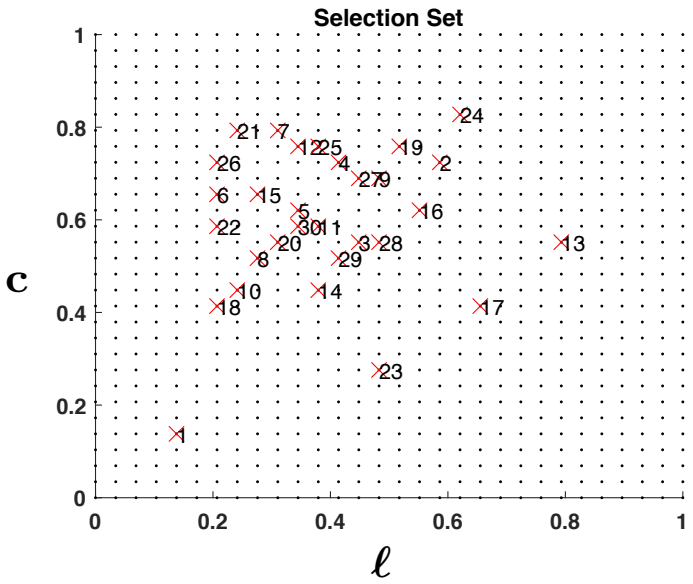
4.6.2 Single-Pixel Camera with a Wide Field of View

In contrast to prior art, our occluder assisted imaging method does not require collection of TR measurements, offering a potential reduction in equipment complexity. We additionally show here that our method can operate with a wide field-of-view single-pixel camera, offering several advantages such as reduced equipment cost (no lens required) and a dramatically increased signal to noise ratio as more photons can be collected per measurement. To the best of our knowledge, this is the first demonstration of NLOS imaging with a wide field-of-view detector.

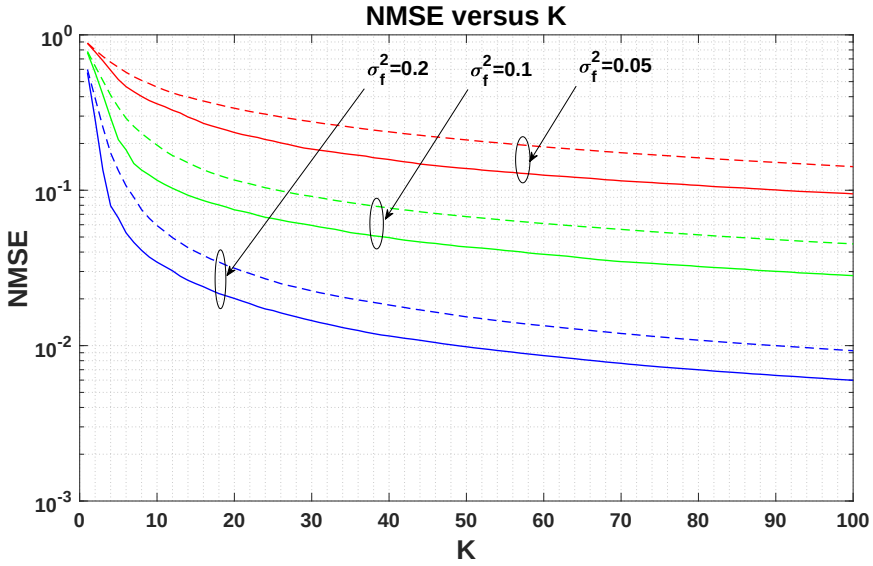
With the setup as before, consider a camera that is configured for a wide field of view, detecting light reflected from multiple positions $\mathbf{c}$ on the illumination wall, capturing more of the backscattered photons from the hidden scene. This modifies the forward measurement model as explained next. Let $\mathcal{C}$ represent the surface of the illumination wall that is in the camera field of view, so that the camera integrates photon measurements from all points $\mathbf{c} \in \mathcal{C}$, while the laser source raster scans the illumination wall as before. This procedure yields measurements that are now parametrized only by $\boldsymbol{\ell}$, as follows (cf. (4.10)):

$$\begin{aligned} y_{\boldsymbol{\ell}} &= \int_{\mathcal{C}} \frac{y_{\boldsymbol{\ell},\mathbf{c}}}{\|\mathbf{c}-\boldsymbol{\Omega}\|^2} \cos(\boldsymbol{\Omega}-\mathbf{c},\mathbf{n}_{\mathbf{c}}) \mathrm{d}\mathbf{c} \\ &= \int_{\mathcal{S}} f(\mathbf{x}) \frac{V(\mathbf{x},\boldsymbol{\ell})}{\|\mathbf{x}-\boldsymbol{\ell}\|^2} \left[\int_{\mathcal{C}} \frac{V(\mathbf{x},\mathbf{c}) G(\mathbf{x},\boldsymbol{\ell},\mathbf{c}) \cos(\boldsymbol{\Omega}-\mathbf{c},\mathbf{n}_{\mathbf{c}})}{\|\mathbf{x}-\mathbf{c}\|^2 \|\mathbf{c}-\boldsymbol{\Omega}\|^2} \mathrm{d}\mathbf{c} \right] \mathrm{d}\mathbf{x}. \end{aligned} \tag{4.15}$$

In deriving (4.15) we use (4.10) and further explicitly account for the quadratic power decay from the illumination wall to the camera positioned at Ω, and the BRDF term that accounts for the reflection from $\mathbf{c}$ to Ω. As before, the measurements are linear in the unknown reflectance, suggesting that the same reconstruction techniques can be utilized in this setup. In the presence of occluders, the nontrivial visibility function $V(\mathbf{x},\mathbf{z})$ results in improved con-



(a) Coordinates of virtual laser ($\boldsymbol{\ell}$) and camera ($\mathbf{c}$) positions. The set $\mathcal{D} \times \mathcal{D}$ of all possible locations is marked with black dots. The set $\mathcal{P}$ selected by the greedy algorithm for a budget constraint $K = 30$ is delineated with red crosses. The numbers indicate the order of selection.



(b) NMSE reconstruction performance versus the number of measurements for the random (dashed lines) and optimized (solid lines) configurations for various values of spatial correlation σ_f^2.

Figure 4-6: Experimental design for choosing informative measurements in occluded NLOS imaging.

ditioning for the measurement operator and successful image reconstruction. In particular, our experimental demonstration in Section 4.9 is based on the forward model in (4.15). We mention in passing that the dual setting, where a wide field-of-view light projector is utilized instead of a focused laser illumination, with measurements collected at multiple positions $\mathbf{c}$ on the illumination wall, might also be of interest.

4.6.3 Aligned Illumination and Detection

Finally, we mention a specific configuration that reduces the dimensionality of the parameter space by imposing the restriction $\boldsymbol{\ell} = \mathbf{c}$ on the measurement configurations[9]. This results in a strict subset of the entire measurement set $\mathcal{D} \times \mathcal{D}$ that is convenient for analytic purposes and for drawing insights about the features of the imaging system, and will be useful for our analysis in Section 4.7.

4.7 Model Misspecification

In this section, we study in more detail the structural properties of the visibility function, which we use in turn to study the robustness of reconstruction with respect to a misspecified description of the location of the occluder.

4.7.1 Parameterizing the Visibility Function

In what follows we consider flat horizontal occluders, i.e. occluders aligned horizontally at some fixed distance from the illumination wall (Figure 4-2). This family of occluders is useful as any occluder that is small compared to the size of the room may be well approximated as being flat and horizontal. We show that the visibility function V associated with a flat horizontal occluder has a simple structure. Specifically, suppose that occluder $\mathcal{O}$ lies on a horizontal plane at distance $H = \alpha D$ with $\alpha \equiv H/D$ from the visible wall, and define the occupancy function $s(\mathbf{x})$ such that for all points $\mathbf{x}$ on that plane set $s(\mathbf{x}) = 0$ if $\mathcal{O}$ occupies $\mathbf{x}$

[9] When $\boldsymbol{\ell} = \mathbf{c}$, the camera focused at $\mathbf{c}$ measures a first-bounce response in addition to the informative third-bounce. We assume here that the dimensions of the hidden scene are such that it is possible to use time-gating to reject this first-bounce signal.

and $s(\mathbf{x}) = 1$ otherwise[10]. A point $\mathbf{x}$ on the hidden wall is not visible from a point $\mathbf{z}$ on the illumination wall if and only if the line that connects them intersects with the occluder, or equivalently, if at the point of intersection it holds that $s(\alpha\mathbf{x}+(1-\alpha)\mathbf{z}) = 0$. This translates to:

$$V(\mathbf{x}, \mathbf{z}) = s(\alpha\mathbf{x} + (1 - \alpha)\mathbf{z}), \tag{4.16}$$

In particular, when $\boldsymbol{\ell} = \mathbf{c}$, it follows from (4.12) and (4.16)

$$(\mathbf{V})_{(\boldsymbol{\ell},\mathbf{c}),\mathbf{x}} = s(\alpha\mathbf{x} + (1 - \alpha)\boldsymbol{\ell}) \tag{4.17}$$

and the visibility matrix $\mathbf{V}$ has a band-like structure. Ignoring edge-effects, this corresponds to a convolution matrix, which is favorable since the convolution structure makes possible deriving analytic conclusions regarding the effect of the parameters of the occluder on the image reconstruction as shown next.

4.7.2 Misspecified Reconstruction

Thus far we have assumed perfect knowledge of occluder parameters for image reconstruction. However, in practice these parameters may be inaccurate, leading to reconstruction errors. In this subsection we study scene reconstruction under a misspecified model for the position of the occluders. Figure 4-7a illustrates our setup where the true position of the flat horizontal occluder appears in black, and our mismatched model assumes the occluder is positioned as appears in red, with δ_x and δ_H vertical and horizontal shifts, respectively.

We study the resulting image reconstruction under the following simplifications: (i) measurements are noiseless, (ii) measurements are taken with parameters satisfying $\boldsymbol{\ell} = \mathbf{c}$, (iii) continuous measurements are collected, i.e. $y_{\boldsymbol{\ell}}$ is available for all points $\boldsymbol{\ell}$ on the visible wall, and (iv) we assume that the hidden wall is far from the illumination wall such that $\|\mathbf{x} - \boldsymbol{\ell}\|^2\|\mathbf{x} - \mathbf{c}\|^2$ and $G(\mathbf{x}, \boldsymbol{\ell}, \mathbf{c})$ are approximately constant, i.e. this is the far-field approximate model of Section 4.2.3.

[10]Here, occluder $\mathcal{O}$ is allowed to be composed of several patches as long as they all lie on the same plane. Equivalently, the support of the function $s(\mathbf{x})$ on the plane can be disjoint.

With these simplifications, the measurements $\mathbf{y}_{\boldsymbol{\ell}}$ are expressed (up to a constant) as

$$y_{\boldsymbol{\ell}} = \int f(\mathbf{x}) s(\alpha \mathbf{x} + (1-\alpha)\boldsymbol{\ell}) \mathrm{d}\mathbf{x}, \tag{4.18}$$

where we have used (4.16), and $f(\mathbf{x})$ is the true reflectance of the hidden wall.

In the presence of errors δ_x, δ_H, the misspecified visibility function can be expressed as $\tilde{V}(\mathbf{x}, \mathbf{z}) = s(\alpha'(\mathbf{x} - \delta_x) + (1-\alpha')(\boldsymbol{\ell} - \delta_x))$, where $\alpha' \equiv \frac{H+\delta_H}{D} = \alpha + \frac{\delta_H}{D}$. This results in a misspecified model:

$$y_{\boldsymbol{\ell}} = \int \hat{f}(\mathbf{x}) s(\alpha'(\mathbf{x} - \delta_x) + (1-\alpha')(\boldsymbol{\ell} - \delta_x)) \mathrm{d}\mathbf{x}. \tag{4.19}$$

In order to study how $\hat{f}(\mathbf{x})$ relates to $f(\mathbf{x})$ it is convenient to work in the Fourier domain[11]. Taking Fourier transforms of the right-hand-side expressions of both (4.18) and (4.19), and equating with each other, it can be shown that[12],

$$\hat{F}(\boldsymbol{\omega}) = \frac{1-\alpha'}{1-\alpha} \frac{S(-\frac{1-\alpha'}{1-\alpha}\frac{\boldsymbol{\omega}}{\alpha'})}{S(-\frac{\boldsymbol{\omega}}{\alpha'})} e^{j\boldsymbol{\omega}\frac{\delta_x}{\alpha'}} F\left(\frac{\alpha}{\alpha'}\frac{1-\alpha'}{1-\alpha}\boldsymbol{\omega}\right), \tag{4.20}$$

where $G(\boldsymbol{\omega})$ denotes the Fourier transform of a function $g(\mathbf{x})$. Of course, this holds for frequencies at which $S(\boldsymbol{\omega})$ is non-vanishing.

The following conclusions regarding reconstruction distortion under misspecified position of the occluder are drawn from (4.20):

- Under no errors ($\delta_x = 0,\ \delta_H = 0$), the reflectivity function is perfectly reconstructed for those frequencies for which the shape-function of the occluder is non-zero.

- Horizontal occluder translation errors ($\delta_x \neq 0,\ \delta_H = 0$) result in simple shifts of the true reflectance.

- Vertical occluder translation errors ($\delta_x = 0,\ \delta_H \neq 0$) result in two kinds of distortions. The first is a scaling effect, while the other is a distortion that depends on the shape

[11] The variable of integration $\mathbf{x}$ in (4.18) and (4.19) ranges over the finite surface of the hidden wall. Correspondingly, $f(\mathbf{x})$ and $s(\mathbf{x})$ are only defined over this region. Formally, when it comes to taking Fourier transforms, we extend the functions on the rest of the space by zero-padding.

[12] Recall $\mathcal{F}[f(t)] = F(\boldsymbol{\omega}) \rightarrow \mathcal{F}[f(at+b)] = \frac{1}{|a|} e^{-j\boldsymbol{\omega}\frac{b}{a}} F(\frac{\boldsymbol{\omega}}{a})$.

of the occluder through the term $S(-\frac{1-\alpha'}{1-\alpha}\frac{\boldsymbol{\omega}}{\alpha'})/S(-\frac{\boldsymbol{\omega}}{\alpha'})$. For this latter term, observe that its effect becomes diminishing for a spectrum $S(\boldsymbol{\omega})$ that is mostly flat over a large range of frequencies. A very narrow occluder has (approximately) this property. (The approximation here is because of the finite support of $s(\mathbf{x})$, see Footnote 11)

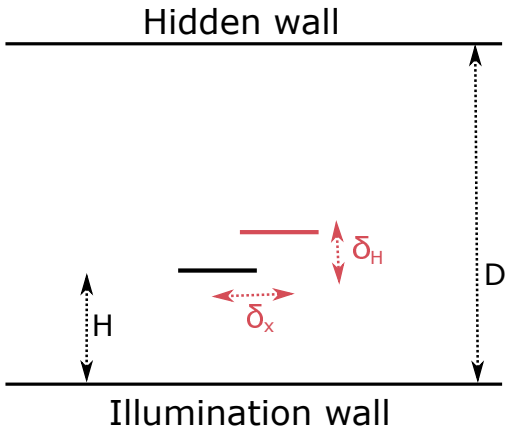
Recall that the conclusions above hold analytically in the limit of a far-field scene and a continuum of noiseless measurements. However, the conclusions are also suggestive and insightful for practical scenarios as illustrated by the numerical study shown in Figure 4-7, where we illustrate high SNR (35dB) reconstruction with a mispositioned occluder. The room setup is as usual with $D = 5$m, and a single (far-field) occluder of width 0.25m positioned at $[.5, 2]$m. Measurements are collected with random $\boldsymbol{\ell}$ and random $\mathbf{c} \neq \boldsymbol{\ell}$. Black solid lines show the true reflectance $f(\mathbf{x})$ whereas dashed green lines depict reconstruction with perfect occluder knowledge. The red curves show reconstructions with horizontally and vertically misspecified occluders. The misspecification is larger in the right subplot. It is evident from the images that horizontal misspecification mostly results in a shifted reconstruction, whereas vertical misspecification results in axis-scaling of the reconstructed scene. Our analytical analysis seems to mostly be valid for the middle section of the reflectivity function whereas edge effect appearing close to the boundaries $x = 0, 1$ are not captured by the analysis.

The robustness of our imaging method with respect to occluder positioning errors is further supported by the experimental demonstration in Section 4.9, where occluder model inaccuracies are unavoidable, yet the reconstruction results we demonstrate are satisfactory.

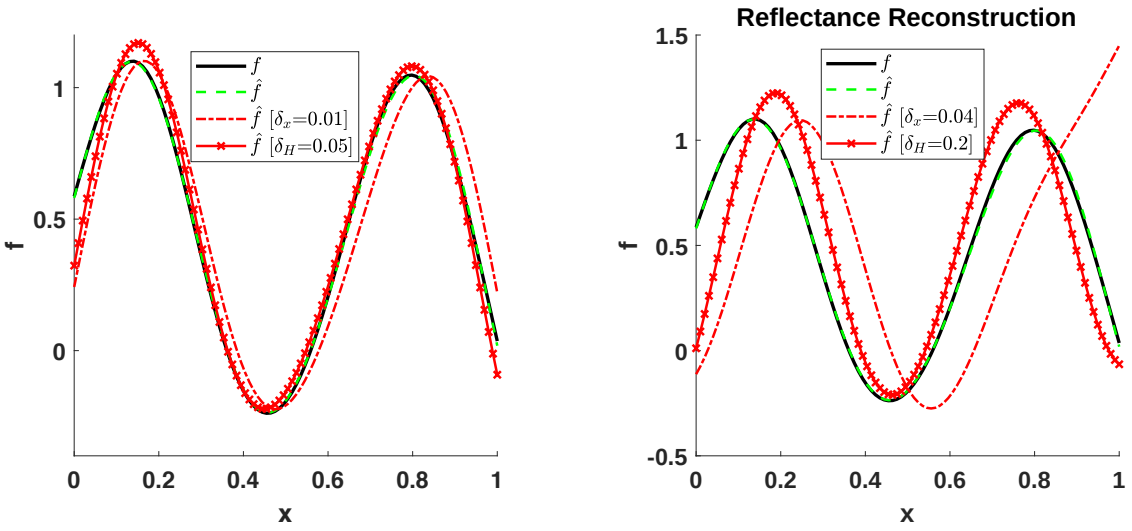
4.8 TR-Measurements in Occluded Settings

Thus far we have studied imaging systems that use either TR measurements in an unoccluded setting, or non-TR measurements in occluded settings. In this section we consider imaging systems that exploit TR measurements in occluded settings, and study their potential benefits. We present initial numerical simulations and leave a full study of this topic to future work.

To be concrete, consider the familiar setting of Figure 4-5a, but now assume the detector supports a non-trivial temporal resolution Δt. We sweep Δt over a range of values, and plot



(a) A shifted occluder setup. The occluder appears in its actual position in black. We perform reconstruction under imperfect knowledge of its position, taken to be as appears in red.



(b) Reconstruction with a misspecified occluder position: (Left) Small, and (Right) Large vertical and horizontal occluder shifts in a far field setup.

Figure 4-7: Occluded NLOS imaging with model inaccuracies.

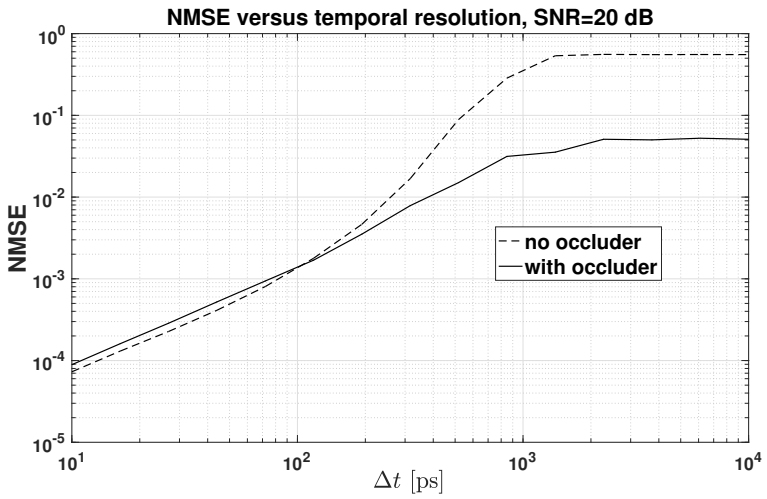


Figure 4-8: NLOS scene reconstruction performance versus temporal resolution in the presence and absence of occluders.

the resulting reconstruction NMSE in Figure 4-8 (solid curve). For comparison, we also plot in dashed line the NMSE performance in the absence of an occluder (this corresponds exactly to the plot in Figure 4-4a). For a large range of temporal resolutions (here, $\Delta t \gtrsim 150$ps) the presence of occlusions leads to a substantial increase in reconstruction performance, allowing the same level of performance to be maintained at inferior temporal resolution levels. When very high temporal resolution is available reconstruction performance with occlusions is slightly degraded with respect to the non occluded setting, due to the occluder blocking of some of the reflected signal.

Note here that TR measurements can be further utilized to improve on other aspects of the system. For instance, one might imagine using coarse TR measurements to find the position of the occluder, which has been up to now assumed known. A study of such possibilities might be an interesting direction for future research.

4.9 Experimental Demonstration

In this section we report experimental results demonstrating our methods and formulations. These results validate our forward model for light propagation in NLOS imaging and further inform future theoretical developments.

4.9.1 Experimental Setup[†]

A schematic illustration of our experimental setup is shown in Figure 4-9. A pulsed 640-nm laser source illuminates a nearly Lambertian visible wall (1st bounce). The light propagates to the hidden wall where it is scattered back towards the illumination wall (2nd bounce). Finally, the backscattered light is collected by a SPAD detector[13] (3rd bounce). In front of the SPAD, an interference filter centered at 640 nm is used to remove most of the background light. In the experiment, the SPAD is operated without a lens to achieve a wide field of view and it is configured for the left side of the visible wall to minimize the direct first bounce.

The occluder is a black circular patch without any back reflections. During the experiment, we turned off all ambient room light to minimize background noise.

4.9.2 Computational Processing

We assume that the geometry of the setup is known, including the locations of raster-scanned laser illumination, SPAD, visible wall and occluder, but the reflectance of the hidden wall is unknown. We use the forward model in Equation (4.15) and obtain an estimate $\hat{\mathbf{f}}$ of the true reflectance by solving the following non-smooth convex optimization problem:

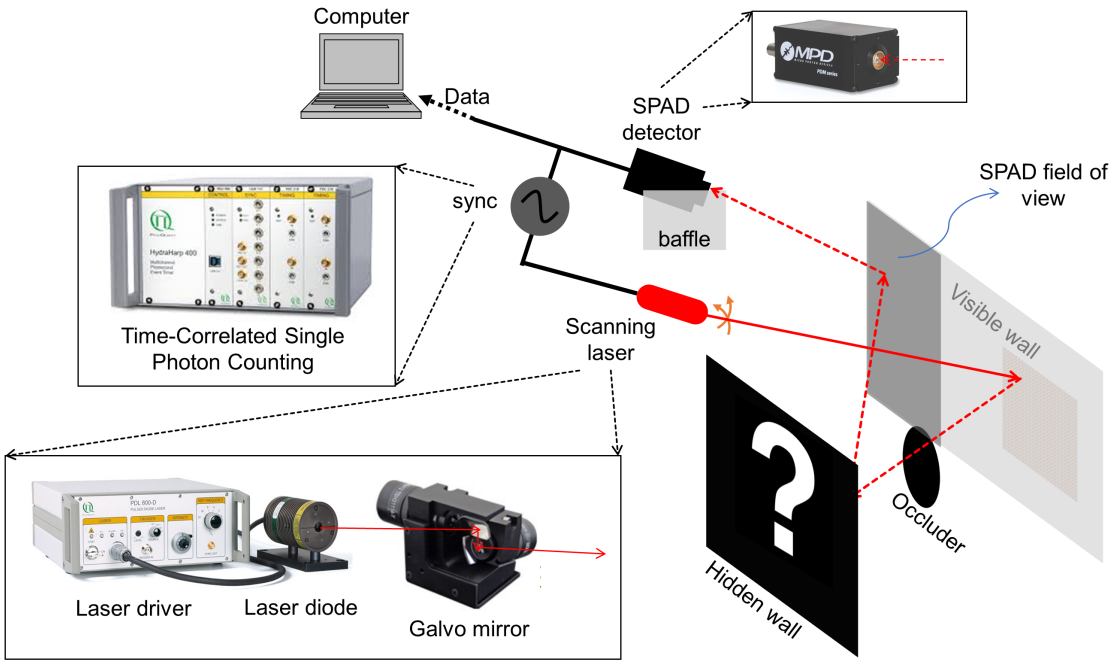
$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{f}\|_2^2 + \lambda\|\mathbf{f}\|_{\mathrm{TV}}, \tag{4.21}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the Total-Variation (TV)-norm and $\lambda > 0$ is a regularization parameter. To solve (4.21) we use an efficient dedicated iterative first-order solver [39], which is based on the popular FISTA algorithm [6].

TV-norm penalization is a standard technique that has been successfully applied in other image reconstruction tasks (e.g., image restoration [68, 6, 62]). Its use is motivated by

[†]The experimental setup was built and operated by Feihu Xu, who also collected the raw measurements.

[13]A SPAD is capable of providing time-resolved measurements. However, for the purpose of this experiment we operate the SPAD as a regular camera, essentially integrating the response over time. To be precise, we only use the time resolved measurements of the SPAD to gate-out the first-bounce response from the illumination wall. Beyond that, no TR measurements are recorded. Notice that the illumination wall is in the direct line of sight of the imaging equipment, thus its location can be well-estimated based on standard imaging techniques. With this information, the time window that corresponds to the first-bounce response is a-priori known. Hence, the same operation achieved here with a SPAD camera can be performed using a time-gated CCD camera.



[Illustration by Feihu Xu] The distances in the experimental setup are as follows: visible wall to hidden wall: ∼106 cm; visible wall to SPAD: ∼156 cm; visible wall to occluder: 37 cm; The diameter of the circular occluder is 3.4 cm.

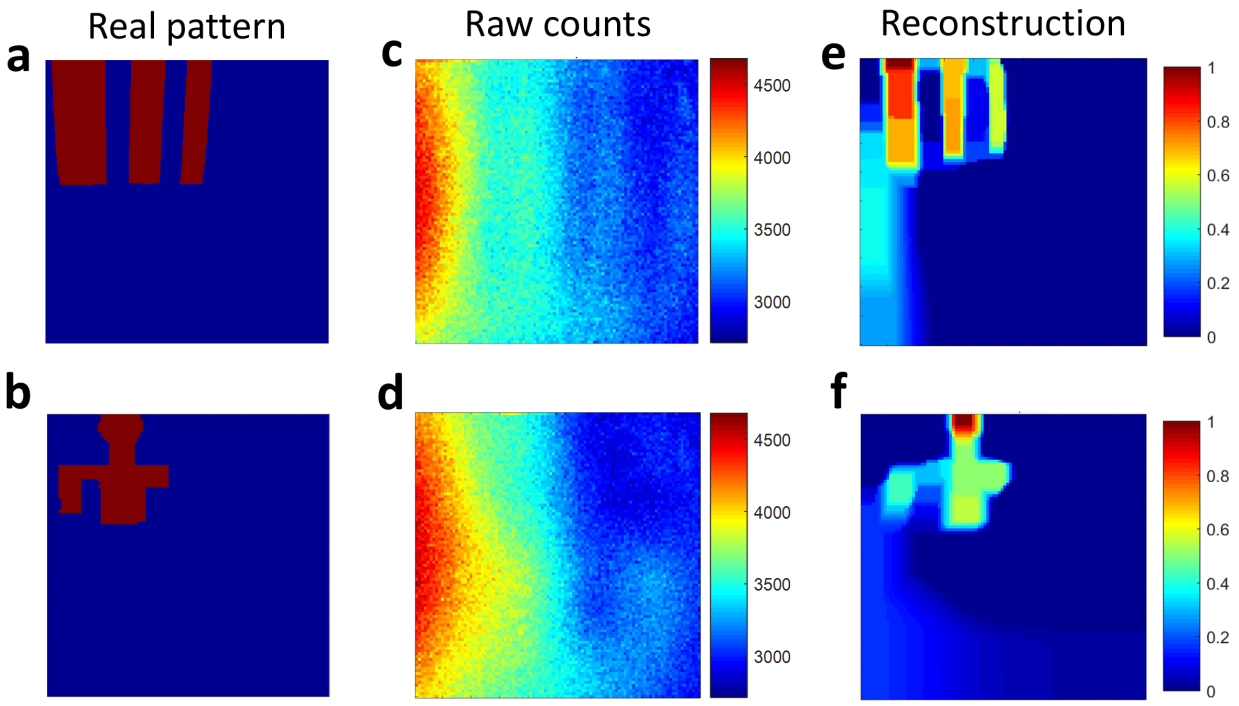
Figure 4-9: Experimental setup for demonstrating occluded NLOS imaging.

the observation that the derivatives of natural images obey heavy-tailed prior distributions [68, 62]. In Figure 4-11 we compare the nonlinear TV-based reconstruction to the linear reconstruction in (4.7) that assumes a Gaussian prior on $f(\mathbf{x})$ (see Section 4.3) with $\sigma_f^2 = 0.02$ and σ^2 tuned to achieve good results. As can be seen from this figure, TV regularization is more accurate and emphasizes edges as expected. The linear reconstruction is blurry but satisfactory and yields a reconstruction that is easily interpretable by the human eye. One should also note that the linear reconstruction is much more efficient in terms of computation. Both methods require tuning of one parameter (λ for TV and σ_f^2 for GP, in addition to σ^2 which can be analytically determined based on the SNR).

For this experiment we operate in a high-photon regime, under which the noise is well-modeled by an additive Gaussian vector $\boldsymbol{\epsilon}$. This motivates the least-squares term in Equation (4.21). In low photon count regimes the measurement model becomes Poisson noise rather than additive Gaussian noise, and the noise is signal-dependent rather than signal-independent [126].

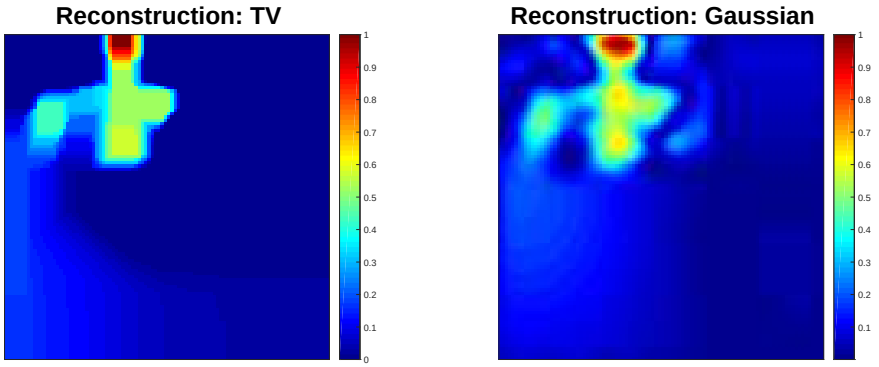
4.9.3 Experimental Results

Reconstruction results using the optimization method in (4.21) are shown in Figure 4-10. The regularization parameter λ was tuned independently for each algorithm to yield a reconstruction that is empirically closest to the ground truth. In Figure 4-10 two different scene patterns on the hidden wall were tested. The laser light was raster scanned on a 100×100 grid and at each point the SPAD detector was turned on for a fixed dwell time such that a total number of $\sim$9 million laser pulses ware emitted and $\sim$3500 back-reflected 3rd bounce photons were recorded on average. The raw measurement counts for each of the hidden patterns are shown in Figures 4-10(c,d): each of the 100×100 entries corresponds to a single measurement $\mathbf{y}_\ell$ for the corresponding virtual laser position $\boldsymbol{\ell}$. The raw counts are processed by the optimization algorithm (4.21) to obtain an estimate of the hidden patterns as shown in Figures 4-10(e,f). These results validate the forward model and the performance of the reconstruction algorithm.



(**a**,**b**) Ground truth of the tested scene patterns on the hidden wall. The patterns are placed in the upper-left corner of the hidden wall. (**c**,**d**) The raw measurement counts for a 100×100 raster-scanning laser points. At each laser point, we turn on the SPAD for a fixed dwell time such that $\sim$3500 photon counts are recorded on average. (**e**,**f**) Reconstruction results from Eq. (4.21).

Figure 4-10: Experimental NLOS imaging results.



(Left) Reconstruction with TV regularization, and (Right) reconstruction obtained via the Gaussian prior model, both shown in absolute value.

Figure 4-11: NLOS reconstruction results in the experimental setup.

4.10 Discussion

Our goal in this chapter was to study the challenging problem of optical NLOS imaging. State of the art techniques employed to tackle this problem involve very sensitive and costly time-resolved optical detectors that are able to record the temporal variation of reflected light and image the hidden scene. Our study shows that high temporal resolution is crucial for obtaining measurements that are informative and allow scene reconstruction, driving up system cost and complexity.

Our motivation to develop a more efficient method for collecting informative measurements in the NLOS setting led us to introduced and study the occluded NLOS imaging setup. We claimed and demonstrated that non-TR measurement of optical reflections collected in this setup are informative about the hidden scene, suggesting an entirely new imaging modality in such environments. We studied various aspects of this newly proposed imaging system, and in particular we developed an efficient technique to choose experimental parameters for collecting informative empirical data, shortening the scene acquisition time and driving down the overall system cost.

Future Research

This work introduced a novel occluded NLOS imaging modality. The initial studies reported here focus on a restricted setting where the geometry of a hidden flat scene and a collection of occluders is assumed known and the imaging goal is to retrieve the reflectivity on the hidden surface from diffuse optical reflections. While serving as a useful testing ground for demonstrating basic principles in occluder assisted NLOS imaging, our study framework and setup were simplified and suggest multiple directions for more in-depth future research, which we illuminate in this section.

Extending our imaging modality to problems involving hidden scene geometry recovery in addition to reflectivity estimation is of major interest in many practical applications. A first step in this direction is to consider restricted settings where geometry recovery can be treated as a parametric problem, for example in a setting similar to ours but with the room size, or the exact occluder location unknown, with its shape otherwise assumed given. Problems such as this can be tackled using generic maximum likelihood parameter estimation methods and it is interesting to fully characterize the circumstances in which accurate estimation is possible.

One simple variant of such problems is to consider a setting where the reflectivity function on the back wall is known and the goal is to retrieve the location of occluders in the room, which may represent, e.g. people or other objects of interest. Inspecting (4.2) this is immediately seen to lead to a quadratic problem in the visibility function, which may be solved in some circumstances by applying lifting techniques. We leave a detailed study of this to future work.

A more challenging extension would be to consider full non-parametric 3D scene reconstruction problems where not much is known about the hidden environment. While this seems like a much more challenging problem it is interesting to consider in this setting hybrid systems that employ limited resolution TR optical measurements in occluded environments, utilizing the structure the occluder endows on the measurements to facilitate high quality reconstruction.

We also mention in passing that our setting manifests a form of opportunistic imaging,

where we utilize some knowledge on the environment to enable image reconstruction that would have otherwise been impossible. Other setups may offer similar advantages, for example utilizing coincidental bumps or edges on the visible surface itself and the occlusions they introduce [110] to enhance imaging. Finally, it is natural to attempt extensions of the discussed methods to non-static environments. For instance a setting with a moving occluder following a known trajectory that facilitates collection of diverse measurements as it traverses the scene.

Chapter 5

Nonlinear MIMO Radar System Design

In this chapter we study Multiple-input multiple-output (MIMO) radars. Such systems are used for interrogating distant scenes by transmitting a probing field from a Tx array of antennas and measuring the returns at an Rx array of sensors. We focus our attention on the Direction of Arrival (DOA) estimation problem where the distant scene consists of a finite number of point targets and the goal is to estimate their azimuths. This setup is illustrated in Figure 5-1.

Multiple-input multiple-output (MIMO) radar systems have been shown to offer superior performance in direction of arrival (DOA) estimation applications compared to their phased array counterparts. The performance of these systems has been studied under various probing field-target interaction mechanisms. However, to the best of our knowledge, these have been restricted to linearized models.

Motivated by various nonlinear imaging modalities that have emerged in recent years

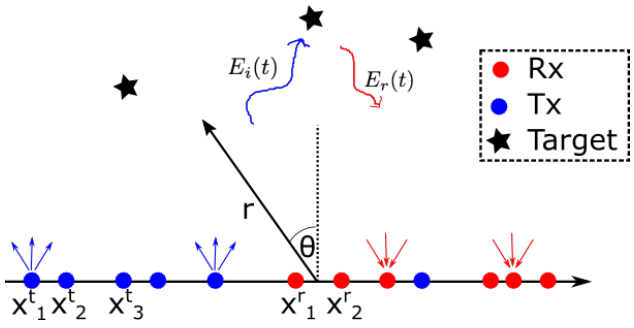


Figure 5-1: MIMO Radar DOA estimation setup.

we study DOA estimation in far field MIMO radar systems in conjunction with a nonlinear probing field-target interaction mechanism. Specifically, we consider simplified toy models of nonlinear targets exhibiting power-law reflection properties and study the role this interaction plays in carrying information about the scene and how to design signal sets and transmitter and sensor placement configurations that capture this information efficiently. We derive theoretical results that demonstrate how power law type interaction can lead to enhanced target identifiability with a fixed number of sensors, and suggest corresponding array topologies and signal sets that support this enhanced level of performance.

5.1 Introduction

Conventional phased array radar systems limit the number of degrees of freedom associated with the signal set by only allowing correlated transmission from different antenna elements [113]. MIMO systems, which have been the focus of research over the last decade, allow transmission of uncorrelated signals. Stoica et al. [69] explored MIMO radar with co-located antennas. They considered configurations with N_t transmitting and N_r receiving antennas in conjunction with suitable signal sets and array configurations and have shown that in several respects performance gains may be attained compared to their conventional phased array counterparts. Specifically in terms of the number of identifiable targets, the performance of a MIMO system with N_t transmit and N_r receive elements is comparable to that of a phased array system employing $O(N_tN_r)$ elements. Other studies [28] have explored MIMO radar systems with distant antennas in conjunction with targets exhibiting reflection fluctuations and have shown that spatial diversity may be utilized to overcome deep fading conditions.

While various deterministic and stochastic probing field-target interaction models have been considered in the past, to the best of our knowledge all such studies have assumed a linearized response where the reflected field scales in proportion to the probing field.

In practice, the interaction mechanism between the transmitted probing field and a distant scene may exhibit complicated characteristics. As a motivating example, in recent years the use of micro-bubbles as a nonlinear contrast agent in medical ultrasound applications has become wide-spread [72]. The highly nonlinear interaction manifests itself in the reflected

signal as harmonics of the incoming signal. Similar nonlinear phenomena have been observed in electromagnetic reflections and have been utilized in microscopy applications [75].

In this work we consider the consequences of a hypothetical memoryless, kth order power-law nonlinear target reflectance model on the design and performance of MIMO radar systems used in DOA estimation applications. We show that in conjunction with a specialized probing signal set and array design a MIMO radar system with N_t transmit and N_r receive elements can attain target identification performance comparable to that of an $O(N_t^k N_r)$ elements phased array setup, offering substantial performance gains with respect to the MIMO setup with linear reflectors.

Reflection models such as the one we consider here may be naturally occurring in some specialized settings or deliberately introduced into reflectors, e.g. by exploiting naturally occurring phenomena, or by introduction of active elements exhibiting desired nonlinear characteristics.

5.2 Propagation Model

In this section we present the radiation propagation model which lies at the foundation of MIMO radar. The signal transmitted from the Tx antennas propagates towards the targets. Upon hitting the targets the signal reflects back towards the receiver antennas where it is recorded and further processed. The model we describe here captures the generic linear reflections model as well as our nonlinear power-law reflection model as we detail next.

5.2.1 Setup

Consider a far field scene distributed along the azimuthal and radial axes θ and r respectively. For convenience define the normalized azimuth $\psi \equiv \frac{1}{2}\sin\theta, \psi \in [-\frac{1}{2}, \frac{1}{2})$ and propagation time $\tau \equiv \frac{r}{c}$ where c is the propagation velocity, such that we can parametrize the scene on the coordinate system (ψ, τ). A Tx antenna array illuminates the scene while an Rx array records the returns. The Tx and Rx arrays consist of N_t and N_r antennas, positioned at $\left\{x_0^t, \cdots, x_{N_t-1}^t\right\}$, $\left\{x_0^r, \cdots, x_{N_r-1}^r\right\}$, respectively, as depicted in Fig. 5-1.

5.2.2 From Tx to Target

The Tx transmits a narrow-band signal at frequency ω and wavelength $\lambda = \frac{2\pi c}{\omega}$. The complex envelope of the signal modulating the nth Tx antenna is $a_n(t)$ such that the resulting far field $E_i(\psi, \tau, t)$ at spatial location (ψ, τ) and time t is given by[1] [113]:

$$E_i(\psi, \tau, t) = \mathrm{Re}\left[\tilde{E}_i(\psi, \tau, t) e^{j\omega t}\right] \tag{5.1}$$

where we have defined:

$$\tilde{E}_i(\psi, \tau, t) \equiv \sum_{n'=0}^{N_t-1} a_{n'}(t-\tau) \exp\left[j\frac{4\pi}{\lambda} x_{n'}^t \psi\right] e^{-j\omega\tau} \tag{5.2}$$

The last equations reflect the temporal delay and phase shifts accrued by the transmitted signal as it propagates in space away from the transmitting antennas.

5.2.3 Target Interaction

The scene is comprised of M point targets at a common temporal delay τ_0 away from the transmitters and azimuths $\psi_1, \cdots, \psi_M$. In the sequel, focusing on the DOA problem, our estimation goal is to retrieve these parameters from the return signal recorded at the receiving antennas.

The signal emanating from the transmitting antennas propagates in space until it reaches the targets and reflects back towards the receiving antennas. The reflection process is the focus of our study in this work. Whereas, to the best of our knowledge, traditional studies have strictly considered simplified linear reflections occurring as a result of the incoming radiation impounding on the targets, here, inspired by practical application, we consider a power-law reflection model with a nonlinear interaction mechanism. In what follows we present the power-low reflection model, which captures both the nonlinear setting, as well as the conventional linear setting as a special case.

We consider a deterministic, kth order power-law nonlinear reflection model such that

[1] We ignore constant scaling factors such as those arising from the power decay experienced by the propagating radiation, as these do not affect target identifiability.

the reflection generated at the lth target is:

$$E_r^l(\psi_l, \tau_0, t) = \beta_l \left(E_i(\psi_l, \tau_0, t)\right)^k \tag{5.3}$$

where $E_r^l(\psi_l, \tau_0, t)$ is the reflected signal that emanates from the lth target and propagates back towards the receiving antennas, and β_l is the coupling coefficient of the lth target.

Notice that the nonlinear reflection model (5.3) captures the conventional linear reflection model by taking $k = 1$, when the returned signal is just a scaled copy of the incoming signal impounding on the target. Also, substituting (5.1) in (5.3), notice that the highest frequency of the reflected signal $E_r^l(\psi_l, \tau_0, t)$ is $k\omega$, which means that a reflection resulting from a kth order power-law interaction is strictly frequency separated from a reflection resulting from lower, order power-law interactions, including linear reflections, rendering signal separation at the receiver easy.

The motivation behind the model (5.3) is its simplicity and the fact that any memoryless nonlinearity respecting the generic functional form $E_r^l(\psi_l, \tau_0, t) = f\left(E_i(\psi_l, \tau_0, t)\right)$ where $f(\cdot)$ is a scalar function centered around 0 may be expanded using a Taylor series: $E_r^l(\psi_l, \tau_0, t) = \sum_{k=1}^{\infty} f_k \left(E_i(\psi_l, \tau_0, t)\right)^k$, which suggests that studying the power-low reflection model may be useful in deriving insights for the generic memoryless nonlinearity case.

5.2.4 From Target to Rx

Next, we develop expressions for the signal recorded at the Rx array. Using (5.1) and (5.3) we have for the complex envelope of the reflected signal component centred around $k\omega$:

$$\begin{aligned}\tilde{E}_r^l(\psi_l, \tau_0, t) &= \beta_l \left(\sum_{n'=0}^{N_t-1} a_{n'}(t-\tau_0) \exp\left[j\frac{4\pi}{\lambda} x_{n'}^t \psi_l\right]\right)^k e^{-j\omega k\tau_0} \\ &= \beta_l e^{-j\omega k\tau_0} \sum_{n''=0}^{N_t'-1} \hat{a}_{n''}(t-\tau_0) \exp\left[j\frac{4\pi}{\lambda} \hat{x}_{n''}^t \psi_l\right]\end{aligned} \tag{5.4}$$

Where the second equation is retrieved from the first by application of the multinomial expansion[2], such that the sum over n'' runs over $N_t' = \left(\!\binom{N_t}{k}\!\right) \equiv \binom{N_t+k-1}{k}$ unique (multinomial) solutions $\boldsymbol{\gamma}^{(n)} = [\gamma_0^{(n)}, \gamma_1^{(n)}, \cdots, \gamma_{N_t-1}^{(n)}]$ of the equation $\sum_{i=0}^{N_t-1} \gamma_i = k$, and the corresponding virtual Tx locations $\hat{x}_n^t$ and transmission functions $\hat{a}_n(t)$ are given according to:

$$\begin{aligned} \hat{a}_n(t) &= \sqrt{c_n} \prod_{i=0}^{N_t-1} a_i^{\gamma_i^{(n)}}(t) \\ \hat{x}_n^t &= \sum_{i=0}^{N_t-1} \gamma_i^{(n)} x_i^t \\ \sqrt{c_n} &\equiv \binom{k}{\gamma_0^{(n)}, \gamma_1^{(n)}, \ldots, \gamma_{N_t-1}^{(n)}} \end{aligned} \tag{5.5}$$

Importantly, notice that the nonlinear reflected signal as given in (5.4) for any integral k is analogous to the reflected signal in a conventional setup with linear reflections $k = 1$, virtual element positions $\hat{x}_n^t$ and signal set $\hat{a}_n(t)$.

The reflected signal propagates towards the Rx array with wavelength $\lambda_k = \frac{\lambda}{k}$ corresponding to the higher frequency. The received signal at the mth antenna element after down-converting to baseband is given according to:

$$\tilde{s}_m(t) = \sum_{l=1}^{M} \sum_{n''=0}^{N_t'-1} \beta_l \hat{a}_{n''}(t - 2\tau_0) \exp\left[j \frac{4\pi}{\lambda} \left(\hat{x}_{n''}^t + \hat{x}_m^r \right) \psi_l \right] e^{-j2\omega k \tau_0} \tag{5.6}$$

where $\hat{x}_m^r \equiv k x_m^r$ are the virtual Rx locations.

5.3 Direction of Arrival Estimation

In this section we consider the DOA estimation problem in the presence of nonlinear reflectors and analyze the fundamental limits of target identifiability. Our final goal is to process the received signals $\{\tilde{s}_m(t)\}$ to retrieve the target parameters $(\boldsymbol{\beta}, \boldsymbol{\psi}) \equiv (\beta_1, \ldots, \beta_M, \psi_1, \ldots, \psi_M)$, and derive theoretical results for how many such targets M are uniquely identifiable from

[2] $\left(\sum_{i=0}^{N_t-1} x_i \right)^k = \sum_{\gamma_0+\gamma_1+\cdots+\gamma_{N_t-1}=k} \binom{k}{\gamma_0, \gamma_1, \ldots, \gamma_{N_t-1}} \prod_{t=0}^{N_t-1} x_t^{\gamma_t}$, where the sum is over all possible solutions of $\gamma_0 + \gamma_1 + \cdots + \gamma_{N_t-1} = k$

the set of received signals, with the transmission set $\{a_n(t)\}$ under our control.

We adapt the analysis of [69] for discrete-time MIMO radar configurations with linear reflectors to a continuous time formulation in conjunction with nonlinear reflectors and show that under ideal conditions it is possible to identify $O(N_t^k N_r)$ targets when utilizing the kth order nonlinearity, an order of magnitude improvement over results derived for the linear reflection case $k = 1$.

Define modified received signals as $\breve{s}_m(t) \equiv \tilde{s}_m(t + 2\tau_0)e^{j2\omega k\tau_0}$. We have, using (5.6):

$$\breve{s}_m(t) = \sum_{n''=0}^{N_t'-1} c_{m,n''}(\boldsymbol{\beta}, \boldsymbol{\psi})\hat{a}_{n''}(t) \tag{5.7}$$

where $c_{m,n''}(\boldsymbol{\beta}, \boldsymbol{\psi}) \equiv \sum_{l=1}^{M} \beta_l \exp\left[j\frac{4\pi}{\lambda}\left(\hat{x}_{n''}^t + \hat{x}_m^r\right)\psi_l\right]$.

For the sequel define $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ to be a vector stacking the elements of $\{c_{m,n''}(\boldsymbol{\beta}, \boldsymbol{\psi})\}$ in some fixed order.

Definition 5.1. *M targets are uniquely identifiable from the receive signals $\{\breve{s}_m(t)\}$ if there exists a transmission set $\{a_n(t)\}$ such that for every combination of M targets or less we have that $\forall m : \breve{s}_m^1(t) = \breve{s}_m^2(t)$ implies $(\boldsymbol{\beta}_1, \boldsymbol{\psi}_1) = (\boldsymbol{\beta}_2, \boldsymbol{\psi}_2)$.*

The problem of parameter identifiability is about determining the maximal number M of uniquely identifiable targets from $\{\breve{s}_m(t)\}$. Our main result for this section is the following one:

Theorem 5.1. *There exists a signaling set $\{a_n(t)\}$ and a selection of N_t Tx and N_r Rx antenna locations such that any $M < \frac{1}{2}\left(\left\lfloor\frac{N_t+k-1}{k}\right\rfloor^k N_r + 1\right) = O(N_t^k N_r)$ targets are identifiable from $\{\breve{s}_m(t)\}$.*

We will prove a series of useful lemmas and end this section with the proof of Theorem 5.1. Notice, comparing to [69] that the maximal possible number of uniquely identifiable targets for a linear $k = 1$ MIMO radar system is given according to $\frac{2}{3}N_t N_r = O(N_t N_r)$ such that the results of Theorem 5.1 suggest orders of magnitude improvement in target identifiability for non-trivial nonlinear settings with k strictly larger than 1.

To prove Theorem 5.1 we start with a definition:

Definition 5.2. *We say that M targets are uniquely identifiable from $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ if for every combination of M targets or less we have that $C(\boldsymbol{\beta}_1, \boldsymbol{\psi}_1) = C(\boldsymbol{\beta}_2, \boldsymbol{\psi}_2)$ implies $(\boldsymbol{\beta}_1, \boldsymbol{\psi}_1) = (\boldsymbol{\beta}_2, \boldsymbol{\psi}_2)$.*

The next lemma regarding the signal set design is useful for proving subsequent claims:

Lemma 5.1. *There exists a set of N_t Tx functions $\{a_n(t)\}$ such that the corresponding effective signal set $\{\hat{a}_n(t)\}$ generated according to (5.5) is orthogonal: $\int_t \hat{a}_{n'}(t)\hat{a}_n^*(t)dt = \delta_{nn'}\hat{g}_n$, with $\hat{g}_n$ non-zero constants.*

Proof. See Section 5.5. □

The next lemma is a restatement of results in [70] with appropriate adaptations to continuous time:

Lemma 5.2. *A necessary condition for parameter identifiability from the received signals $\{\breve{s}_m(t)\}$ is parameter identifiability from $C(\boldsymbol{\beta}, \boldsymbol{\psi})$. It is also sufficient if the signaling set satisfies lemma 5.1.*

Proof. Given $\{a_n(t)\}$ the $\{\breve{s}_m(t)\}$ are determined from the elements of $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ as per (5.7). We trivially have that if M parameters are not identifiable from $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ they are also not identifiable from $\{\breve{s}_m(t)\}$.

Conversely, assume that M targets are uniquely identifiable from $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ and that the signaling set satisfies Lemma 5.1. We show that the target parameters are uniquely identifiable from $\{\breve{s}_m(t)\}$. Indeed, using (5.7) we have that the following holds at the receiver: $\int_t \breve{s}_m(t)\hat{a}_{n''}^*(t)dt = \hat{g}_{n''}c_{m,n''}(\boldsymbol{\beta}, \boldsymbol{\psi})$. Repeating this for every m and n'' we can extract $C(\boldsymbol{\beta}, \boldsymbol{\psi})$, and since it allows unique identification of the parameters so does $\{\breve{s}_m(t)\}$. □

Using Lemma 5.2 we have that the number of identifiable targets from $\{\breve{s}_m(t)\}$ is equal to the number of identifiable targets from $C(\boldsymbol{\beta}, \boldsymbol{\psi})$. The next lemma is useful for the proof of the main theorem.

Lemma 5.3. *There exists a configuration of N_t Tx and N_r Rx antenna locations $\{x_n^t\}, \{x_n^r\}$ such that the set $\{\hat{x}_{n''}^t + \hat{x}_m^r\}$ contains $N_s = \left\lfloor \frac{N_t+k-1}{k} \right\rfloor^k N_r = O(N_t^k N_r)$ contiguous points on a uniform $\frac{\lambda}{2}$-spaced grid (starting at 0 without loss of generality).*

Proof. See Section 5.4. □

Finally, we can prove Theorem 5.1:

Proof. For the sensor locations choose a setup that satisfies Lemma 5.3 with $N_s = \left\lfloor \frac{N_t+k-1}{k} \right\rfloor^k N_r$ the number of contiguous samples $\{\hat{x}^t_{n''} + \hat{x}^r_m\}$. Define the vector $\tilde{C}(\boldsymbol{\beta}, \boldsymbol{\psi})$ as a sub-vector of $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ according to:

$$\tilde{C}(\boldsymbol{\beta}, \boldsymbol{\psi}) \equiv \left[\sum_{l=1}^{M} \beta_l e^{j2\pi 0 \psi_l}, \cdots, \sum_{l=1}^{M} \beta_l e^{j2\pi(N_s-1)\psi_l} \right]^T \tag{5.8}$$

the number of identifiable targets from $\tilde{C}(\boldsymbol{\beta}, \boldsymbol{\psi})$ is not greater than the number of identifiable targets from $C(\boldsymbol{\beta}, \boldsymbol{\psi})$ as the former contains a subset of the elements of the latter. Given L targets define:

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{\psi}) &\equiv [B'(\psi_1), \cdots, B'(\psi_L)] \\ B'(\psi) &\equiv [\exp(j2\pi 0\psi), \cdots, \exp(j2\pi(N_s-1)\psi)]^T \\ \boldsymbol{\beta} &\equiv [\beta_1, \cdots, \beta_L]^T \end{aligned} \tag{5.9}$$

and notice that with these definitions we have $\tilde{C}(\boldsymbol{\beta}, \boldsymbol{\psi}) = \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\beta}$.
As any N_s distinct vectors $\{B'(\psi_1), \cdots, B'(\psi_{N_s})\}$ are linearly independent (stacked side by side they form a Vandermonde matrix), we use the result from [82], [121] (Theorem 1) to claim that a sufficient condition for parameter identifiability from $\tilde{C}(\boldsymbol{\beta}, \boldsymbol{\psi})$ is $L < \frac{N_s+1}{2}$, such that plugging the expression for N_s we have that we can uniquely identify any $M < \frac{1}{2}\left(\left\lfloor \frac{N_t+k-1}{k} \right\rfloor^k N_r + 1 \right) = O(N_t^k N_r)$ point targets, which is our key result. □

5.4 Antenna Array Design

In this section we design antenna arrays in conjunction with Lemma 5.3. The goal is to choose positions $\{x^t_n, x^r_n\}$ such that the resulting virtual positions $\{\hat{x}^t_n, \hat{x}^r_n\}$ satisfy the conditions of the lemma, with $\{\hat{x}^t_n + \hat{x}^r_m\}$ covering a contiguous uniform $\frac{\lambda}{2}$-spaced grid of $\left\lfloor \frac{N_t+k-1}{k} \right\rfloor^k N_r = O(N_t^k N_r)$ elements.

The virtual Tx antenna locations are determined from the physical antenna locations according to (5.5). Our construction will result in virtual arrays such that the virtual Tx array will span a uniform grid of spacing $N_r\frac{\lambda}{2}$ while the virtual Rx array will span a grid with spacing $\frac{\lambda}{2}$.

For the Rx array choose $\{x_n^r\}$ on a uniform grid with spacing $\frac{\lambda}{2k}$:

$$x_n^r = n\frac{\lambda}{2k} \qquad\qquad n = 0, \cdots, N_r - 1 \tag{5.10}$$

which, coupled with the definition $\hat{x}_m^r \equiv kx_m^r$ results in the desired virtual Rx array.

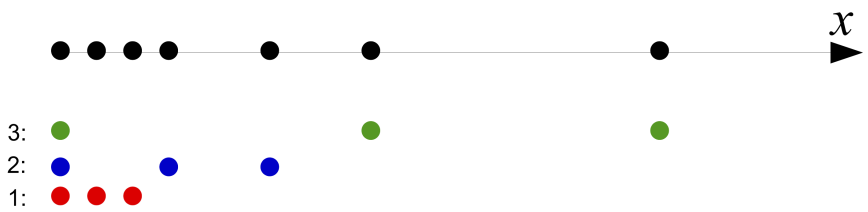
As for the Tx array, our design problem hints at the one studied in [87] where the authors considered the diversity of the co-array formed according to position differences between pairs of physical elements. They showed that a nested geometry maximizes the number of degrees of freedom available for DOA estimation with a given number of elements.

For nonlinear imaging, virtual locations $\hat{x}_n^t$ are formed as k-sums of the set $\{x_n^t\}$ according to (5.5). We use nested arrays similar to those proposed in [87] and obtain Tx diversity of $O(N_t^k)$ virtual elements covering a uniform $N_r\frac{\lambda}{2}$-spaced grid as required.

With kth order nonlinearity we design a nested Tx array partitioned into k hierarchies. The ith hierarchy is a uniformly spaced array with N_t^i elements and spacing d_t^i, such that all hierarchies share a common element at location 0. An example for $k = 3$ is depicted in Figure 5-2 with the three hierarchies in color and the resulting array in black.

The first $i = 1$ Tx hierarchy is designed with spacing $d_t^1 = N_r\frac{\lambda}{2}$ and yet unspecified number of elements N_t^1. Subsequent hierarchies are designed according to the following iterative rule: For the $(i+1)$th Tx hierarchy choose spacing $d_t^{i+1} = N_t^i d_t^i$ and again, a yet unspecified number of elements N_t^i. For simplicity, taking into account the k multiplicity of the common 0 element, populate all hierarchies with an equal number of $N_t^i = \left\lfloor \frac{N_t+(k-1)}{k} \right\rfloor$ elements discarding the remaining antennas.

With this design, we show that the virtual Tx array covers an $\left\lfloor \frac{N_t+(k-1)}{k} \right\rfloor^k$ elements uniform contiguous grid with spacing $N_r\frac{\lambda}{2}$. Indeed, the virtual Tx array contains every k-sum of element positions. Specifically, it contains any such sum with exactly one element from each of the k hierarchies, and these result in unique virtual elements due to the geometric spacing



A Tx array (black) and its three constituent hierarchies with three elements each. The first element is shared between hierarchies such that the overall number of elements is $N_t = 7$.

Figure 5-2: Transmitter array design for nonlinear MIMO radar.

of the sub-arrays. The overall number of such combinations equals $(N_t^i)^k = \left\lfloor \frac{N_t+(k-1)}{k} \right\rfloor^k$.

Combining with the N_r Rx elements we end up with $\{\hat{x}_n^t + \hat{x}_m^r\}$ covering a contiguous uniform $\frac{\lambda}{2}$-spaced grid of $\left\lfloor \frac{N_t+(k-1)}{k} \right\rfloor^k N_r$ elements.

5.5 Signal Set Synthesis

In this section we provide a proof for Lemma 5.1. Namely, we design a signal set $\{a_n(t)\}$ such as to satisfy the correlation property $\int_t \hat{a}_{n'}(t)\hat{a}_n^*(t)dt = \delta_{nn'}\hat{g}_n$. In what follows we present a construction technique that results in constant modulus signals, which is desirable for practical applications.

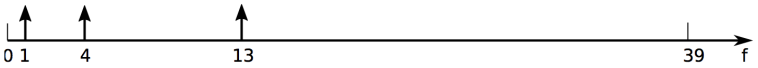
For kth order nonlinearity and N_t transmitters the Tx signals $\{a_n(t)\}, n = 0, \cdots, N_t - 1$ are defined to be windowed pure discrete tones amplitude modulating a rectangular shaping function:

$$a_n(t) = \sum_{m=0}^{\Omega_{N_t}-1} \exp(j2\pi \frac{\Omega_n m}{\Omega_{N_t}}) h(t - mT_c) \tag{5.11}$$

where $h(t) = \mathbb{1}_{0\le t\le T_c}(t)$ is a rectangular shaping function and T_c the chip length. The nth discrete tone frequency Ω_n is defined recursively according to:

$$\begin{cases} \Omega_0 = 1 \\ \Omega_n = k\Omega_{n-1} + 1, \qquad n \ge 1 \end{cases} \tag{5.12}$$

The discrete tones are windowed to a finite time record of length Ω_{N_t}. This is schematically



Signal set frequency occupation for $k = 3$ nonlinearity and $N_t = 3$ transmitters: $\Omega_0 = 1$, $\Omega_1 = 4$, $\Omega_2 = 13$.

Figure 5-3: Signal set design for nonlinear MIMO radar.

depicted in Figure 5-3 for $k = 3$ and $N_t = 3$.

With the definition above and using (5.5), the virtual signal set becomes:

$$\hat{a}_{n'}(t) = \sqrt{c_{n'}} \sum_{m=0}^{\Omega_{N_t}-1} \exp(j2\pi \frac{m}{\Omega_{N_t}} \sum_{j=0}^{N_t-1} \gamma_j^{(n')} \Omega_j) h(t - mT_c) \tag{5.13}$$

We now show that the signal set as defined satisfies Lemma 5.1. The next lemma is useful for proving the orthogonality relations:

Lemma 5.4. *for every* $\{\gamma_i \geq 0\}$*, such that* $\sum_{i=0}^{N_t-1} \gamma_i = k$*:*

1. $\sum_{i=0}^{R-1} \gamma_i \Omega_i < \Omega_R$ *for every* $R \leq N_t$

2. $\{\gamma_i\}$ *are uniquely determined from* $u = \sum_{i=0}^{N_t-1} \gamma_i \Omega_n$

Proof. The first claim follows immediately from the construction. To prove the second claim, use the first claim with $R = N_t - 1$ to show $\gamma_{N_t-1} = \left\lfloor \frac{u}{\Omega_{N_t-1}} \right\rfloor$. Then, apply the same procedure on $u - \gamma_{N_t-1}\Omega_{N_t-1}$ get γ_{N_t-2} and continue similarly for all following coefficients. □

Finally, using $\int_t h(t - mT_c)h^*(t - mT_c) = T_c$ we have:

$$\int_t \hat{a}_{n'}(t)\hat{a}_n^*(t)dt = T_c\sqrt{c_{n'}c_n} \sum_{m=0}^{\Omega_{N_t}-1} \exp(j2\pi \frac{m}{\Omega_{N_t}^k} \left[\sum_{j=0}^{N_t-1} \gamma_j^{(n')} \Omega_j^k - \sum_{j=0}^{N_t-1} \gamma_j^{(n)} \Omega_j^k \right]) = \delta_{nn'} \hat{g}_n \tag{5.14}$$

with $\hat{g}_n = T_c c_n \Omega_{N_t}^k$, where in the last equality we have used Lemma 5.4 to claim that the term in brackets is zero if and only if $n = n'$. Thus, our signal set adheres to Lemma 5.1 as required.

5.6 Numerical Experiments

We complement our analysis with the results of a numerical experiment. The setup is comprised of 13 far-field reflecting targets with angles as depicted in red in Figure 5-5 and in Figure 5-4. The targets exhibit nonlinear reflectance with $k = 3$ and unit coupling coefficients $\beta = 1$. We design a Tx array with $N_t = 7$ elements according to the scheme of Section 5.4 and the example given there as the union of three constituent equi-populated sub-arrays. In units of $\frac{\lambda}{2}$ the constituent sub-arrays are positioned at $\{0, 1, 2\}, \{0, 3, 6\}, \{0, 9, 18\}$ such that the Tx locations are $\{0, 1, 2, 3, 6, 9, 18\}$ as in Figure 5-2. The receiver array is degenerate with a single element at $x_0^r = 0$. For the signaling set we implemented the design scheme of Section 5.5 with tones $\Omega \in \{1, 4, 13, 40, 121, 364, 1093\}$ and a sequence length equal to $\Omega_7 = 3280$ chips. To probe the stability of the estimation problem we have included the effect of a complex AWGN impairing the received signal. For the simulation described here we have assumed SNR = 10dB. With the setup as defined above the virtual array strictly covers the contiguous section $\{0, \cdots, 26\}$ which was used in the estimation procedure.

For DOA estimation we implemented a single-shot MUSIC algorithm [71]. In Figure 5-5 we plot the score function vs. ψ where it is evident that the algorithm gives excellent estimates for the location of all 13 targets, in accord with Theorem 5.1 which guarantees identifiability of up to 13 targets under these conditions. Also notice that with seven transmitting elements and one receiving element conventional MIMO radar techniques cannot support DOA estimation for more than 6 targets under any circumstances, such that the above experiment exemplifies the additional degrees of freedom supported by the nonlinear interaction between the probing field and the reflecting targets.

5.7 Discussion

Most published studies analyzing imaging or DOA estimation problems in radar like settings consider a simplified linear probing field-target interaction model. We have identified a seemingly important theoretical gap in that while many real world scenes exhibit far more complicated non-linear interaction mechanisms the theory for modeling these interactions

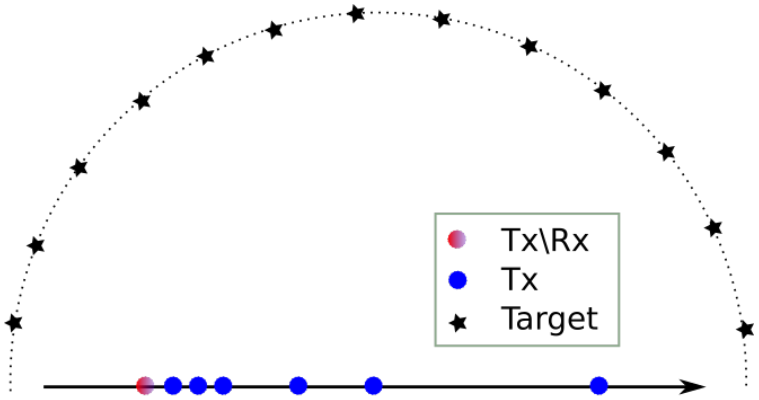
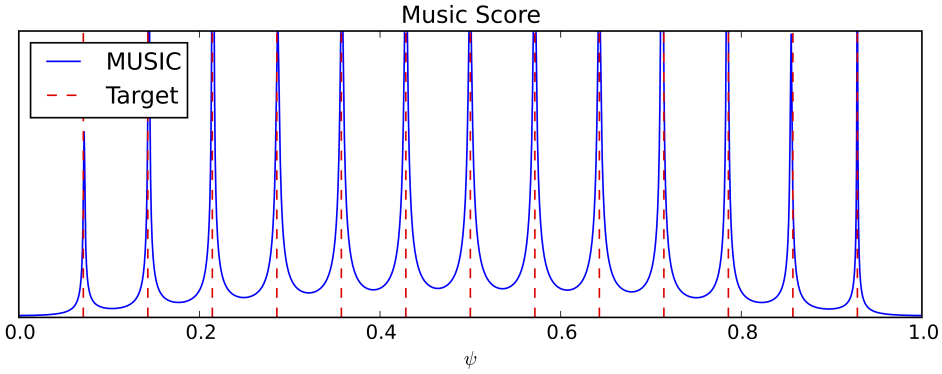


Figure 5-4: Numerical experiment setup for nonlinear MIMO radar.



Single-shot MUSIC score for 13 $k = 3$ nonlinear reflecting targets and a setup as described in the text.

Figure 5-5: DOA estimation via MUSIC algorithm for nonlinear MIMO radar.

and analyzing its implications is lacking. We are thus interested in understanding how non-conventional, i.e. nonlinear, interaction mechanisms come into play in determining fundamental bounds concerning the performance of such systems.

Specifically, we wish to understand whether or not unconventional interaction models can help improve the trade-off curve between imaging or estimation quality and available system resources as quantified by various parameters such as the number of antenna elements.

We have introduced the notion of target probing through nonlinearities as a means to enhance the number of identifiable targets in applications utilizing antenna arrays in MIMO configurations. We have shown that a virtual Tx array emerges in conjunction with the nonlinearities such that the effective number of degrees of freedom available for DOA estimation is asymptotically orders of magnitude larger than available in conjunction with linear reflectors.

Our result reveals an inherent asymmetry between the Tx and Rx arrays under such nonlinearities as is evident from our expression for the number of identifiable targets which scales as $O(N_t^k N_r)$. An effective way to reap the most benefit from the proposed scheme would be to introduce a single element for the Rx while transmitting with multiple antennas at the Tx, which would lead to the biggest impact under a constraint on the total number of antennas.

With respect to DOA estimation performance in noisy environments our nonlinear scheme inherits performance bounds from conventional results pertaining to DOA estimation with linear targets with corresponding virtual arrays and signal sets replacing physical ones.

A topic we have only briefly alluded to is the applicability of the nonlinear reflectors model to practical applications. Sophisticated probing-field-target interaction models have yet to be fully exploited by conventional radar systems. Further research is required to evaluate if such models can be implemented in practice.

More generally, in future research we will be interested in considering a wide class of general nonlinear interaction models, potentially additionally introducing memory effects and analyzing the corresponding performance. A general approach for that matter might be to investigate information-theoretic aspects of the problem, that is to study how the probing field-scene interaction can be viewed as a channel and considering how information evolves

as it traverses the channel, specifically asking what channels enable extraction of the most amount of information at the receiver.

On the more technical side of things we are interested in developing good signal sets for the transmitter side in such applications. For our power-law nonlinear MIMO radar it is important to develop such sets with good temporal auto-correlation properties facilitating estimation of both distance as well as azimuthal features of the scene.

Finally, we suggest connecting the submodular optimization techniques for array design developed in the previous chapters to efficiently design effective array configurations in non traditional imaging and estimation settings.

Chapter 6

Concluding Remarks

The proliferation of cheap computational resources and communication systems in recent years and decades has led to an explosion in the variety and quantity of readily available data. In this era of big data many applications consider data to be an unlimited resource and focus on methods and techniques to utilize large quantities of it to achieve desirable outcomes, such as enabling high quality inference or facilitating optimal decision making.

In contrast, in many applications, especially those related to physical environments where collecting additional data points involves deploying specialized sensing equipment or running sensitive experiments, taking measurements or collecting data is very much still a complicated and costly endeavour which takes time and requires careful planning.

The focus of this thesis is on efficient data collection strategies for learning in physical environments. We consider various systems and environments where data collection is challenging, and suggest models that capture the essence of the data collection mechanism in these setups. We then analyze and try to understand how informative data and measurements can be collected efficiently, enabling a reduction in the amount of resources required to achieve a prescribed level of system performance.

As the various systems we consider embody very different physical phenomena, from laser reflections recorded on a detector to measurements of the position of a satellite orbiting a planet, specialized models need to be tailored to capture important features that define the performance envelope relevant in each scenario. At the same time, the motivation to reduce the amount of resources needed to achieve some specified performance level is shared

between these system instances such that in many cases we can apply similar mathematical principles to tackle these challenging problems.

We briefly survey the various systems we studied and analyzed in this thesis.

In Chapter 2 we considered a large class of dynamical systems. We were interested in situations where the system model was misspecified and our goal was to devise a strategy to conduct experiments in order to collect informative data points that can be used to fine tune the system model, increasing its accuracy and enabling high fidelity predictions. The algorithmic approach we developed enables a reduction in the number of experiments needed to achieve a specified level of system model accuracy.

In Chapter 3 our focus was in far field imaging in one or multiple wavelengths. Collecting spatial measurements in this setting requires deploying antenna elements, whose complexity, weight and size are main determinants of system cost. In this setting, we developed algorithms to design antenna array configurations under various constraints on the position and number of antenna elements, as well as novel designs that allow robust imaging in multiple wavelengths. Our designs enable reductions in the number of deployed antenna elements while adhering to specified constraints.

In Chapter 4 we put our focus on NLOS imaging where existing systems rely on costly detectors that enable hidden scene reconstruction by recording high resolution temporal optical measurements. In lieu of these sensitive time resolved measurements, our goal was to identify situations where it is possible to collect informative measurements by exploiting structure in non-time-resolved measurements. We have identified the role of occluders in NLOS imaging as endowing structure on the measurements, enabling high fidelity hidden scene recovery with fundamentally lower quality measurements, and developed an algorithmic approach to determine how to perform efficient scene interrogation. Our methods and designs are an instance of a novel NLOS imaging modality that may be of importance in future applications.

In Chapter 5 we considered the problem of DOA estimation in MIMO radar. While current systems assume a linear reflectance model, several practical applications suggest that this may not be an accurate representation of all reflectors. We suggested a theoretical nonlinear reflection model and studied its implications on the number of point targets that could

be identified by a MIMO radar system. We showed that our novel signal set and array design, tailored for such environments, enable an order of magnitude increase in the number of identifiable targets compared to systems assuming a linear reflectance model suggesting that, perhaps counter intuitively, complicated nonlinear interaction mechanisms could increase the informativeness of measurements and correspondingly, system performance.

6.1 Future Research

While the ideas we explore in this thesis illuminate certain facets of efficient data collection strategies in physical environments, the broader topic is very general and offers many more questions that warrant further research and study. We list here a few topics that seem particularly interesting and promising:

Online Learning Throughout this work our data collection strategies were designed off-line, i.e. the decision on which data to collect happened before any actual data from the system became available. In specific circumstances, e.g. when the Gaussian representation is exact, this off-line strategy may in fact be shown to be optimal. However, for non-ideal models there is value in adapting the data collection process in accordance with previously collected measurements, especially in settings where data collection naturally occurs sequentially in time, such as when experimenting with dynamical systems. Studying adaptive online strategies for data collection is a very interesting research topic that can offer substantial benefits in terms of acquisition time and cost.

Data Collection for Decision Making Most of the learning we discussed in this thesis was designed with an ultimate goal of inferring a model or a representation (e.g. an image) for a system under study. The systems we considered were passive in the sense that the state of the system was not of importance to us. It is interesting to consider the problem of efficiently learning systems that are active in the sense that they respond to our actions or queries with outcomes that may be of differing value to us. For example, imagine a physical system governed by some unknown rules where our goal is to induce some sort of deliberate desirable change, such as to determine the value of some observable parameter.

The problem of learning in this setup is referred to as reinforcement learning and it has gained some renewed attention in recent years in the context of artificial intelligence. It would be very interesting to study the problem of efficiently interacting with these systems to learn their interaction model and achieve desirable results quickly.

NLOS Imaging Systems The NLOS imaging modality we have suggested and studied in this work demonstrated that occluders can increase the informativeness of optical measurements by endowing them with a structure that can be exploited for high fidelity hidden scene reconstruction. It would be interesting to extend this idea to other forms of structure that may be present in optical measurements. For example, in settings where we have prior knowledge about some features of the hidden space, or what occupies it, we can hope to utilize it to extract more information from the measurements and improve the performance of the imaging system.

Nonlinear Radar Our investigation of nonlinear MIMO radars represents an initial effort in the sense that we explored a very basic nonlinear reflectance model. Our results suggest that such nonlinear interactions can offer advantages for DOA estimation performance in such environments. A broader question that naturally arises is then what are the limits of such advantages that can be offered by nonlinear interaction models. Would we be able to extract much more spatial information in environments with rich interactions, and if so, how could we efficiently introduce such phenomena to existing environments in order to increase the information that can be inferred by radar systems with fixed resources.

Array Design The array design paradigms we introduced and analyzed focused on designing efficient sensor configurations for scene estimation in multiple wavelengths under structural array constraints. Namely, when the sensors are constrained to be placed in specific locations, or in specific densities over prescribed areas. While these represented useful scenarios, there are additional resources whose efficient utilization may be of interest when designing such arrays. In particular we mention here computational resources. In the limit when the number of antennas is large and computations are to be performed in real time using simple computational units the amount of required computational effort or data com-

munication between antennas for scene reconstruction may be a limiting factor in designing the array. As such, it would be interesting to explore array design paradigms where the available computational resources or communication throughput between sensing elements are constrained and the goal is to find designs that allow computationally efficient high quality inference.

Misspecified Dynamical Systems Our study of experimental design for learning misspecified dynamical systems demonstrated the merits of formulating the problem of efficient data collection in this context. The formulations and results we discussed form an exposition to this topic, which can be further explored taking various approaches. One such direction is extending the basic model describing the misspecified setup. In particular, the additive misspecified driving term with known prior statistics model we have assumed was limiting. In some applications the misspecified component of the system model may be better represented using a different functional form, e.g. multiplicative or other, and it is interesting to classify what model fits different systems and what are the tradeoffs of using one model over another in face of uncertainty. Another interesting question that arises in this context is understanding the information value of having an accurate prior for the misspecified system term, that is, quantifying ,or at least bounding, the gain in designing near-optimal experiments according to our criteria and its dependence on the certainty we have about the prior. It will also be interesting to explore the problem of online learning the statistics of the misspecified term with the progression of experiments. Another interesting related topic is studying misspecified driven dynamical systems where the driving term is under our control and allows an additional degree of freedom in manipulating the system.

Appendix A

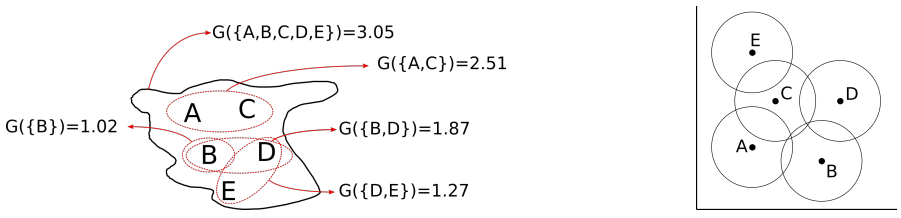
Submodular Maximization

A set function $G : 2^{\mathcal{V}} \to \mathbb{R}$ defined over the power set of a given set $\mathcal{V}$ assigns a real number to each subset of $\mathcal{V}$ (see illustration in Figure A-1). Following [83] we define two useful properties of set functions:

Definition A.1. *Let $G : 2^{\mathcal{V}} \to \mathbb{R}$ be a set function.*
*(a) G is **submodular** if it satisfies the property of decreasing marginals: $\forall \mathcal{S}, \mathcal{T} \subseteq \mathcal{V}$ such that $\mathcal{S} \subseteq T$ and $x \in \mathcal{V} \backslash T$ it holds that $G(\mathcal{S} \cup \{x\}) - G(\mathcal{S}) \geq G(\mathcal{T} \cup \{x\}) - G(\mathcal{T})$.*
*(b) G is **monotonic** (increasing) if $\forall \mathcal{S}, \mathcal{T} \subseteq \mathcal{V}$ s.t. $\mathcal{S} \subseteq \mathcal{T}$ we have $G(\mathcal{S}) \leq G(\mathcal{T})$.*

Submodular and monotonic functions are prevalent as these two properties naturally hold in many practical applications.

Example A.1. *Let $\mathcal{V} = \{A, B, C, D, E\}$ such that each $x \in \mathcal{V}$ corresponds to some disc in the 2D space illustrated in Figure A-1. Define the function $\forall \mathcal{S} \subseteq \mathcal{V} : G(S) = |\bigcup_{x \in \mathcal{S}} area(x)|$,*



(Left) A set function $G(\cdot)$ operating on the power set of $\mathcal{V} = \{A, B, C, D, E\}$, where only some of the assignments are shown. (Right) The set cover function.

Figure A-1: Submodular function maximization.

i.e. for each $\mathcal{S}$, $G(\mathcal{S})$ measures the area covered by the discs in $\mathcal{S}$. The function $G(\cdot)$ is an example of a set cover function. A set cover function is always monotonic and submodular.

In this work we are often interested in solving the following constrained maximization problem:

$$S^\star = \underset{\mathcal{S}:|\mathcal{S}|\leq K, \mathcal{S}\subseteq\mathcal{A}}{\operatorname{argmax}} G(\mathcal{S}) \tag{A.1}$$

where the function $G(\cdot)$ is submodular and monotonic. The following lemma implies that (A.1) is an NP-hard optimization problem such that no computationally tractable solver can retrieve $S^\star$. In lieu of an optimal solver, it is reasonable to consider the greedy solver, as delineated in Algorithm 3.

Algorithm 3 Greedy Submodular Maximization

function S=GreedyMax($G(\cdot), \mathcal{A}, K$)
  $\mathcal{S} \leftarrow \emptyset$
  for $i = 1$ to K **do**
    $x^\star \leftarrow \operatorname{argmax}_{x\in\mathcal{A}\setminus\mathcal{S}} G(\mathcal{S}\cup\{x\})\}$
    $\mathcal{S} \leftarrow \mathcal{S}\cup\{x^\star\}$
  end for
  Return $\mathcal{S}$
end function

We have the following classical result [83]:

Lemma A.1. *Let $G(\cdot)$ be a monotonic, submodular set function and $\mathcal{S}^\star$ defined according to* (A.1)*. Let $\mathcal{S}^{gr}$ be the set retrieved by the greedy maximization Algorithm 3. We have the following guarantee for the performance of the greedy algorithm:*

$$G(\mathcal{S}^{gr}) \geq (1-(1-\frac{1}{K})^K)G(\mathcal{S}^\star) \geq (1-\frac{1}{e})G(\mathcal{S}^\star)$$

Moreover, no polynomial time algorithm can provide a better approximation guarantee unless P=NP [26].

The previous lemma guarantees that the efficient greedy solver Algorithm 3 retrieves an approximately optimal solution to (A.1).

Algorithm 3 runs in time $O(|\mathcal{A}|K)$, linear in the size of the set $\mathcal{A}$ and the number of selected elements K [78]. However, more efficient variants of the algorithm have been suggested and analyzed. The 'lazy greedy' variant, which was introduced in [78] and appears here as Algorithm 4 was shown to offer substantial running-time improvements in practice.

Algorithm 4 Lazy Greedy Submodular Maximization

function S=LazyGreedyMax($G(\cdot), \mathcal{A}, K$)
  $\mathcal{S} \leftarrow \emptyset$
  $\forall x \in \mathcal{A}: \quad M[x] \leftarrow \infty$
  for $i = 1$ to K **do**
    stop $= 0$
    while *stop* $= 0$ **do**
      $x^\star = \underset{x \in \mathcal{A}}{\text{argmax}}\ M[x]$
      $M[x^\star] = G(S \cup \{x^\star\}) - G(S)$
      if $M[x^\star] \geq \underset{x \in \mathcal{A}}{\text{argmax}}\ M[x]$ **then**
        stop $= 1$
      end if
    end while
    $M[x^\star] \leftarrow -\infty$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x\}$
  end for
  Return $\mathcal{S}$
end function

The lazy variant of the greedy algorithm works by keeping an array M of size $|\mathcal{A}|$ that keeps estimates of the current marginal values for each of the elements of $\mathcal{A}$ with respect to the current candidate set $\mathcal{S}$. Instead of updating the full array M after each new addition into $\mathcal{S}$ at cost $|\mathcal{A}|$ we just target the highest marginal elements, revising our estimate only for that element until we see that even after updating the estimate it is the highest marginal value. With $G(\cdot)$ being submodular we are guaranteed that at this point we have found the current best greedy addition into $\mathcal{S}$ and the iterations progress until K elements are added.

The constant $(1 - 1/e) \approx 63\%$ guarantee from Lemma A.1 is non-adaptive in the sense that it applies to every problem of the form (A.1). It is possible to provide a tighter guarantee once the optimization has been performed [67]:

Lemma A.2. *Let $\mathcal{S}^\star$ and $\mathcal{S}^{gr}$ be as in Lemma A.1. Let $\mathcal{B}$ be the set of top K maximal arguments of the function $f(x) = G(\mathcal{S}^{gr} \cup \{x\}) - G(\mathcal{S}^{gr})$, i.e. the top marginals of $G(\cdot)$ on*

top of the set $\mathcal{S}^{gr}$. Then we have:

$$G(\mathcal{S}^{\star}) \leq G(\mathcal{S}^{gr}) + \sum_{x \in \mathcal{B}} [G(\mathcal{S}^{gr} \cup \{x\}) - G(\mathcal{S}^{gr})] \tag{A.2}$$

Calculating the set $\mathcal{B}$ and evaluating (A.2) takes $O(|\mathcal{A}| \log |\mathcal{A}|)$ time as all we need to do is evaluate the marginals of $G(\cdot)$ with respect to all $|\mathcal{A}|$ elements of $\mathcal{A}$, sort them and sum the top K of which. The result is a bound on the gap between $G(\mathcal{S}^{\star})$ and $G(\mathcal{S}^{\text{gr}})$ which in practice tends to be smaller then that prescribed by Lemma A.1.

Appendix B

Inference in a Gaussian Process

Here we briefly review inference in a Gaussian Processes (GP) and its connection to kernel regression.

Definition B.1 ([92]). *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Let $f(x)$ be a GP defined over some region $x \in \mathcal{D} \subseteq \mathbb{R}^d$, with $m(\cdot) : \mathbb{R}^d \to \mathbb{R}$ the mean function[1] and $k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the covariance function, denoted $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$. Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be a finite set of sampling points and $f(\mathcal{X}) \equiv [f(x_1), \ldots, f(x_N)]$ a corresponding set of Gaussian process random variables stacked in vector form. Then, the GP assumption implies that $f(\mathcal{X})$ is consistently distributed as a Gaussian random vector with mean $m(\mathcal{X}) \equiv [m(x_1), \ldots, m(x_N)]^\top$ and a $N \times N$ covariance matrix defined according to $[K(\mathcal{X}, \mathcal{X})]_{i,j} = k(x_i, x_j)$.

Often, we do not have direct access to the GP sample values themselves $f(x)$ but rather to noisy measurements $y(x) = f(x) + \epsilon$ with the noise usually taken to be distributed as a Gaussian $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. We will be interested in collecting such noisy measurements at some process locations and estimating the underlying process values at other locations. Concretely, let $\mathcal{A}$ be a set of locations where noisy samples are collected, i.e. we have access to $y(\mathcal{A}) \equiv \{f(x) + \epsilon | x \in \mathcal{A}\}$ stacked as a vector. We are interested in estimating process values over a different set $\mathcal{B}$, i.e. making inference over $f(\mathcal{B}) = \{f(x) | x \in \mathcal{B}\}$ stacked as

[1] Usually we assume, without loss of generality, $m(x) \equiv 0$ to streamline our analysis

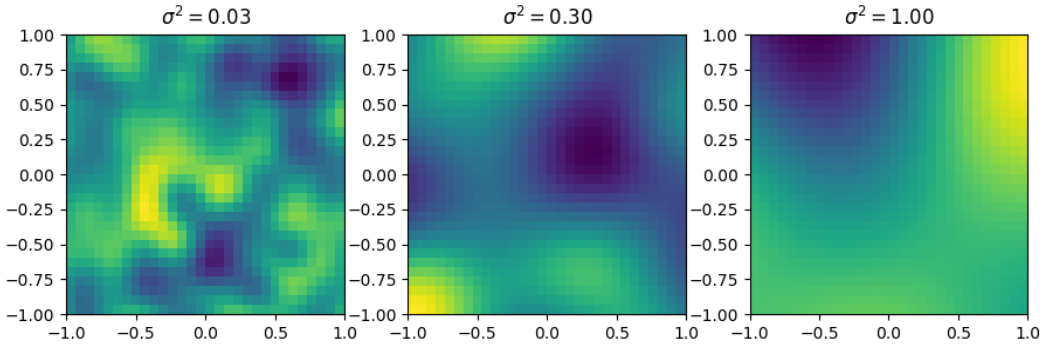


Figure B-1: Gaussian RBF kernel sample functions for various σ_f^2.

a vector. With all random variables in the problem distributed as Gaussians the posterior distribution $f(\mathcal{B})|y(\mathcal{A})$ may be retrieved using generic rules for Gaussian inference [92]. The final result, assuming zero mean function for the GP, reads:

$$f(\mathcal{B})|y(\mathcal{A}) \sim \mathcal{N}(\mu_{\mathcal{B}|\mathcal{A}}, \Sigma_{\mathcal{B}|\mathcal{A}})$$

$$\mu_{\mathcal{B}|\mathcal{A}} = K(\mathcal{B},\mathcal{A})[K(\mathcal{A},\mathcal{A}) + \sigma_n^2 I]^{-1} y(\mathcal{A}) \tag{B.1}$$

$$\Sigma_{\mathcal{B}|\mathcal{A}} = K(\mathcal{B},\mathcal{B}) - K(\mathcal{B},\mathcal{A})[K(\mathcal{A},\mathcal{A}) + \sigma_n^2 I]^{-1} K(\mathcal{A},\mathcal{B}) \tag{B.2}$$

where we have additionally defined $K(\mathcal{A},\mathcal{B})$ to be a $|\mathcal{A}| \times |\mathcal{B}|$ matrix with elements $k(x,x')$ for all $x \in \mathcal{A}, x' \in \mathcal{B}$.

Kernel functions The specific choice for the kernel $k(x,x')$ determines the characteristics of the GP, in terms of the structure and smoothness of typical sample functions. Here we mention two popular kernel choices. The Gaussian Radial Basis Function (RBF) $k(x,x') = \exp(-\frac{1}{2\sigma_f^2}\|x-x'\|^2)$ is both stationary, i.e. depends solely on $x - x'$ and more specifically isotropic, i.e. depends just on the distance $\|x - x'\|$. The kernel variance parameter σ_f^2 controls the typical correlation distance for sample functions, i.e. the larger it is, the larger the point wise separation needed to decorrelate process point samples. In Figure B-1 we illustrate three samples drawn from a zero-mean Gaussian RBF GP with varying levels of spatial correlation. While these figures illustrate single drawings from the GP they are indicative of the variation of the typical correlation distance with the kernel variance.

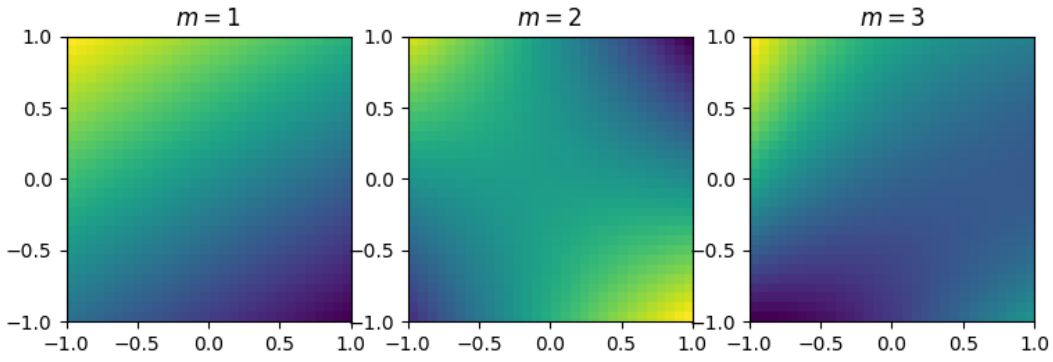


Figure B-2: Polynomial kernel sample functions for various m.

Another kernel of interest is the polynomial kernel $k(x, x') = (1 + \langle x, x' \rangle)^m$ with m the order parameter. In Figure B-2 we plot draws from the polynomial kernel for various orders. Notice how the structural complexity of the sample functions varies with the order.

The polynomial kernel is closely connected to linear regression with polynomial features of degree m as we detail in the next paragraph (and is visible in the sample plot).

Kernel regression Assume there exists a feature space transformation $\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^D$ with D the feature space dimension, such that $\forall x, x',\ k(x, x') = \phi(x)^\top \phi(x')$. Define $f(x) = c^\top \phi(x)$ with $c \in \mathbb{R}^D,\ c \sim \mathcal{N}(0, I_{D \times D})$ and $I_{D \times D}$ the D-dimensional unit matrix. Noting that $f(x)$ as defined is zero-mean Gaussian distributed with $\mathbb{E}\left[f(x)f(x')\right] = \phi(x)^\top \mathbb{E}\left[cc^\top\right] \phi(x) = k(x, x')$ we have $f(x) \sim \mathcal{GP}(0, k(x, x'))$ and we see that a GP with kernel $k(x, x')$ is closely tied to a linear regression in the feature space $\phi(x)$. The GP sample functions are expected to look like linear functions in the feature space according to $f(x) = c^\top \phi(x)$.

As an example, for the non-biased polynomial kernel $k(x, x') = \langle x, x' \rangle^m$ it may be shown that there exists a feature space transformation $\phi(x)$ in dimension $D = O(d^m)$, with the features proportional to multinomial factors $x_1^{m_1} \cdot x_2^{m_2} \cdots x_d^{m_d}$ and $\sum_i m_i = m$ and the GP sample functions tend to look like polynomials, as in Figure B-2 [92].

Appendix C

Proofs

C.1 Proof of Theorem 2.2

Using the definition of mutual information[1] we have

$$G(\mathcal{Y}_0) = I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))) = H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))|\boldsymbol{\Theta}) \tag{C.1}$$

$$\tilde{G}(\mathcal{Y}_0) = I(\boldsymbol{\Theta}; \tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) = H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))|\boldsymbol{\Theta}) \tag{C.2}$$

Conditioned on $\boldsymbol{\Theta}$ the remaining uncertainty in the measurements is just the random noise and we have $H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))|\boldsymbol{\Theta}) = H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))|\boldsymbol{\Theta}) = H(\boldsymbol{\epsilon})$ such that:

$$G(\mathcal{Y}_0) - \tilde{G}(\mathcal{Y}_0) = H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))) - H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) \tag{C.3}$$

Notice that both $\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))$ and $\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))$ are collections of $\tilde{K}$ Gaussian random variables as noisy samples from the GP. Now apply the generic formula for the entropy of a Gaussian random vector[2]:

$$\begin{aligned} H(\tilde{\mathbf{F}}(\mathcal{Y}_m(\mathcal{Y}_0))) &= \log((\pi e)^{\tilde{K}} \det \boldsymbol{\Sigma}_m) \\ H(\tilde{\mathbf{F}}(\mathcal{Y}_g(\mathcal{Y}_0))) &= \log((\pi e)^{\tilde{K}} \det \boldsymbol{\Sigma}_g) \end{aligned} \tag{C.4}$$

[1] $I(x;y)=H(x)-H(x|y)=H(y)-H(y|x)$

[2] $\mathbf{x} \in \mathbb{R}^k, \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow H(\mathbf{x}) = \log((\pi e)^k \det \boldsymbol{\Sigma})$

where $\tilde{K} \equiv d\tilde{K}$ and:

$$\begin{aligned}\boldsymbol{\Sigma}_m &= \mathbf{k}(\mathcal{Y}_m, \mathcal{Y}_m) + \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_{\tilde{K}} \\ \boldsymbol{\Sigma}_g &= \mathbf{k}(\mathcal{Y}_g, \mathcal{Y}_g) + \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_{\tilde{K}}\end{aligned} \tag{C.5}$$

So,

$$\tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0) = \log(\det(\boldsymbol{\Sigma}_g)) - \log(\det(\boldsymbol{\Sigma}_m)) \tag{C.6}$$

Now define $\mathbf{X} \equiv \frac{1}{\delta}(\boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_m) \leftrightarrow \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_m + \delta\mathbf{X}$ with $\mathbf{X} \in \mathbb{R}^{\tilde{K}\times\tilde{K}}$ satisfying $\forall i, j\ |X_{ij}| \leq 1$ according to our assumption of bounded covariance differences.

$\boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_m$ are both positive-definite and invertible such that we can write:

$$\det(\boldsymbol{\Sigma}_g) = \det(\boldsymbol{\Sigma}_m + \delta\mathbf{X}) = \det(\boldsymbol{\Sigma}_m)\det(\mathbf{I} + \delta\boldsymbol{\Sigma}_m^{-1}\mathbf{X}) \tag{C.7}$$

Substituting (C.7) in (C.6) we have:

$$\tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0) = \log(\det(\mathbf{I} + \delta\boldsymbol{\Sigma}_m^{-1}\mathbf{X})) = \log(\det(\tilde{\mathbf{X}})) \tag{C.8}$$

with $\tilde{\mathbf{X}} \equiv \mathbf{I} + \delta\boldsymbol{\Sigma}_m^{-1}\mathbf{X}$. We turn next to bounding $\log(\det(\tilde{\mathbf{X}}))$. First notice:

$$\begin{aligned}\left|[\delta\boldsymbol{\Sigma}_m^{-1}\mathbf{X}]_{ij}\right| &= \delta\left|\sum_r \Sigma^{-1}_{m,ir}X_{rj}\right| \leq \delta\sum_r \left|\Sigma^{-1}_{m,ir}X_{rj}\right| \leq \delta\sum_r \left|\Sigma^{-1}_{m,ir}\right| \leq \delta\|\boldsymbol{\Sigma}_m^{-1}\|_\infty \\ &\leq \delta\sqrt{\tilde{K}}\|\boldsymbol{\Sigma}_m^{-1}\|_2 = \delta\sqrt{\tilde{K}}\sigma_{\max}(\boldsymbol{\Sigma}_m^{-1}) = \frac{\delta\sqrt{\tilde{K}}}{\sigma_{\min}(\boldsymbol{\Sigma}_m)} \leq \frac{\delta\sqrt{\tilde{K}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)}\end{aligned} \tag{C.9}$$

where we used the matrix norm inequality $\|\mathbf{A}\|_\infty \leq \sqrt{\tilde{K}}\|\mathbf{A}\|_2$ for $\mathbf{A} \in \mathbb{R}^{\tilde{K}\times\tilde{K}}$ and $\sigma_{\max}(\cdot)$.

Thus we have that $\tilde{\mathbf{X}}$ has diagonal elements centered around 1, i.e. for all i

$$\left|\tilde{X}_{ii} - 1\right| \leq \frac{\delta\sqrt{\tilde{K}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)}$$

and the row-sums over non-diagonal entries satisfy for all i

$$\sum_{r \neq i} \left| \tilde{X}_{ir} \right| \leq \frac{\delta \sqrt{\tilde{K}} (\tilde{K} - 1)}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)}$$

. Designating the eigenvalues of $\tilde{\mathbf{X}}$ as $\{\lambda_i\}$ and applying the Gershgorin circle theorem, we have:

$$1 - \frac{\delta \tilde{K}^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)} \leq |\lambda_i| \leq 1 + \frac{\delta \tilde{K}^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)} \,. \tag{C.10}$$

Using $\log(\det(\tilde{\mathbf{X}})) = \sum_i \log(|\lambda_i|)$, this implies that

$$\tilde{K} \log \left(1 - \frac{\delta \tilde{K}^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)} \right) \leq \log(\det(\tilde{\mathbf{X}})) \leq \tilde{K} \log \left(1 + \frac{\delta \tilde{K}^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)} \right) \tag{C.11}$$

where the left hand side is to be interpreted as minus infinity when the argument of the log function is negative. Finally, using $\tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0) = \log(\det(\tilde{\mathbf{X}}))$ and $\log(1+x) \leq -\log(1-x)$ we have:

$$|\tilde{G}(\mathcal{Y}_0) - G(\mathcal{Y}_0)| \leq -\tilde{K} \log \left(1 - \frac{\delta \tilde{K}^{\frac{3}{2}}}{\sigma_{\min}(\boldsymbol{\Sigma}_\epsilon)} \right) \tag{C.12}$$

C.2 Proof of Lemma 3.2

For simplicity of notation, we suppress the subscript $\mathcal{S}$ throughout. Begin by expanding the mutual information expressions:

$$\begin{aligned} I(\tilde{\boldsymbol{f}}; \{\beta_m\}) &= H(\tilde{\boldsymbol{f}}) - H(\tilde{\boldsymbol{f}} | \{\beta_m\}) \\ I(\hat{\boldsymbol{f}}; \boldsymbol{\beta}) &= H(\hat{\boldsymbol{f}}) - H(\hat{\boldsymbol{f}} | \boldsymbol{\beta}) \end{aligned} \tag{C.13}$$

Examining (3.15) and (3.21) we have $H(\tilde{\boldsymbol{f}} | \{\beta_m\}) = H(\hat{\boldsymbol{f}} | \boldsymbol{\beta}) = H(\boldsymbol{w})$ and so:

$$I(\tilde{\boldsymbol{f}}; \{\beta_m\}) - I(\hat{\boldsymbol{f}}; \boldsymbol{\beta}) = H(\tilde{\boldsymbol{f}}) - H(\hat{\boldsymbol{f}}) \tag{C.14}$$

Next, notice that $\tilde{\boldsymbol{f}}$ and $\hat{\boldsymbol{f}}$ are both circular, complex, Gaussian random N-length vectors, such that their entropies are given according to [81]:

$$\begin{aligned} H(\tilde{\boldsymbol{f}}) &= \log((\pi e)^N \det(\tilde{\boldsymbol{\Sigma}})) & \tilde{\Sigma}_{ij} &= \mathbb{E}[\tilde{\boldsymbol{f}}_i \tilde{\boldsymbol{f}}_j^*] \\ H(\hat{\boldsymbol{f}}) &= \log((\pi e)^N \det(\hat{\boldsymbol{\Sigma}})) & \hat{\Sigma}_{ij} &= \mathbb{E}[\hat{\boldsymbol{f}}_i \hat{\boldsymbol{f}}_j^*] \end{aligned} \tag{C.15}$$

Further expand using the independence between $\boldsymbol{w}$ and β_m and $\mathbb{E}[\beta_m \beta_{m'}^*] = \delta_{mm'}\sigma_m^2$:

$$\tilde{\Sigma}_{ij} = \mathbb{E}[(\sum_m K_{im}\beta_m + w_i)(\sum_{m'} K_{jm'}^*\beta_{m'}^* + w_j^*)] = [\boldsymbol{\Sigma}_{ww}]_{ij} + \sum_m K_{im}K_{jm}^*\sigma_m^2 \tag{C.16}$$

$$\hat{\Sigma}_{ij} = \mathbb{E}[(\sum_{m\in\mathcal{M}} K_{im}\beta_m + w_i)(\sum_{m'\in\mathcal{M}} K_{jm'}^*\beta_{m'}^* + w_j^*)] = [\boldsymbol{\Sigma}_{ww}]_{ij} + \sum_{m\in\mathcal{M}} K_{im}K_{jm}^*\sigma_m^2 \tag{C.17}$$

Comparing (C.16) and (C.17) and using:

$$\left|\sum_{m\notin\mathcal{M}} K_{im}K_{jm}^*\sigma_m^2\right| \leq \sum_{m\notin\mathcal{M}} \left|K_{im}K_{jm}^*\sigma_m^2\right| \leq \sum_{m\notin\mathcal{M}} \sigma_m^2 = \epsilon \tag{C.18}$$

we have $|\tilde{\Sigma}_{ij} - \hat{\Sigma}_{ij}| \leq \epsilon$ such that we may write:

$$\hat{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Sigma}} + \epsilon \boldsymbol{X} \tag{C.19}$$

for some $N \times N$ matrix $\boldsymbol{X}$ satisfying $|X_{ij}| \leq 1$. We use (C.19) to bound the determinants. $\tilde{\boldsymbol{\Sigma}}$ is positive-definite and invertible such that we can write:

$$\det(\hat{\boldsymbol{\Sigma}}) = \det(\tilde{\boldsymbol{\Sigma}} + \epsilon\boldsymbol{X}) = \det(\tilde{\boldsymbol{\Sigma}})\det(\boldsymbol{I}_N + \epsilon\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}) \tag{C.20}$$

Substituting (C.20) in (C.15) we have:

$$H(\hat{\boldsymbol{f}}) - H(\tilde{\boldsymbol{f}}) = \log(\det(\boldsymbol{I}_N + \epsilon\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})) = \log(\det(\tilde{\boldsymbol{X}})) \tag{C.21}$$

with $\tilde{\boldsymbol{X}} \equiv \boldsymbol{I}_N + \epsilon\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}$. We turn next to bounding the term $\log(\det(\tilde{\boldsymbol{X}}))$. First notice:

$$\left|[\epsilon\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]_{ij}\right| = \epsilon\left|\sum_m \tilde{\Sigma}^{-1}_{im}X_{mj}\right| \leq \epsilon\sum_m \left|\tilde{\Sigma}^{-1}_{im}X_{mj}\right| \leq \epsilon\sum_m \left|\tilde{\Sigma}^{-1}_{im}\right|$$
$$\leq \epsilon\|\tilde{\boldsymbol{\Sigma}}^{-1}\|_\infty \leq \epsilon\sqrt{N}\|\tilde{\boldsymbol{\Sigma}}^{-1}\|_2 = \epsilon\sqrt{N}\sigma_{\max}(\tilde{\boldsymbol{\Sigma}}^{-1}) = \epsilon\sqrt{N}\frac{1}{\sigma_{\min}(\tilde{\boldsymbol{\Sigma}})} \leq \frac{\epsilon\sqrt{N}}{\sigma_w^2} \quad \text{(C.22)}$$

Where we have used the matrix norm equivalence $\|\boldsymbol{A}\|_\infty \leq \sqrt{N}\|\boldsymbol{A}\|_2$ (for $N\times N$ matrices) and $\sigma_{\max}(\cdot)$ ($\sigma_{\min}(\cdot)$) is the maximal (minimal) singular value such that $\sigma_{\min}(\tilde{\boldsymbol{\Sigma}}) \geq \sigma_w^2$. Thus we have that $\tilde{\boldsymbol{X}}$ has diagonal elements centered around 1: $\left|\tilde{X}_{ii} - 1\right| \leq \frac{\epsilon\sqrt{N}}{\sigma_w^2}$ and the row-sums over non-diagonal entries satisfy $\sum_{m\neq i}\left|\tilde{X}_{im}\right| \leq \frac{\epsilon\sqrt{N}(N-1)}{\sigma_w^2}$.

Applying the Gershgorin circle theorem we have for the eigenvalues of $\tilde{\boldsymbol{X}}$:

$$1 - \frac{\epsilon\sqrt{N}N}{\sigma_w^2} \leq |\lambda_i| \leq 1 + \frac{\epsilon\sqrt{N}N}{\sigma_w^2}$$

$\det(\tilde{\boldsymbol{X}})$ is a positive real number as the quotient of the determinants of two positive definite matrices $\det(\tilde{\boldsymbol{X}}) = \frac{\det(\hat{\boldsymbol{\Sigma}})}{\det(\tilde{\boldsymbol{\Sigma}})}$ such that we can write $\det(\tilde{\boldsymbol{X}}) = \prod_i \lambda_i = \prod_i |\lambda_i|$ and consequently:

$$N\log(1 - \frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2}) \leq \log(\det(\tilde{\boldsymbol{X}})) \leq N\log(1 + \frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2}) \quad \text{(C.23)}$$

which finally leads to:

$$-N\log(1+\frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2}) \leq I(\tilde{\boldsymbol{f}};\{\beta_m\}) - I(\hat{\boldsymbol{f}};\boldsymbol{\beta}) \leq -N\log(1-\frac{\epsilon N^{\frac{3}{2}}}{\sigma_w^2}) \quad \text{(C.24)}$$

C.3 Proof of Lemma 3.3

The left inequality is trivial. We have $I(\hat{\boldsymbol{f}}_{\mathcal{S}_d^\star};\boldsymbol{\beta}) \leq I(\hat{\boldsymbol{f}}_{\mathcal{S}^\star};\boldsymbol{\beta})$ as the second optimization is over a larger set.

To prove the right inequality we show that for every $\mathcal{S} \subseteq \mathcal{A}$ there is $\mathcal{S}_d \subseteq \mathcal{V}$ such that

$$I(\hat{\boldsymbol{f}}_{\mathcal{S}};\boldsymbol{\beta}) \leq I(\hat{\boldsymbol{f}}_{\mathcal{S}_d};\boldsymbol{\beta}) + N\log(1+\frac{4\delta P(1+\delta)N^{\frac{3}{2}}}{\lambda\sigma_w^2}) \quad \text{(C.25)}$$

we will show this for $|\mathcal{S}|{=}N$ but the proof for other cardinalities is identical.

With distance δ between adjacent elements of $\mathcal{V}$, for every $\mathcal{S}{=}\{x_1,\dots,x_N\}\subset\mathcal{A}$ there is a set $\mathcal{S}_d{=}\left\{x_1^d,\dots,x_N^d\right\}$ such that $\mathcal{S}_d\subseteq\mathcal{V}$ and $|x_i-x_i^d|\leq\frac{\delta}{2}$ for all i. We have, similarly to Appendix C.2:

$$I(\hat{\boldsymbol{f}}_{\mathcal{S}};\boldsymbol{\beta})-I(\hat{\boldsymbol{f}}_{\mathcal{S}_d};\boldsymbol{\beta})=H(\hat{\boldsymbol{f}}_{\mathcal{S}})-H(\hat{\boldsymbol{f}}_{\mathcal{S}_d}) \tag{C.26}$$

Using the model (3.21):

$$\begin{aligned}H(\hat{\boldsymbol{f}}_{\mathcal{S}})&=\log((\pi e)^N\det(\hat{\boldsymbol{\Sigma}}^{\mathcal{S}}))\\ H(\hat{\boldsymbol{f}}_{\mathcal{S}_d})&=\log((\pi e)^N\det(\hat{\boldsymbol{\Sigma}}^{\mathcal{S}_d}))\end{aligned} \tag{C.27}$$

where:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}^{\mathcal{S}}&\equiv\mathbb{E}[\hat{\boldsymbol{f}}_{\mathcal{S}}\hat{\boldsymbol{f}}_{\mathcal{S}}^{\dagger}]=\boldsymbol{K}_{\mathcal{S}}\boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{K}_{\mathcal{S}}^{\dagger}+\boldsymbol{\Sigma}_{ww}\\ \hat{\boldsymbol{\Sigma}}^{\mathcal{S}_d}&\equiv\mathbb{E}[\hat{\boldsymbol{f}}_{\mathcal{S}_d}\hat{\boldsymbol{f}}_{\mathcal{S}_d}^{\dagger}]=\boldsymbol{K}_{\mathcal{S}_d}\boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{K}_{\mathcal{S}_d}^{\dagger}+\boldsymbol{\Sigma}_{ww}\\ \boldsymbol{\Sigma}_{\beta\beta}&\equiv\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^{\dagger}]\end{aligned} \tag{C.28}$$

and $H(\hat{\boldsymbol{f}}_{\mathcal{S}})-H(\hat{\boldsymbol{f}}_{\mathcal{S}_d})=\log(\det(\hat{\boldsymbol{\Sigma}}^{\mathcal{S}}))-\log(\det(\hat{\boldsymbol{\Sigma}}^{\mathcal{S}_d}))$.
Both $\boldsymbol{K}_{\mathcal{S}}$ and $\boldsymbol{K}_{\mathcal{S}_d}$ are size $N\times|\mathcal{M}|$ matrices defined as per the definition in (3.14), such that:

$$|[\boldsymbol{K}_{\mathcal{S}}]_{nm}{-}[\boldsymbol{K}_{\mathcal{S}_d}]_{nm}|=|\mathrm{sinc}(m{+}\frac{2}{\lambda}x_n){-}\mathrm{sinc}(m{+}\frac{2}{\lambda}x_n^d)|\leq 2\frac{2}{\lambda}|x_q{-}x_q^d|\leq\frac{2}{\lambda}\delta \tag{C.29}$$

where for the first inequality we have used the fact that $\mathrm{sinc}(\cdot)$ is Lipschitz with constant smaller than 2. We can thus define $\boldsymbol{\Delta}\equiv\boldsymbol{K}_{\mathcal{S}}-\boldsymbol{K}_{\mathcal{S}_d}$ and we have $|\Delta_{nm}|\leq\frac{2}{\lambda}\delta$. Substitution in (C.28) yields:

$$\hat{\boldsymbol{\Sigma}}^{\mathcal{S}}=\hat{\boldsymbol{\Sigma}}^{\mathcal{S}_d}+\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{\Delta}^{\dagger}+\boldsymbol{\Delta}\boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{K}_{\mathcal{S}_d}^{\dagger}+\boldsymbol{K}_{\mathcal{S}_d}\boldsymbol{\Sigma}_{\beta\beta}\boldsymbol{\Delta}^{\dagger} \tag{C.30}$$

We bound the perturbation terms by noticing:

$$\begin{aligned}\left|[\boldsymbol{\Delta\Sigma}_{\beta\beta}\boldsymbol{\Delta}^\dagger]_{ij}\right| &= \left|\sum_m \Delta_{im}[\boldsymbol{\Sigma}_{\beta\beta}]_{mm}\Delta_{jm}\right| \leq (\frac{2}{\lambda}\delta)^2 \sum_m [\boldsymbol{\Sigma}_{\beta\beta}]_{mm} \leq \frac{4}{\lambda^2}\delta^2 P \\ \left|[\boldsymbol{\Delta\Sigma}_{\beta\beta}\boldsymbol{K}^\dagger_{\mathcal{S}_d}]_{ij}\right| &= \left|\sum_m \Delta_{im}[\boldsymbol{\Sigma}_{\beta\beta}]_{mm}[\boldsymbol{K}_{\mathcal{S}_d}]_{jm}\right| \leq \frac{2}{\lambda}\delta \cdot 1 \cdot \sum_m [\boldsymbol{\Sigma}_{\beta\beta}]_{mm} \leq \frac{2}{\lambda}\delta P \end{aligned} \tag{C.31}$$

and overall we have:

$$|\hat{\Sigma}^{\mathcal{S}}_{ij} - \hat{\Sigma}^{\mathcal{S}_d}_{ij}| \leq \frac{4}{\lambda}\delta P + \frac{4}{\lambda^2}\delta^2 P = \frac{4}{\lambda}\delta P(1+\delta) \tag{C.32}$$

define: $\epsilon' \equiv \frac{4}{\lambda}\delta P(1+\delta)$ and we have

$$\hat{\boldsymbol{\Sigma}}^{\mathcal{S}} = \hat{\boldsymbol{\Sigma}}^{\mathcal{S}_d} + \epsilon' \boldsymbol{X} \tag{C.33}$$

with $N \times N$ matrix $\boldsymbol{X}$ satisfying $|X_{ij}| \leq 1$ which is akin to (C.19). We thus port the results from Appendix C.2 here (we only need the lower bound):

$$-N\log(1 + \frac{\epsilon' N^{\frac{3}{2}}}{\sigma_w^2}) \leq I(\hat{\boldsymbol{f}}_{\mathcal{S}_d};\boldsymbol{\beta}) - I(\hat{\boldsymbol{f}}_{\mathcal{S}};\boldsymbol{\beta}) \tag{C.34}$$

which, upon substitution of ϵ' is equivalent to (C.25).

C.4 Proof of Theorem 3.1

The greedy algorithm sequentially selects elements according to the rule $x^\star = \text{argmax}_{x\in\mathcal{V}\setminus\mathcal{S}} I(\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}};\boldsymbol{\beta})$ where $\mathcal{S}$ is the set of elements selected so far. We recursively show that the added elements can be selected on a $\frac{\lambda}{2}$-spaced grid centered around $x = 0$. Expanding the mutual information as in Appendix C.2 we have:

$$\text{argmax}_{x\in\mathcal{V}\setminus\mathcal{S}} I(\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}};\boldsymbol{\beta}) = \text{argmax}_{x\in\mathcal{V}\setminus\mathcal{S}} H(\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}}) \tag{C.35}$$

We begin by showing that the first selected element is $x_1 = 0$. Indeed, using the results from Appendix C.2:

$$\underset{x\in\mathcal{V}}{\operatorname{argmax}}\, H(\hat{\boldsymbol{f}}_{\{x\}}) = \underset{x\in\mathcal{V}}{\operatorname{argmax}}\, \det(\hat{\boldsymbol{\Sigma}}_{11}) = \underset{x\in\mathcal{V}}{\operatorname{argmax}}\, \hat{\boldsymbol{\Sigma}}_{11} \tag{C.36}$$

where again using Appendix C.2 and under the assumptions of the theorem (high SNR):

$$\hat{\boldsymbol{\Sigma}}_{11} = \sum_m K_{1m}K_{1m}^*\sigma_m^2 = \sum_m \operatorname{sinc}^2(m+\frac{2}{\lambda}x)\sigma_m^2 \leq \sigma_0^2 \sum_m \operatorname{sinc}^2(m+\frac{2}{\lambda}x) = \sigma_0^2 \tag{C.37}$$

where we used $\sigma_m^2 \leq \sigma_0^2, \quad \forall m \neq 0$ and the identity:

$$\sum_m \operatorname{sinc}(m+a)\operatorname{sinc}(m+b) = \operatorname{sinc}(b-a) \tag{C.38}$$

It easy to see that in (C.37) equality is achieved for the choice $x_1 = 0$ which is the claim.

Next, assume that the greedy algorithm has already picked a set $\mathcal{S} = \{x_1, \ldots, x_{|\mathcal{S}|}\}$ of adjacent elements on a $\frac{\lambda}{2}$-spaced grid centered around $x = 0$ and show that the next element to be added is an adjacent location on the same $\frac{\lambda}{2}$-spaced grid. We have using $H(x,y)=H(x)+H(y|x)$:

$$\underset{x\in\mathcal{V}\setminus\mathcal{S}}{\operatorname{argmax}}\, H(\hat{\boldsymbol{f}}_{\mathcal{S}\cup\{x\}}) = \underset{x\in\mathcal{V}\setminus\mathcal{S}}{\operatorname{argmax}}\, H(\hat{\boldsymbol{f}}_{\{x\}}|\hat{\boldsymbol{f}}_{\mathcal{S}}) = \underset{x\in\mathcal{V}\setminus\mathcal{S}}{\operatorname{argmax}}\, \sigma_{x|\mathcal{S}}^2 \tag{C.39}$$

where $\sigma_{x|\mathcal{S}}^2$ is the conditional variance of the Gaussian observation collected at x given the Gaussian observations made at the set $\mathcal{S}$:

$$\sigma_{x|\mathcal{S}}^2 = \sigma_x^2 - \boldsymbol{\Sigma}_{x\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{x\mathcal{S}}^\dagger \tag{C.40}$$

with the usual definitions for the covariance matrices (as in Appendix C.2):

$$\begin{aligned}
[\boldsymbol{\Sigma}_{x\mathcal{S}}]_{1i} &= \sum_m \operatorname{sinc}(m+\frac{2}{\lambda}x)\operatorname{sinc}(m+\frac{2}{\lambda}x_i)\sigma_m^2 = \operatorname{sinc}(m(i)+\frac{2}{\lambda}x)\sigma_{m(i)}^2 \\
[\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}]_{ij} &= \sum_m \operatorname{sinc}(m+\frac{2}{\lambda}x_i)\operatorname{sinc}(m+\frac{2}{\lambda}x_j)\sigma_m^2 = \delta_{ii}\sigma_{m(i)}^2
\end{aligned} \tag{C.41}$$

where in the last equations almost all sinc$(\cdot)$ functions nulled out as the x_i's are situated on a $\frac{\lambda}{2}$ grid, and we have defined $m(i) \equiv -\frac{2}{\lambda} x_i$, such that $I \equiv \{m(i)\}$ is a set of consecutive integers. Additionally, we have:

$$\sigma_x^2 = \sum_m \text{sinc}^2(m + \frac{2}{\lambda} x) \sigma_m^2 \tag{C.42}$$

Substituting back into (C.40) we have:

$$\sigma_{x|\mathcal{S}}^2 = \sum_m \text{sinc}^2(m + \frac{2}{\lambda} x) \sigma_m^2 - \sum_{i \in I} \text{sinc}^2(m(i) + \frac{2}{\lambda} x) \sigma_{m(i)}^2 = \sum_{m \notin I} \text{sinc}^2(m + \frac{2}{\lambda} x) \sigma_m^2 \tag{C.43}$$

and this is similar to the optimization over the selection of the first location in (C.37) with the optimum achieved by selecting x such that $\frac{2}{\lambda} x = m$ for the first $m \notin I$ which is the next adjacent location on the $\frac{\lambda}{2}$ grid which completes the proof.

Bibliography

[1] Juan-Felipe PJ Abascal, Simon R Arridge, David Atkinson, Raya Horesh, Lorenzo Fabrizi, Marzia De Lucia, Lior Horesh, Richard H Bayford, and David S Holder. Use of anisotropic modelling in electrical impedance tomography; description of method and preliminary assessment of utility in imaging brain function in the adult human head. *Neuroimage*, 43(2):258–268, 2008.

[2] Tülay Adali, Peter J Schreier, and Louis L Scharf. Complex-valued signal processing: The proper way to deal with impropriety. *IEEE Transactions on Signal Processing*, 59(11):5101–5125, 2011.

[3] Mauricio A Álvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard Baraniuk. Flatcam: Thin, bare-sensor cameras using coded aperture and computation. *arXiv preprint arXiv:1509.00116*, 2015.

[6] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[7] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[8] José M Bernardo. Expected information as expected utility. *The Annals of Statistics*, pages 686–690, 1979.

[9] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

[10] George EP Box, William G Hunter, and Stuart J Hunter. *Statistics for Experimenters*. John Wiley & Sons, 1978.

[11] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

[12] David J Brady, Nikos P Pitsianis, and Xiaobai Sun. Reference structure tomography. *JOSA A*, 21(7):1140–1147, 2004.

[13] Niv Buchbinder, Moran Feldman, Joseph Seffi, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5):1384–1402, 2015.

[14] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-Line-of-Sight imaging using a time-gated single photon avalanche diode. *Optics express*, 23(16):20997–21011, 2015.

[15] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *Integer programming and combinatorial optimization*, pages 182–196. Springer, 2007.

[16] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

[17] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[18] David R Chase, Lee-Yin Chen, and Robert A York. Modeling the capacitive nonlinearity in thin-film BST varactors. *IEEE transactions on microwave theory and techniques*, 53(10):3215–3220, 2005.

[19] Adam Lloyd Cohen. Anti-pinhole imaging. *Journal of Modern Optics*, 29(1):63–67, 1982.

[20] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

[21] Ernesto De Vito, Veronica Umanità, and Silvia Villa. An extension of Mercer theorem to matrix-valued measurable kernels. *Applied and Computational Harmonic Analysis*, 34(3):339–351, 2013.

[22] Marco F Duarte, Mark A Davenport, Dharmpal Takbar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

[23] Yonina C Eldar. *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 2015.

[24] Hooman Fatoorehchi, Hossein Abolghasemi, and Reza Zarghami. Analytical approximate solutions for a general nonlinear resistor–nonlinear capacitor circuit model. *Applied Mathematical Modelling*, 39(19):6021–6031, 2015.

[25] Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

[26] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[27] El E Fenimore and TM Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied optics*, 17(3):337–347, 1978.

[28] E. Fishler, A. Haimovich, R. S. Blum, L. J. Cimini, D. Chizhik, and R. A. Valenzuela. Spatial diversity in radars-models and detection performance. *IEEE Trans. on Signal Process.*, 54(3):823–838, 2006.

[29] L. Formaggia, J. F. Gerbeau, F. Nobile, and A. Quarteroni. Numerical treatment of defective boundary conditions for the Navier–Stokes equations. *SIAM Journal on Numerical Analysis*, 40(1):376–401, 2002.

[30] Gero Friesecke. Lecture notes in Fourier analysis [MA4064], 2013.

[31] Satoru Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

[32] Genevieve Gariepy, Francesco Tonolini, Robert Henderson, Jonathan Leach, and Daniele Faccio. Detection and tracking of moving objects hidden from view. *Nature Photonics*, 2015.

[33] Houcem Gazzah and Sylvie Marcos. Cramer-Rao bounds for antenna array design. *IEEE Transactions on Signal Processing*, 54(1):336–345, 2006.

[34] C William Gear. *Numerical initial value problems in ordinary differential equations*. Prentice Hall PTR, 1971.

[35] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[36] Herbert Goldstein. *Classical mechanics*. Pearson Education India, 1965.

[37] E Haber, UM Ascher, DA Aruliah, and DW Oldenburg. Fast simulation of 3D electromagnetic problems using potentials. *Journal of Computational Physics*, 163(1):150–171, 2000.

[38] Eldad Haber, Lior Horesh, and Luis Tenorio. Numerical methods for experimental design of large-scale linear ill-posed inverse problems. *Inverse Problems*, 24(5):055012, 2008.

[39] Zachary T Harmany, Roummel F Marcia, and Rebecca M Willett. This is SPIRAL-TAP: Sparse poisson intensity reconstruction algorithms, theory and practice. *IEEE Transactions on Image Processing*, 21(3):1084–1096, 2012.

[40] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The Elements of Statistical Learning*, volume 2. Springer, Dordrecht, Germany, 2009.

[41] Randy L Haupt. Thinned arrays using genetic algorithms. *Antennas and Propagation, IEEE Transactions on*, 42(7):993–999, 1994.

[42] Randy L Haupt and Douglas H Werner. *Genetic algorithms in electromagnetics*. John Wiley & Sons, 2007.

[43] Xiaoyi He and Li-Shi Luo. Lattice boltzmann model for the incompressible Navier–Stokes equation. *Journal of statistical Physics*, 88(3):927–944, 1997.

[44] Felix Heide, Lei Xiao, Wolfgang Heidrich, and Matthias B Hullin. Diffuse mirrors: 3D reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3222–3229, 2014.

[45] HD Helms and JB Thomas. Truncation error of sampling-theorem expansions. *Proceedings of the IRE*, 50(2):179–184, 1962.

[46] Lior Horesh, Eldad Haber, and Luis Tenorio. Optimal experimental design for the large-scale nonlinear ill-posed problem of impedance imaging. *Large-Scale Inverse Problems and Quantification of Uncertainty*, pages 273–290, 2010.

[47] Lior Horesh, Leo Liberti, Haim Avron, and David Nahamoo. Globally convergent system and method for automated model discovery, April 6 2015. US Patent Office Serial Number 14/755,942.

[48] D Jagerman. Bounds for truncation error of the sampling expansion. *SIAM Journal on Applied Mathematics*, 14(4):714–723, 1966.

[49] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[50] M.C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[51] Majid M Khodier and Christos G Christodoulou. Linear array geometry synthesis with minimum sidelobe level and null control using particle swarm optimization. *Antennas and Propagation, IEEE Transactions on*, 53(8):2674–2679, 2005.

[52] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using ultrafast transient imaging. *International journal of computer vision*, 95(1):13–28, 2011.

[53] Jonathan Klein, Christoph Peters, Jaime Martín, Martin Laurenzis, and Matthias B Hullin. Tracking objects outside the line of sight using 2D intensity images. *Scientific Reports*, 6, 2016.

[54] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 3:19, 2012.

[55] Andreas Krause and Carlos Guestrin. Submodularity and its applications in optimized information gathering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(4):32, 2011.

[56] Andreas Krause and Carlos E Guestrin. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

[57] Andreas Krause, H Brendan McMahan, Carlos Guestrin, and Anupam Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9:2761–2801, 2008.

[58] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

[59] James D Krieger, Yuval Kochman, and Gregory W Wornell. Design and analysis of multi-coset arrays. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3781–3785. IEEE, 2013.

[60] James D Krieger, Yuval Kochman, and Gregory W Wornell. Multi-coset sparse imaging arrays. *IEEE Transactions on Antennas and Propagation*, 62(4):1701–1715, 2014.

[61] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *Signal Processing Magazine, IEEE*, 13(4):67–94, 1996.

[62] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Advances in Neural Information Processing Systems*, pages 1033–1041, 2009.

[63] B Preetham Kumar and GR Branner. Design of unequally spaced arrays for performance improvement. *IEEE Transactions on Antennas and Propagation*, 47(3):511–523, 1999.

[64] Yury A Kutoyants. *Identification of dynamical systems with small noise*, volume 300. Springer Science & Business Media, 2012.

[65] Remi Lam, Lior Horesh, Haim Avron, and Karen Willcox. An optimization framework for hybrid first principles data-driven modeling. In *Copper Mountain Conference on Iterative Methods*, 2016.

[66] Remi Lam, Karen Willcox, and David H Wolpert. Bayesian optimization with a finite budget: An approximate dynamic programming approach. In *Advances in Neural Information Processing Systems*, pages 883–891, 2016.

[67] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[68] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70, 2007.

[69] J. Li and P. Stoica. MIMO radar with colocated antennas. *IEEE Signal Processing Mag.*, 24(5):106–114, 2007.

[70] J. Li, P. Stoica, L. Xu, and W. Roberts. On parameter identifiability of MIMO radar. *IEEE Signal Process. Lett.*, 14(12):968–971, 2007.

[71] W. Liao and A. Fannjiang. MUSIC for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 2014.

[72] J. R. Lindner. Microbubbles in medical imaging: current applications and future directions. *Nature Reviews Drug Discovery*, 3(6):527–533, 2004.

[73] Juan A Martinez and Bruce A Mork. Transformer modeling for low-and mid-frequency transients-a review. *IEEE Transactions on Power Delivery*, 20(2):1625–1632, 2005.

[74] Kevin McGoff, Sayan Mukherjee, Natesh Pillai, et al. Statistical inference for dynamical systems: A review. *Statistics Surveys*, 9:209–252, 2015.

[75] J. Mertz. Nonlinear microscopy: new techniques and applications. *Current opinion in neurobiology*, 14(5):610–616, 2004.

[76] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[77] M Kivanc Mihcak, Igor Kozintsev, Kannan Ramchandran, and Pierre Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6(12):300–303, 1999.

[78] Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, pages 234–243. Springer, 1978.

[79] Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization: Identifying representative elements in massive data. In *Advances in Neural Information Processing Systems*, pages 2049–2057, 2013.

[80] Paul M Muller and William L Sjogren. Mascons: Lunar mass concentrations. *Science*, 161(3842):680–684, 1968.

[81] Fredy D Neeser and James L Massey. Proper complex random processes with applications to information theory. *Information Theory, IEEE Transactions on*, 39(4):1293–1302, 1993.

[82] A. Nehorai, D. Starer, and P. Stoica. Direction-of-Arrival estimation in applications with multipath and few snapshots. *Circuits, Syst. and Signal Process.*, 10(3):327–342, 1991.

[83] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[84] Giacomo Oliveri, Matteo Carlin, and Andrea Massa. Complex-weight sparse linear array synthesis by Bayesian compressive sampling. *Antennas and Propagation, IEEE Transactions on*, 60(5):2309–2326, 2012.

[85] Alan V Oppenheim, Alan S Willsky, and Syed Hamid Nawab. *Signals and systems*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1983.

[86] James G Oxley. *Matroid theory*, volume 3. Oxford University Press, USA, 2006.

[87] P. Pal and P. P. Vaidyanathan. Nested arrays: a novel approach to array processing with enhanced degrees of freedom. *IEEE Trans. on Signal Process.*, 58(8):4167–4181, 2010.

[88] Benjamin Peherstorfer and Karen Willcox. Dynamic data-driven reduced-order models. *Computer Methods in Applied Mechanics and Engineering*, 291:21 – 41, 2015.

[89] Matthias Poloczek, Jialei Wang, and Peter I Frazier. Multi-information source optimization. *arXiv preprint arXiv:1603.00389*, 2016.

[90] Friedrich Pukelsheim. *Optimal design of experiments*, volume 50. siam, 1993.

[91] Ramesh Raskar, Amit Agrawal, and Jack Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 795–804. ACM, 2006.

[92] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[93] A Rossi, F Marzari, and P Farinella. Orbital evolution around irregular bodies. *Earth, planets and space*, 51(11):1173–1180, 1999.

[94] Z Rosson, F Hall, and T Vogel. Orbital behavior around a nonuniform celestial body. In *Journal of Physics: Conference Series*, volume 750, page 012022. IOP Publishing, 2016.

[95] Guy Satat, Matthew Tancik, and Ramesh Raskar. Lensless imaging with compressive ultrafast sensing. *arXiv preprint arXiv:1610.05834*, 2016.

[96] RJ Sault and MH Wieringa. Multi-frequency synthesis techniques in radio interferometric imaging. *Astronomy and Astrophysics Supplement Series*, 108:585–594, 1994.

[97] Dan R Scholnik and Jefrey O Coleman. Optimal design of wideband array patterns. In *Radar Conference, 2000. The Record of the IEEE 2000 International*, pages 172–177. IEEE, 2000.

[98] Cosma Rohilla Shalizi et al. Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074, 2009.

[99] Manohar Shamaiah, Siddhartha Banerjee, and Haris Vikalo. Greedy sensor selection: Leveraging submodularity. In *49th IEEE conference on decision and control (CDC)*, pages 2572–2577. IEEE, 2010.

[100] Gal Shulkind, Lior Horesh, and Haim Avron. Experimental design for non-parametric correction of misspecified dynamical models. *arXiv:1705.00956*, 2017.

[101] Gal Shulkind, Stefanie Jegelka, and Gregory W Wornell. Multiple wavelength sensing array design. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 3424–3428. IEEE, 2017.

[102] Gal Shulkind, Stefanie Jegelka, and Gregory W Wornell. Sensor array design through submodular optimization. *arXiv:1705.06616*, 2017.

[103] Gal Shulkind, Gregory W Wornell, and Yuval Kochman. Direction of arrival estimation in MIMO radar systems with nonlinear reflectors. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 3016–3020. IEEE, 2016.

[104] M Skolnik, G Nemhauser, and J Sherman. Dynamic programming applied to unequally spaced arrays. *IEEE Transactions on Antennas and Propagation*, 12(1):35–43, 1964.

[105] Christian Soize, Evangéline Capiez-Lernout, J-F Durand, C Fernandez, and L Gagliardini. Probabilistic model identification of uncertainties in computational models for dynamical systems and experimental validation. *Computer Methods in Applied Mechanics and Engineering*, 198(1):150–163, 2008.

[106] L Tenorio, C Lucero, V Ball, and L Horesh. Experimental design in the context of Tikhonov regularized inverse problems. *Statistical Modelling*, 13(5-6):481–507, 2013.

[107] Leon Thomsen. Weak elastic anisotropy. *Geophysics*, 51(10):1954–1966, 1986.

[108] Christos Thrampoulidis, Gal Shulkind, Feihu Xu, William T Freeman, Jeffrey H Shapiro, Antonio Torralba, Franco NC Wong, and Gregory W Wornell. Exploiting occlusion in Non-Line-of-Sight active imaging. *arXiv:1711.06297*, 2018. First two authors contributed equally.

[109] Andrew Tizzard, Lior Horesh, Rebecca J Yerworth, David S Holder, and RH Bayford. Generating accurate finite element meshes for the forward model of the human head in eit. *Physiological measurement*, 26(2):S251, 2005.

[110] Antonio Torralba and William T Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 374–381. IEEE, 2012.

[111] Lloyd N Trefethen. *Approximation theory and approximation practice*. Siam, 2013.

[112] Palghat P Vaidyanathan and Piya Pal. Sparse sensing with co-prime samplers and arrays. *IEEE Transactions on Signal Processing*, 59(2):573–586, 2011.

[113] Harry L Van Trees. *Detection, estimation, and modulation theory, optimum array processing*. John Wiley & Sons, 2004.

[114] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Transactions on Graphics (TOG)*, 26(3):69, 2007.

[115] Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature communications*, 3:745, 2012.

[116] A Veneziani and C Vergara. Flow rate defective boundary conditions in haemodynamics simulations. 2005.

[117] Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

[118] Jan Vondrák. Symmetry and approximability of submodular maximization problems. *SIAM Journal on Computing*, 42(1):265–304, 2013.

[119] H Wang and M Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(4):823–831, 1985.

[120] M Wax, T.J. Shan, and T Kailath. Spatio-temporal spectral analysis by eigenstructure methods. *IEEE transactions on acoustics, speech, and signal processing*, 32(4):817–827, 1984.

[121] M. Wax and I. Ziskind. On unique localization of multiple sources by passive sensor arrays. *IEEE Trans. on Acoust., Speech and Signal Process.*, 37(7):996–1000, 1989.

[122] Daming Wei, Osamu Okazaki, Kenichi Harumi, Eishi Harasawa, and Hidehiro Hosaka. Comparative simulation of excitation and body surface electrocardiogram with isotropic and anisotropic computer heart models. *IEEE Transactions on biomedical engineering*, 42(4):343–357, 1995.

[123] J Andre C Weideman. Computation of the complex error function. *SIAM Journal on Numerical Analysis*, 31(5):1497–1518, 1994.

[124] Christopher KI Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.

[125] Duncan J Wingham. The reconstruction of a band-limited function and its Fourier transform from a finite number of samples at arbitrary locations by singular value decomposition. *Signal Processing, IEEE Transactions on*, 40(3):559–570, 1992.

[126] Feihu Xu, Gal Shulkind, Christos Thrampoulidis, Jeffrey H Shapiro, Antonio Torralba, Franco N C Wong, and Gregory W Wornell. Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *arXiv:1802.03529*, 2018. First three authors contributed equally.

[127] Jui L Yen. On nonuniform sampling of bandwidth-limited signals. *Circuit Theory, IRE Transactions on*, 3(4):251–257, 1956.