ON BAYESIAN APPROACH TO EXPERIMENTAL DESIGN

Lappeenranta-Lahti University of Technology LUT

Master's Program in Computational Engineering, Master's Thesis

2022

Abdur Rahman

Examiner: Associate Professor Tapio Helin
Professor Heikki Haario

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering

Abdur Rahman

On Bayesian Approach to Experimental Design

Master's thesis
2022
44 pages
Examiners: Associate Professor Tapio Helin and Professor Heikki Haario

Keywords: Bayesian approach, optimal experimental design, A-optimality, D-optimality

Design of experiment has been widely applied in the fields of science and industry but an excellent design of experiment diverges the ratio of extracted information to invested resources, so it is necessary to obtain an optimal design. The Bayesian approach plays an important role to solve this problem, as it treats model parameters as random variables rather than constants. An optimal design can be obtained by optimizing the predicted utility of the experiment. Two design models, A- and D-optimal designs are presented in this thesis. A-optimality minimized the sum of the main diagonal elements of the information matrix and D-optimality maximized the determinant of the information matrix.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Tapio Helin for supervising, and Heikki Haario for examining my thesis. It was a great experience to work with my thesis supervisor, Tapio Helin. I need to thank Robert Rockenfeller, who gave me an opportunity to come to LUT University, Finland from Universität Koblenz.Landau, Germany. I can't forget Thomas Götz, Michael Hinze, Christian Kahle and Karunia Putra Wijaya from Universität Koblenz.Landau and Matylda Jablonska-Sabuka, Satu-Pia Reinikainen, Lassi Roininen, Lasse Lensu and Veli-Matti Taavitsainen from LUT University.

Finally, I am grateful to my beautiful wife, Mashiat Madiha, who gave me support to come to LUT University and her sincere patience for me.

Lappeenranta, August 8, 2022

Abdur Rahman

LIST OF SYMBOLS AND ABBREVIATIONS

$\langle ., . \rangle$	Inner product in a function space
$\|.\|$	Determinant of a matrix
$\perp$	Independent
μ	Mean
σ^2	Variance
Ω	Probability space
$\mathbb{C}$	Complex numbers
$\mathbb{E}$	Expected value
f_X	Probability density function
F_X	Cumulative distribution function
$\mathcal{L}(., .)$	Loss function
$\mathcal{N}$	Normal distribution
$\mathbb{N}$	Natural numbers
$\mathbb{P}$	Probability
$\mathbb{R}$	Set of real numbers
Tr	Trace
U	Utility function
DOE	Design of Experiment
GP	Gaussian Process
GPR	Gaussian Process Regression
i.i.d.	Independent and Identically Distributed
KL	Kullback-Leibler
MLE	Maximum Likelihood Estimate
OED	Optimal Experimental Design
PDF	Probability Density Function

CONTENTS

1 INTRODUCTION

1.1 Background

Design of experiment (DOE) is a statistical method has introduced first by British statistician Sir Ronald Fisher in the $1920s$ [1] for the purposes of agriculture but has been widely applied in the fields of science and industry to gain insights into physical and social circumstances [2] to support the design, development and optimization [3]. For example, machine learning algorithms need to optimize the parameter (e.g. weights) in the sense of training data. In general, we select a set of optimal hyperparameters to benchmark the process of data training to get a better outcome.

Despite the advancement of new quantitative experimental techniques, data is frequently limited and for that, the modeller is confronted with a situation in which large regions of parameter space can adequately describe the measured data [4–8]. This isn't an issue if the predictions required to test the hypothesis are well constrained [9–11]. More information will be required if this is not the case. Optimal experiment design (OED) methods can be used to determine which experiments are most useful for statistical inference [12]. Classical design criteria are frequently based on linearization around a best fit parameter set [11] and are concerned with effectively constraining the parameters [13, 14]. However, when data is scarce due to model complexity or the model is strongly non-linear, such methods are inapplicable [11, 15]. As a result, researching the role of parameter uncertainty in OED is an intriguing topic to research.

To select field experiments, the first experimental design methods relied primarily on heuristics based on concepts such as space-filling and blocking [1, 16–19]. Some of these methods work well, but they could be better if they took into account what we know about the physical processes being inferred or measured [12]. For a range of models based on ordinary differential equations [20–22], partial differential equations [23], and differential algebraic equations [24], it has been shown that physical model-guided experiment selection greatly improves the cost-effectiveness of experimental designs. The alphabetic optimality criterion is typically applied when model observables are linear with regard to model parameters [25, 26]. A-optimality, for example, is used to minimise the average variance of parameter estimates, while D-optimality is used to maximise the differential Shannon entropy [27]. These criteria were created in both Bayesian and non-Bayesian contexts [25, 28–31].

Shannon information leads to Bayesian D-optimality in the normal linear model [27], which can be used for both prediction and mixed utility functions that describe multiple concurrent goals for an experiment. By selecting appropriate utility functions, it is also possible to derive the Bayesian equivalents of some other popular optimality criteria. Some of the alphabetical optimality criteria have utility-based Bayesian counterparts, but not all. When planning an experiment, it may be thought that a prediction is more important than an inference. This could happen, for example, in quality control, where the future level of output must be maintained. The predictive Bayesian method is appropriate for both the design and analysis of these types of situations [32]. Other utility functions may be developed for the design of experiments that take into account more specific concerns. Randomization, for instance, is not required for inference in a Bayesian experiment; it is ”merely useful" [33] but randomization is an essential component of design. This problem is considered within the Bayesian optimal experimental design theory for linear models [34, 35].

OED in Bayesian approach is to optimize the design of the experiment in such a way, that we can obtain the most informative data for parameter estimation or response predictions [36, 37] to address the aims of the analysis [38]. The problem which need to be solved in experimental design optimization is to identify an experimental design includes underlying model, among a set of candidate models of interest using less steps [2, 39, 40]. Maximization (or minimization of cost) of the information extracted from the data is an expected objective of an experiment [41].

Our aim is to find an optimal experimental design x^* using utility function $f : \Omega \rightarrow \mathbb{R}$, that describes the value of executing an experiment in the space of possible design. Mathematically,

$$x^* = \underset{y \in \Omega \subset \mathbb{R}^d}{\arg\max} f(x) \tag{1}$$

where Ω is any design space of interests, often a compact subset of $\mathbb{R}^d$ and furthermore, we assume the function f has no simple closed form but can be evaluated at any arbitrary point in the domain assumed to be expensive to evaluate [40].

1.2 Objectives

The goal of the thesis is to cover the mathematical theory behind Bayesian approach in an experimental design to find the best experimental design.

Research questions:

1. What are the criteria for optimal experimental design?
2. How can we obtain an optimal experimental design on Bayesian approach?

1.3 Structure of the thesis

The thesis tells everything need to know about optimal design on Bayesian approach to find the best way to design an experiment. The theory behind Gaussian distribution, Bayesian inference, and Gaussian process regression is explained in Chapter 2. Prior distribution, posterior distribution, and Bayesian estimators discussed in Chapter 3. In Chapter 4, there are theory about how to design experiments and to obtain an optimal design. In the section on Bayesian optimal design, A- and D-optimality are discussed. In the Chapter 5, there is conclusion of this thesis and at the end, there are references.

2 GAUSSIAN RANDOM VARIABLES

A real valued Gaussian (random) process (GP) to be a random field on a parameter set, for which the (finite dimensional) distributions of all random field are multivariate Gaussian.

GP is a stochastic process (a collection of random variables indexed by some mathematical set) assuming that every random variables has a multivariate normal distribution whose mean function and covariance function are fully described [43]. GP is a joint distribution of those random variables over a continuous domain i.e. time or space. This is a natural generalization of the Gaussian distribution using a vector for the mean and a matrix for the covariance [44]. When no better information of the unknowns is available, the Gaussian distribution is a popular choice as a prior information. The fact, that uncorrelated Gaussian random variables are independent is a notable property of Gaussian random variables [45]. Another reason for the Gaussian distribution's wide use is that its properties can be changed and used in an analytical form without much effort.

2.1 Gaussian distribution

Let $X_1, X_2, ..., X_n$ be independent and identically distributed (i.i.d.) random variables.

Definition 2.1.1 (Sample spaces)**.** *A sample space Ω can be define as the set of all possible outcomes of an experiment.*

Subsets of Ω are called events.

Definition 2.1.2 (Independent events)**.** *Events A and B are called independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B),$$

and we express $A \perp B$ means A and B are independent.

We define $\mathbb{P}$ is the probability and $\mathbb{R}$ is the real number in this thesis.

Lemma 2.1.1. *For any events A and B,*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Proof. Write $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$ and the events are disjoint. $\mathbb{P}$ is additive for disjoint events, such that

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)) \\
&= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) \\
&= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) + \mathbb{P}(A \cap B) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}((A \cap B^c) \cup (A \cap B)) + \mathbb{P}((A^c \cap B) \cup (A \cap B)) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).
\end{aligned}$$

□

Here, A^c is a complement set, that contains all of the universal set's elements which are not present in the given set.

Theorem 2.1.1. *Let A_n be monotone increasing and $A = \cup_{i=1}^{\infty} A_i$. Then*

$$\mathbb{P}(A_n) \to \mathbb{P}(A)$$

as $n \to \infty$.

Proof. Define $B_1 = A_1$, $B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}$, $B_3 = \{\omega \in \Omega : \omega \in A_3, \omega \notin A_2, \omega \in A_1\}$, ... It can be shown that $B_1, B_2, ...$ are disjoint, $A_n = \cup_{i=1}^{n} A_i = \cup_{i=1}^{n} B_i$ for each n and $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$. By definition,

$$\mathbb{P}(A_n) = \mathbb{P}\Big(\cup_{i=1}^{n} B_i \Big) = \sum_{i=1}^{n} \mathbb{P}(B_i).$$

Hence,

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \mathbb{P}\Big(\cup_{i=1}^{\infty} B_i \Big) = \mathbb{P}(A).$$

□

Definition 2.1.3 (Random field)**.** *Let Ω be a parameter space, X be a stochastic process over Ω is a collection of a random variables $\{X(t) : t \in \Omega\}$. If Ω is a set of N dimension and the random variables $X(t)$ are all vector valued of dimension d, then vector valued random field X is called a (N, d) random field.*

A random variable X is a real-valued function with domain Ω, such that

$$X(t) \in \mathbb{R} = \{y : -\infty < y < +\infty\}, \quad \text{for all} \quad t \in \Omega.$$

Definition 2.1.4 (Probability density function)**.** *Consider the random variable X with probability density function (pdf) $f_X(x)$ is an integration function. We have*

- $f_X(x) \geq 0, \quad$ *for all* $\quad x \in \mathbb{R}$.
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.
- $\mathbb{P}(a < X \leq b) = f(a) - f(b) = \int_a^b f_X(x)\, dx$.
- *For a set A,* $\mathbb{P}(X \in A) = \int_A f_X(x)\, dx$.

Here, A is some intervals, a and b are some constant.

Definition 2.1.5 (Characteristic function)**.** *The characteristic function of a real valued random variable X can be defined as*

$$\phi_X(t) = \mathbb{E}(\exp(itX)), \tag{2}$$

where i is the imaginary unit i.e. $i^2 = -1$.

It can be shown that

$$\phi_{aX+b}(t) = \exp(itb)\phi_X(at). \tag{3}$$

If X has the density $f_X(x)$ then the characteristic function is it's Fourier transform, such that

$$\phi(t) = \int_{-\infty}^{\infty} \exp(itx) f_X(x)\, dx. \tag{4}$$

If $\phi(t)$ is integrable, then we get

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)\phi(t)\, dt,$$

which is the inverse Fourier transform.

Definition 2.1.6 (Location and scale parameters)**.** *Let X be a real-valued random variable, with density*

$$f_X(x|\mu, \sigma) = \frac{1}{\sigma} g\Big(\frac{x-\mu}{\sigma}\Big),$$

where g is also a density, $-\infty < \mu < \infty$ and $\sigma > 0$. Then μ and σ are called location and scale parameter.

Definition 2.1.7 (Gaussian distribution)**.** *A continuous random variable X is said to be Gaussian (or normally distributed) if it has the pdf*

$$f_X(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2) \quad \textit{for all} \quad x \in \mathbb{R}, \tag{5}$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$.

Here μ and σ^2 is the mean and variance of X respectively, and the characteristic function is given by

$$\phi_X(t) = \mathbb{E}\big\{\exp(itX)\big\} = \exp(it\mu - \sigma^2 t^2/2). \tag{6}$$

We abbreviate this by writing $X \sim \mathcal{N}(\mu,\, \sigma^2)$. When $\mu = 0$ and $\sigma = 1$, we say that X has a standard normal distribution and if a random variable has zero mean simply, we call it centered.

The pdf of a standard normal random variable X is given by,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \quad \text{for all} \quad x \in \mathbb{R}. \tag{7}$$

A normal distribution with random variable $X \sim \mathcal{N}(0,\, 1)$, that is $\mu = 0$ and $\sigma = 1$.

Univariate normal variable X can be written as

$$X = \sigma Z + \mu, \tag{8}$$

where $Z \sim \mathcal{N}(0, 1)$ is the standard random variable with the characteristic function $e^{-\frac{t^2}{2}}$.

Proposition 2.1.1. *If $f : \mathbb{R}_+ \to \mathbb{R}$ be a function, such that*

$$f(x) = f(0) + \int_0^x g(t)\, dt,\ \mathbb{E}\{|f(X)|\} < \infty \quad \textit{and} \quad X \geq 0,$$

then

$$\mathbb{E}\{f(X)\} = f(0) + \int_0^\infty g(t)\mathbb{P}(X \geq t)\, dt. \tag{9}$$

If $g \geq 0$ and right hand side of equation (9) is finite, then $\mathbb{E}\{.\} < \infty$.

Proof. Using Fubini's theorem, for $X \geq 0$

$$\begin{aligned}\int_{\Omega} f(X)\, d\mathbb{P} &= \int_{\Omega} \Big(f(0) + \int_0^{\infty} 1_{t \leq X} g(t)\, dt \Big) d\mathbb{P} \\ &= f(0) + \int_0^{\infty} g(t) \Big(\int_{\Omega} 1_{t \leq X}\, d\mathbb{P} \Big)\, dt \\ &= f(0) + \int_0^{\infty} g(t) \mathbb{P}(X \geq t)\, dt.\end{aligned}$$

□

Corollary 2.1.1. *If $\mathbb{E}(|X|^r) < \infty$ for an integer $r > 0$, then*

$$\mathbb{E}(X^r) = r \int_0^{\infty} t^{r-1} \mathbb{P}(X \geq t)\, dt - r \int_0^{\infty} t^{r-1} \mathbb{P}(-X \geq t)\, dt. \tag{10}$$

If $\mathbb{E}(|X|^r) < \infty$ for real $r > 0$, then

$$\mathbb{E}(|X|^r) = r \int_0^{\infty} t^{r-1} \mathbb{P}(|X| \geq t)\, dt. \tag{11}$$

Left hand side of equation (11) is finite if and only if the right hand side is also finite.

Proof. Equation (10) follows from proposition (2.1.1) with $f(x) = x^r$ and $g(t) = \frac{d}{dt} f(t) = rt^{r-1}$. Since $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$, where $X^+ = \max\{X, 0\}$ and $X^- = \min\{X, 0\}$, now applying proposition (2.1.1) separately we will get corollary (2.1.1). □

Theorem 2.1.2. *If X_1 and X_2 are two independent random variables such that $X_1 + X_2$ follows normal distribution, then each of the variable X_1 and X_2 is normal.*

Proof. We assume that $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$ without the loss of generality. We know that $\mathbb{E}(aX_j^2) < \infty$, $j = 1, 2$ for any constant a and therefore, the characteristic functions $\phi_1(.)$ and $\phi_2(.)$ are analytic. By the uniqueness of the analytic extension, $\phi_1(s)\phi_2(s) = \exp(-s^2/2)$, for all $s \in \mathbb{C}$, set of complex numbers. Thus $\phi_j(z) \neq 0$ for all $z \in \mathbb{C}, j = 1, 2$. Thus both characteristic function correspond to normal distribution. □

Definition 2.1.8 (Covariance matrices)**.** *Let X and Y be two random variables. The covariance of X and Y is defined as $cov(X,Y) = \mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$ assuming that these expectations exist.*

The variance of X is $varX = cov(X,X)$ and the expectation is linear such that $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$. The covariance is symmetric and bilinear such that $cov(aX + bY, Z) = acov(X,Z) + bcov(Y,Z)$.

Definition 2.1.9 (Positive semi-definite)**.** *A matrix (covariance matrix) $C_{d\times d}$ is called positive semi-definite (positive definite) if $x^\top Cx \geq 0 (> 0)$ for all $x \in \mathbb{R}^d$.*

A function $C : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called positive definite if the matrices $(C(t_i,t_j))_{i,j=1}^n$ are positive definite for all $1 \leq n < \infty$ and all $(t_1,...,t_n) \in \mathbb{R}^n$.

Note that, a (vector valued) random variable X is symmetric if X and $-X$ has the same distribution.

Lemma 2.1.2. *For every matrix A, vector b, and random vector X, we have*

1. $\mathbb{E}(AX + b) = A\mathbb{E}(X) + b$,
2. $Cov(AX) = A(CovX)A^T$,
3. $CovX$ *is symmetric and positive definite,*
4. $\mathbb{P}(X \in \mathbb{E}\{.\} + range(CovX)) = 1$.

Lemma 2.1.3. *Each bilinear form of C has the dot product representation*

$$C(x,y) = \mathcal{C}\langle x,y\rangle,$$

where $\mathcal{C}$ is a linear mapping, represented by a $d \times d$ matrix, $\mathcal{C} = [c_{i,j}]$. Furthermore, we have $\mathcal{C} = \mathcal{C}^\top$.

Definition 2.1.10 (Gaussian random variable)**.** *A real-valued random variable X is called a Gaussian random variable, if it has characteristic function*

$$\mathbb{E}(e^{i\lambda X}) = \exp\left(im\lambda - \frac{\sigma^2\lambda^2}{2}\right).$$

Here m and σ are some real numbers. From the above characteristic function, using differentiation with respect to λ and setting $\lambda = 0$, we get

$$\mathbb{E}(X) = m \quad and \quad Var(X) = \sigma^2.$$

An $\mathbb{R}^n$ valued random variable ξ is a Gaussian random variable if (y, ξ) is a real valued Gaussian random variable for each $y \in \mathbb{R}^n$, that it has characteristic function

$$\phi_\xi(y) = \mathbb{E}(\exp(i(y,\xi))) = \exp\Big(i\mathbb{E}(y,\xi) - \frac{Var(y,\xi)}{2}\Big), \tag{12}$$

for each $y \in \mathbb{R}^n$. Setting $m = (m_1, ..., m_n)$, $\mathbb{E}(\xi_j) = m_j$ and $\mathbb{E}(\xi_j - m_j)(\xi_k - m_k) = \Sigma_{j,k}$, we can rewrite equation (12) as

$$\phi_\xi(y) = \exp\Big(imy - \frac{y^\top \Sigma y}{2}\Big), \tag{13}$$

where $\Sigma = \{\Sigma_{j,k}\}_{j,k=1}^n$ is a symmetric $n \times n$ matrix with real components. Here m and Σ are the mean and covariance matrix of ξ and the rank of Σ is the dimension of the subspace of $\mathbb{R}^n$.

Lemma 2.1.4. *Let ξ be an $\mathbb{R}^n$ valued Gaussian random variable with mean vector m and assume that*

$$\mathbb{E}((\xi_j - m_j)(\xi_k - m_k)) = 0, \quad j \neq k.$$

Then $\xi_1, ..., \xi_n$ are independent.

Definition 2.1.11 (Multivariate gaussian)**.** *A real-valued random variable X is said to be multivariate Gaussian if, for all $a = (a_1, a_2, ..., a_d) \in \mathbb{R}^d$, the real valued variable $\langle a, X\rangle = \sum_{i=1}^d a_i X_i$ is Gaussian. Here, $\mu \in \mathbb{R}^d$ with $\mu_j = \mathbb{E}\{X_j\}$ is a mean vector and C is a positive semi-definite $d \times d$ covariance matrix with elements $c_{i,j} = \mathbb{E}\{(X_i - \mu_i)(X_j - \mu_j)\}$, such that the probability density of X is given by*

$$\varphi(x) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^\top C^{-1}(x-\mu)), \tag{14}$$

where $|C| = \det C$ is the determinant of C. We write this as $X \sim \mathcal{N}_d(\mu, C)$ with dimension d.

Corollary 2.1.2. *Let X and Y has a (joint) normal distribution on $\mathbb{R}^{d_1+d_2}$ and $||.||_X, ||.||_Y$ are $\langle ., .\rangle$-orthogonal (X and Y are uncorrelated), then X, Y are independent.*

Here $||.||$ is some space.

Theorem 2.1.3. *Let X, Y be a real valued random variables have the joint characteristics function $\phi(t, s)$. Let $\mathbb{E}(|X|^m) < \infty$ for some $m \in \mathbb{N}$ and $g(y)$ be such that $g(Y) =$*

$\mathbb{E}\{X^m|Y\}$. *Then for all real* s

$$(-i)^m \frac{\partial^m}{\partial t^m}\phi(t,s)\Big|_{t=0} = \mathbb{E}(g(Y)\exp(isY)). \tag{15}$$

If $g(y) = \sum c_k y^k$ *is a polynomial, then*

$$(-i)^m \frac{\partial^m}{\partial t^m}\phi(t,s)\Big|_{t=0} = \sum_k (-i)^k c_k \frac{d^k}{ds^k}\phi(0,s). \tag{16}$$

Proof. $\mathbb{E}(|X|^m) < \infty$, then the joint characteristic function $\phi(t,s) = \mathbb{E}(\exp(itX + isY))$ can be differentiated m times with respect to t and

$$\frac{\partial^m}{\partial t^m}\phi(t,s) = i^m \mathbb{E}(X^m \exp(itX + isY)).$$

Putting $t = 0$ gives us equation (15). To prove equation (16), we need to prove $\mathbb{E}(|Y|^r) < \infty$, where r is the degree of the polynomial $g(y)$. Using Jensen's inequality, we get $\mathbb{E}(|g(Y)|) \leq \mathbb{E}(|X|^m) < \infty$, and since $|g(y)/y^r| \to constant \neq 0$ as $|y| \to \infty$, therefore there is $constant > 0$ such that $|y|^r \leq constant\,|g(y)|$, for all y. Therefor, $\mathbb{E}(|Y|^r) < \infty$. □

Definition 2.1.12 (Maximum likelihood estimate)**.** *The maximum likelihood estimate (MLE)* $\hat{\theta}$ *is a value of* θ*, where the likelihood function* $\mathcal{L}(\theta) = f(x|\theta)$ *attains it's supremum, i.e.,*

$$\sup_\theta f(x|\theta) = f(x|\hat{\theta}).$$

The MLE can be obtain to solve the likelihood equation

$$\frac{\partial}{\partial \theta_j} \log\, f(x|\theta) = 0,\ j = 1, ..., n,$$

if f is differentiable with respect to θ and $f > 0$.

Theorem 2.1.4. *Given a random sample of size* n *from a normal distribution with mean* μ *and variance* Σ*, the log-likelihood is given by*

$$\mathcal{L}(\mu, \Sigma) = -\frac{n}{2}(\overline{X} - \mu)^\top \Sigma^{-1}(\overline{X} - \mu) - \frac{n}{2}\operatorname{Tr}(\Sigma^{-1}S) - \frac{n}{2}\log|\Sigma|.$$

The MLE is,

$$\hat{\mu} = \overline{X} \quad \textit{and} \quad \hat{\Sigma} = \Big(\frac{n-1}{n}\Big)S.$$

Here, S is a variance-covariance matrix.

Proof. By taking logarithm in the likelihood function of normal distribution we get log-likelihood, differentiate depending on the parameter of μ and Σ we get the MLE. □

Lemma 2.1.5. *The random vector $X \sim \mathcal{N}_d(\mu, \Sigma)$, if and only if for every $a \in \mathbb{R}^d$, $a^\top X$ has distribution $\mathcal{N}_1(a^\top \mu, a^\top \Sigma a)$.*

Proof. Since $X \sim \mathcal{N}_d(\mu, \Sigma)$, the parameters $a^\top \mu$ and $a^\top \Sigma a$ are the expectation and covariance of the variable $a^\top X$. So, it is enough to show that $a^\top X$ is normally distributed. Since the distribution of X and $\mu + LZ$ are same, then the variable $a^\top X$ has the same distribution as $a^\top \mu + (L^\top a)^\top Z$. The second term is a constant as well as a linear combination of independent variables, $b^\top Z$ ($b = L^\top a$) and normally distributed with zero mean and unique variance.

Again, let $a^\top X \sim \mathcal{N}(a^\top, a^\top \Sigma a)$, then $a^\top X$ and $a^\top Y$ has the same distribution for an $\mathcal{N}_d(\mu, \Sigma)$-distributed vector Y. If this is true for every a, then X and Y also have the same distribution; hence $X \sim \mathcal{N}_d(\mu, \Sigma)$. □

Corollary 2.1.3. *If the vector $X = (X_1, ..., X_d)$ has the distribution $\mathcal{N}_d(\mu, \Sigma)$ and $A : \mathbb{R}^d \to \mathbb{R}^m$ is an arbitrary matrix, then AX has distribution $\mathcal{N}_m(A\mu, A\Sigma A^\top)$.*

Proof. $A\mu$ and $A\Sigma A^\top$ are the expectation and covariance matrix of AX. It's enough to prove, that AX is normally distributed. We have $a^\top(AX) = (A^\top a)^\top X$. According to lemma (2.1.5), this variable has a one-dimensional normal distribution, which implies that, AX has a multivariate normal distribution. □

Above lemma and corollary imply that the marginal distributions of a multivariate normal distribution are normally distributed [46].

Theorem 2.1.5. *Let a be a vector and X be a random vector of the same length d with mean μ and variance Σ. Then $\mathbb{E}(a^\top X) = a^\top \mu$ and $Var(a^\top X) = a^\top \Sigma a$. If A is a matrix with d columns, then $\mathbb{E}(AX) = A\mu$ and $Var(AX) = A\Sigma A^\top$.*

Proof.

$$\begin{aligned}\mathbb{E}(a^\top X) &= \mathbb{E}(a^\top \frac{1}{d}\sum_{i=1}^{d} X_i) = a^\top \frac{1}{d}\mathbb{E}(\sum_{i=1}^{d} X_i) \\ &= a^\top \mu.\end{aligned}$$

Again

$$\begin{aligned}Var(a^\top X) &= \mathbb{E}(\sum_{i=1}^{d}(a^\top X_i - \mu)^2) = a^\top \mathbb{E}(\sum_{i=1}^{d}(X_i - \mu)^2)a \\ &= a^\top \Sigma a.\end{aligned}$$

□

Lemma 2.1.6. *The vector $X = (X_1, ..., X_d)$ has a multivariate normal distribution with a diagonal matrix Σ, if and only if $X_1, ..., X_d$ are independent and has normal marginal distributions.*

Proof. A symmetric, positive definite diagonal matrix Σ can be written as $\Sigma = LL^\top$ for L the diagonal matrix with entries the square roots of the diagonal entries of Σ. By definition, X has distribution $\mathcal{N}_d(\mu, \Sigma)$ if it has the same distribution as $\mu + LZ = (\mu_1 + L_{11}Z_1, ..., \mu_d + L_{dd}Z_d)$ for independent standard normal variables $Z_1, ..., Z_d$. Hence $X = (X_1, ..., X_d)$ are independent and normally distributed. □

Lemma 2.1.7. *Let ξ be a real valued Gaussian random variable with zero mean and variance σ^2. Then, for all $a > 0$,*

$$\mathbb{P}(|\xi| > a) \leq \exp\Big((-\frac{a^2}{2\sigma^2})\Big), \tag{17}$$

and $a/\sigma \geq 1$,

$$(\sigma/a)\phi(a/\sigma) \leq \mathbb{P}(|\xi| > a) \leq 2(\sigma/a)\phi(a/\sigma). \tag{18}$$

Proof. We divide $|\xi|$ using σ and when $\sigma = 1$ we get,

$$\begin{aligned}\mathbb{P}(|\xi| > a) &= \frac{2}{\sqrt{2\pi}} \int_a^\infty \exp(-u^2/2)\, du \\ &\leq \frac{2}{a\sqrt{2\pi}} \int_a^\infty u \exp(-u^2/2)\, du \\ &= \frac{2}{a\sqrt{2\pi}} \exp(-a^2/2)\end{aligned}$$

Also, using derivatives, $\exp(-a^2/2) - \mathbb{P}(|\xi| > a)$ is increasing for $a < 2/\sqrt{2\pi}$. Now,

$$\mathbb{P}(|\xi| > a) = \frac{2}{\sqrt{2\pi}} \int_{a/\sigma}^\infty \exp(-u^2/2)\, du \leq \frac{2\sigma}{\sqrt{2\pi}a} \int_{a/\sigma}^\infty u \exp(-u^2/2)\, du.$$

□

Definition 2.1.13 (Gaussian process)**.** *A real-valued stochastic process $\{X(t), t \in T\}$ (T is some index set) is a GP, if it's finite dimensional distribution are Gaussian which is characterized by it's mean function m and it's covariance kernel Σ, given by*

$$m(t) = \mathbb{E}(X(t)) \quad \textit{and} \quad \Sigma(s,t) = \mathbb{E}(X(t) - m(t))(X(s) - m(s)).$$

A GP is defined as a distribution across functions, according to one definition [44]. We can sample a function at the point x from a GP that has been completely described by a mean and covariance function according to,

$$f(x) \sim \mathcal{GP}(m, k),$$

where $f(.)$ is a covariance function, which is a subclass of kernel functions, and $m(.)$ is the function we sample from the GP.

The GP approach to nonparametric regression uses a Gaussian stochastic process prior to perform Bayesian inference directly on the space of functions f.

Gaussian process regression (GPR) is a kernel technique, but being derived in a completely different way [47, 48]. The GPR model commonly known as the Kriging model, is a Bayesian nonparametric approach that implements GP for regression analysis [49]. The fact, that functions can be readily defined by a mean function $m(x)$ and a covariance function $k(x, x')$ is a major advantage of utilizing the Gaussian prior assumption [50],

such that

$$m(x) = \mathbb{E}(f(x)) \quad and \quad k(x, x') = \mathbb{E}((f(x) - m(x))(f(x') - m(x'))).$$

3 BAYESIAN INFERENCE

Thomas Bayes is the father of Bayesian inference but the credit goes to Pierre-Simon Laplace [51] for driving the formula that we use now-a-days [52]. The primary difference between Bayesian inference and classical inference (or frequentist inference) is that it treats model parameters as random variables rather than constants [53]. Prior information can be explicitly considered using the Bayesian framework (or paradigm). It can also be used to create a complex statistical model that is difficult to solve using traditional approaches. One disadvantage of Bayesian inference is that it always requires a prior distribution to be defined, even in the absence of any prior knowledge.

However, suitable uninformative prior distributions [54] have been constructed to overcome this issue, and in many circumstances, a good feature of Bayesian inference is that these priors lead to exactly the same point and interval estimates as classical inference. When there is at least a considerable amount of data available, the issue becomes even less pressing. The Bayesian technique often converges to the same inferential results as sample size grows, regardless of the stated prior distribution.

If the likelihood is Gaussian, for example, using a Gaussian distribution as a prior is advantageous since the product of two Gaussian probability density distributions is also Gaussian. In that situation, calculating the posterior covariance is also straightforward. Furthermore, a Gaussian distribution's conditional mean and maximum a posteriori values are the same.

3.1 Prior distribution

The prior distribution is a crucial component of Bayesian inference, representing knowledge about an uncertain parameter θ, that is coupled with the probability distribution of new data to produce the posterior distribution, which is utilized for future inferences and decisions concerning θ [54, 55]. Axioms of decision theory can be used to justify the presence of a prior distribution for any problem.

Definition 3.1.1 (Conditional probability)**.** *If $\mathbb{P}(B) > 0$, then the conditional probability of A given B is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{19}$$

When B is fixed, $\mathbb{P}(B) > 0$, $\mathbb{P}(.|B)$ is a probability. The probability $\mathbb{P}(A \cap B)$ is the joint

probability of A and B.

Lemma 3.1.1. *Let A and B are two independent events, then $\mathbb{P}(A|B) = \mathbb{P}(A)$. When A and B are any pair events, then*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

Definition 3.1.2 (Prior distribution)**.** *A prior is expressed as a probability distribution and can be determined in a variety of ways (e.g., previous information, subjective evaluation, maximization of entropy under constraints), and is typically combined with the likelihood function using Bayes' theorem to obtain a posterior distribution.*

3.2 Posterior distribution

Posterior distribution holds all the information related to the unknown parameter θ [56] after the observation of data X.

Definition 3.2.1 (Posterior distribution)**.** *Let $X = x$ be the observed realization of a random variable X with density function $f(x|\theta)$ (from now we will use density f(.) instead of $f_X(.)$). Prior distribution with density function $f(\theta)$ allows to compute the density function $f(\theta|x)$ of the posterior distribution using Bayes theorem*

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)\,d\theta}. \tag{20}$$

For discrete parameters θ, the integral in the denominator has to be replaced with a sum.

The term $f(x|\theta)$ is the likelihood function $\mathcal{L}(.)$. The denominator can be rewrite as

$$\int f(x|\theta)f(\theta)\,d\theta = \int f(x,\theta)\,d\theta = f(x),$$

that is it does not depend on θ.

Definition 3.2.2 (Bayesian point estimates)**.** *The posterior mean $\mathbb{E}(\theta|x)$ is the expectation of the posterior distribution, such that*

$$\mathbb{E}(\theta|x) = \int \theta f(\theta|x)\,d\theta.$$

The posterior mode is the mode of the posterior distribution, such that

$$Mod(\theta|x) = \arg\max_{\theta} f(\theta|x).$$

The posterior median is the median of the posterior distribution, i.e. any number a that satisfies

$$\int_{-\infty}^{a} f(\theta|x)\, d\theta = 0.5 \; and \; \int_{a}^{\infty} f(\theta|x)\, d\theta = 0.5. \tag{21}$$

3.3 Bayesian estimators

Using the Bayes theorem, the posterior density can be given

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{m(x)}, \tag{22}$$

where $m(x) = \int f(x|\theta)f(\theta)\, d\theta$ is the marginal distribution of X. Then, we can define the posterior risk of an estimator $\hat{\theta}(x)$ by

$$r(\hat{\theta}|x) = \int \mathcal{L}(\theta, \hat{\theta}(x))f(\theta|x)\, d\theta, \tag{23}$$

where $\mathcal{L}(.,.)$ is the loss function.

Definition 3.3.1 (Risk)**.** *The risk of an estimator $\hat{\theta}$, is*

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta}\Big(\mathcal{L}(\theta, \hat{\theta})\Big) = \int \mathcal{L}(\theta, \hat{\theta}(x))f(x|\theta)\, dx.$$

Definition 3.3.2 (Bayes risk)**.** *The maximum risk is given by*

$$\overline{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta}),$$

then the Bayes risk is defined as

$$r(f, \hat{\theta}) = \int R(\theta, \hat{\theta})f(\theta)\, d\theta.$$

Here, $f(\theta)$ is a prior for θ.

Definition 3.3.3 (Bayes rule)**.** *An estimator $\hat{\theta}$ is called a Bayes rule with respect to the*

prior f, if

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta}).$$

Here the infimum is over all estimators of $\tilde{\theta}$.

Theorem 3.3.1. *The Bayes risk $r(f, \hat{\theta})$ satisfies*

$$r(f, \hat{\theta}) = \int r(\hat{\theta}|x) m(x)\, d\theta,$$

where $m(x) = \int f(x|\theta) f(\theta)\, d\theta$ and $r(\hat{\theta}|x)$ is the posterior risk. If $\hat{\theta}(x)$ is the value of θ that minimizes $r(\hat{\theta}|x)$, then $\hat{\theta}$ is the Bayes estimator.

Proof. The Bayes risk can be rewrite as follows

$$\begin{aligned}
r(f, \hat{\theta}) &= \int R(\theta, \hat{\theta}) f(\theta)\, d\theta \\
&= \int \Big(\int \mathcal{L}(\theta, \hat{\theta}(x)) f(x|\theta)\, dx \Big) f(\theta)\, d\theta \\
&= \int \int \mathcal{L}(\theta, \hat{\theta}(x)) f(x, \theta)\, dx d\theta \\
&= \int \int \mathcal{L}(\theta, \hat{\theta}(x)) f(\theta|x) m(x)\, dx d\theta \\
&= \int \Big(\int \mathcal{L}(\theta, \hat{\theta}(x)) f(\theta|x)\, d\theta \Big) m(x)\, dx \\
&= \int r(\hat{\theta}|x) m(x)\, dx.
\end{aligned}$$

If $\hat{\theta}(x)$ be the value of θ, which minimizes $r(\hat{\theta}|x)$, then we will minimize the integrand at every x, which minimize the integral $\int r(\hat{\theta}|x) m(x)\, dx$. □

Theorem 3.3.2. *If $\mathcal{L}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then the Bayes estimator is*

$$\hat{\theta}(x) = \int \theta f(\theta|x)\, d\theta = \mathbb{E}(\theta|X = x). \tag{24}$$

When $\mathcal{L}(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, then the Bayes estimator is the median of the posterior $f(\theta|x)$ and $\mathcal{L}(\theta, \hat{\theta})$ is zero-one loss, then the Bayes estimator is the mode of the posterior $f(\theta|x)$.

Here, zero-one loss literally counts the number of errors made by a hypothesis function on the training set.

Proof. We will consider the theorem in the context of squared error loss. The Bayes rule $\hat{\theta}(x)$ minimizes $r(\hat{\theta}|x) = \int(\theta - \hat{\theta}(x))^2 f(\theta|x)\, d\theta$. The derivative of $r(\hat{\theta}|x)$ with respect to $\hat{\theta}(x)$ and setting it equal to zero gives the equation, $\int(\theta - \hat{\theta}(x)) f(\theta|x)\, d\theta = 0$. Solving for $\hat{\theta}(x)$ we will get the estimate. □

Definition 3.3.4 (Minimax rule)**.** *An estimator $\hat{\theta}$ is called minimax, if*

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta}),$$

where the infimum is over all estimators $\tilde{\theta}$ and the supremum is over all admissible values of θ.

Theorem 3.3.3. *Let $\hat{\theta}^f$ be the Bayes rule for some prior f, such that*

$$r(f, \hat{\theta}^f) = \inf_{\hat{\theta}} r(f, \hat{\theta}). \tag{25}$$

Let

$$R(\theta, \hat{\theta}^f) \leq r(f, \hat{\theta}^f) \quad \textit{for all} \quad \theta. \tag{26}$$

Then $\hat{\theta}^f$ is called minimax and f is called a least favorable prior.

Proof. Let $\hat{\theta}^f$ is not minimax. Then there exist another rule $\hat{\theta}_0$, such that $\sup_{\theta} R(\theta, \hat{\theta}_0) < \sup_{\theta} R(\theta, \hat{\theta}^f)$. We know that the average of a function is always less or equal to it's maximum. We have $r(f, \hat{\theta}_0) \leq \sup_{\theta} R(\theta, \hat{\theta}_0)$. Hence

$$r(f, \hat{\theta}_0) \leq \sup_{\theta} R(\theta, \hat{\theta}_0) < \sup_{\theta} R(\theta, \hat{\theta}^f) \leq r(f, \hat{\theta}^f),$$

which is contradicts. □

Theorem 3.3.4. *Let $\hat{\theta}$ is the Bayes rule with respect to some prior f. Suppose $\hat{\theta}$ has constant risk, such that $R(\theta, \hat{\theta}) = constant$. Then $\hat{\theta}$ is minimax.*

Proof. The Bayes risk is, $r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta)\, d\theta = constant$. So, $r(\theta, \hat{\theta}) \leq r(f, \hat{\theta})$ for all θ. Using theorem (3.3.3), we can get the proof. □

From above we can say, that with a constant risk function, the Bayes estimators are minimax [57].

Definition 3.3.5 (Admissible)**.** *An estimator $\hat{\theta}$ is called inadmissible if there exists another rule $\hat{\theta}'$, such that*

$$R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta}) \quad \textit{for all} \quad \theta$$

and

$$R(\theta, \hat{\theta}') < R(\theta, \hat{\theta}) \quad \textit{for at least one} \quad \theta.$$

Otherwise, $\hat{\theta}$ is called admissible.

Theorem 3.3.5. *If X is a random variable follows normal distribution with mean θ and variance σ^2, then $aX + b$ is inadmissible as an estimator of θ for squared error loss if*
(a) a < 1
(b) a > 0
(c) a = 1, b $\neq$ 0.

Proof. For any a and b, the risk of the rule $aX + b$ is given by

$$R(\theta, aX + b) = a^2\sigma^2 + \{(a-1)\theta + b\}^2 \equiv \rho(a, b).$$

(a) If $a > 1$

$$\rho(a, b) \geq a^2\sigma^2 > \sigma^2 = \rho(1, 0),$$

so, $aX + b$ is dominated by X.
(b) If $a < 0$, then $(a-1)^2 > 1$ and

$$\begin{aligned}\rho(a, b) &\geq \{(a-1)\theta + b\}^2 = (a-1)^2\{\theta + \frac{b}{a-1}\}^2 \\ &> \{\theta + \frac{b}{a-1}\}^2 = \rho(0, \frac{-b}{a-1}).\end{aligned}$$

(c) If $a = 1, b \neq 0$,

$$\rho(1, b) = \sigma^2 + b^2 > \sigma^2 = \rho(1, 0),$$

so, $X + b$ is dominated by X.

□

Theorem 3.3.6. *Suppose that, $R(\theta, \hat{\theta})$ is a continuous function of θ for every $\hat{\theta}$. Let f be a prior density with full support, meaning that, for every θ and every $\epsilon > 0$, $\int_{\theta-\epsilon}^{\theta+\epsilon} f(\theta)\, d\theta > 0$. Let $\hat{\theta}^f$ be the Bayes rule. If the Bayes risk is finite then $\hat{\theta}^f$ is admissible.*

Proof. If $\hat{\theta}^f$ is inadmissible, then there exists a better rule $\hat{\theta}$, such that $R(\theta, \hat{\theta}) \leq R(\theta, \hat{\theta}^f)$ for all θ and $R(\theta_0, \hat{\theta}) < R(\theta_0, \hat{\theta}^f)$ for some θ_0. Let $v = R(\theta_0, \hat{\theta}^f) - R(\theta_0, \hat{\theta}) > 0$. R is continuous, so there exists some constant $\epsilon > 0$ and then $R(\theta, \hat{\theta}^f) - R(\theta, \hat{\theta}) > v/2$ for all $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Then,

$$\begin{aligned} r(f, \hat{\theta}^f) - r(f, \hat{\theta}) &= \int R(\theta, \hat{\theta^f}) f(\theta)\, d\theta - \int R(\theta, \hat{\theta}) f(\theta)\, d\theta \\ &= \int \Big[R(\theta, \hat{\theta}^f) - R(\theta, \hat{\theta}) \Big] f(\theta)\, d\theta \\ &\geq \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} \Big[R(\theta, \hat{\theta}^f) - R(\theta, \hat{\theta}) \Big] f(\theta)\, d\theta \\ &\geq \frac{v}{2} \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} f(\theta)\, d\theta \\ &> 0. \end{aligned}$$

Hence $r(f, \hat{\theta}^f) > r(f, \hat{\theta})$. This is contradiction. □

Theorem 3.3.7. *Let $\hat{\theta}$ has constant risk and is admissible. Then it is minimax.*

Proof. The risk $R(\theta, \hat{\theta})$ is $constant$. Let $\hat{\theta}$ is not minimax, so there exists a rule $\hat{\theta}'$ such that

$$R(\theta, \hat{\theta}') \leq \sup_{\theta} R(\theta, \hat{\theta}') < \sup_{\theta} R(\theta, \hat{\theta}) = constant.$$

Then, $\hat{\theta}$ is inadmissible. □

Theorem 3.3.8. *Let $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$. Then, under squared error loss, $\hat{\theta} = \overline{X}$ is minimax.*

Proof. We know, that under squared error loss, $\hat{\theta}$ is admissible. The risk of $\hat{\theta} = 1/n$ is a constant. The proof can be draw from theorem (3.3.7). □

4 EXPERIMENTAL DESIGN

The core of this thesis is the theory related to obtain an OED. When it comes to experimental design, every part of the experiment is planned out in advance [58]. To get the most out of the given resources, it is critical to plan the optimal experiments based on recognized restrictions. It is also critical to plan the experiments such that they are centered around the goal.

4.1 Optimal design

The main purpose of optimal design is to choose a design d^* from a set of possible designs $\mathcal{D}$, which produces the best estimate of the desired parameters [59–61]. The optimal design is contingent upon how the term "best" is defined.

Let d be a design and Y be the vector of observations obtain under design d and X is $n \times k$ design matrix with known entries specified by design d. Let introduce noise $\epsilon \sim \mathcal{N}(\mu, \sigma)$, then

$$\mathbb{E}(Y + \epsilon) = X\theta,\ Cov(Y + \epsilon) = \sigma^2 I_n, \tag{27}$$

where θ is a $k \times 1$ vector of unknown parameter and I_n is the identity matrix of order n. Here, $\mathbb{E}(\epsilon) = 0$. Note that $C_d = X^\top X$ is an information matrix for the design d.

Definition 4.1.1 (Uniformly optimal)**.** *A design d^* is said to be uniformly optimal among a class $\mathcal{D}$ of designs if for any design $d \in \mathcal{D}$, $C_{d^*} - C_d$ is non-negative.*

Now we will introduce some criteria [60, 62, 63] for optimality.

Definition 4.1.2 (Φ_p-criteria)**.** *The Φ_p-criteria are defined as*

$$\Phi_p(C_d) = \left[\frac{1}{k}\operatorname{Tr}(C_d^{-p})\right]^{\frac{1}{p}} = \left[\frac{1}{k}\sum_{i=1}^{k}\lambda_i^{-p}(C_d)\right]^{\frac{1}{p}},\ 0 < p < \infty.$$

$$\Phi_0(C_d) = \lim_{p\to 0}\Phi_p(C_d) = \prod_{i=1}^{k}\lambda_i^{-1/k},$$

$$\Phi_{+\infty}(C_d) = \lim_{p\to +\infty}\Phi_p(C_d) = \max\lambda_i^{-1}(C_d),$$

$$\Phi_{-\infty}(C_d) = \lim_{p\to -\infty}\Phi_p(C_d) = \min\lambda_i^{-1}(C_d).$$

Definition 4.1.3 (Φ_p-optimal)**.** *A design d^* is said to be Φ_p-optimal if it holds that*

$$\Phi_p(C_{d^*}) \leq \Phi_p(C_d),$$

for all $d \in \mathcal{D}$.

We need to clarify the optimality criteria [63, 64] before to find the optimal design. Majorization and Schur convexity [65] are important ideas for figuring out how two or more criteria relate to each other.

Definition 4.1.4 (Convex)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it's domain is a convex set for all x, y in it's domain, and all $\lambda \in [0, 1]$, we have*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

If we take any two points x and y, then f evaluated at any convex combination of these two points should be not larger than the same convex combination of $f(x)$ and $f(y)$.

Corollary 4.1.1. *Consider an unconstrained optimization problem*

$$\min_{x \in \mathbb{R}^n} f(x)$$

where f is convex and differentiable. Then, any point $\overline{x}$ that satisfies $\nabla f(\overline{x}) = 0$ is a global minimum.

Proof. Using first order characterization of convexity, we have

$$f(y) \geq f(x) + \nabla f^\top(x)(y - x), \quad for\, all \quad (x, y)$$

Particularly,

$$f(y) \geq f(\overline{x}) + \nabla f^\top(\overline{x})(y - x), \quad for\, all \quad y.$$

Since $\nabla f(\overline{x}) = 0$, we get

$$f(y) \geq f(\overline{x}), \quad for\, all \quad y.$$

□

Now we introduce MS-optimality [66] depending on the Φ_p-criteria.

Definition 4.1.5 (MS-optimal)**.** *A design d^* of a class of designs $\mathcal{D}$ is called MS-optimal if it minimizes Φ_{-1} and if it maximizes Φ_{-2} among the subclass of designs minimizing Φ_{-1}.*

There are some situations where a design is better than others in terms of multiple criteria, not just one. Using this as a starting point, Kiefer [67] proposes the concept of "universal optimality".

Definition 4.1.6 (Universally optimal)**.** *A design $d^* \in \mathcal{D}$ is universally optimal design if d^* is Φ-optimal for all the criteria $\Phi(C)$ from C to $]-\infty, +\infty]$ satisfying*
(a) Φ is invariant under each permutation of rows and columns,
(b) $\Phi(aC)$ is non-increasing in the scalar $a > 0$,
(c) Φ is convex.

Let P_σ denote the (d, d)-matrix that permutes the components of a vector according to the permutation σ lying in S_t, where S_t is the symmetric group on $\{1, ..., d\}$.

Definition 4.1.7 (Symmetric)**.** *A function ϕ on $\mathbb{R}^d$ is symmetric if for all $x \in \mathbb{R}^d$ and for all $\sigma \in S_d$, $\phi(P_\sigma x) = \phi(x)$.*

Proposition 4.1.1. *A design d^* is universally optimal among a class $\mathcal{D}$ of designs if it satisfies*
(i) $\mathrm{Tr}\ C_{d^*} = \max_{d\in\mathcal{D}} \mathrm{Tr}\ C_d$,
(ii) for all $d \in \mathcal{D}$, there exist scalers $a_{d\sigma} \geq 0$ such that

$$C_{d^*} = \sum_{\sigma\in S_t} a_{d\sigma} P_\sigma C_d P_\sigma^\top.$$

Proof. Condition (i) is necessary as $C \to -\mathrm{Tr}\ C$ satisfies condition (a), (b) and (c) in definition 4.1.6. Let d^* be a universally optimal design, and further assume that there exists a design d_1 such that

$$C_{d^*} = \sum_{\sigma\in S_t} P_\sigma C_{d_1} {P_\sigma}^\top.$$

Let $\mathcal{A}$ be the convex cone generated by the matrices $\{P_\sigma\, C_{d_1}\, P_\sigma^\top\}_{\sigma\in S_t}$. Consider the criterion Φ defined by:

$$\Phi(C_d) = \begin{cases} 0 & \text{if} \quad C_d \in \mathcal{A} \\ +\infty & \text{if} \quad C_d \notin \mathcal{A} \end{cases}$$

For all $\sigma \in S_t$, $P_\sigma \mathcal{A} P_\sigma^\top = \mathcal{A}$, thus $\Phi(P_\sigma C_d P_\sigma^\top) = \Phi(C_d)$. Also, for any constant $a > 0$, $\Phi(aC_d) = \Phi(C_d)$. Hence Φ satisfies the conditions from the definition 4.1.6. Then, we have $\Phi(C_{d1}) < \Phi(C_{d^*})$, which is a contradiction. □

Definition 4.1.8 (Majorize)**.** *Let $x, y \in \mathbb{R}^d$, we denote $x_{\downarrow i}$ be the ith largest component of x. We say x is majorized by y, and write $x \prec y$, if*

$$\sum_{i=1}^{d} x_i = \sum_{i=1}^{d} y_i \quad \textit{and} \quad \textit{for all} \quad k = 1, ..., d-1 : \sum_{i=1}^{k} x_{\downarrow i} \leq \sum_{i=1}^{k} y_i$$

The term strict majorization is used when $\sum_{i=1}^{k} x_{\downarrow i} < \sum_{i=1}^{k} y_i$.

Corollary 4.1.2. *A design d^* is Φ_p-optimal among a class $\mathcal{D}$ of design for $p \geq -1$ if for all $d \in \mathcal{D}$ it hold that*

$$\lambda(C_{d^*}) \prec \lambda(C_d),$$

where $\lambda(C_d)$ is the d-vector of the decreasing ordered eigenvalues of C_d.

Lemma 4.1.1. *Let A and B be two (n, n) symmetric matrices, then*

$$\lambda(A + B) \prec \lambda(A) + \lambda(B).$$

Proof. From Ky Fan's maximum principle [68] we know that

$$\lambda(A) = \sum_{j=1}^{k} \lambda_j(A) = \max \sum_{j=1}^{k} \langle x_j, A x_j \rangle$$

So

$$\begin{aligned} \lambda(A + B) &= \max \sum_{j=1}^{k} \langle x_j, (A + B) x_j \rangle \\ &\prec \max \sum_{j=1}^{k} \langle x_j, A x_j \rangle + \max \sum_{j=1}^{k} \langle x_j, B x_j \rangle \\ &\prec \lambda(A) + \lambda(B). \end{aligned}$$

□

Proposition 4.1.2. *A design d^* is universally optimal (with condition a′) among a class*

$\mathcal{D}$ of designs if and only if

$$(a) \quad \text{Tr } C_d = \max_{d \in \mathcal{D}} \text{Tr } C_d,$$
$$(b) \quad \lambda\Big(\frac{C_{d^*}}{\text{Tr } C_{d^*}}\Big) \prec \lambda\Big(\frac{C_d}{\text{Tr } C_d}\Big).$$

Proof. Let the conditions (a) and (b) holds, then for all $d \in \mathcal{D}$, condition (b) implies that $\lambda(C_{d^*}) \prec \lambda(\frac{\text{Tr } C_{d^*}}{\text{Tr } C_d} C_d)$ and $\Phi(C_{d^*}) \leq \Phi(\frac{\text{Tr } C_{d^*}}{\text{Tr } C_d} C_d)$. By the definition 4.1.6 and by condition (a), $\Phi(C_{d^*}) \leq \Phi(C_d)$.

In an opposite way, let say d^* be universally optimal (with condition a'). Then the condition (a) is true. If the condition (b) is not true, then we assume, that there exists a design d_1 such that

$$\lambda\Big(\frac{C_{d^*}}{tr\, C_{d^*}}\Big) \not\prec \lambda\Big(\frac{C_{d_1}}{\text{Tr } C_{d_1}}\Big).$$

Now we define the set $\mathcal{A}$ is a cone and using lemma 4.1.1, we can show that $\mathcal{A}$ is convex, which give the identical proof of proposition 4.1.1. □

Corollary 4.1.3. *A design d^* is MS-optimal among a class $\mathcal{D}$ if it minimizes Φ_{-1} and if $\lambda(C_{d^*}) \prec \lambda(C_d)$ for all the designs minimizing Φ_{-1}.*

Definition 4.1.9 (Schur-convex)**.** *A real function ϕ on $\mathbb{R}^d$ is Schur-convex if*

$$x \prec y \Rightarrow \phi(x) \leq \phi(y).$$

and Schur-concave if

$$x \prec y \Rightarrow \phi(x) \geq \phi(y).$$

P_σ denote by the (k, k)-matrix that permutes the components of a vector according to the permutation σ lying in S_t, where S_t is the symmetric group on $\{1, ..., k\}$.

Corollary 4.1.4. *A convex symmetric function ϕ on $\mathbb{R}^d$ is Schur-convex.*

For all symmetric real matrices,

$$\delta(C) \prec \lambda(C),$$

where $\delta(C)$ is the vector of diagonal terms of C. This is sometimes referred as Schur's theorem.

We denote $\lambda(C)$ as the vector of the decreasing ordered eigenvalues of C.

Definition 4.1.10 (Schur-convex criterion)**.** *A criterion $C \mapsto \phi(C)$ is Schur-convex on the eigenvalues if*

$$\lambda(C) \prec \lambda(D) \Rightarrow \phi(C) \leq \phi(D).$$

Proposition 4.1.3. *Let Φ be a Schur convex criterion on the eigenvalues. So that $\Phi(C) = \phi(\lambda(C))$. Then d^* is Φ-optimal among a class $\mathcal{D}$ of designs if*

$$\Phi(C_{d^*}) \leq \phi(\delta(C_d)) \quad for\, all \quad d \in \mathcal{D},$$

where $\delta(C_d)$ is the vector of diagonal terms of C_d in decreasing order.

Proof. Using Schur's theorem

$$\delta(C) \prec \lambda(C),$$

Then

$$\Phi(C_{d^*}) \leq \phi(\delta(C_d)) \quad for\, all \quad d \in \mathcal{D}.$$

Hence the proof. □

4.2 Bayesian optimal design

Experimental design has always taken prior knowledge into account when determining the type of experiment to conduct. The Bayesian technique is distinct in that it considers prior information to act as a probability function, adding some degree of uncertainty into the model. BODs are a subclass of experimental designs whose objective is to produce optimal decisions in the face of uncertainty. The optimization is accomplished by selecting the optimal optimization criteria based on some statistical criteria. When dealing with many experimental sets, rather than utilizing the same model for each set, we can use the results of prior tests to change our model for the subsequent experimental sets. This can be employed in standard experimental design, but optimum designs yield more precise results due to their optimization.

The optimal design can be achieved by optimizing the experiment's predicted utility [69]. If a design plan is chosen from the possible plan set $\mathcal{D}$, then the sample data x is gath-

ered, and a decision rule Ψ is picked from the possible rule set $\mathcal{H}$ using the given d^* and observed x. This utility function can be denoted as $U(\Psi, d^*, x, y)$; then for any design d^*, the expected utility can be given as:

$$U(d^*) = \mathbb{E}[U(\Psi, d^*, x, y] = \iint U(\Psi, d^*, x, y)\xi(y|x, d^*)\xi(x|d^*)\, dy\, dx$$

where $\xi(x|d^*)$ is the likelihood function under the given d^* and $\xi(y|x, d^*)$ is the posterior distribution of y under the given d^* and observed x.

There are other modes, but we will concentrate on A- and D-optimality in this thesis as the information matrix's trace is minimized when A-optimality (average) is used and the determinant of the information matrix is maximized by D-optimality (determinant).

4.2.1 A-optimality

Let $\hat{\eta}_d$ is the best linear unbiased estimate of η using a design d with variance $Var(\hat{\eta}_d) = V_d$. The posterior covariance matrix can be used to figure out the A-optimal design.

Definition 4.2.1 (A-optimality)**.** *A design $d^* \in \mathcal{D}$ is said to be A-optimal in $\mathcal{D}$ iff*

$$\mathrm{Tr}(V_{d^*}) \leq \mathrm{Tr}(V_d),$$

for any other design $d \in \mathcal{D}$.

In the A-optimality criterion, the trace of V_d is minimized, which implies that the average variance of the BLUE of the components of η is minimized.

Lemma 4.2.1. *If A and B are square matrices of the same size then*

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

Proof. Let $A = (a_{ij})$ and $B = (b_{ij})$ be $n \times n$ matrices. Then

$$\mathrm{Tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ji}a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} = \mathrm{Tr}(BA).$$

□

When studying the trace of a matrix product, it is widely established that the matrix product is invariant under cyclic permutations [70]. $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA)$ for matrices A, B, and C, where CAB and BCA are cyclic permutations of ABC. As a result, when evaluating their traces, these three permutations are equal.

Theorem 4.2.1. *Let $A_1, ..., A_n$ be $n \times n$ matrices. Then*

$$\mathrm{Tr}(A_1...A_{n-1}A_n) = \mathrm{Tr}(A_nA_1...A_{n-1})$$

Proof. Using the lemma 4.2.1, for matrices A_1 and A_2, $\mathrm{Tr}(A_1A_2) = \mathrm{Tr}(A_2A_1)$.
Let $B = A_1A_2...A_{n-1}$, then we get

$$\mathrm{Tr}(BA_n) = \mathrm{Tr}(A_nB)$$

or

$$\mathrm{Tr}(A_1A_2...A_n) = \mathrm{Tr}(A_nA_1...A_{n-1}).$$

□

This is not true in the case of more generic permutations. In general, $\mathrm{Tr}(ABC) \neq \mathrm{Tr}(CBA)$.

Theorem 4.2.2. *Let $U(\Psi, d^*, x, y) = -||A(x - x_{CM}(y))||_2^2$. It follows that $\Phi_A(d^*) = \mathrm{Tr}(A\Gamma_{post}(d^*)A^\top)$.*

Proof. A-optimality corresponds to defining the utility function U as

$$U(\Psi, d^*, x, y) = -||A(x - x_{CM}(y))||_2^2,$$

where $x_{CM}(y)$ is the posterior mean and A is a weight matrix. Now we get the minimiza-

tion target $\Phi_A(d^*)$ as [71]

$$\begin{aligned}
\Phi_A(d^*) &= -\iint U(\Psi, d^*, x, y)\xi(y|x, d^*)\xi(x|d^*)\,dy\,dx \\
&= \iint (x - x_{CM}(y))^\top A^\top A(x - x_{CM}(y))\xi(y|x, d^*)\xi(x|d^*)\,dy\,dx \\
&= \mathrm{Tr}\Big[\iint (x - x_{CM}(y))^\top A^\top A(x - x_{CM}(y))\xi(y|x, d^*)\xi(x|d^*)\,dy\,dx\Big] \\
&= \iint \mathrm{Tr}\Big[A(x - x_{CM}(y))(x - x_{CM}(y))^\top A^\top\Big]\xi(y|x, d^*)\xi(x|d^*)\,dy\,dx \\
&= \mathrm{Tr}\Big[A\iint (x - x_{CM}(y))(x - x_{CM}(y))^\top \xi(y|x, d^*)\,dy\,\xi(x|d^*)\,dx A^\top\Big] \\
&\quad [\text{where we applied theorem (4.2.1)}] \\
&= \mathrm{Tr}\Big[A\int \mathbb{E}_{Y|x,d^*}\big[(x - x_{CM}(y))(x - x_{CM}(y))^\top\big]\xi(x|d^*)\,dx A^\top\Big] \\
&= \mathrm{Tr}\Big[A\int \Gamma_{post}(d^*)\xi(x|d^*)\,dx A^\top\Big] \\
&= \mathrm{Tr}\Big[A\Gamma_{post}(d^*)A^\top\Big]
\end{aligned}$$

since for Gaussian prior, the posterior covariance matrix $\Gamma_{post}(d^*)$ is independent of the measurement x. □

4.2.2 D-optimality

The logarithmic determinant of the posterior covariance matrix can be used to calculate the D-optimal design.

Definition 4.2.2 (D-optimality)**.** *A design $d^* \in \mathcal{D}$ is said to be D-optimal in $\mathcal{D}$ iff*

$$\det(V_{d^*}) \leq \det(V_d),$$

for any other design $d \in \mathcal{D}$.

D-optimal design corresponds to defining the utility function U as

$$U(x, d^*) = KL\Big[Y|x, d^*\Big|\Big|Y|d^*\Big],$$

which is independent of the unknown x. Here, $KL\Big[Y|x, d^*\Big|\Big|Y|d^*\Big]$ is the Kullback-

Leibler (KL) divergence [72] can be defined as

$$\begin{aligned} KL\Big[Y|x,d^*\Big|\Big|Y|d^*\Big] &= \int \xi(y|x,d^*)\,\log\Big(\frac{\xi(y|x,d^*)}{\xi_{prior}(y)}\Big)\,dy \\ &= \int \xi(y|x,d^*)\,\log(\xi(y|x,d^*)\,dy - \int \xi(y|x,d^*)\,\log\xi_{prior}(y)\,dy \\ &= \int \xi(y|x,d^*)\,\log(\xi(y|x,d^*)\,dy + H(Y|x,d^*), \end{aligned}$$

where $H(Y|x,d^*)$ is the differential entropy of the distribution $\xi(y|x,d^*)$. Since the utility function $U(x,d^*)$ is independent of y, then the D-optimality can be expressed as [73]

$$\begin{aligned} \Phi_D(d^*) &= -\int U(x,d^*)\xi(x|d^*)\,dx \\ &= \int H(Y|x,d^*)\xi(x|d^*)\,dx + \iint \xi(x,y|d^*)\log\xi_{prior}(y)\,dy\,dx. \end{aligned}$$

Here

$$\begin{aligned} \iint \xi(x,y|d^*)\log\xi_{prior}(y)\,dy\,dx &= \iint \xi(x|y,d^*)\,dx\xi_{prior}(y)\log\xi_{prior}(y)\,dy \\ &= \int \xi_{prior}(y)\log\xi_{prior}(y)\,dy \\ &= -H(Y). \end{aligned}$$

Theorem 4.2.3. *Let* $U(x,d^*) = \log(\frac{\xi_{post}(y|x,d^*)}{\xi_{prior}(y)})$. *It follows, that*

$$\Phi_D(d^*) = \frac{1}{2}\log\{\det\Gamma_{post}(d^*)\} + constant.$$

Proof. For a Gaussian random variable Y, $H(Y)$ has a closed form [74] given as

$$H(Y) = \frac{n}{2} + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log(\det\Gamma_{post}(d^*)).$$

Then the D-optimality finally becomes

$$\begin{aligned} \Phi_D(d^*) &= H(Y|x,d^*) - H(Y) \\ &= \frac{1}{2}\log\Big(\frac{\det\Gamma_{post}(d^*)}{\det\Gamma_{prior}}\Big) \\ &= \frac{1}{2}\log(\det\Gamma_{post}(d^*) - \det\Gamma_{prior}) \\ &= \frac{1}{2}\log(\det\Gamma_{post}(d^*) + constant \end{aligned}$$

Here $\det\Gamma_{prior}$ is constant with respect to d^*, which is neglectable. □

5 CONCLUSION

Due to the properties of Gaussian distribution, it is easy to use for likelihood function in our posterior calculation although depending on data the distribution need to be fixed. In this thesis, Bayesian approach has been introduced from a modeling perspective to optimize an optimal experimental design. We have highlighted the relevant theory and criteria for obtaining A- and D-optimal designs. A-optimality minimizes the cost while D-optimality maximizes the cost and for that we have used both optimality for our optimal experimental design. Despite the fact that the underlying principles of Bayesian approach are quite old, the field is experiencing a renaissance, aided by new problems, models, theory, and software implementations.

REFERENCES

[1] R.A. Fisher. *The Design of Experiments*. The Design of Experiments. Oliver and Boyd, 1935.

[2] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[3] Benjamin Durakovic. Design of experiments application, concepts, examples: State of the art. *Periodicals of Engineering and Natural Sciences (PEN)*, 5(3), 2017.

[4] Ben Calderhead and Mark Girolami. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus*, 1(6):821–835, Dec 2011.

[5] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[6] J Hasenauer, S Waldherr, K Wagner, and F Allgöwer. Parameter identification, experimental design and model falsification for biological network models using semidefinite programming. *IET Syst Biol*, 4(2):119–130, Mar 2010.

[7] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 06 2009.

[8] J. Vanlier, C. A. Tiemann, P. A. J. Hilbers, and N. A. W. van Riel. A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, 02 2012.

[9] David Gomez-Cabrero, Albert Compte, and Jesper Tegner. Workflow for generating competing hypothesis from models with parameter uncertainty. *Interface Focus*, 1(3):438–449, Jun 2011.

[10] David J. Klinke. An empirical bayesian approach for model-based inference of cellular signaling networks. *BMC Bioinformatics*, 10(1):371, 2009.

[11] Clemens Kreutz and Jens Timmer. Systems biology: experimental design. *FEBS J*, 276(4):923–942, Feb 2009.

[12] Scott N. Walsh, Tim M. Wildey, and John D. Jakeman. Optimal Experimental Design Using a Consistent Bayesian Approach. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 4(1), 09 2017. 011005.

[13] Daniel Faller, Ursula Klingmüller, and Jens Timmer. Simulation methods for optimal experimental design in systems biology. *Simulation*, 79(12):717–725, 2003.

[14] Maria Rodriguez-Fernandez, Pedro Mendes, and Julio R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2):248–265, 2006. 5th International Conference on Systems Biology.

[15] Aleksandar Jankovic, Gaurav Chaudhary, and Francesco Goia. Designing the design of experiments (doe) – an investigation on the influence of different factorial designs on the characterization of complex systems. *Energy and Buildings*, 250:111298, 2021.

[16] David Roxbee Cox and Nancy Reid. *The theory of the design of experiments*. Chapman and Hall/CRC, 2000.

[17] Andrej. Pázman. *Foundations of optimum experimental design*. D. Reidel ; Distributors for the U.S.A. and Canada, Kluwer Academic Publishers, Dordrecht, Holland; Boston; Hingham, MA, U.S.A., 1986.

[18] Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

[19] Dariusz Ucinski. *Optimal measurement methods for distributed parameter system identification*. CRC press, 2004.

[20] Stefan Körkel *, Ekaterina Kostina, Hans Georg Bock, and Johannes P. Schlöder. Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes. *Optimization Methods and Software*, 19(3-4):327–338, 06 2004.

[21] Matthias Chung and Eldad Haber. Experimental design for biological systems. *SIAM Journal on Control and Optimization*, 50(1):471–489, 2012.

[22] Hans Georg Bock, Stefan Körkel, and Johannes P Schlöder. Parameter estimation and optimum experimental design for differential equation models. *Model based parameter estimation*, pages 1–30, 2013.

[23] 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.

[24] Irene Bauer, Hans Georg Bock, Stefan Körkel, and Johannes P. Schlöder. Numerical methods for optimum experimental design in dae systems. *Journal of Computational and Applied Mathematics*, 120(1):1–25, 2000.

[25] Eldad Haber, Zhuojun Magnant, Christian Lucero, and Luis Tenorio. Numerical methods for a-optimal designs with a sparsity constraint for ill-posed inverse problems. *Computational Optimization and Applications*, 52(1):293–314, 2012.

[26] E Haber, L Horesh, and L Tenorio. Numerical methods for experimental design of large-scale linear ill-posed inverse problems. *Inverse Problems*, 24(5):055012, sep 2008.

[27] JoséM. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 2022/08/06/ 1979.

[28] Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 38(1):A243–A272, 2016.

[29] Anthony Curtis Atkinson and Alexander N Donev. *Optimum experimental designs*, volume 5. Clarendon Press, 1992.

[30] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273 – 304, 1995.

[31] Quan Long, Marco Scavino, Raúl Tempone, and Suojin Wang. A laplace method for under-determined bayesian optimal experimental designs. *Computer Methods in Applied Mechanics and Engineering*, 285:849–876, 2015.

[32] Seymour Gkisser. *Predictive inference: an introduction*. Chapman and Hall/CRC, 2017.

[33] Dennis V Lindley and Melvin R Novick. The role of exchangeability in inference. *The annals of statistics*, pages 45–58, 1981.

[34] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[35] Frank G Ball, Adrian FM Smith, and Isabella Verdinelli. Biased coin designs with a bayesian bias. *Journal of Statistical Planning and Inference*, 34(3):403–421, 1993.

[36] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, abs/1012.2599, 2010.

[37] James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for bayesian optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[38] Christopher C. Drovandi and Minh-Ngoc Tran. Improving the Efficiency of Fully Bayesian Optimal Design of Experiments Using Randomised Quasi-Monte Carlo. *Bayesian Analysis*, 13(1):139 – 162, 2018.

[39] Jay I. Myung, M Pitt, Yun Tang, and Daniel R. Cavagnaro. Bayesian adaptive optimal design of psychology experiments. In *Proceedings of the 2nd International Workshop in Sequential Methodologies (IWSM2009)*, pages 1–6. Citeseer, 2009.

[40] Ziyu Wang, Babak Shakibi, Lin Jin, and Nando Freitas. Bayesian multi-scale optimistic optimization.

[41] Costas Papadimitriou and Costas Argyris. Bayesian optimal experimental design for parameter estimation and response predictions in complex dynamical systems. *Procedia Engineering*, 199:972–977, 2017. X International Conference on Structural Dynamics, EURODYN 2017.

[42] P. Frazier. A tutorial on bayesian optimization. *ArXiv*, abs/1807.02811, 2018.

[43] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[44] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[45] Michael B Marcus and Jay Rosen. *Markov Processes, Gaussian Processes, and Local Times*, volume 100 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

[46] Fetsje Bijma, Marianne Jonker, and Aad Vaart. *An Introduction to Mathematical Statistics*. Amsterdam University Press, 2017.

[47] Ulisses Braga-Neto. *Fundamentals of pattern recognition and machine learning*. Springer, Cham, Switzerland, 1st ed. 2020. edition, 2020.

[48] Matthew F. Dixon. *Machine Learning in Finance From Theory to Practice*. Springer International Publishing, Cham, 1st ed. 2020. edition, 2020.

[49] Renjun Qiu, Liang Yan, and Xiaojun Duan. Solving fredholm integral equation of the first kind using gaussian process regression. *Applied Mathematics and Computation*, 425:127032, 2022.

[50] Abbas Rohani, Morteza Taki, and Masoumeh Abdollahpour. A novel soft computing model (gaussian process regression with k-fold cross validation) for daily and monthly solar radiation forecasting (part: I). *Renewable Energy*, 115:411–422, 2018.

[51] J. M Bernardo and Adrian F. M Smith. *Bayesian theory / Jose M. Bernardo, Adrian F.M. Smith.* Wiley series in probability and mathematical statistics. Wiley, Chichester, England, 1993.

[52] Jeff. Grover. *Strategic Economic Decision-Making Using Bayesian Belief Networks to Solve Complex Problems*. SpringerBriefs in Statistics, 9. Springer New York, New York, NY, 2013.

[53] BOREK PUZA. *Bayesian Methods for Statistical Analysis*. ANU Press, 2015.

[54] Andrew Gelman. *Bayesian data analysis*. Texts in statistical science. Chapman and Hall/CRC, an imprint of Taylor and Francis, Boca Raton, FL, 3rd ed. edition, 2013.

[55] Andrew Gelman. Prior distribution. *Encyclopedia of environmetrics*, 3(4):1634–1637, 2002.

[56] Leonhard Held and Daniel Sabanés Bové. *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*. Statistics for Biology and Health. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2020.

[57] Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer texts in statistics. Springer, New York, corrected second print. edition, 2010 - 2005.

[58] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

[59] Luc Pronzato. Optimal experimental design and some related control problems. *Automatica*, 44(2):303–325, 2008.

[60] Pierre Druilhet. Conditions for optimality in experimental designs. *Linear Algebra and its Applications*, 388:147–157, 2004. Tenth Special Issue (Part 1) on Linear Algebra and Statistics.

[61] Ashish Das. An introduction to optimality criteria and some results on optimal block design. In *Design Workshop Lecture Notes*, pages 1–21. Citeseer, 2002.

[62] J. Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2(5):849–879, 1974.

[63] A Hedayat. Study of optimality criteria in design of experiments. Technical report, ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF MATHEMATICS, 1980.

[64] Kirti R Shah and BIKAS Sinha. *Theory of optimal designs*, volume 54. Springer Science & Business Media, 2012.

[65] Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

[66] J. A. Eccleston and A. Hedayat. On the Theory of Connected Designs: Characterization and Optimality. *The Annals of Statistics*, 2(6):1238 – 1255, 1974.

[67] J. Kiefer. Balanced Block Designs and Generalized Youden Designs, I. Construction (Patchwork). *The Annals of Statistics*, 3(1):109 – 118, 1975.

[68] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[69] D.V. Lindley. *Bayesian Statistics, A Review*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1972.

[70] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2000.

[71] M Burger, A Hauptmann, T Helin, N Hyvönen, and J-P Puska. Sequentially optimized projections in x-ray imaging. *Inverse Problems*, 37(7):075006, jun 2021.

[72] Ahmed Attia, Alen Alexanderian, and Arvind K Saibaba. Goal-oriented optimal design of experiments for large-scale bayesian linear inverse problems. *Inverse Problems*, 34(9):095009, jul 2018.

[73] Alen Alexanderian, Philip J. Gloor, and Omar Ghattas. On Bayesian A- and D-Optimal Experimental Designs in Infinite Dimensions. *Bayesian Analysis*, 11(3):671 – 695, 2016.

[74] C Shannon. A mathematical theory of communication. *Mobile computing and communications review*, 5(1):3–55, 2001.