Symmetric Kullback Leibler divergence-based design of experiments with estimation of unspecified values

Brijesh Kumar * **Mani Bhushan** **

* *Department of Chemical Engineering, Indian Institute of Technology Bombay, India (e-mail: brijesh19@iitb.ac.in)*
** *Department of Chemical Engineering, Indian Institute of Technology Bombay, India (e-mail: mbhushan@iitb.ac.in)*

Abstract: In this work, we propose a Symmetric Kullback Leibler divergence (SKLD)-based approach for optimal Design of Experiments (DOE) along with estimation of unspecified values in the design of experiments data matrix. Using SKLD as optimality criteria as opposed to various existing alphabetic optimality criteria, facilitates the incorporation of end-user desired performance of estimates. For the case when experimental noise is Gaussian and uncorrelated, the proposed approach results in a Mixed Integer Non-Linear Programming (MINLP) problem. This problem is NP-hard to solve. Hence, a novel heuristic solution strategy is also proposed which solves the proposed problem iteratively and sequentially. In particular, the MINLP problem is split into two sub-problems: (i) Non-Linear Programming (NLP) problem: to estimate optimal unspecified values, and (ii) Non-Linear Integer Programming (IP) problem: to obtain optimal DOE. These two subproblems are solved sequentially and iteratively until convergence is reached. The proposed solution strategy guarantees the decreasing behaviour of SKLD value. The efficacy of the proposed solution strategy is tested on an illustrative example and a Material synthesis problem, and performance is compared with Fedorov exchange algorithm, Forward Greedy search algorithm, and some of the popular MINLP solvers available in GAMS environment. Results demonstrate that the proposed solution approach outperforms most other methods.

Keywords: Design of experiments, Symmetric Kullback Leibler divergence, Estimation of unspecified value, Convex reformulation

1. INTRODUCTION

The Design of Experiments (DOE) problem has been investigated since the early 19th century (Pukelsheim, 2006). It arises in numerous applications across diverse disciplines such as agriculture, pharmaceutical, chemical processing, manufacturing, and foods ((Velicheti et al., 2022), (Yining et al., 2017)). These disciplines widely use parameter-based models for modelling, optimization, and control. The parameters of these models are estimated using experimental data. In literature, a typical DOE problem setting is to strategically select a small subset of experiments from a given large pool of n experiments (data points) to maximize the statistical efficiency of regression (Montgomery (2017), Velicheti et al. (2022)). This is needed since performing all n experiments is tedious and economically infeasible. The problem of selection of a subset of experiments is similar to problems arising in other contexts, for instance, sensor placement design (Joshi and Boyd, 2008), and feature selection (Velicheti et al., 2022)).

In literature, the DOE problem is formulated as an optimization problem using various alphabetical optimality criteria, such as A-, D-, E-, G-, V-, and T- optimality criteria (Montgomery, 2017). These optimality criteria may lead to different solutions of the Optimal Design of Experiment (ODOE) problem for the Gaussian case. It is not clear which optimal criteria should be then used for ODOE. In the recent literature on sensor placement design, information theory-based Kullback Leibler Divergence (KLD) has been proposed. KLD incorporates the user's desired estimation performance (Prakash and Bhushan (2023), Arjun and Jan (2024)).

In many situations, some of the values in the set of n given experiments may be unspecified by the end-user. In such a scenario, apart from selecting an appropriate subset of experiments, the corresponding unspecified values must also be estimated. In most existing literature, the DOE and estimation of the unspecified value problems are treated as separate problems. The alphabetic-optimality criteria based DOE problems with all specified values have been solved using different approaches such as Fedorov exchange methods (Fedorov, 2013), Genetic algorithms (Langner et al., 2003), and Mixed integer Convex methods (Hendrych et al., 2023). In the machine learning domain, various methods for estimation of missing values have been proposed, for instance, surrogates imputation, k-nearest neighbours imputation, and machine learning-based imputation (Malarvizhi and Thanamani, 2012). These methods use patterns and relationships with known (specified) data

to estimate the missing data (Rubin, 2004). However, these may not be the best for estimation of unspecified values in DOE context since the aim of DOE is to maximize efficiency of regression and not any statistical similarity. The recent work of Velicheti et al. (2022) considered coupled problems of DOE and estimation of unspecified value using A-optimality criteria in a maximum entropy principle-based framework. To the best of our knowledge, there are no other works which simultaneously tackle the problem of DOE along with specification of unspecified values.

In the current work, we propose Symmetric Kullback Leibler Divergence (SKLD) based DOE with unspecified value estimation. This formulation allows the user to incorporate the desired performance of the estimates in the objective function. While this formulation can be used when estimates are non-Gaussian, in the current work we restrict to the Gaussian and uncorrelated experimental noise case. The proposed formulation is a Mixed-integer Non-Linear Programming (MINLP) problem. The proposed MINLP problem is an NP-hard problem. Hence, in this work, we propose a novel solution strategy to solve the proposed formulation iteratively and sequentially. In particular, we break the problem into two parts: (i) DOE, and (ii) unspecified value estimation, and solve these problems sequentially till convergence while guaranteeing a decreasing behaviour of objective function value (SKLD value) with iterations. The efficacy of the proposed solution strategy is tested on an illustrative example and a Material synthesis data set (low-temperature microwave-assisted thin film crystallization process). The results are compared with some other methods: (i) Heuristic algorithms such as (ia) Forward Greedy search, and (ib) Fedorov exchange algorithms, as well as (ii) state-of-the-art popular MINLP solvers in *GAMS* (using inbuilt solvers *SBB* and *Alpha ECP*) (Bussieck and Meeraus, 2004).

The main contributions of the current work are:

- DOE with estimation of unspecified value is formulated as minimization of symmetric KLD problem.
- A novel solution strategy is proposed for solving the resulting MINLP formulation.
- The efficacy of the proposed novel solution strategy is demonstrated by comparing performance with other optimization approaches on: (i) an illustrative example and (ii) a material synthesis problem.

The rest of the paper is organized as: Section (2) provides relevant background. Section (3) presents our proposed formulation. Section (4) proposes a solution strategy for solving the resulting MINLP problem. Section (5) compares proposed solution strategy with a few other implementations on case studies. Section (6) concludes the work.

2. RELEVANT BACKGROUND

2.1 Symmetric Kullback Leibler Divergence (SKLD)

Symmetric Kullback Leibler Divergence measures the statistical distance between two probability density functions (pdf) (Kullback, 1997). Let $\mathbf{z} \in \mathbb{R}^{\eta}$ be a continuous random variable, and $f(\mathbf{z})$, and $g(\mathbf{z})$ be different pdfs of $\mathbf{z}$.

***Definition** 1.* SKLD, denoted as $\{\mathcal{K}_s(g(\mathbf{z})||f(\mathbf{z}))\}$, is defined as (Arjun and Jan, 2024):

$$\mathcal{K}_s(g(\mathbf{z})||f(\mathbf{z})) = \frac{1}{2}\int_{-\infty}^{\infty}\left\{g(\mathbf{z})\ ln\left(\frac{g(\mathbf{z})}{f(\mathbf{z})}\right) + f(\mathbf{z})\ ln\left(\frac{f(\mathbf{z})}{g(\mathbf{z})}\right)\right\}d\mathbf{z} \quad (1)$$

SKLD satisfies following properties (Arjun and Jan, 2024):

(1) $\mathcal{K}_s(g(\mathbf{z})||f(\mathbf{z})) \geq 0$ for all $f(\mathbf{z})$ and $g(\mathbf{z})$
(2) $\mathcal{K}_s(g(\mathbf{z})||f(\mathbf{z})) = 0$ if and only if $f(\mathbf{z}) = g(\mathbf{z})$
(3) $\mathcal{K}_s(g(\mathbf{z})||f(\mathbf{z})) \leq \mathcal{K}_s(g(\mathbf{z})||r(\mathbf{z})) + \mathcal{K}_s(r(\mathbf{z})||f(\mathbf{z}))$ for all $f(\mathbf{z})$, $g(\mathbf{z})$ and $r(\mathbf{z})$, where $r(\mathbf{z})$ is another pdf.

For the case when random variables $\mathbf{z}$ follow Gaussian distribution, i.e. $f(\mathbf{z}) \equiv \mathcal{N}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$ and $g(\mathbf{z}) \equiv \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean and covariance and $\mathcal{N}(.)$ represents Gaussian distribution, equation (1) simplifies to:

$$\mathcal{K}_s(g(\mathbf{z})||f(\mathbf{z})) = \frac{1}{4}\Big[tr\{\boldsymbol{\Sigma}_f\boldsymbol{\Sigma}_g^{-1}\} + tr\{\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_f^{-1}\} + (\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T[\boldsymbol{\Sigma}_g^{-1} + \boldsymbol{\Sigma}_f^{-1}](\boldsymbol{\mu}_f - \boldsymbol{\mu}_g) - 2\eta\Big] \quad (2)$$

where, $tr(.)$ represent the trace of matrix $(.)$.

We next discuss the DOE problem with unspecified value estimation in a linear regression setting.

2.2 Linear Regression and Design of Experiments

Parameter Estimation for Linear Regression Models

Linear regression models are widely used since they permit simpler implementation and tractable solutions for DOE (Langner et al. (2003), Ravi et al. (2016), Velicheti et al. (2022)). A linear regression model in presence of unspecified values can be represented by:

$$\mathbf{y} = \mathbf{X(m)}\ \boldsymbol{\beta} + \boldsymbol{\nu} \quad (3)$$

Here $\boldsymbol{\beta} \in \mathbb{R}^N$ is the parameter-vector to be estimated, data matrix $\mathbf{X(m)} \in \mathbb{R}^{n\times N}$ is such that i^{th} row represents i^{th} experiment. Further, vector $\mathbf{m} \in \mathbb{R}^p$ represents the unspecified elements in data matrix $\mathbf{X(m)}$. The left hand side vector $\mathbf{y} \in \mathbb{R}^n$ will be the observed response when experiments are performed corresponding to $\mathbf{X(m)}$. The experimental noise ($\boldsymbol{\nu} \in \mathbb{R}^n$) is assumed to follow a Gaussian distribution with mean zero and known diagonal covariance matrix $\boldsymbol{\Sigma_\nu}$ (i.e. $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_\nu})$). Diagonal nature of $\boldsymbol{\Sigma_\nu}$ implies that experimental noise terms across different experiments are independent.

In many situations, the user has some prior knowledge about the parameters which can be expressed as:

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{e} \quad (4)$$

where, $\widetilde{\boldsymbol{\beta}}$ is the prior estimate of $\boldsymbol{\beta}$, and error $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_e})$. Augmenting equations (3) and (4), we get:

$$\mathbf{z} = \widetilde{\mathcal{X}}\boldsymbol{\beta} + \mathbf{f} \quad (5)$$

where, augmented experimental noise $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_f})$ with $\boldsymbol{\Sigma_f} = \begin{bmatrix}\boldsymbol{\Sigma_e} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma_\nu}\end{bmatrix}$. Further, augmented experimental observation ($\mathbf{z} \in \mathbb{R}^{(n+N)}$) and augmented input data matrix ($\widetilde{\mathcal{X}} \in \mathbb{R}^{(n+N)\times N}$) is represented by $\mathbf{z} = \begin{bmatrix}\widetilde{\boldsymbol{\beta}}\\ \mathbf{y}\end{bmatrix}$ and $\widetilde{\mathcal{X}} = \begin{bmatrix}\mathbf{I}\\ \mathbf{X(m)}\end{bmatrix}$, respectively. Considering equation (5), the generalized least square estimate of ($\boldsymbol{\beta}$) is obtained as,

$$\hat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}} ||\boldsymbol{\Sigma_f}^{\frac{-1}{2}}(\mathbf{z} - \widetilde{\mathcal{X}}\boldsymbol{\beta})||_2^2 = (\widetilde{\mathcal{X}}^T\boldsymbol{\Sigma}_\mathbf{f}^{-1}\widetilde{\mathcal{X}})^{-1}\widetilde{\mathcal{X}}^T\boldsymbol{\Sigma}_\mathbf{f}^{-1}\mathbf{z} \tag{6}$$

The corresponding covariance of estimate ($\hat{\boldsymbol{\beta}}$) is given by:

$$cov(\hat{\boldsymbol{\beta}}) = (\widetilde{\mathcal{X}}^T\boldsymbol{\Sigma}_\mathbf{f}^{-1}\widetilde{\mathcal{X}})^{-1} = (\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}(\mathbf{m})^T\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1} \tag{7}$$

where, $cov(.)$ represents covariance.

Design of Experiments for Linear Regression Models

The Design of experiments problem is an optimization problem in which r number of experiments are to be selected from the given n number of experiments ($r \leq n$). Further, the unspecified values, if any, in the selected r experiments also have to be estimated. This problem is to be solved for an appropriate design criteria. Let $q_i,\ i = 1, 2, ...n$ be binary variables defined as:

$$q_i = \begin{cases} 1 & \text{if } i^{th} \text{ experiment is selected} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Now, $(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1})$ in equation (7) can be re-written to reflect the covariance of estimated parameters corresponding to the selected experiments as:

$$\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1} = diag\left(\left\{\frac{q_i}{\sigma_i^2}\right\}_{i=1,2,....n}\right) \tag{9}$$

where, $diag(.)$ represents a diagonal matrix and σ_i^2 is the variance of noise which will corrupt the i^{th} experimental observation. For ease of readability, in the rest of the paper the DOE problem with no unspecified values will be referred as DOE, while DOE problem with unspecified values will be labeled as DOE-UV.

***Remark** 1.* In literature, number of experiments selected (r) is assumed to be greater than or equal to the number of features (N) (i.e $r \geq N$) (Velicheti et al., 2022). This will ensure invertibility of $\mathbf{X}(\mathbf{m})^T\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m})$. However, given the prior information of the estimate as represented in equation (4), the assumption of $r \geq N$ is not made in the current work.

3. PROPOSED FORMULATION

In this work, we propose to formulate the DOE problem in the presence of unspecified values for linear regression models using SKLD optimality criteria. In the current work, we assume the experimental noises to be Gaussian and uncorrelated. The SKLD computes the distance between two pdfs of a random variable (Section 2). In the current work, we consider SKLD for the estimated parameters with the two pdfs being: (i) design pdf of estimates i.e. pdf corresponding to a particular combination of chosen r experiments with specifications of the corresponding unspecified values, and (ii) reference pdf of estimates i.e. pdf provided by the end-user. In the current work, the reference pdf corresponds to selecting all experiments (the best-case scenario) with no unspecified values. Denote reference and design pdfs as $\widetilde{f}(\hat{\boldsymbol{\beta}}) \sim \mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_f)$ and $\widetilde{g}(\hat{\boldsymbol{\beta}}) \sim \mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_g)$ where, $\widetilde{\boldsymbol{\Sigma}}_f$ and $\widetilde{\boldsymbol{\Sigma}}_g$ are the reference and design covariance of estimate $\hat{\boldsymbol{\beta}}$. The reference covariance of estimate ($\hat{\boldsymbol{\beta}}$) is $\widetilde{\boldsymbol{\Sigma}}_f = (\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\nu}}^{-1}\mathbf{X})^{-1}$ where, $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\nu}}^{-1} = diag\left(\left\{\frac{1}{\sigma_i^2}\right\}_{i=1,2,....n}\right)$. The DOE problem with estimation of unspecified values is then formulated as:

***Formulation** 1.*

$$\min_{\mathbf{q},\mathbf{m}} \frac{1}{4}\Big[tr\{(\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1}\widetilde{\boldsymbol{\Sigma}}_f^{-1}\} + tr\{\widetilde{\boldsymbol{\Sigma}}_f(\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))\} - 2N\Big]$$

$$s.t \quad \sum q_i = r, \quad lb \leq \mathbf{m} \leq ub$$
$$\mathbf{m} \in \mathbb{R}^p, \quad q_i \in \{0,1\},\ \forall\ i = 1, 2, .., n$$

In the above, lb and ub are the lower and upper bounds of respective unspecified values ($\mathbf{m}$) in the data matrix ($\mathbf{X}(\mathbf{m})$). Solving the above problem will result in choice of r experiments and the corresponding specifications of unspecified values, such that the resulting density function of the estimated parameters is as close as possible to the reference density function. Formulation 1 is a Mixed Integer Non-Linear Programming (MINLP) problem.

4. SOLUTION STRATEGY

In this section, we propose a novel approach to solve the MINLP problem Formulation 1. The approach relies on splitting the original problem Formulation 1 into two subproblems as discussed next:

(1) Non-Linear programming (NLP) problem for estimation of unspecified values (continuous variables) only, given the chosen experiments (integer variables) as:

***Formulation** 2.*

$$\min_{\widetilde{\mathbf{m}}_\Phi} \mathcal{L} = \frac{1}{4}\Big[tr\{(\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1}\widetilde{\boldsymbol{\Sigma}}_f^{-1}\} + tr\{\widetilde{\boldsymbol{\Sigma}}_f(\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))\} - 2N\Big]$$

$$s.t \quad lb \leq \mathbf{m} \leq ub, \quad \boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1} = \boldsymbol{\Sigma}_\Phi^{-1}, \quad \mathbf{m} \in \mathbb{R}^p$$

Formulation 2 is solved for a selected set of experiments (Φ) where $\boldsymbol{\Sigma}_\Phi^{-1}$ is a known, diagonal matrix defined as:

$$[\boldsymbol{\Sigma}_\Phi^{-1}]_{i,i} = \begin{cases} \frac{1}{\sigma_i^2} & \text{if } q_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The decision variables $\widetilde{\mathbf{m}}_\Phi$ for Formulation 2 is the vector of unspecified elements of $\mathbf{X}(\mathbf{m})$ corresponding to Φ. Formulation 2 is an NLP problem.

(2) Non-Linear Integer Programming (IP) problem for selecting r out of n experiments for a specified choice of vector m. The resulting optimization Formulation, labeled Optimum Design of Experiment (ODOE) with known data matrix ($\mathbf{X}$) is:

***Formulation** 3.*

$$\min_{\mathbf{q}} \frac{1}{4}\Big[tr\{(\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1}\widetilde{\boldsymbol{\Sigma}}_f^{-1}\} + tr\{\widetilde{\boldsymbol{\Sigma}}_f(\boldsymbol{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))\} - 2N\Big]$$

$$s.t \quad \sum q_i = r,\ \mathbf{X}(\mathbf{m}) = \mathbf{X},\ q_i \in \{0,1\},\ \forall\ i = 1, ..n$$

The proposed solution approach requires sequential solving of NLP and IP problems as posed above. To facilitate this, the IP problem (Formulation 3) is reformulated as Mixed Integer Semidefinite Programming (MISDP) problem. This ensures that the integer relaxation of the IP problem is convex in nature.

Theorem 1. The equivalent MISDP problem formulation for IP problem Formulation 3 is:

Formulation 4.

$$\begin{aligned}\min_{\mathbf{q},U,\mathbf{H}}\ \mathcal{J} =& \frac{1}{4}\Big[U + tr\{\widetilde{\mathbf{\Sigma}}_f(\mathbf{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m})\} - 2N\Big] \\ s.t\ \ & \sum q_i = r,\ \ \mathbf{X}(\mathbf{m}) = \mathbf{X} \\ & tr\{\mathbf{H}\widetilde{\mathbf{\Sigma}}_f^{-1}\} \le U,\ \ q_i \in \{0,1\},\ \forall\ i = 1,2,..,n \\ & \begin{bmatrix}\mathbf{H} & \mathbf{I} \\ \mathbf{I} & \mathbf{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m})\end{bmatrix} \succeq 0\end{aligned}$$

where $\mathcal{J}$ is the objective function obtained by solving Formulation 4 for a chosen $\mathbf{m}$.

Proof: The proof is motivated from Arjun and Jan (2024). Introducing auxiliary variables $\mathbf{H}$ and corresponding epigraph constraints (Boyd and Vandenberghe, 2004) to Formulation 3, leads to the following equivalent formulation:

Formulation 4.1.

$$\begin{aligned}\min_{\mathbf{q},U,\mathbf{H}}\ & \frac{1}{4}\Big[U + tr\{\widetilde{\mathbf{\Sigma}}_f(\mathbf{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m})\} - 2N\Big] \\ s.t\ \ & \sum q_i = r,\ \ \mathbf{X}(\mathbf{m}) = \mathbf{X} \\ & tr\{\mathbf{H}\widetilde{\mathbf{\Sigma}}_f^{-1}\} \le U,\quad q_i \in \{0,1\},\ \forall\ i = 1,2,..,n \\ & \mathbf{H} = (\mathbf{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1}\end{aligned}$$

Constraint $\mathbf{H} = (\mathbf{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1}$ in above formulation can be relaxed to

$$\mathbf{H} \succeq (\mathbf{\Sigma}_e^{-1} + \mathbf{X}^T(\mathbf{m})\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1}\mathbf{X}(\mathbf{m}))^{-1} \tag{11}$$

This relaxation does not affect the optimal solution since the relaxed constraint will be active at optimality. Further, using Schur's complement (Boyd and Vandenberghe, 2004), constraint (11) can be rewritten in a Linear matrix inequality (LMI) form. This results in Formulation 4 thereby proving equivalence of Formulations 3 and 4.

The advantage of Formulation 4 is that since its integer relaxation is convex, the formulation can be solved to global optimality within any IP framework which relies on solving the integer relaxed subproblems, such as in a branch and bound framework.

The proposed solution approach involves fixing integer variables $\mathbf{q}$ and continuous variables $\mathbf{m}$ in a sequential manner. To fix $\mathbf{q}$ for a given $\mathbf{m}$, the LMI in Formulation 4 is solved. However, to fix $\mathbf{m}$ for a given $\mathbf{q}$, Formulation 2 is solved twice, once for the selected experiments (Φ) and once for the remaining experiments ($\Omega \setminus \Phi$). It should be noted that for a given choice of $\mathbf{q}$ solving Formulation 2 for the full $\mathbf{m}$ vector will not be able to uniquely fix the components of $\mathbf{m}$ corresponding to experiments which are not selected. Hence, we proposed solving Formulation 2 twice as indicated. The sequence of alternating between $\mathbf{m}$ and $\mathbf{q}$ is continued iteratively till convergence.

The pseudo-code of the developed solution strategy solving the proposed DOE with unspecified values problem (Formulation 1) is listed in Algorithm 1 (Algo. 1).

The inputs to Algo. 1 are problem specific information, namely data matrix ($\mathbf{X}(\mathbf{m})$), prior covariance matrix ($\mathbf{\Sigma_e}$), experimental noise variances ($\{\sigma_i^2\}_{i=1,2,...n}$), reference covariance ($\widetilde{\mathbf{\Sigma}}_{\boldsymbol{f}}$), number of experiments to be selected (r),

Algorithm 1 : Sequential and iterative approach for DOE with estimation of unspecified value

Input: $\mathbf{X}(\ \mathbf{m})$, $\mathbf{\Sigma_e}$, $\{\sigma_i^2\}_{i=1,2,...n}$, $\widetilde{\mathbf{\Sigma}}_{\boldsymbol{f}}$, r, $\mathbf{m}^{inital}$, $iter$
Output: Φ, $\mathcal{J}^l$, $\mathbf{X}$

1: $\Omega = \{1, 2, ..., n\}$
2: $\mathbf{m}^0 \leftarrow \mathbf{m}^{initial}$
3: $\mathbf{X} \leftarrow \mathbf{X}(\mathbf{m}^0)$
4: $[\mathcal{J}^0,\ \Phi^0] \leftarrow$Solve Formulation 4
5: **for** $l = 1 : iter$ **do**
6: $\quad\mathbf{\Sigma}_{\boldsymbol{\nu}}^{-1} \leftarrow \mathbf{\Sigma}_{\Phi^{l-1}}^{-1}$
7: $\quad[\mathcal{L}^l,\ \widetilde{\mathbf{m}}_{\Phi^{l-1}}] \leftarrow$ Solve Formulation 2
8: $\quad[\mathcal{L}^{*l},\ \widetilde{\mathbf{m}}_{(\Omega\setminus\Phi^{l-1})}] \leftarrow$ Solve Formulation 2
9: $\quad\mathbf{m}^l \leftarrow \widetilde{\mathbf{m}}_{\Phi^{l-1}} \cup \widetilde{\mathbf{m}}_{(\Omega\setminus\Phi^{l-1})}$
10: $\quad\mathbf{X} \leftarrow \mathbf{X}(\mathbf{m}^l)$
11: $\quad[\mathcal{J}^l,\ \Phi^l] \leftarrow$Solve Formulation 4
12: $\quad$**if** $\Phi^l \neq \Phi^{l-1}$ **then**
13: $\quad\quad$*continue*
14: $\quad$**else**
15: $\quad\quad$*terminate*
16: $\quad$**end if**
17: **end for**

and optimization relevant information namely initial guess ($\mathbf{m}^{inital}$) for vector $\mathbf{m}$, and maximum number of iterations denoted by *iter*. The outputs of Algo. 1 are Optimal DOE (Φ), Optimal SKLD value ($\mathcal{J}^l$), and data matrix with estimates of unspecified values ($\mathbf{X}$). In particular, in line 4 in Algo. 1 initial ODOE is obtained for given initial guess of unspecified values ($\mathbf{m}^0$) by solving problem Formulation 4. The iterative procedure to solve both subproblems starts from line 5. In line 7, the NLP problem Formulation 2 is solved for a given experiment set (Φ^{l-1}) at l^{th} iteration such that the unspecified values correspond to only those given experiments and only these unspecified values are considered as decision variables for optimization. In line 8, the NLP problem Formulation 2 is solved for the remaining experiments (i.e. $\Omega\setminus\Phi^{l-1}$) at l^{th} iteration. The unspecified values $\mathbf{m}^l$ at l^{th} iteration are updated from the solutions obtained in lines 7 and 8 in line 9, while the MISDP problem Formulation 4 for known data matrix is solved in line 11. Line 12 checks if the DOE sets obtained from l^{th} and $(l-1)^{th}$ iterations are same to either terminate or continue the sequential solving of subproblems until specified number of iterations ($iter$) are reached.

Theorem 2. Algo. 1 guarantees that SKLD value (objective function of Formulation 1) decreases with iterations.

Proof: Consider that iterations upto $l-1$ have been completed. Thus, optimal SKLD value ($\mathcal{J}^{l-1}$), optimal selected experiments set (Φ^{l-1}), and vector $\mathbf{m}^{l-1}$ is available. The following steps are undertaken in Algo. 1 to complete the next iteration:

(1) Formulation 2 (line 7 of Algo. 1) is solved to update the elements of $\mathbf{m}$ corresponding to Φ^{l-1}. The optimal SKLD value $\mathcal{L}^l$ resulting from this formulation satisfies

$$\mathcal{L}^l \le \mathcal{J}^{l-1} \tag{12}$$

since $\mathbf{m}^{l-1}$ is a feasible solution of Formulation 2 with $\Phi = \Phi^{l-1}$.

(2) After obtaining $\widetilde{\mathbf{m}}_{\Phi^{l-1}}$ from Formulation 2, at the next step (line 8 of Algo. 1) Formulation 2 is re-solved to update the remaining part of vector $\mathbf{m}$ namely

$\widetilde{\mathbf{m}}_{\Omega\setminus\Phi^{l-1}}$. At the end of this step we have the full updated $\mathbf{m}$ vector as $\mathbf{m}^l$ (line 9 of Algo. 1).

(3) The $\mathbf{m}^l$ vector is now used in Formulation 4 to obtain updated DOE Φ^l with corresponding SKLD value being $\mathcal{J}^l$. This SKLD value cannot be higher than $\mathcal{L}^l$ since $\Phi = \Phi^{l-1}$ is a feasible solution of Formulation 4 with the corresponding SKLD value being $\mathcal{L}^l$. Thus,

$$\mathcal{J}^l \leq \mathcal{L}^l \leq \mathcal{J}^{l-1} \tag{13}$$

where the right-most inequality follows from equation (12).

This shows that the SKLD value decreases with iterations.

5. CASE STUDIES

We now apply the proposed SKLD-based DOE with estimation of unspecified values approach as implemented in Algo. 1 to two case studies: (i) Illustrative Example and (ii) Material Synthesis (MS). In the current work, Formulation 2 (NLP) in Algo. 1 is solved using interior point optimization (inbuilt function *fmincon* in *MATLAB*), while the MISDP in Formulation 4 is solved using MOSEK solver in *MATLAB* using *YALMIP* toolbox (MATLAB, 2025).

To benchmark performance of Algo. 1, we compare its performance with the following methods:

(1) **Fedorov exchange (FE) algorithm:**
FE method has been used in literature (Yining et al., 2017) to solve DOE problems with all specified values. It starts with a randomly selected set of experiments. FE finds the best experiment set by exchanging each index of randomly selected experiment sets with the remaining experimental sets until no exchange can give a better result than the previous best result obtained by exchange. We adapt it to solve the MINLP problem Formulation 3 similar to that in Algo. 1. In particular, at l^{th} iteration in Algo. 1, the set Φ^l (line 11 in Algo. 1) is obtained by solving Formulation 3 using FE method with starting set being Φ^{l-1}. Since the results are sensitive to the user-specified initial set (at $l = 0$), we apply the FE approach for t-trial runs with random choices for the initial set.

(2) **Forward Greedy search (FGS) algorithm:**
Adapted from sensor placement design literature (Prakash and Bhushan, 2023), FGS starts with an empty set and greedily selects an experiment to be conducted till r experiments are selected. We extend this idea to solve the DOE with unspecified values problem similar to Algo. 1. In particular, we replace Formulation 4 for selecting the set of experiments at the l^{th} iteration in Algo. 1 with Formulation 3 which is solved by FGS.

(3) **Popular MINLP solvers:**
In this work, the default solvers of *GAMS*, namely *SBB* and *Alpha ECP*) (Bussieck and Meeraus, 2004) are considered to solve Formulation 1.

The results are presented for two scenarios (A) DOE and (B) DOE-UV. For scenario (A) the results are compared with FE and FGS algorithms, while for scenario (B) the results are compared with FE, FGS, and popular MINLP solvers. For the FE approach, $t = 10$ is considered.

Case study (I): Illustrative Example
Consider regression model (Table 1) as:

$$\mathcal{Y} = \mathcal{X}_1\beta_1 + \mathcal{X}_2\beta_2 + \mathcal{X}_3\beta_3 + \boldsymbol{\nu} \tag{14}$$

where, $\mathcal{Y}$ is response vector, and $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ correspond to columns of the $\mathbf{X}$ matrix. Thus, $n = 5$ and $N = 3$. The corresponding unspecified values vector for scenario B is $\mathbf{m} = \{m_1,\ m_2,\ m_3,\ m_4,\ m_5\}$. The data matrix

Table 1. Illustrative Example: Data matrix

#	$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{X}_3$
1	3.4442	28.1680	7.3642
2	m_1	52.5973	m_2
3	1.3810	52.5973	m_3
4	3.4442	m_4	8.7923
5	m_5	30.2500	7.3642

for scenario A is same as scenario B with known $\mathbf{m} = \{2.034,\ 5.024,\ 9.875,\ 35.347,\ 3.812\}$ The corresponding inverse covariance matrix of prior estimate ${\boldsymbol{\Sigma}_e}^{-1} = 0.01 * \mathbf{I}_{5\times5}$. The variance (σ_i^2) of experimental noise for the five experiments are: 0.8649, 0.2468, 0.1865, 0.5263, 0.3826. The considered percentage of unspecified data in the data matrix is 33.33% ($p = 5$ elements of the data matrix). The reference covariance $(\widetilde{\boldsymbol{\Sigma}}_f)$ is computed with all 5 experiments with known data matrix as for scenario A.

(1) Scenario A (DOE): This scenario involves solution of only the DOE problem. This was achieved by solving Formulation 4 using MOSEK, and Formulation 3 for FE and FGS methods. Table 2 presents the performance of these approaches. It was observed that MOSEK outperformed FGS and FE methods because MOSEK solves Formulation 4 whose integer relaxation leads to a convex problem.

(2) Scenario B (DOE-UV): Table 3 presents the performance of different methods. It was observed that Algo. 1 and FE gave identical performance which was better than FGS for all values of r. The performance was also superior to popularly used MINLP solvers, except for $r = 2$.

***Remark** 2.* The advantage of being able to optimally fix unspecified values along with selecting optimal experiments is seen by comparing results from Tables 2 and 3. In particular, it is observed that the experiments chosen for various values of r are identical in columns 2 of the two tables (Table 2: MOSEK, Table 3: Algo. 1). However, the SKLD values in Table 2 are much higher lower than the corresponding values in Table 3. The ability to select the unspecified values in Table 3 results leads to design pdf being much closer to the reference pdf.

Table 2. Illustrative Example: Scenario A

r	**MOSEK**	**FGS**	**FE**
1	**3961.1** [3]	**3961.1** [3]	**3961.1** [3]
2	**795.5988** [2 4]	1292.5 [2 3]	1277.6 [1 5]
3	**0.1182** [2 3 5]	**0.1182** [2 3 5]	**0.1182** [2 3 5]
4	**0.0119** [2 3 4 5]	**0.0119** [2 3 4 5]	**0.0119** [2 3 4 5]
5	**0** [1:5]	**0** [1:5]	**0** [1:5]

Case study (II): Material Synthesis (MS)
We now consider a Material Synthesis of low-temperature microwave-assisted thin film crystallization case study to

Table 3. Illustrative Example: Scenario B

r	Algo. 1	FGS	FE	Alpha ECP	SBB
1	3616.1 [3]	3616.1 [3]	3616.1 [3]	14033.025 [1]	**2470.5** [3]
2	550.16 [2 4]	555.62 [2 3]	550.16 [2 4]	529.181 [3 5]	**529.181** [3 5]
3	**0.0524** [2 3 5]	**0.0524** [2 3 5]	**0.0524** [2 3 5]	2.279 [1 4 5]	0.054 [2 3 5]
4	**0.0011** [2 3 4 5]	**0.0011** [2 3 4 5]	**0.0011** [2 3 4 5]	0.586 [1 2 4 5]	0.003 [2 3 4 5]
5	**0** [1:5]	**0** [1:5]	**0** [1:5]	0.002 [1:5]	0.033 [1:5]

Table 4. Material synthesis results

r	Scenario	SKLD	Optimum DOE
1	**A**	832620	[1]
	B	**83336**	[1]
17	**A**	0.1227	[1 3 6 8 11 16 22 26 28 30 32 33 34 37 40 42 45]
	B	**0.1129**	[4 8 9 11 15 16 17 19 21 22 28 31 40 41 42 44 46]

demonstrate the utility of our proposed algorithm for DOE-UV (Nakamura et al., 2017). The data set contains 103 experiments and 5 variables (Nakamura et al., 2017). In the current study, we considered the first $n = 50$ experiments from that set in all the $N = 5$ variables. These variables are (1) X1: Tri-Ethyl-Gallium (TEG) volume ratio, (2) X2: Temperature ($^{\circ}C$), (3) X3: Heat to a temperature in time (minute), (4) X4: Heat as soon as possible (minute), and (5) X5: Constant Power (W). The regression model used to predict the Percentage coverage (Y) of material is:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \boldsymbol{\nu} \quad (15)$$

We randomly generate experimental noise variance between 0 and 1 for the 50 experiments. The inverse covariance matrix of prior estimate is assumed to be $\boldsymbol{\Sigma_e}^{-1} = 0.01 * \mathbf{I}_{5\times5}$. The reference covariance ($\widetilde{\boldsymbol{\Sigma}}_f$) is computed with all 50 experiments. The considered percentage of unspecified data is 6% ($p = 15$ elements of the data matrix). Table 4 presents the SKLD values and ODOE for $r = 1$ and $r = 17$ for both Scenarios A and B. From Table 4, it can be observed that the SKLD values decrease as the number of experiments increases for both scenarios. Additionally, the ODOE is different for $r = 17$. However, Scenario B exhibits a lower SKLD value, indicating that the corresponding DOE sets are closer to the reference.

6. CONCLUSION

In this work, we presented symmetric Kullback-Leibler-Divergence based approach for solving problem of design of experiments with unspecified values in the data matrix. The SKLD-based problem formulation allows incorporation of the user's desired performance specification in the design procedure. Additionally, we proposed a novel solution strategy for solving the resulting MINLP problem. This solution strategy sequentially solved NLP and MISDP problems and guaranteed decreasing behavior of objective function value. Results on two case studies showed that the proposed solution approach performed better than heuristic algorithms, and similar to SBB solver in GAMS. SKLD-based DOE-UV problems for non-Gaussian noise scenarios and nonlinear regression models can be investigated in future.

REFERENCES

Arjun, M. and Jan, N.M. (2024). Convex optimization approach to design sensor networks using information theoretic measures. *AIChE Journal*, 70(2), e18267.

Boyd, S.P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bussieck, M.R. and Meeraus, A. (2004). General algebraic modeling system (GAMS). In *Modeling languages in mathematical optimization*, 137–157. Springer.

Fedorov, V.V. (2013). *Theory of optimal experiments*. Elsevier.

Hendrych, D., Besançon, M., and Pokutta, S. (2023). Solving the optimal experiment design problem with mixed-integer convex methods. *arXiv preprint arXiv:2312.11200*.

Joshi, S. and Boyd, S.P. (2008). Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2), 451–462.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Langner, A.H., Carlyle, W.M., Montgomery, D.C., Borror, C.M., and Runger, G.C. (2003). Genetic algorithms for the construction of D-optimal designs. *Journal of Quality Technology*, 35, 28–46.

Malarvizhi, R. and Thanamani, A.S. (2012). K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev*, 5(1), 5–7.

MATLAB (2025). Mathwork. `https://in.mathworks.com/products/matlab.html`.

Montgomery, D.C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Nakamura, N., Seepaul, J., Kadane, J.B., and Reeja-Jayan, B. (2017). Design for low-temperature microwave-assisted crystallization of ceramic thin films. *Applied Stochastic Models in Business and Industry*, 33(3), 314–321.

Prakash, O. and Bhushan, M. (2023). Kullback Leibler divergence based sensor placement in linear processes for efficient data reconciliation. *Computers & Chemical Engineering*, 173, 108181.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Ravi, S.N., Ithapu, V., Johnson, S., and Singh, V. (2016). Experimental design on a budget for sparse linear models and applications. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, 583–592. PMLR.

Rubin, D.B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Velicheti, R.K., Srivastava, A., and Salapaka, S.M. (2022). Design of experiments with imputable feature data. In *2022 Eighth Indian Control Conference (ICC)*, 25–30. IEEE.

Yining, W., Adams, W.Y., and Aarti, S. (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research*, 18, 1–41.