Sensitivity-preserving of Fisher Information Matrix through random data down-sampling for experimental design

Kathrin Hellmuth · Christian Klingenberg · Qin Li

Abstract The quality of numerical reconstructions of unknown parameters in inverse problems heavily relies on the chosen data. It is crucial to select data that is sensitive to the parameters, which can be expressed through a sufficient conditioning of the Fisher Information Matrix. We propose a general framework that provides an efficient down-sampling strategy that can select experimental setups that preserves this conditioning, as opposed to the standard optimization approach in optimal experimental design. Matrix sketching techniques from randomized linear algebra is heavily leaned on to achieve this goal. The method requires drawing samples from a sensitivity-informed distribution, and gradient free sampling methods are integrated to execute the data selection. Numerical experiments demonstrate the effectiveness of this method in selecting sensor locations for Schrödinger potential reconstruction.

Kathrin Hellmuth
Department of Computing and Mathematical Sciences, California Institute of Technology, CA, USA
E-mail: hellmuth@caltech.edu

Christian Klingenberg
Department of Mathematics, University of Würzburg, Germany
E-mail: christian.klingenberg@uni-wuerzburg.de

Qin Li
Department of Mathematics, University of Wisconsin-Madison, WI, USA
E-mail: qinli@math.wisc.edu

1 Introduction

Inverse problems are ubiquitous. A system in the forward setting maps the parameter to data:

$$y(u) = \mathcal{F}(u, p) + \eta(u)\,, \tag{1.1}$$

where $\mathcal{F} : \Omega \times \mathbb{R}^K \to \mathbb{R}$ is the map, and u is the design variable in the design space Ω, the set that collects all accessible experimental specifications, with is typically of infinite size $|\Omega|$. p is the to-be-inferred parameter, assumed to be K-dimensional, and $\eta \sim \mathcal{N}(0, \Gamma)$ describes Gaussian measurement noise. When the parameter p is fixed, the forward problem returns the solution $y(u)$ for every chosen experimental design variable u.

The associated inverse problem is to revert the process: given the reading of y, we are to infer parameter p. There are many approaches to execute this inversion, such as cost function minimization, maximum likelihood estimation or Bayesian maximum-a-posteriori estimation. In the following, let

$$\hat{p} = \mathcal{R}(y, \Omega) \tag{1.2}$$

be an unbiased estimator resulting from a generic reconstruction strategy $\mathcal{R}$.

Often there are abundant choices in the design set, namely $|\Omega| \gg K$. In this case, it is natural to suspect that one does not need the full data set of $\{y(u)\}_{u \in \Omega}$ to characterize $p \in \mathbb{R}^K$. The task at hand is to select a down-sampled y that can give an almost equally good recovery of p. This reduces experimental as well as computational cost, and sometimes renders the problem computationally or experimentally tractable [12]. More specifically, one aims to design a small finite subset $\Omega_c \subset \Omega$, either through a deterministic or random selection process, and define the down-sampled data:

$$|\Omega_c| = c \ll |\Omega|\,, \quad \text{and accordingly define} \quad y_c = y|_{\Omega_c}\,, \quad \mathcal{F}_c = \mathcal{F}|_{\Omega_c}\,, \tag{1.3}$$

so that

$$\hat{p} \underbrace{\approx}_{\text{hopefully}} \hat{p}_c = \mathcal{R}(y_c, \Omega_c) \tag{1.4}$$

and thus recovering (1.2) using a smaller set of data.

There are many perspectives to take to compare (1.2) and (1.4). One frequently encountered quantity is the Fisher Information Matrix (FIM). When the full set of data in the design space Ω is used, the FIM is defined as:

$$\begin{aligned}
\mathbb{R}^{K \times K} \ni \mathcal{I}(\Omega) = \mathcal{I}(\Omega, p_*) &= G^\top(p_*)\Gamma^{-1}G(p_*) \\
&= \int_\Omega G^\top(p_*, u)\Gamma^{-1}(u)G(p_*, u)du\,,
\end{aligned} \tag{1.5}$$

where $G(p_*)$ is the Frechet derivative of $\mathcal{F}$ at a background parameter p_*:

$$G(p_*, u) = \nabla_p \mathcal{F}(p_*, u) \,.$$

Accordingly, when the data is down-sampled as is done in (1.3), the associated FIM is

$$\mathbb{R}^{K \times K} \ni \mathcal{I}(\Omega_c) = \int_{\Omega_c} G^\top(p_*, u) \Gamma^{-1}(u) G(p_*, u) du \,. \tag{1.6}$$

FIM is an important quantity that can describe the amount of information contained in data through characterizing the local sensitivity of the data with respect to (w.r.t.) the parameter around a presumed groundtruth parameter p_*. Indeed, according to the Cramer-Rao inequality [29], the inverse of FIM bounds the variance of the unbiased estimator $\hat{p}$ in the reconstruction. If a FIM has a good conditioning with high eigenvalues, we obtain low variance, and thus high confidence in the reconstruction. We note a global characterization of solvability is hardly feasible for any generic non-linear inverse problems, and localization as is done by FIM is conventional. Many classical optimal experimental design methods are about manipulating this FIM matrix. This is typically rephrased as an optimization task: One is to examine a weighing strategy so that the re-weighted $\mathcal{I}(\Omega)$ presents the optimal eigenvalue structure. The standard quantities to consider are its trace (A-optimal) and determinant (D-optimal) [22, 28, 1, 3, 31, 2]. We also refer readers to a very nice review in [19].

In this work, we would like to approach the problem from a different perspective: instead of searching for an *optimal* subset of experiments Ω_c, we are content when finding a *sufficient* one whose data adheres similar parameter sensitivity as the full setup Ω. Mathematically, this amounts to finding those designs $\Omega_c = \{u_1, ..., u_c\} \subset \Omega$ so that $\mathcal{I}(\Omega_c)$ is as informative as $\mathcal{I}(\Omega)$, or

Design Ω_c to ensure $Eig(\mathcal{I}(\Omega_c)) \approx Eig(\mathcal{I}(\Omega))$, so that (1.4) holds.

This strategy is in significant contrast to the "optimal design" where one looks the best weighing/selection strategy that provides the optimal eigenvalue structure of FIM. We relax this optimality requirement. This relaxation provides us some flexibility in developing algorithms. More specifically, since we do not look for optimizing the eigenvalue structure, techniques typically unemployed to experimental design can now be leveraged on, and one can potentially avoid deploying iterative solvers. The newly involved technics can also expand the breath of conclusion.

Indeed, we will spell out a generic condition for c, and a generic down-sample strategy that still yields sensitive data for a very general class of problems. These strategies are independent of the source of the inverse problem, nor do they require a specific structure of the original FIM. The proposed sampling strategy is probabilistic in nature, and thus sensitivity can only be guaranteed with a high probability. This sampling strategy, when applied to any specific problem, leads to a specific distribution for constructing the mask Ω_c. This distribution incorporates the property of $\mathcal{F}$, and thus integrates the knowledge from the underlying model.

The technical preparation of our approach comes from a seemingly unrelated research area of randomized linear algebra (RNLA) [24]. Indeed, the sensitivity of the data is coded in its FIM (1.6), which enjoys a special tensor structure, in case of Gaussian noise. This special structure allows us to deploy random sketching techniques from RNLA, to pin the conditions for preserving the eigenvalue structure. Specifically in this context, we can spell out a probability distribution to draw Ω_c, and show that with high probability, the associated down-sampled FIM is well-conditioned, and thus (1.4) holds true with high probability.

As such, the novelty of our work lies in establishing this new perspective on qualitative experimental design in a rather general framework. With this new perspective, we propose a concrete algorithmic pipeline to numerically execute this data selection through sampling.

The integration of probabilistic methods to design tasks is currently in its fancy and has been studied for instance in [7,26] for matrix sketching techniques for the input-to-output map or a low rank basis representation of the data, respectively. In a Bayesian optimal design setting, a data and model adapted random mask for MRI data acquisition could be constructed in [30]. Further interesting applications can be found in elliptic solution operator learning [6,5] from random input data on the basis of the randomized singular value decomposition. In [21], the authors examined the same question in a different light, where they used the sketching methods to study how many variables can be stably recovered when the experiments are fixed.

The two main technical pillars of our proposed method is the matrix sketching, and probability sampling method. We briefly review them in Section 2.1 and Section 2.2 respectively. In Section 3 we turn back to the problems (1.2)-(1.4), and examine their FIMs' relation around the global minimum. The problem will be cast in one that invites direct use of random sketching. Such application to our context is discussed in Section 3.3 that will lead to a very concrete down-sample strategy. Theoretical guarantees will be provided also in this section. To execute this strategy, practical considerations about sampling choices also play a vital role, and they are discussed in Section 3.4. In Section 4, we apply this general program to the potential reconstruction problem for the Schrödinger equation, and we conclude the article in Section 5.

2 Preview of technical preparations

Two main bodies of technical preparation for the current work are matrix sketching techniques rooted in randomized numerical linear algebra (RNLA), and sampling algorithms, rooted in Bayesian problems, that will be leveraged to implement the downsampling. We recall these tools in this section and unify notations.

2.1 Matrix Sketching in RNLA

RNLA sees its biggest impact in big data applications, where large data sets, that usually exceed RAM capacities, need to be stored and analyzed quickly. Techniques developed within the domain of RNLA typically target at accessing and assessing a subset of data that is reduced in size but still representative, through "sketching", see [24, 27, 38, 20] and references therein.

The technique most relevant to our context is the simple computation of matrix-matrix product: how to compute $B := A^\top A \in \mathbb{R}^{K \times K}$ efficiently? In the regime where A is a tall but skinny matrix, id est (i.e.) when it attains significantly many more rows than its number of columns, matrix A significantly outsizes matrix B, suggesting information is condensed. A Monte Carlo based method can be proposed as a sketching mechanism that sketches rows of A. In the following, we lay out an obvious generalization of this well-established result from RNLA [24, 27] which generalizes the treatment of A, a tall skinny matrix to a quasimatrix defined on measures (hence potentially having uncountable infinitely many rows).

Definition 2.1 Let μ be a probability measure on Ω, and consider a function $A : \Omega \times \mathbb{R}^K \to \mathbb{R}$ such that $A(u, \cdot)$ is linear in $\cdot$ for all u, and that $\|A(\cdot, p)\|_{L^2(\Omega;\mu)} < \infty$ for all $p \in \mathbb{R}^K$. Furthermore, at any fixed $u \in \Omega$, denote $A_{u,:} \in \mathbb{R}^{1 \times K}$ the function A's evaluation ('row') as $A_{u,:} : p \mapsto A_{u,:} \cdot p = A(u, p)$. The Frobenius norm for this matrix is defined as $\|A\|_F^2 = \int_\Omega \|A_{u,:}\|_2^2 \, d\mu(u)$ and the quasimatrix product is defined in the typical manner

$$\mathbb{R}^{K \times K} \ni B = A^\top A = \int_\Omega A_{u,:} \otimes A_{u,:} \, d\mu(u).$$

By definition, B takes on an integral form, and thus can be rewritten into an expectation. More precisely, define the random variable

$$X = \frac{1}{\pi(u)} A_{u,:}^\top A_{u,:}, \quad \text{with} \quad u \sim \pi\mu \text{ for a probability density } \pi \text{ on } (\Omega, \mu).$$

Then $B = \mathbb{E}(X)$. It is then a standard Monte Carlo technique to replace the integral by sample averages:

$$B \approx \frac{1}{c} \sum_{j=1}^{c} X_j, \quad \text{where} \quad X_j \sim X \quad \text{is a drawing.} \tag{2.1}$$

We can summarize this proposal in the following algorithm:

Algorithm 1 BasicMatrixMultiplication; extended from [24, Algorithm 3]

Input: $\Omega \times K$ quasimatrix A, a finite sample size $c \ll |\Omega|$ and probability measure $\pi\mu$ on Ω.
Output: Matrix $\mathsf{C} \in \mathbb{R}^{c \times K}$ such that $\mathsf{C}^\top \mathsf{C} \approx \mathsf{A}^\top \mathsf{A}$.
1: **for** $j = 1, ..., c$ **do**
2:     Sample $u_j \sim \pi\mu$ i.i.d.;
3:     Set the j-th row of C as $\mathsf{C}_{j:} = \mathsf{A}_{u_j,:}/\sqrt{c\pi(u_j)}$.
4: **end for**
5: **return** C and $\mathsf{C}^\top \mathsf{C}$.

Clearly, this algorithm is arrived simply by setting $\mathsf{X}_j = \mathsf{C}_{j:}^\top \mathsf{C}_{j:}$ in (2.1). To justify the algorithm, the approximation sign in (2.1) can be made more precise, and the dependence on c and π can be spelled out explicitly, largely by deploying central limit theorem and various application of the Chernoff estimate. It is worth noting that the random variable here X is a matrix instead of a scalar, so the application of concentration inequality needs caution. Nevertheless, a clever choice of π allows derivation of the following theorem. The proof tightly follows the (finite dimensional) matrix setting in RNLA literature, especially in [24, Theorem 7]:

Theorem 2.1 *Let A be a $\Omega \times K$ quasimatrix with the space Ω with a probability measure μ. Fix a small, finite number $c \ll |\Omega|$ and consider a probability density π on $(\Omega; \mu)$, for which there exists a $\beta \in (0, 1]$ with*

$$\pi(u) \geq \beta \frac{\|\mathsf{A}_{u,:}\|_2^2}{\|\mathsf{A}\|_F^2},$$

and let the matrix $\mathsf{C} \in \mathbb{R}^{c \times K}$ be constructed by Algorithm 1. Then $\mathsf{C}^\top \mathsf{C}$ approximates $\mathsf{A}^\top \mathsf{A}$ with high precision and high probability:

$$\mathbb{P}\left(\|\mathsf{A}^\top \mathsf{A} - \mathsf{C}^\top \mathsf{C}\|_F \leq \frac{1 + \sqrt{8\beta^{-1}\log(\delta^{-1})}}{\sqrt{\beta c}}\|\mathsf{A}\|_F^2\right) \geq 1 - \delta. \qquad (2.2)$$

Here δ is any prescribed failure rate, and $\mathbb{P}$ is taken over all drawings of C.

The theorem states that if the rows of A are chosen proportional to its "volume" – the L_2 norm of the row – then with high probability $(1 - \delta)$, the approximation of B by $\mathsf{C}^\top \mathsf{C}$ is accurate, with the error of the Frobenius norm decaying in the format of $\sqrt{\log(\delta^{-1})/c}$, where c is the chosen number of columns.

The optimal choice of the sampling strategy is to set $\pi(u) = \|\mathsf{A}_{u,:}\|_2^2/\|\mathsf{A}\|_F^2$. Then one has $\beta = 1$, and the error term in (2.2) achieves its minimum. Suppose we set $\delta = 0.01$, then noting $\log(\delta^{-1})$ only gives 2 and is an $O(1)$ number, having the error to be $\epsilon\|\mathsf{A}\|_F^2$ requires $c \geq \frac{O(1)}{\epsilon^2}$. This is an expected MC sampling rate.

2.2 Sampling Algorithms

Sampling is the class of tasks aimed at drawing representative samples from a desired distribution (sometimes referred to as target distribution). Throughout this section, we denote $\tilde{\mu}$ as our desired distribution. This task is frequently called upon in the context of Bayesian sampling, where the target distribution is the posterior distribution $\tilde{\mu}(u) \doteq \mu_{\mathrm{pos}}(u|y) \propto \mu_{\mathrm{pr}}(u)l(y|u)$, and thus a drawing from this distribution provides one solution to the associated inverse problem. In general, due to the positivity of a probability measure, we denote the target distribution

$$\tilde{\mu}(u) \propto e^{-\Phi(u)}, \tag{2.3}$$

where Φ is sometimes referred to as the potential, and $\propto$ means that $\tilde{\mu}$ is normalized to be integrable to 1.

Classical methods are predominantly of Markov Chain Monte Carlo (MCMC) type. The strategy is to design a Markov chain whose invariant measure is the target distribution. When a sample walks through this Markov chain, in time, the distribution of the sample converges to the target distribution. Most well-known examples include Langevin Monte Carlo, Hamiltonian Monte Carlo, and Metropolis-Hasting LMC, and so on [13,34,11,9,25,4,16].

Another sampling paradigm that has recently attracted significant research interest is the ensemble-type method. Originally developed in the context of data assimilation [32,17], this approach has since been adapted to address sampling problems. Notable examples include the Ensemble Kalman Sampler (EKS) [18] or the Consensus Based Sampler (CBS) [8]. These methods evolve an entire ensemble of samples simultaneously through interactive dynamics. The interaction mechanism encodes communication among particles and is carefully crafted to ensure desirable properties, such as being gradient-free or affine-invariant. This remains an active area of research, with non-asymptotic convergence theory still under development.

In our setting, we have the flexibility to choose among various sampling methods, making both classical MCMC approaches and more recently developed ensemble-based methods potentially valuable. Since our goal involves selecting a subset of samples $u \in \Omega_c$, methods that evolve the entire ensemble are directly relevant. We provide further discussion of the EKS and the CBS below.

EKS Sampling. EKS can be viewed as an ensemble version of the Langevin dynamics. It allocates computational resources to update c samples of $\{u_j\}_{j=1}^c$ simultaneously:

$$\mathrm{d}u_j = -C(U)\nabla\Phi(u_j)\,\mathrm{d}t + \sqrt{2C(U)}\,\mathrm{d}W_j, \tag{2.4}$$

where $C(U) = c^{-1}\sum_j(u_j - \bar{u}) \otimes (u_j - \bar{u})$ is the empirical covariance matrix between the particles, and $\bar{u} = c^{-1}\sum_{j'} u_{j'}$ is the mean. W_j are independent

and identically distributed Brownian motions. Often Φ takes on a quadratic form: $\Phi(u) = \frac{1}{2}\|f(u) - d\|^2$, then if f is mildly nonlinear,

$$C(U)\nabla\Phi(u_j) = \frac{f(u_j) - d}{c}\sum_{j'}(u_{j'} - \bar{u}) \otimes (u_{j'} - \bar{u}) \cdot \nabla f(u_j)$$

$$\approx \frac{f(u_j) - d}{c}\sum_{j'}(u_{j'} - \bar{u}) \otimes (f(u_{j'}) - \bar{f}), \qquad (2.5)$$

where the mild nonlinearity of f allows us to approximate $\nabla f(u_j)$ by a constant for all u_j. $\bar{f} = \frac{1}{c}\sum_j f(u_j)$ is used to denote the ensemble average of f evaluation. Under this weakly nonlinear assumption, the implementation of (2.4) is gradient free, and thus achieves a desired property.

When Φ is Lipschitz-smooth, it was shown in [14] and [37] that the mean-field limit of (2.4) when $c \to \infty$ is:

$$\partial_t\rho = \nabla \cdot (\rho C(\rho)\nabla\Phi) + \text{tr}(C(\rho)D^2\rho),$$

and for this equation, it is straightforward to check that $\rho \propto e^{-\Phi}$ is an invariant measure. When Φ is strongly convex, it was also shown in [18] that this PDE converges exponentially fast.

In summary, when denoting $\rho_c = \frac{1}{c}\sum \delta_{u_j}$ the empirical distribution, then for large enough c and t one has $\rho_c \approx \tilde{\mu}$, and $\{u_j\}$ are regarded as samples drawn from the target distribution $\tilde{\mu} \propto e^{-\Phi}$.

CBS Sampling. CBS was introduced in [8] as another method to draw a set of samples simultaneously from a target distribution. It relies on the Laplace principle [35]. A set of c particles $\{u_j\}_{j=1}^c$ evolve according to

$$\mathrm{d}u_j = -(u_j - \mathcal{M}_\beta(\rho_t^c))\,\mathrm{d}t + \sqrt{2(1+\beta)\Gamma_\beta(\rho_t^c)}\,\mathrm{d}W_j, \qquad (2.6)$$

where $\rho_t^c = \frac{1}{c}\sum_{j=1}^c \delta_{u_j(t)}$ is the empirical distribution. $\mathcal{M}_\beta(\rho)$ is the weighted mean parameterized by β: $\mathcal{M}_\beta(\rho) := \mathcal{M}(L_\beta\rho) = \int u\,(L_\beta\rho)(\mathrm{d}u)$ with $L_\beta\rho = \frac{\rho e^{-\beta\Phi}}{\int \rho e^{-\beta\Phi}\,\mathrm{d}u}$ being the weighted version of ρ and $\mathcal{M}$ operator takes the mean of a probability distribution. In the $\beta \to \infty$ limit, $L_\beta\rho$ converges to a Dirac delta centered on the global minimum of Φ over the support of ρ, and thus $\mathcal{M}_\beta(\rho) \to \text{argmin}_u \left.\Phi\right|_{\text{supp}(\rho)}$. The second term introduces stochastic deviations in proportion to the covariance of the weighted distribution

$$\Gamma_\beta(\rho) := \Gamma(L_\beta\rho) := \int (u - \mathcal{M}(L_\beta\rho)) \otimes (u - \mathcal{M}(L_\beta\rho))\,(L_\beta\rho)(\mathrm{d}u)$$

and allows exploration of the distribution landscape. In the mean field limit $c \to \infty$, the particle distribution follows

$$\partial_t\rho = \nabla \cdot ((u - \mathcal{M}(L_\beta\rho))\,\rho + (1+\beta)\Gamma_\beta(\rho)\nabla\rho) .$$

Under certain conditions [8], one can show the steady state of this equation is a Gaussian approximation of the target distribution around its global maximum, and the PDE solution converges to it exponentially fast. Furthermore, in [33] the author links this process with Langevin dynamics, viewing it as a gradient-free relaxation.

Greedy Sampling. All of the sampling strategies discussed above can be further improved. Within the MCMC framework, for instance, sampling algorithms can be enhanced by incorporating a selection mechanism in which proposed samples are accepted or rejected based on a prescribed criterion. A classical example is the use of the Metropolis–Hastings (MH) algorithm as a post-processing step to retain only "good" samples. This added step incurs minimal computational cost but helps mitigate bias introduced by the MCMC procedure.

Similar strategies can also be applied to ensemble-based methods. However, in contrast to the well-established use of MH in MCMC, the development of such correction mechanisms for ensemble methods is still limited. One notable example is the recent introduction of a Metropolis adjustment to correct bias in ensemble-based sampling [36].

It is important to note that the introduction of the Metropolis–Hastings (MH) step is primarily aimed at correcting sampling bias. However, other acceptance criteria tailored to the specific problem can also be employed. In our case, for instance, we evaluate the convexity of the down-sampled Hessian and retain or discard samples based on whether they lead to an improvement in convexity—measured by a selection criterion such as the inverse condition number of the Gauss–Newton Hessian, or its minimum eigenvalue. This simple yet effective strategy is summarized in Algorithm 2 and serves to guide the ensemble evolution toward more favorable configurations through early stopping of the sampling algorithm.

Algorithm 2 Greedy Sampling

Input: initial sample $\{u_j\}_{j=1,...,c}$, sample update rule $R : \Omega^c \to \Omega^c$, number of iterations $I > 0$, a quantity of interest to be maximized Q
Output: updated sample $\{u_j\}_{j=1,...,c}$ with improved evaluation criterion.

1: **for** $i = 1, ..., I$ **do**
2:     Generate sample update: $\{v_j\}_{j=1,...,c} = R(\{u_j\}_{j=1,...,c})$.
3:     **if** $\mathsf{Q}(\{v_j\}) > \mathsf{Q}(\{u_j\})$, **then** Update $\{u_j\}_{j=1,...,c} \leftarrow \{v_j\}_{j=1,...,c}$
4:     **end if**
5: **end for**
6: **return** sample $\{u_j\}_{j=1,...,c}$.

---

3 The general program

Having reviewed both matrix sketching techniques and sampling algorithms, we now return to our qualitative experimental design problem and apply these

tools to address it. Specifically, our goal is to identify suitable experimental setups that preserve data sensitivity for parameter reconstruction (1.2), even when the data is down-sampled (1.4). To achieve this, we reformulate the task as a sketching problem over the FIM, allowing us to leverage results such as Theorem 2.1 for theoretical guarantees. This reformulated problem, when executed numerically, is coupled with a sampling strategy. In particular, ensemble-based sampling methods—such as those described in (2.4) and (2.6)—are employed to guide the selection process.

In the following sections, we begin by analyzing the structure of the FIM, which lays the foundation for applying sketching techniques. This is followed by the introduction of sampling methods as an algorithmic strategy for selecting informative data points. Additional practical considerations are discussed in subsection 3.4.

3.1 FIM structure

Without any specifics, it is impossible to characterize the global behavior of the landscape of the loss function for a generic nonlinear inverse problem in (1.2). Nevertheless, when we confine ourselves to the vicinity of a fixed parameter value p_*, data sensitivity can be quantified by the conditioning of the FIM (1.6). Drawing from linear algebra, FIM that has small conditioning number and relatively big eigenvalues are tied to problems that are sensitive to data, and are thus preferred.

In order to define the FIM for very general sets Ω, we require the following technical assumptions in accordance to [29]:

Assumption 1 *Equip the space Ω of admissible experimental designs with a probability measure μ. Then let that the additive measurement error $\eta \sim \mathcal{N}(0, \Gamma)$ follows a centered Gaussian distribution error with self-adjoint, positive covariance operator $\Gamma : L^2(\Omega, \mu) \to L^2(\Omega, \mu)$ of trace class.*

Moreover, assume that F is Fréchet differentiable, and the image of its Fréchet derivative is contained in the Cameron-Martin space corresponding to Γ.

In the generic form of (1.2), the formula for FIM, denoted as $\mathcal{I}(\Omega)$, can be recasted. Denoting $J := \Gamma^{-1/2} G(p_*)$:

$$\mathcal{I}(\Omega) = \int_{\Omega} (\Gamma^{-1/2} G(p_*))_{u,:} (\Gamma^{-1/2} G(p_*))_{u,:}^{\top} \, \mathrm{d}\mu(u) =: \int_{\Omega} J_{u,:} J_{u,:}^{\top} \, \mathrm{d}\mu(u), \quad (3.1)$$

3.2 Setup and Sampling Perspective

To proceed we now make two assumptions that outline the setting in which we operate:

Assumption 2

(A2.1) There is an underlying ground truth parameter p_* such that $y = \mathcal{F}(p_*) + \eta$.
(A2.2) The FIM $\mathcal{I}(\Omega)$ at the ground truth parameter p_* is positive definite with reasonably high inverse condition number $c_{\mathrm{inv}}^{\Omega}$ and minimal eigenvalue $\lambda_{\min}^{\Omega}$.

These assumptions outline the setting in which we operate. Assumption (A2.1) states that the measurements are generated by the true model that we aim to recover. Assumption (A2.2) is introduced to ensure that the data contains sufficient information to enable successful reconstruction when all measurements are used. A violation of this assumption indicates that the problem is intrinsically ill-posed, in which case further structural improvements are necessary before addressing experimental design questions, and is out of the scope of the current paper.

This formulation places us within the framework described in Section 2.1. Preserving data sensitivity under down-sampling now translates into selecting rows from the matrix $G_* := G(p_*)$ such that

$$\int_{\Omega} J_{u,:} J_{u,:}^{\top} d\mu(u) \approx \frac{1}{c} \sum_{u \in \Omega_c} J_{u,:} J_{u,:}^{\top} \tag{3.2}$$

holds with high probability.

3.3 Experimental Design through Sampling

In light of (3.2), we deploy Algorithm 1 and obtain the following down-sampling strategy:

$$y_c = \left\{ \frac{y(u)}{\sqrt{c\pi(u)}} \right\}_{u \in \Omega_c} \quad \text{with} \quad |\Omega_c| = c, \quad \text{and} \quad u \sim \pi\mu. \tag{3.3}$$

The associated FIM at the global ground truth parameter then becomes:

$$\mathcal{I}(\Omega_c) = \sum_{u \in \Omega_c} \frac{1}{c\pi(u)} J_{u,:} J_{u,:}^{\top}.$$

As suggested in Theorem 2.1, there is an optimal sampling strategy with each row being selected with a rate proportional to its volume. In our context, this optimal strategy is:

$$\tilde{\mu} := \tilde{\pi}\mu, \quad \text{with} \quad \tilde{\pi}(u) \propto \|J_{u,:}\|_2^2.$$

A specific case is when the design space $\Omega \subset \mathbb{R}^L$ is continuously parameterized and of finite Lebesgue measure. Assuming for simplicity of the presentation that μ is the uniform distribution over Ω. Then $\tilde{\mu}$ can be characterized as:

$$\tilde{\mu}(u) = \frac{1}{Z} e^{-\Phi(u)} \quad \text{with} \quad \Phi(u) := -\log(\|J_{u,:}\|_2^2). \tag{3.4}$$

We show below that if π is close to the optimal $\tilde{\pi}$, having enough samples ensures the local sensitivity of down-sampled data with high probability.

Theorem 3.1 *Consider an inverse problem that satisfies assumptions (A2.1)–(A2.2) and let the re-weighted data be constructed as in (3.3), where the sampling probability density $\pi(u)$ on Ω satisfies*

$$\pi \geq \beta\tilde{\pi}, \tag{3.5}$$

for some $\beta \in (0,1]$. Furthermore, assume that $\|J_{u,:}\|_2$ is bounded for every $u \in \Omega$, then with a sufficiently large c, the forward map $F|_{\Omega_c}$ is locally sensitive w.r.t. p at p_* with a high probability. More specifically, for any failure probability $\delta \in (0,1)$ and any error tolerance $\varepsilon \in \left(0, \lambda_{\min}^{\Omega}\right)$, a choice of the sample size

$$c \geq \|J\|_F^4 \frac{(1 + \sqrt{8\beta^{-1}\log(\delta^{-1})})^2}{\beta\varepsilon^2} \tag{3.6}$$

assures that with probability at least $1 - \delta$, the inverse condition numbers $c_{\mathrm{inv}}^{\Omega}, c_{\mathrm{inv}}^c$ and minimum eigenvalues $\lambda_{\min}^{\Omega}, \lambda_{\min}^c$ of $\mathcal{I}(\Omega)$ and $\mathcal{I}(\Omega_c)$, respectively, satisfy

$$c_{\mathrm{inv}}^c \geq c_{\mathrm{inv}}^{\Omega} \frac{\lambda_{\min}^{\Omega} - \varepsilon}{\lambda_{\min}^{\Omega} + \varepsilon} \qquad and \qquad \lambda_{\min}^c \geq \lambda_{\min}^{\Omega} - \varepsilon > 0.$$

Proof Noting that

$$\lambda_{\min}^c \geq \lambda_{\min}^{\Omega} - |\lambda_{\min}^{\Omega} - \lambda_{\min}^c| \geq \lambda_{\min}^{\Omega} - \|\mathcal{I}(\Omega) - \mathcal{I}(\Omega_c)\|_F, \tag{3.7}$$

we are to bound the second term. Using Theorem 2.1, it is straightforward to see that with probability at least $1 - \delta$

$$\|\mathcal{I}(\Omega) - \mathcal{I}(\Omega_c))\|_F = \|J^\top J - C^\top C\|_F \leq \frac{1 + \sqrt{8\beta^{-1}\log(\delta^{-1})}}{\sqrt{\beta c}}\|J\|_F^2.$$

To achieve $\lambda_{\min}^c \geq \lambda_{\min}^{\Omega} - \varepsilon$, according to (3.7), we need to bound the term above by ε, which yields the choice of c as given by Theorem 2.1. Similar estimations of the maximum eigenvalue yields the bound for the inverse condition number $c_{\mathrm{inv}}^c = \frac{\lambda_{\min}^c}{\lambda_{\max}^c}$.

3.4 Practical considerations

According to Theorem 3.1, we are looking for c i.i.d samples from the optimal probability distribution $\tilde{\mu} := \tilde{\pi}\mu$.

A natural application of EKS provides us the following sampling strategy: Set c interactive samples $U = \{u_j\}_{j=1,\dots,c}$ uniformly at initial time, noting that the strategy can readily be adapted for Gaussian μ. We then evolve them according to

$$\mathrm{d}u_j = \sum_{j'} D_{j,j'} u_{j'}\,\mathrm{d}t + \sqrt{2C(U)}\,\mathrm{d}W_j,$$

where the first term contains the approximation to $C(U) \cdot \nabla_u \Phi(u_j)$:

$$
\begin{aligned}
-C(U) \cdot \nabla_u \Phi(u_j) &= -\left(\frac{1}{c} \sum_{j'} (u_{j'} - \bar{u}) \otimes (u_{j'} - \bar{u}) \right) \cdot \nabla_u \Phi(u_j) \\
&= \frac{2}{c \| J_{u_j,:} \|_2^2} \sum_{j'} \left(D_u J_{u_j,:}(u_{j'} - \bar{u}) \right)^\top J_{u_j,:}(u_{j'} - \bar{u}) \\
&\approx \frac{2}{c \| J_{u_j,:} \|_2^2} \sum_{j'} \left(J_{u_{j'},:} - \overline{J(U)} \right)^\top J_{u_j,:} u_{j'} \\
&=: \sum_{j'} D_{j,j'} u_{j'} .
\end{aligned}
$$

The approximation in the second to last line originates from approximating the gradient term by a difference in analogy to (2.5), with $\overline{J(U)} = \frac{1}{c} \sum_j J_{u_j,:}$, and the fact that the $\bar{u}$ term vanishes. Running this SDE forward in time using the classical Euler-Maruyama method gives:

$$
u_j^{t_{k+1}} = u_j^{t_k} + \Delta t_k \sum_{j'} D_{j,j'}^{t_k} u_{j'}^{t_k} + \sqrt{2 \Delta t_k C(U^{t_k})} \zeta_j^{t_k},
$$

with $\zeta_j^{t_k} \sim N(0, I)$ independent and identically distributed and adaptive time step $\Delta t_k = \frac{\Delta t_0}{\| D^{t_k} \|_F + \varepsilon}$ in dependence of the difference matrix $D^{t_k} = (D_{j,j'}^{t_k})_{j,j'}$ for some $\varepsilon > 0$ as proposed in [23,18].

Application of CBS is straightforward. As in [8] we deploy the forward in time discretization using an exponential integrator:

$$
u_j^{t_{k+1}} = e^{-\Delta t} u_j^{t_k} + (1 - e^{-\Delta t}) \mathcal{M}_\beta(\rho_{t_k}^c) + \sqrt{(1 - e^{-2\Delta t})(1 + \beta)} \Gamma_\beta(\rho_{t_k}^c) \zeta_j^{t_k} .
$$

Remark 3.1 (On sampling accuracy) One key drawback of ensemble based method is the lack of non-asymptotic convergence rate. The samples provided by these methods are not necessarily the best samples drawn from the optimal distribution. Meanwhile, though the bound for c in Theorem 3.1 is explicit, the constants depend on quantities that are not known a-priori ($\| J \|_F^2 = \int \| J_{u,:} \|_2^2 d\mu(u)$ or the minimum eigenvalue $\lambda_{\min}^\Omega$), bringing another uncertainty to set parameters.

However, it is important to note that accurately sampling from the target distribution $\tilde{\mu}$ is not our ultimate goal; rather, our primary objective is to improve the conditioning of the FIM. As a result, we are willing to tolerate imperfect sampling if it leads to better conditioning.

Remark 3.2 The optimal sampling density is $\tilde{\pi}(u) \propto \| J_{u,:} \|_2^2$, but Theorem 3.1 does allow us to set $\beta < 1$. In certain situations, the underlying inverse problem structure and some prior knowledge of $\mathcal{F}$ could potentially give some insights. For instance, in certain cases, one can show $J_{u,:}$ is uniformly bounded above and below for all u, and the bounds are tight enough. When this happens, choosing a uniform density π may already give a satisfying sampling result. This is confirmed in our numerical test, seen in Figure 4.6.

4 Application to the Schrödinger potential reconstruction

In this section, we demonstrate the potential of the proposed algorithm on a specific example[1]: inverse steady state Schrödinger equation. The spatial domain is set to be $X = [-1, 1]^2 \subset \mathbb{R}^2$ and the time-independent PDE with non-negative source term $\gamma \in C_+^\infty(X)$ writes as:

$$(-\Delta + p)u_p = \gamma \quad x \in X, \tag{4.1}$$
$$u_p = 0 \quad x \in \partial X.$$

In the forward problem with a fixed source $\gamma \neq 0$, existence of a positive solution u_p for a fixed non-negative parameter $p \in C^\infty(X)$ follows from standard elliptic theory.

The inverse problem is to reconstruct the potential p from measurements of the observable solution u_p. Clearly inferring p becomes trivial when the full noise free u_p is known for only one source $\gamma > 0$: one has $p = \frac{\gamma + \Delta u_p}{u_p}$ pointwise in X. The problem arises in the finite dimensional setting: Only a finite number of potentially noisy measurements of u_p is taken and p is parameterized by finite many parameters. The goal is to find the optimal experimental setting (measuring location) for best inferring p.

Parameter Discretization. Let $\{\phi_k : X \to \mathbb{R}\}_{k=1,...,K}$ be a given finite set of basis functions on X, and our admissible set for p is assumed to be:

$$\mathcal{A} := \left\{ p : X \to \mathbb{R}_0^+, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto p(x) = \sum_{k=1}^K p_k \phi_k(x_1, x_2) \right.$$
$$\left. \text{for some } p_k \in \mathbb{R}, k = 1, ..., K \right\}.$$

In the numerical examples in Section 4.1.1, we used $K = 9$ with corresponding basis

$$\{\phi_{k_1, k_2}(x_1, x_2) = \cos(k_1 \pi x_1) \cos(k_2 \pi x_2)\}_{k_1, k_2 = 0, 1, 2}.$$

Space discretization. To numerically realize the PDE solution, we use its numerical solution computed on equidistant Cartesian grid $\{\xi_n, n = 1, ..., (N_x + 1)^2\}$, where we set N_x cells in every direction.

4.1 Fixed Source Term γ

In the first set of experiments, we fix the source term at $\gamma = 10^4$ for all trials. Based on the above considerations, this well-controlled setting should be sufficient for successful reconstruction, provided that the data is appropriately chosen.

---

[1] Code for generating the examples is available upon request.

Data. Without loss of generality, we assume all possible measurements are point-wise measurements, meaning $\mathcal{F}(x,p) = u_p(x)$ for all $x \in X$. We denote the ground truth data generated by the ground truth media $p_* \in \mathcal{A}$ and independent and identically distributed (i.i.d.) standard normal noise $\eta(x) \sim \mathcal{N}(0,1)$ by

$$\{y(x) = \mathcal{F}(x,p_*) + \eta(x) = u_{p_*}(x) + \eta(x)\}_{x \in \Omega},$$

so Assumption (A2.1) is satisfied. The question related to experimental design now translates to a search for the number c and locations $\Omega_c \subset X$ so to make the associated down-sampled optimization problem locally strictly convex.

Numerical full measurement setup. The full measurement setup considers uniformly weighted measurements taken at all inner vertices, i.e.

$$\Omega = \{\xi_n\}_{n=1}^{(N_x+1)^2} \backslash \partial X. \qquad \text{and} \qquad \mu = \frac{1}{|\Omega|} \quad \text{with} \quad |\Omega| = N = (N_x - 1)^2. \tag{4.2}$$

Computation of $J_{x,:}$. Evaluation of $\tilde{\pi}$ requires computation of the gradient $J_{x,:} = \nabla_p u_{p_*}(x)$ for all $x \in \Omega$. In Appendix A we spell out the details of deploying an adjoint based method to compute the gradient. For example, the k-th entry of the gradient reveals

$$[J_{x,:}]_k = [\nabla_p u_{p_*}(x)]_k = \langle g^{(x)}, \phi_k u_{p_*} \rangle_{L^2(X)}, \tag{4.3}$$

where $g^{(x)}$ satisfies the adjoint equation

$$-\Delta g^{(x)} + p_* g^{(x)} = -\delta_x \text{ on } X, \quad g^{(x)} = 0 \text{ on } \partial X. \tag{4.4}$$

This demonstrates Assumption 1. Computationally both the forward and adjoint solvers are conducted by a finite element approach with nodal basis defined on an equidistant Cartesian grid $\{\xi_n\}$.

4.1.1 Importance Sampling Distributions

As a numerical study, we first run the equation with fine discretization, and plot out the optimal sampling strategy $\tilde{\mu}$. In the four examples shown, the $K = 9$ ground truth parameters are set according to Table 4.1. As shown in Figure 4.1, the optimal sampling distribution $\tilde{\mu} := \tilde{\pi}\mu \propto \tilde{\pi}$ shows significant dependence on the underlying ground truth parameter.
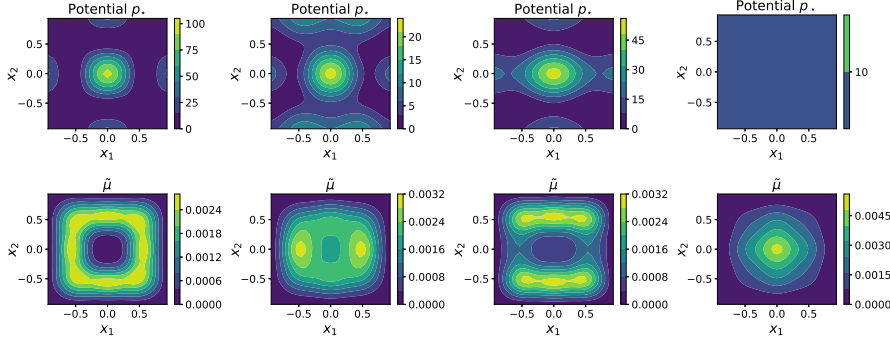
Fig. 4.1: Top row shows four different ground-truth media p_*, and the bottom row shows the optimal sampling distribution $\tilde{\pi}$ for each of them.

System	ground truth parameter
A	$p_*^A = \begin{pmatrix} 13.6 & 10 & 10 \\ 10 & 10 & 10 \\ 10 & 10 & 10 \end{pmatrix}$
B	$p_*^B = \begin{pmatrix} 5.856 & 0.103 & 3.168 \\ 3.7441 & 2.493 & 1.124 \\ 0.9902 & 3.803 & 0.846 \end{pmatrix}$
C	$p_*^C = \begin{pmatrix} 11 & 8.889 & 7.778 \\ 6.667 & 5.556 & 4.444 \\ 3.333 & 2.222 & 1.111 \end{pmatrix}$
D	$p_*^D = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

Table 4.1: Test scenarios to study the optimal sampling strategy $\tilde{\pi}$. The (i, j) entry of the matrix is the coefficient for p_k with $(k_1 = i, k_2 = j)$.

We then scale the parameters by multiplying p_* with a scaling parameter α. Varying the amplitude of α, we observe very different pattern for $\tilde{\mu}$ as well, as shown in Figure 4.2. In this plot, we scale the ground truth distribution by constant ($\alpha = 10$ or 0.1) and we observe very different optimal distribution. Drawn from this numerical observation, we expect $\tilde{\mu}$ to be more centered in the middle when p_* takes on small values, but develop interesting patterns when p_* has a large scaling.

4.1.2 Effect of Sampling

As a proof of concept, we now study the performance of the sensitivity based sampling strategy for its recovery of optimal sensor locations. The inverse
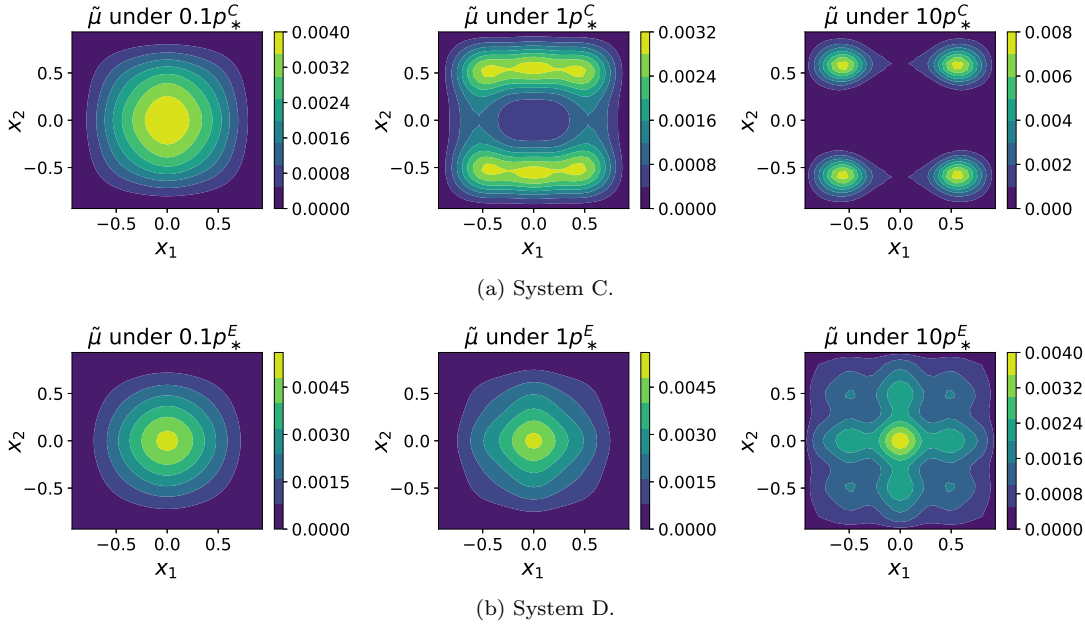
(a) System C.

(b) System D.

Fig. 4.2: Optimal importance sampling distributions $\tilde{\mu}$ for scaling parameters αp_* with $\alpha = 0.1$ (left), $\alpha = 1$ (center) and $\alpha = 10$ (right). The ground truth parameters p_* from System C and D from Table 4.1 are taken.

condition number c_{inv} and the minimal eigenvalue λ_{min} of the FIM are key quantities to be examined, and superscripts will refer to the specific design under investigation.

Effect on sensor locations and minimal FIM eigenvalue. We choose the ground truth parameter of System C in Table 4.1 and use an adapted greedy version, based on the condition number, of EKS in [18] as described in Section 3.4 and a similar adaptation of CBS in [8].

To start, we evaluate the FIM given by the full dataset. In Figure 4.3, with $N_x = 30$, we mark $N = (N_x - 1)^2 = 841$ red dots as the sensor locations and computed the optimal distribution $\tilde{\mu}$. The inverse condition number and minimum eigenvalue of the FIM in this setting are $c_{\mathrm{inv}}^{\mathrm{full}} = 8e{-}4$ and $\lambda_{\mathrm{min}}^{\mathrm{full}} = 0.8 > 0$, and the problem is locally sensitive.

To proceed with down-sampling, we allow only $c = 18 = 2K$ sensor locations. The initial guess was a uniformly weighted normal distribution over Ω and the output is severely worse, with the inverse conditioning and minimal eigenvalue degenerated to $c_{\mathrm{inv}}^{\mathrm{init}} = 1.54e{-}7$ and $\lambda_{\mathrm{min}}^{\mathrm{init}} = 1.48e{-}4$. Both EKS and CBS with greedy selection, after a running of 25 iterations, move the samples to new locations, and improve the conditioning of the weighted FIM to $c_{\mathrm{inv}}^{\mathrm{EKS}} = 2.25e{-}3$ with $\lambda_{\mathrm{min}}^{\mathrm{EKS}} = 2.06$, and $c_{\mathrm{inv}}^{\mathrm{CBS}} = 1.56e{-}3$ with $\lambda_{\mathrm{min}}^{\mathrm{CBS}} = 1.41$,
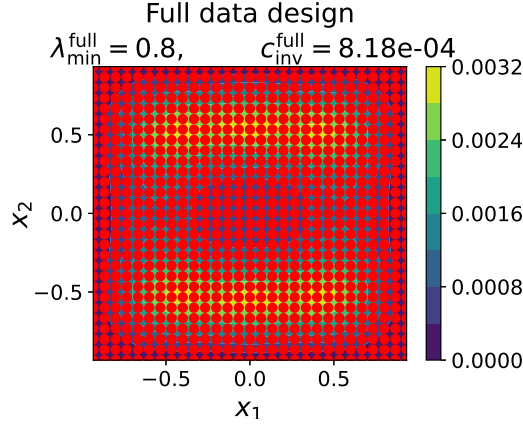
Fig. 4.3: Full Data Setup: Measurement locations (red dots) are located in all grid points. The optimal importance sampling distribution $\tilde{\mu}$ is drawn in the background.

respectively. We also compare these results to a repeated greedy random sampling from the normal distribution that was used to produce the initial guess, with 25 iterations. This yields a design that is informed by the same number of intermediate sensor locations, but attains worse sensitivity values of $(\lambda_{\min}^{\text{normal}}, c_{\text{inv}}^{\text{normal}}) = (5.08e{-}4, 4.05e{-}7)$. The samples drawn from the initial distribution, the iterated solution according to EKS and CBS are all plotted in Figure 4.4, and the evolution of the smallest eigenvalues and the approximation error in the FIM along iteration are plotted in Figure 4.5.

We observe that the inverse condition number and the minimum eigenvalue of the FIM corresponding to designs generated by EKS and CBS are even larger than those obtained using the full dataset. This suggests that incorporating a large number of data points can, in fact, dilute the information—by averaging highly sensitive sensors with less informative ones. In contrast, our experimental design strategy focuses on selecting informative data points and emphasizing them more heavily, thereby amplifying the overall information content.

An interesting numerical discovery is that in this case, the uniformly distributed sensor locations, as depicted in Figure 4.6, also perform well, attaining a minimum eigenvalue and conditioning $(\lambda_{\min}^{\text{init},u}, c_{\text{inv}}^{\text{init},u})$ being $(0.36, 3.4e{-}4)$. Indeed the optimal importance sampling distribution $\tilde{\mu}$ is bounded from above by 0.0031 (in comparison to $\frac{1}{N} \approx 0.0012$ for a uniform distribution). Hence, the uniform distribution in this particular case is a good approximation (with $\beta \leq 0.383$). Starting from uniform distribution, we once again apply greedy EKS, CBS for 25 iterations and can further improve $(\lambda_{\min}^{\text{EKS}}, c_{\text{inv}}^{\text{EKS}})$ and $(\lambda_{\min}^{\text{CBS}}, c_{\text{inv}}^{\text{CBS}})$ to $(1, 77, 1.91e{-}3)$ and $(1.17, 1.24e{-}3)$ respectively. We find that a greedy repeated sampling w.r.t. the uniform distribution improves the conditioning to
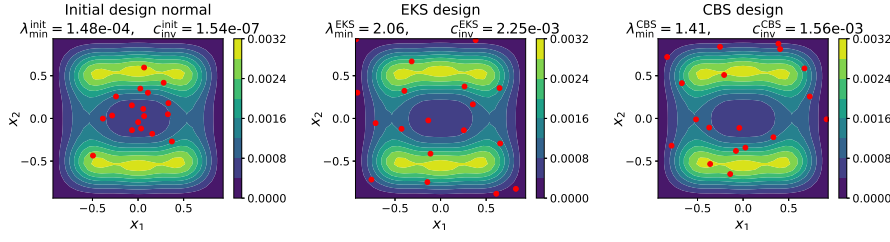
Fig. 4.4: Red markers demonstrate the location of the sensors, with the background plotted as the optimal distribution. The left panel shows the distribution of the initial samples. The middle and the right panel show, respectively, the EKS and CBS samples after 25 iterations. The minimum eigenvalues of the FIM change from $1.48e-4$ to 2.06 and 1.41 and the inverse condition numbers from order $1e-7$ to $1e-3$, ensuring local sensitivity.
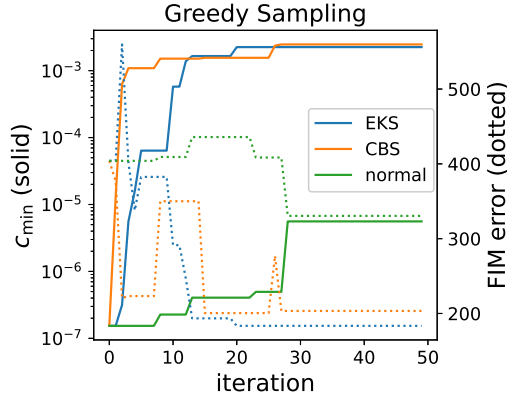

Fig. 4.5: Evolution of minimum eigenvalue (solid lines) and deviation of the down-sampled FIMs from the full data setup in Frobenius norm (dashed lines). Three sampling methods are used: EKS (blue), CBS (orange) and repeated sampling from the initial guess distribution (green), all used in greedy mode. Initial distribution is shared across three sampling methods.

$c_{inv}^{rand} = 9.88e-4$ and $\lambda_{min}^{rand} = 0.84$ in comparison to the initial guess, but does not reach the sensitivity of our proposed designs, as summarized in Table 4.2.

Effect on the square loss function. The sensitivity of the data w.r.t. the parameter is reflected, for example, in the strong convexity of the quadratic cost function $\mathcal{C}(p) = \|y(\cdot) - \mathcal{F}(\cdot, p)\|_{L^2(\Omega, \mu)}^2$, whose minimization serves as a commonly used inversion technique. In what follows, we visualize the landscape of this cost function across the parameter space for different experimental designs, in order to assess the impact of our sampling strategy on data sensitivity and the difficulty of the full nonlinear inversion problem.

Design D	$\lambda_{\min}^D$	c_{inv}^D
full data D^{full}	0.8	$8.18 \cdot 10^{-4}$
normal initial guess	$1.48 \cdot 10^{-4}$	$1.54 \cdot 10^{-7}$
EKS sample	2.06	$2.25 \cdot 10^{-3}$
CBS sample	1.41	$1.56 \cdot 10^{-3}$
greedy normal sampling	$5.08 \cdot 10^{-4}$	$4.05 \cdot 10^{-7}$
uniform initial guess	0.36	$3.4 \cdot 10^{-4}$
EKS sample	1.77	$1.91 \cdot 10^{-3}$
CBS sample	1.17	$1.24 \cdot 10^{-3}$
greedy uniform sampling	0.84	$9.88 \cdot 10^{-4}$

Table 4.2: Comparison of sensitivity measures associated to different designs. Rows below an initial guess refer to sampling starting from this initial design.

For visualization, we confine ourselves to a two-dimensional admissible set with $\mathcal{A} = \{p : X \to \mathbb{R} \mid p(x) = p_1 \cos(x_1) + p_2 \cos(x_2) + 12\}$ and the ground truth parameter $p_*(x_1, x_2) = 1 \cos(x_1) + 10 \cos(x_2) + 12$ and work with noise free synthetic data in the following. The profile of p_* and the optimal importance sampling distribution are depicted in Figure 4.7. The scaling for p_* in the x_1 and x_2 direction is very different, and p_* changes its profile in x_2 direction significantly more. This is in alignment with the extension of the sampling probability.

When the full dataset is used, the loss function is convex, indicating the full dataset contains sufficient information for the recovery, with a conditioning of $c_{\text{inv}}^{\text{full}} = 0.43$ and a minimum eigenvalue being $\lambda_{\min}^{\text{full}} = 47.3$, as shown in Figure 4.8. An initial setup of 8 normally distributed sensor locations shows significantly reduced convexity in the landscape of the objective function, and the inverse condition number becomes 0.01, with a minimum eigenvalue of 2.06. Sampling with a greedy strategy based on the condition number in Figure 4.9 according to EKS and CBS enhances both the convexity and the conditioning dramatically, as plotted in Figure 4.9.

4.2 Source Term Design

In our second set of experiments, we allow the source term to be adjusted as well. In particular, we set:

$$\gamma(x) = \gamma_1 x_1 + \gamma_2 x_2 + 10, \quad \text{with} \quad \vec{\gamma} = (\gamma_1, \gamma_2) \in [-2, 2]^2.$$

Similar to the previous example, the possible measurements are the solution evaluated at points $u^\gamma(x)$. The entire forward map is:

$$\hat{\mathcal{F}}(x, \vec{\gamma}, p) = u^{\vec{\gamma}}(x), \quad \text{and} \quad \hat{y}(x, \vec{\gamma}) = \hat{\mathcal{F}}(x, \vec{\gamma}, p_*) + \eta(x, \vec{\gamma}).$$

The flexibility of x and $\vec{\gamma}$ means we have four dimension of design space: $\hat{\Omega} = \Omega \times [-2, 2]^2$. We endow this space with μ, the uniform distribution,
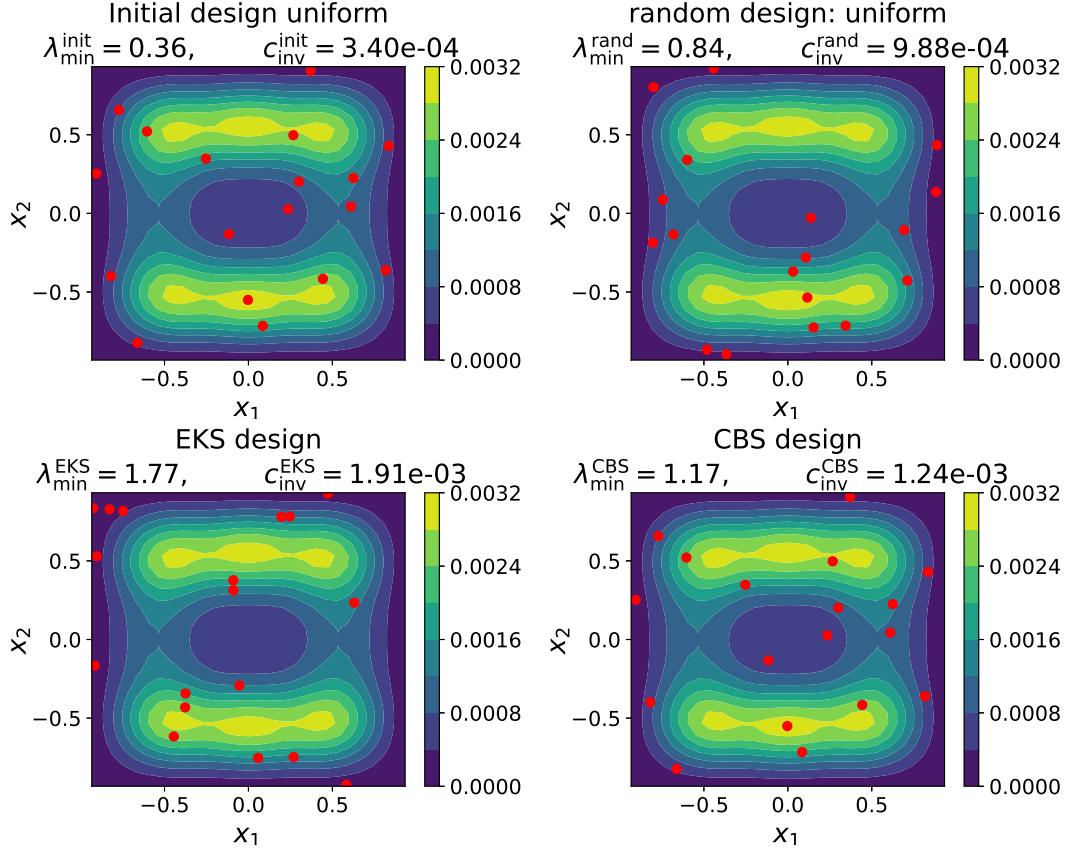
Fig. 4.6: Uniformly distributed initial guess (upper left) of the distribution of the sensors (red dots) in the domain X, where the optimal importance sampling distribution is drawn in the background. Application of greedy EKS (lower left), CBS (lower left) and repeated sampling w.r.t. the uniform distribution changes the sensor distribution and dramatically increases the condition number and minimum eigenvalue of the respective FIM.

and fix the parameter dimension to $K = 9$ again. The continuous $\vec{\gamma}$ space prevents us to compute the full landscape, as well as the normalization constant of $\tilde{\pi}$ exactly, and we use this as an example to demonstrate our method in this setting. Note however, that the eigenvalues of the FIM in our sampling approach depend on the normalization constant through the data reweighting process, which prevents us to utilize the minimum eigenvalue to evaluate and compare sensitivity between reweighted and uniformly weighted designs.

To initiate the program, we sample in the design space using a Gaussian in space Ω and uniform distribution on $[-2, 2]^2$. The sampling strategies, upon a few runs, improve the data sensitivity: the inverse condition number increase from $1e{-}9$ to $1e{-}4$ or $1e{-}3$ depending on the sampling method, and both out-
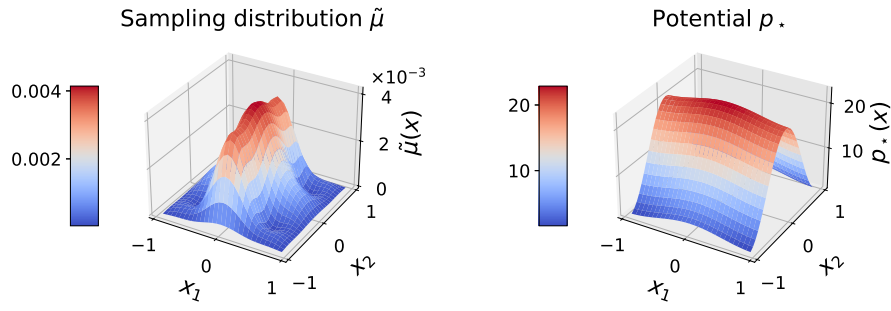
Fig. 4.7: Optimal importance sampling distribution $\tilde{\mu}$ (left) and shape of the ground truth parameter p_* (right) in the two-dimensional setting.
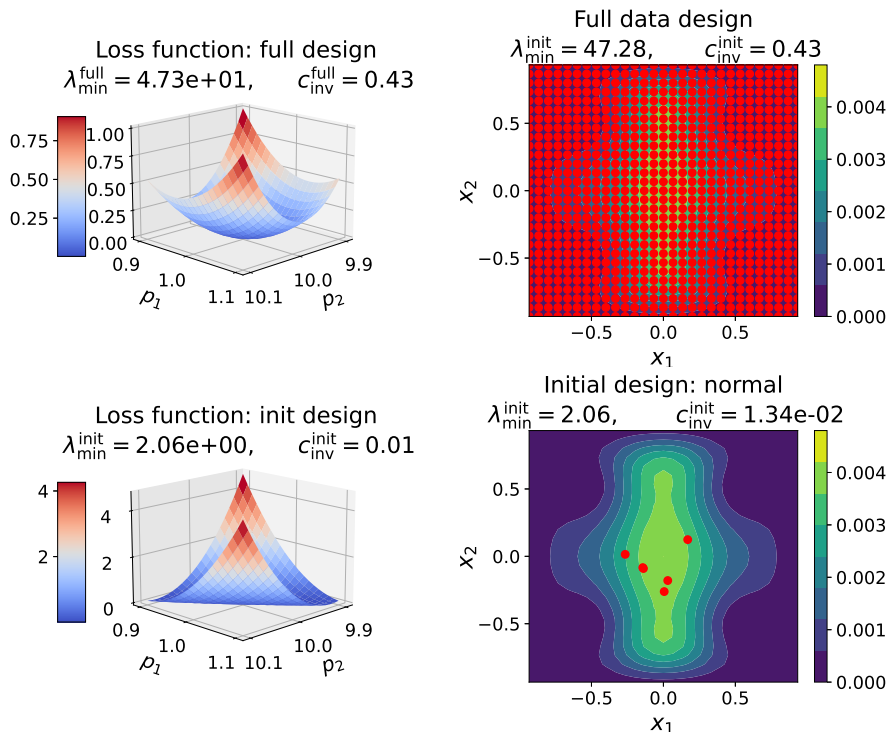

Fig. 4.8: Loss landscapes (left) for different sensor locations (right): full data setup (first row) and normally distributed initial sensor locations (second row).
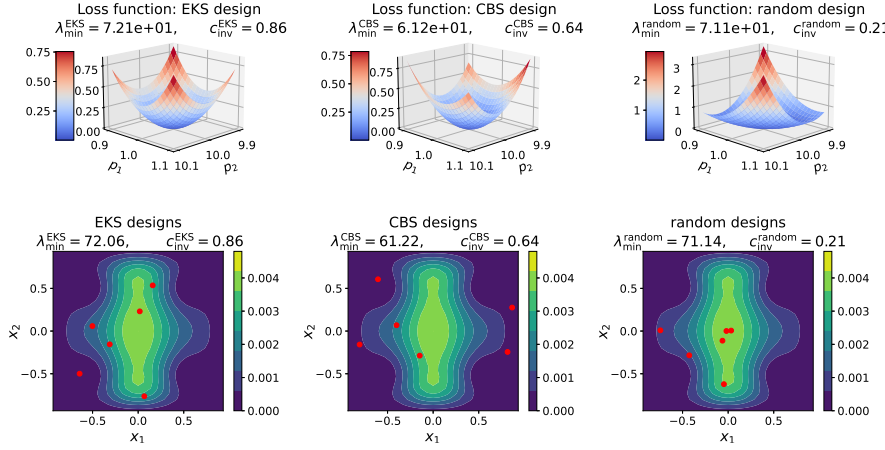
Fig. 4.9: The top row shows the quadratic loss landscapes and the bottom row shows the locations of the samples with the background presenting the optimal distribution. The three panels are results from greedy versions of EKS, CBS and repeated normal sampling from initial guess distribution.

perform the repeated greedy random sampling, as summarized in Table 4.3. The second test is to initiate the program by sampling using uniform distribution on the entire $\hat{\Omega}$. As in the case for a fixed source term, this produces local convexity values $2.25e-4$ for the FIM conditioning, which is already significantly better than the normal initial sampling. Our sampling algorithm, both sampled by EKS and CBS, as well as the greedy uniform sampling can further improve this sensitivity up to an inverse conditioning of order $e-3$. See Table 4.3.

In Figure 4.10, we show the final output of the distribution of the selected data points. The initial guess collects data at $18 = 2K$ experiments with randomly chosen source parameters $\vec{\gamma}$ uniformly sampled in $[-2, 2]$, and corresponding normally distributed sensor locations that are very concentrated in the center. Both EKS and CBS both push these samples to the wider domain, and they return better sensitivity. No clear tendency is observable for the change in the source parameters.

As in the case for a fixed source term, a uniform sensor placement performs well already, yielding local convexity values $2.25e-4$ and for the inverse FIM conditioning, but can be slightly improved by repeated greedy random sampling, or EKS or CBS sampling from the sensitivity based distribution to values exceeding $1e-3$.
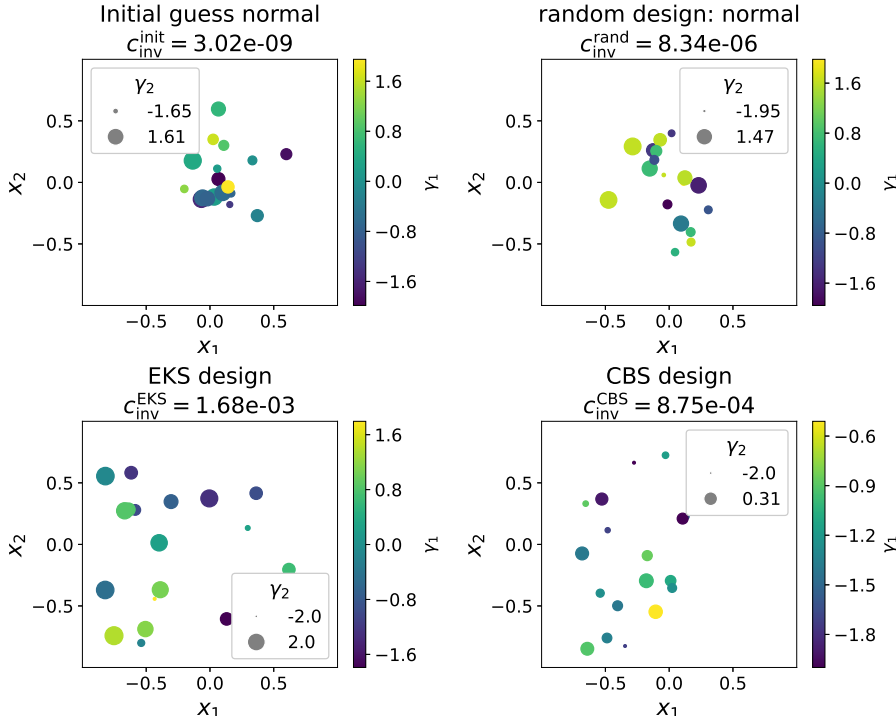
Fig. 4.10: Four different designs, characterized by their sensor locations given by the dot locations, and γ_1, γ_2 values encoded in colour and size of the dots, together with their sensitivity measures: normally distributed initial sensor location guess with uniformly distributed γ_1, γ_2 (upper left), greedy repeated sampling w.r.t. this distribution (upper right), greedy EKS (lower left) and CBS (lower right) w.r.t. the rescaled sensitivity based sampling distribution, after 60 iterations each.

Design D	c_{inv}^D
normal initial guess	$3.02 \cdot 10^{-9}$
EKS sample	$1.68 \cdot 10^{-3}$
CBS sample	$8.75 \cdot 10^{-4}$
greedy normal sampling	$8.34 \cdot 10^{-6}$
uniform initial guess	$2.25 \cdot 10^{-4}$
EKS sample	$1.83 \cdot 10^{-3}$
CBS sample	$1.68 \cdot 10^{-3}$
greedy uniform sampling	$1.16 \cdot 10^{-3}$

Table 4.3: Comparison of the inverse conditioning as a local sensitivity measure emerging from different designs strategies. Rows below an initial guess refer to sampling starting from this initial configuration.

5 Discussion

In this work, we study the question of experimental design of a parameterized
inverse problem with the perspective of preserving the sensitivity of the data
w.r.t. the parameter as a basis for successful numerical reconstruction, at least
locally around a presumed true parameter value. Supposing that the full data
set is sensitive w.r.t. the parameter, we examine how much one can down-
sample the data. Taking the perspective of random sampling rather than an
optimal selection of data, this problem is formulated as a matrix sketching
problem, where a well-studied sketching algorithm from RNLA becomes handy.
The sample size depends on a sampling distribution that reflects the structure
of the forward problem. To draw samples from this distribution, sampling
algorithms such as EKS and CBS are implemented.

The general program described in this article can be applied to a variety
of experimental design / data selection tasks merged from inverse problems.
As a proof of concept, we provide a numerical test using Schrödinger equation
as the forward model. The optimal distribution is problem dependent and is
typically unavailable. In various applications, knowledge of the forward model
can be used to obtain some qualitative estimates.

Following this work, many new questions can be asked. In (Bayesian) op-
timal experimental design [1,39], K- and E-optimality seek to maximize the
inverse condition number or the minimal eigenvalue of inverse of the Bayesian
covariance matrix or the FIM, respectively, i.e. the same quantities we use to
evaluate sensitivity of designs in this work and run the greedy selection. Find-
ing the explicit relation between the two approaches is also one of interesting
future direction.

Our approach suffers from a drawback that is typical for all experimental
design methods: the sampling of designs requires knowledge of the underlying
ground truth parameter p_* to build $\tilde{\pi}$, as demonstrated in Figure 4.1. Sev-
eral strategies have been developed in classical optimal experimental design
literature [1,19] to mitigate this drawback, summarized under sequential ex-
perimental design, some of which can be directly integrated into our approach.
In particular, we see synergies between our approach and the greedy approach
consisting of alternating phases of experimental design and parameter infer-
ence through gradient based optimization.

Finally, we see potential application of our approach to more recently devel-
oped inversion frameworks that rely on a least squares optimization. Examples
of such frameworks can be found in [10,15], where Gaussian processes or neu-
ral networks are incorporated in the inversion process. A detailed derivation
is left for further investigation.

A Appendix: Derivation of the formula for $\nabla_p u_p(x)$

We derive the formula for the gradient $\nabla_p u_p(x)$ of the solution to the Schrödinger equation
w.r.t. the potential p, that we require for the computation of the sampling probabilities.

In the following derivations, all gradients are with respect to x, unless specified otherwise. For a fixed measurement location $x \in X$, we can then define the Lagrange function $\mathcal{L} : \mathcal{A} \times H_0^1(X) \times H_0^1(X) \to \mathbb{R}$ as

$$\mathcal{L}_x(p, u, g) = u(x) + \langle \nabla g, \nabla u \rangle_{L_2(X)} + \langle g, pu \rangle_{L^2(X)} - \langle g, f \rangle_{H^1(X), H^{-1}(X)},$$

where g is the Lagrange multiplier, and $\langle \cdot, \cdot \rangle_{H_0^1(X), H^{-1}(X)}$ denotes the duality bracket in $H_0^1(X) \times H^{-1}(X)$. Using (4.1), one immediately sees $\mathcal{L}_x(p, u_p, g) = u_p(x)$. Therefore, confined on this solution manifold, chain rule gives:

$$\left. \frac{\partial u_p(x)}{\partial p_j} \right|_{p=\hat{p}} = \left. \frac{\partial \mathcal{L}_x}{\partial p_j} \right|_{\substack{p=\hat{p} \\ u=u_{\hat{p}}}} + \left. \frac{\partial \mathcal{L}_x}{\partial u} \right|_{\substack{p=\hat{p} \\ u=u_{\hat{p}}}} \left. \frac{\partial u_p}{\partial p_j} \right|_{p=\hat{p}}.$$

This equation holds valid for arbitrary g, and thus we would like to choose $g = g_x$ such that $\partial \mathcal{L}_x / \partial u = 0$. If so:

$$\left. \frac{\partial u_p(x)}{\partial p_j} \right|_{p=\hat{p}} = \left. \frac{\partial \mathcal{L}_x}{\partial p_j} \right|_{\substack{p=\hat{p} \\ u=u_{\hat{p}}}} = \left. \frac{\partial \langle g_x, pu \rangle_{L^2(X)}}{\partial p_j} \right|_{\substack{p=\hat{p} \\ u=u_{\hat{p}}}}$$

$$= \left. \frac{\partial \langle g_x, \sum_k p_k \phi_k u \rangle_{L^2(X)}}{\partial p_j} \right|_{\substack{p=\hat{p} \\ u=u_{\hat{p}}}} = \langle g_x, \phi_j u_{\hat{p}} \rangle_{L^2(X)}.$$

It remains to compute $g_x \in H_0^1(X)$ for which $\partial \mathcal{L}_x(p, u, g_x) / \partial u = 0$. From integration by parts we see

$$\partial_u \mathcal{L}_x = \partial_u \left[u(x) + \langle \nabla g_x, \nabla u \rangle_{L^2(X)} + \langle g_x, pu \rangle_{L^2(X)} \right]$$

$$= \partial_u \left[u(x) + \langle -\Delta g_x, u \rangle_{H^{-1}(X), H^1(X)} + \langle pg_x, u \rangle_{L^2(X)} \right].$$

Setting this to be zero, we have the condition for g_x:

$$-\Delta g_x + pg_x = -\delta_x \text{ on } X, \quad g_x = 0 \text{ on } \partial X.$$

References

1. Alexanderian, A.: Optimal experimental design for infinite-dimensional bayesian inverse problems governed by pdes: A review. Inverse Problems **37** (2021). DOI 10.1088/1361-6420/abe10c
2. Attia, A., Constantinescu, E.: Optimal experimental design for inverse problems in the presence of observation correlations. SIAM Journal on Scientific Computing **44**(4), A2808–A2842 (2022)
3. Bandara, S., Schlöder, J.P., Eils, R., Bock, H.G., Meyer, T.: Optimal experimental design for parameter estimation of a cell signaling model. PLoS computational biology **5**(11), e1000558 (2009)
4. Bou-Rabee, N., Hairer, M.: Nonasymptotic mixing of the mala algorithm. IMA Journal of Numerical Analysis **33**(1), 80–110 (2013)
5. Boullé, N., Halikias, D., Townsend, A.: Elliptic pde learning is provably data-efficient. Proceedings of the National Academy of Sciences **120**(39), e2303904120 (2023)
6. Boullé, N., Townsend, A.: Learning elliptic partial differential equations with randomized linear algebra. Found. Comput. Math. **23**(2), 709–739 (2023). DOI 10.1007/s10208-022-09556-w
7. Bui-Thanh, T., Li, Q., Zepeda-Núñez, L.: Bridging and improving theoretical and computational electrical impedance tomography via data completion. SIAM Journal on Scientific Computing **44**(3), B668–B693 (2022). DOI 10.1137/21M141703X

8. Carrillo, J.A., Hoffmann, F., Stuart, A.M., Vaes, U.: Consensus-based sampling. Studies in Applied Mathematics **148**(3), 1069–1140 (2022). DOI https://doi.org/10.1111/sapm. 12470

9. Chen, T., Fox, E., Guestrin, C.: Stochastic gradient hamiltonian monte carlo. In: International conference on machine learning, pp. 1683–1691. PMLR (2014)

10. Chen, Y., Hosseini, B., Owhadi, H., Stuart, A.M.: Solving and learning nonlinear pdes with gaussian processes. Journal of Computational Physics **447**, 110668 (2021)

11. Cheng, X., Chatterji, N.S., Bartlett, P.L., Jordan, M.I.: Underdamped langevin mcmc: A non-asymptotic analysis. In: Conference on learning theory, pp. 300–323. PMLR (2018)

12. Chung, J., Chung, M., Slagel, J.T.: Iterative sampled methods for massive and separable nonlinear inverse problems. In: Scale Space and Variational Methods in Computer Vision: 7th International Conference, SSVM 2019, Hofgeismar, Germany, June 30–July 4, 2019, Proceedings 7, pp. 119–130. Springer (2019)

13. Dalalyan, A.S., Karagulyan, A.: User-friendly guarantees for the langevin monte carlo with inaccurate gradient. Stochastic Processes and their Applications **129**(12), 5278–5311 (2019)

14. Ding, Z., Li, Q.: Ensemble kalman sampler: Mean-field limit and convergence analysis. SIAM Journal on Mathematical Analysis **53**(2), 1546–1578 (2021). DOI 10.1137/20M1339507

15. Dong, S., Wang, Y.: A method for computing inverse parametric pde problems with random-weight neural networks. Journal of Computational Physics **489**, 112263 (2023)

16. Dwivedi, R., Chen, Y., Wainwright, M.J., Yu, B.: Log-concave sampling: Metropolis-hastings algorithms are fast. Journal of Machine Learning Research **20**(183), 1–42 (2019)

17. Evensen, G., Vossepoel, F.C., Van Leeuwen, P.J.: Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem. Springer Nature (2022)

18. Garbuno-Inigo, A., Hoffmann, F., Li, W., Stuart, A.M.: Interacting langevin diffusions: Gradient structure and ensemble kalman sampler. SIAM Journal on Applied Dynamical Systems **19**(1), 412–441 (2020). DOI 10.1137/19M1251655

19. Huan, X., Jagalur, J., Marzouk, Y.: Optimal experimental design: Formulations and computations (2024). URL https://arxiv.org/abs/2407.16212

20. Huang, Z., Becker, S.: Spectral estimation from simulations via sketching. Journal of Computational Physics **447**, 110686 (2021)

21. Jin, R., Li, Q., Nair, A., Stechmann, S.: Unique reconstruction for discretized inverse problems: a random sketching approach via subsampling (2024)

22. Kiefer, J.: Optimum experimental designs. Journal of the Royal Statistical Society. Series B (Methodological) **21**(2), 272–319 (1959). URL http://www.jstor.org/stable/2983802

23. Kovachki, N.B., Stuart, A.M.: Ensemble kalman inversion: a derivative-free technique for machine learning tasks. Inverse Problems **35**(9), 095005 (2019). DOI 10.1088/1361-6420/ab1c3a

24. Mahoney, M.W.: Lecture notes on randomized linear algebra (2016)

25. Mangoubi, O., Smith, A.: Mixing of hamiltonian monte carlo on strongly log-concave distributions: Continuous dynamics. The Annals of Applied Probability **31**(5), 2019–2045 (2021)

26. Manohar, K., Brunton, B.W., Kutz, J.N., Brunton, S.L.: Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. IEEE Control Systems Magazine **38**(3), 63–86 (2018). DOI 10.1109/MCS.2018.2810460

27. Martinsson, P.G., Tropp, J.A.: Randomized numerical linear algebra: Foundations and algorithms. Acta Numerica **29**, 403–572 (2020). DOI 10.1017/S0962492920000021

28. Mitchell, T.J.: An algorithm for the construction of "d-optimal" experimental designs. Technometrics **42**(1), 48–54 (2000)

29. Nordebo, S., Gustafsson, M., Khrennikov, A., Nilsson, B., Toft, J.: Fisher information for inverse problems and trace class operators. Journal of mathematical physics **53**(12) (2012)

30. Orozco, R., Herrmann, F.J., Chen, P.: Probabilistic bayesian optimal experimental design using conditional normalizing flows. arXiv preprint arXiv:2402.18337 (2024)

31. Park, S., Kato, D., Gima, Z., Klein, R., Moura, S.: Optimal experimental design for parameterization of an electrochemical lithium-ion battery model. Journal of The Electrochemical Society **165**(7), A1309 (2018)
32. Reich, S.: A dynamical systems framework for intermittent data assimilation. BIT Numerical Mathematics **51**, 235–249 (2011)
33. Riedl, K., Klock, T., Geldhauser, C., Fornasier, M.: Gradient is all you need? (2023). URL `https://arxiv.org/abs/2306.09778`
34. Roberts, G.O., Tweedie, R.L.: Exponential convergence of langevin distributions and their discrete approximations. Bernoulli **2**(4), 341–363 (1996). URL `http://www.jstor.org/stable/3318418`
35. Shun, Z., McCullagh, P.: Laplace approximation of high dimensional integrals. Journal of the Royal Statistical Society Series B: Statistical Methodology **57**(4), 749–760 (1995)
36. Sprungk, B., Weissmann, S., Zech, J.: Metropolis-adjusted interacting particle sampling (2023). URL `https://arxiv.org/abs/2312.13889`
37. Vaes, U.: Sharp propagation of chaos for the ensemble langevin sampler (2024)
38. Woodruff, D.P., et al.: Sketching as a tool for numerical linear algebra. Foundations and Trends® in Theoretical Computer Science **10**(1–2), 1–157 (2014)
39. Ye, J.J., Zhou, J.: Minimizing the condition number to construct design points for polynomial regression models. SIAM Journal on Optimization **23**(1), 666–686 (2013)