A-OPTIMAL DESIGN OF EXPERIMENTS FOR INFINITE-DIMENSIONAL BAYESIAN LINEAR INVERSE PROBLEMS WITH REGULARIZED ℓ_0-SPARSIFICATION[*]

ALEN ALEXANDERIAN[†], NOEMI PETRA[†], GEORG STADLER[†], AND
OMAR GHATTAS[†‡§]

Abstract. We present an efficient method for computing A-optimal experimental designs for infinite-dimensional Bayesian linear inverse problems governed by partial differential equations (PDEs). Specifically, we address the problem of optimizing the location of sensors (at which observational data are collected) to minimize the uncertainty in the parameters estimated by solving the inverse problem, where the uncertainty is expressed by the trace of the posterior covariance. Computing optimal experimental designs (OEDs) is particularly challenging for inverse problems governed by computationally expensive PDE models with infinite-dimensional (or, after discretization, high-dimensional) parameters. To alleviate the computational cost, we exploit the problem structure and build a low-rank approximation of the parameter-to-observable map, preconditioned with the square root of the prior covariance operator. The availability of this low-rank surrogate, relieves our method from expensive PDE solves when evaluating the optimal experimental design objective function, i.e., the trace of the posterior covariance, and its derivatives. Moreover, we employ a randomized trace estimator for efficient evaluation of the OED objective function. We control the sparsity of the sensor configuration by employing a sequence of penalty functions that successively approximate the ℓ_0-"norm"; this results in binary designs that characterize optimal sensor locations. We present numerical results for inference of the initial condition from spatio-temporal observations in a time-dependent advection-diffusion problem in two and three space dimensions. We find that an optimal design can be computed at a cost, measured in number of forward PDE solves, that is independent of the parameter and sensor dimensions. Moreover, the numerical optimization problem for finding the optimal design can be solved in a number of interior-point quasi-Newton iterations that is insensitive to the parameter and sensor dimensions. We demonstrate numerically that ℓ_0-sparsified experimental designs obtained via a continuation method outperform ℓ_1-sparsified designs.

Key words. Optimal experimental design, A-optimal design, Bayesian inference, sensor placement, ill-posed inverse problems, low-rank approximation, randomized trace estimator, randomized SVD.

AMS subject classifications. 62K05, 35Q62, 62F15, 35R30, 35Q93, 65C60.

1. Introduction. Recent advances in theory [32] and numerical algorithms (e.g., [8]) are enabling efficient solution of infinite-dimensional Bayesian inverse problems. This opens the door to consideration of the upstream question: how do we place sensors to *optimally* infer model parameters for large-scale problems? Here we present an efficient method for such optimal experimental design (OED) problems. Specifically, we consider Bayesian linear inverse problems governed by PDEs whose solution is the posterior probability law for a parameter field. The numerical solution of such inverse problems is challenging due to the infinite (or, when discretized, large) dimension of the parameters, ill-posedness of the inverse problem, and expensive-to-compute PDE models. The Bayesian inverse problem is by itself very challenging, but it is merely a subproblem within the OED problem, and must be solved repeatedly when using conventional OED methods. Hence, it is essential to make maximum use

of the problem structure to realize efficient algorithms that are *scalable*, i.e., their performance—measured in number of governing (forward) PDE solves—is independent of the discretized parameter and sensor dimensions, and the discretization of the governing PDE.

When formulating an OED problem, a basic question is the precise meaning of what constitutes the *design*. The present work concerns computation of optimal sensor locations where observational data will be collected. A subsequent question concerns the definition of an *optimal* design, which leads to the choice of the design criterion. For Bayesian inverse problems, a natural choice is to seek a design that minimizes the average posterior variance of the inversion parameters, leading to the Bayesian A-optimal design criterion.

Standard references for optimal experimental design include [2,6,28,30,33]. While most of the classical texts concern problems with small or moderate parameter dimension and focus mainly on well-posed inverse problems, there has been recent interest in optimal design for large-scale ill-posed linear [16,18] and nonlinear [17,20] inverse problems. The numerical methods in the present paper are closest to those in [16,18], where the authors consider finite-dimensional linear inverse problems, and develop a framework to control the mean square error of the regularized Tikhonov estimates. This leads to a design criterion that seeks to minimize the sum of the estimation bias and the variability of the estimator around its mean, which can be related to A-optimal designs. These contributions use ℓ_1-penalties to control the sparsity of the design and, in [18], a low-rank singular value decomposition (SVD) of the parameter-to-observable map is used; the present paper builds on and extends both of these ideas. Further recent work that employs a Bayesian formulation includes [21, 22], where the authors use a decision theoretic design criterion, generalized polynomial chaos surrogates, and stochastic optimization to tackle nonlinear inverse problems, albeit in low to moderate parameter dimension.

In this paper, we devise scalable numerical methods for computing A-optimal designs for infinite-dimensional Bayesian linear inverse problems governed by (time-dependent) PDEs. As suggested in [18], having a low-rank SVD *surrogate* of the parameter-to-observable map relieves the OED method of repeated PDE solves. In the Bayesian context, we can improve on this idea and further exploit problem structure; namely, we construct a low-rank SVD representation of the parameter-to-observable map *preconditioned* by the square root of the prior covariance operator [8,14]. This preconditioning amounts to filtering through the prior the information gained from the data about the model parameters. In the case of smoothing priors usually used in infinite-dimensional problems, this preconditioning results in faster spectral decay and thus allows for a more efficient low-rank approximation. The remaining steps in the solution of the OED problem use this low-rank surrogate and thus do not require additional PDE solves. As a result of a consistent discretization of the problem, i.e., one that respects the infinite-dimensional Hilbert-space structure, the numerical rank of the prior-preconditioned parameter-to-observable map is bounded with respect to the discretized parameter dimension.

We consider a finite number of candidate locations for the placement of sensors; the optimal configuration is a sparse subset of these locations. To each candidate sensor location we assign a non-negative number that weights the observation from that sensor. Finding an optimal design then amounts to choosing an optimal weight vector; a weight of 0 indicates absence of a sensor and a weight of 1 corresponds to a sensor being placed at that location. For computational convenience, we allow the

weights to take on any value in $[0, 1]$, and use a sparsifying penalty to control the number of nonzero weights, and thus, the number of allocated sensors. One option for such a penalty is the ℓ_1-norm; see, e.g., [16, 18, 20]. This approach, however, does not lead to a binary (i.e., 0–1) design. Motivated by continuation methods used in topology optimization [4, 5], we propose to solve a sequence of OED problems with penalty functions that successively approximate the ℓ_0-"norm". This, in contrast to an ℓ_1-penalty approach, does result in a binary design. As a test problem for our OED method, we consider a forward problem in the form of a time-dependent advection diffusion model, in which we infer the probability law of the initial condition. For this problem, we demonstrate the success of our continuation method, and show that the weights found by this continuation approach improve over optimal designs obtained via an ℓ_1-penalty approach.

For typical infinite-dimensional Bayesian inverse problems, the performance of our OED method is insensitive to the number of candidate sensor locations. This is due to the fact that although the dimension of the observations increases with the number of candidate locations, the amount of independent information that can be gained from nearby sensors is typically limited. Thus, increasing the number of sensors beyond a certain point does not significantly increase the numerical rank of the prior-preconditioned parameter-to-observable map. As a consequence, the number of PDE solves required to compute a low-rank SVD surrogate for the (prior-preconditioned) parameter-to-observable map is bounded as the number of candidate sensor locations increases. Moreover, in our computational results we find that the numerical optimization problem to compute the optimal design can be solved in a number of (quasi-Newton) iterations that is independent of the number of candidate locations.

The large-scale nature of the Bayesian inverse problems we target necessitates the use of randomized methods in linear algebra. In particular, we utilize randomized trace estimators [3, 23] to estimate the trace of the posterior covariance operator, and randomized SVD [19] to compute a low-rank surrogate of the prior-preconditioned parameter-to-observable map. Moreover, for computing the application of matrix square roots, as needed in our method, we employ matrix-free iterative methods [11].

The structure of this paper is as follows. After presenting the requisite background material in Section 2, we formulate the optimal design problem in infinite dimensions in Section 3. Then, we detail the components of our OED method in Section 4. Section 5 provides a description of our model problem, namely the inference of the initial condition in a time-dependent advection-diffusion equation. We present a comprehensive numerical study in Section 6. Finally, in Section 7, we draw conclusions, and discuss limitations and possible extensions of our method.

2. Background. In this section, we provide the background material required for the formulation and numerical solution of optimal experimental design problems in the context of infinite-dimensional Bayesian inverse problems. In Section 2.1, we present the Bayesian inverse problem in an infinite-dimensional Hilbert space setting, adopting the framework in [32]. In Section 2.2, we describe a discretization that is consistent with the infinite-dimensional inference problem formulation; this presentation follows [8]. Finally, in Section 2.3, we briefly comment on the randomized SVD and randomized trace estimators, which are used in our numerical method.

2.1. Bayesian inversion in Hilbert spaces. We begin our discussion by first considering a deterministic inverse problem. Given finite-dimensional observations

$\boldsymbol{d} \in \mathbb{R}^q$, we seek the model parameter m that solves

$$\min_{m \in \mathscr{H}} \mathcal{J}(m) := \frac{1}{2} \|\mathcal{F}(m) - \boldsymbol{d}\|_{\boldsymbol{\Gamma}}^2 + \mathcal{R}(m). \tag{2.1}$$

The function $\mathcal{F} : \mathscr{H} \to \mathbb{R}^q$ is the parameter-to-observable map and $\mathcal{R}$ denotes a regularization term. In the applications targeted in this paper $\mathscr{H} = L^2(\mathcal{D})$, where $\mathcal{D} \subset \mathbb{R}^d$ is a bounded domain (with $d = 2, 3$) and an evaluation of $\mathcal{F}$ involves the solution of a PDE, followed by the application of an observation operator. Note that the solution of a deterministic inverse problem can be thought of as a *point estimate* of m. To obtain a full probabilistic description of the parameter m, we are led to a Bayesian formulation of the problem, whose solution is a posterior probability law for m.

In this paper, we consider a parameter m which is modeled as a random-field (random function). To be precise, letting $(\Omega, \Sigma, \mathsf{P})$ be an appropriate probability space, $m : \mathcal{D} \times \Omega \to \mathbb{R}$ is a function such that for each $\boldsymbol{x} \in \mathcal{D}$, $m(\boldsymbol{x}, \cdot)$, is a real-valued random variable; thus, we can view m as an indexed collection of random variables, $\{m(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{D}}$, where, following the common practice, we suppress the dependence on ω. On the other hand, for each $\omega \in \Omega$, $m(\cdot, \omega) : \mathcal{D} \to \mathbb{R}$ is a real-valued function. We consider the case where $m(\cdot, \omega) \in \mathscr{H}$, and thus, we can also view m as a random-variable, $m : (\Omega, \Sigma, \mathsf{P}) \to (\mathscr{H}, \mathfrak{B}(\mathscr{H}))$, where $\mathfrak{B}(\mathscr{H})$ denotes the Borel σ-algebra on $\mathscr{H}$. Recall that the law of m is a probability measure μ on $(\mathscr{H}, \mathfrak{B}(\mathscr{H}))$ given by $\mu(E) = \mathsf{P}(m \in E)$ for $E \in \mathfrak{B}(\mathscr{H})$.

The infinite-dimensional Bayesian inverse problem can then be formulated as using observations to update our knowledge of the law of m, as a probability measure on $(\mathscr{H}, \mathfrak{B}(\mathscr{H}))$. Since, in contrast to the finite-dimensional case, there is no Lebesgue measure on $\mathscr{H}$, the infinite-dimensional Bayes formula is given by

$$\frac{d\mu_{\text{post}}}{d\mu_0} \propto \pi_{\text{like}}(\boldsymbol{d}|m). \tag{2.2}$$

Here, $\frac{d\mu_{\text{post}}}{d\mu_0}$ denotes the Radon-Nikodym derivative [34] of the posterior measure μ_{post} with respect to μ_0, and $\pi_{\text{like}}(\boldsymbol{d}|m)$ denotes the data likelihood. Conditions under which the posterior measure is well defined and (2.2) holds are given in detail in [32]. Note that we consider a finite-dimensional observation vector, motivated by the fact that in practice data are available only at a finite number of sensor locations and a finite number of points in time.

In the present work, we consider the Gaussian-linear case, i.e., the parameter-to-observable map $\mathcal{F}$ is linear. Moreover, we assume an additive noise model, $\boldsymbol{d} = \mathcal{F}m + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}})$ is a centered Gaussian on $\mathbb{R}^q$; the latter implies

$$\pi_{\text{like}}(\boldsymbol{d}|m) \propto \exp\left\{ -\frac{1}{2}(\mathcal{F}m - \boldsymbol{d})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\mathcal{F}m - \boldsymbol{d})\right\}. \tag{2.3}$$

We use a Gaussian prior, $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$, where $m_0 \in \mathscr{H}$ is sufficiently regular, and $\mathcal{C}_0$ is an appropriate covariance operator, i.e., $\mathcal{C}_0$ must be symmetric, positive, and of trace-class. We define the covariance operator as the inverse of an elliptic differential operator. A common alternative choice for statistical inverse problems is to specify a covariance function between any two spatial points, which results in a dense covariance matrix. For large-scale problems, however, the construction and "inversion" of such a dense covariance matrix can be infeasible. On the contrary,

specifying the covariance as the inverse of an elliptic differential operator allows to build on existing fast solvers for elliptic equations. As detailed in [8, 32], the PDE solution operator used as covariance operator $\mathcal{C}_0$ must be sufficiently smoothing and have bounded Green's functions. For example, the biharmonic operator has bounded Green's functions in two and three space dimensions. Therefore, we choose $\mathcal{C}_0 = \mathcal{A}^{-2}$, with $\mathcal{A}$ a Laplacian-like operator in the sense of Assumption 2.9 in [32]. This choice also allows efficient applications of the square root operator $\mathcal{C}_0^{1/2} = \mathcal{A}^{-1}$, as required below. The elliptic PDE corresponding to $\mathcal{A}$ written in weak form is as follows: For $s \in \mathscr{H} = L^2(\mathcal{D})$, the solution $m = \mathcal{A}^{-1}s$ satisfies

$$\int_{\mathcal{D}} \alpha \nabla m \cdot \nabla p + \beta m p \, d\boldsymbol{x} = \int_{\mathcal{D}} sp \, d\boldsymbol{x}, \quad \text{for all } p \in H^1(\mathcal{D}), \qquad (2.4)$$

with $\alpha, \beta > 0$ controlling the variation and the correlation length. Due to the present choice of the prior and the noise model, and the linearity of $\mathcal{F}$, the posterior measure is a Gaussian, $\mathcal{N}(m_{\text{post}}, \mathcal{C}_{\text{post}})$ with [32, Section 6.4],

$$\mathcal{C}_{\text{post}} = (\mathcal{F}^* \mathbf{\Gamma}_{\text{noise}}^{-1} \mathcal{F} + \mathcal{C}_0^{-1})^{-1}, \qquad m_{\text{post}} = \mathcal{C}_{\text{post}}(\mathcal{F}^* \mathbf{\Gamma}_{\text{noise}}^{-1} \boldsymbol{d} + \mathcal{C}_0^{-1} m_0), \qquad (2.5)$$

where $\mathcal{F}^* : \mathbb{R}^q \to \mathscr{H}$ is the adjoint of $\mathcal{F}$.

2.2. Discretization of the infinite-dimensional Bayesian inverse problem. In this section, we describe the discretization of the Bayesian inverse problem (2.2) in the Gaussian linear case. We consider a finite-dimensional subspace $V_h \subset L^2(\mathcal{D})$ given by $V_h = \mathsf{span}\{\phi_1, \ldots, \phi_n\}$, where $\{\phi_j\}_{j=1}^n$ are continuous Lagrange nodal basis functions. Given $m_h \in V_h$, we denote by $\boldsymbol{m}$, the vector of its coordinates in V_h; i.e., for $m_h = \sum_{j=1}^n m_j \phi_j$, we have $\boldsymbol{m} = (m_1, \ldots, m_n)^T$. After this discretization, we replace the task of inferring the parameter $m \in L^2(\mathcal{D})$ with that of inferring the coefficients for the finite-element approximation m_h of m.

Following [8], we state the finite-dimensional Bayesian inverse problem such that it is consistent with the corresponding inference problem in $L^2(\mathcal{D})$. Consequently, we work in $\mathbb{R}^n$, with the weighted inner product, $\langle \cdot, \cdot \rangle_{\text{M}}$ given by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\text{M}} = \langle \mathbf{M}\boldsymbol{x}, \boldsymbol{y} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product, and $\mathbf{M}$ is the (symmetric positive definite) finite-element mass matrix. It is convenient to introduce the notation $\mathbb{R}_{\mathbf{M}}^n$ for $\mathbb{R}^n$ when endowed with the $\langle \cdot, \cdot \rangle_{\text{M}}$ inner product. Note that the mapping $m_h \mapsto \boldsymbol{m}$ is a Hilbert-space isomorphism between V_h (with L^2-inner product) and $\mathbb{R}_{\mathbf{M}}^n$. For a linear operator $\mathbf{A} : \mathbb{R}_{\mathbf{M}}^n \to \mathbb{R}_{\mathbf{M}}^n$, the adjoint operator is given by $\mathbf{A}^* = \mathbf{M}^{-1}\mathbf{A}^T\mathbf{M}$. A linear operator $\mathbf{A}$ on $\mathbb{R}_{\mathbf{M}}^n$ is self-adjoint if $\mathbf{A} = \mathbf{A}^*$; for convenience, we refer to such operators as $\mathbf{M}$-symmetric. In the sequel, we will encounter linear mappings $\mathbf{A}_1 : \mathbb{R}_{\mathbf{M}}^n \to \mathbb{R}^q$ and $\mathbf{A}_2 : \mathbb{R}^r \to \mathbb{R}_{\mathbf{M}}^n$, where $\mathbb{R}^q$ and $\mathbb{R}^r$ are endowed with the Euclidean inner product; the corresponding adjoints are given by [8], $\mathbf{A}_1^* = \mathbf{M}^{-1}\mathbf{A}_1^T$ and $\mathbf{A}_2^* = \mathbf{A}_2^T\mathbf{M}$.

For the discretized problem, the density for the prior (as a measure over the space $\mathbb{R}_{\mathbf{M}}^n$) is characterized by

$$\pi_{\text{prior}}(\boldsymbol{m}) \propto \exp\left\{ -\frac{1}{2} \|\mathbf{A}(\boldsymbol{m} - \boldsymbol{m}_0)\|_{\mathbf{M}}^2 \right\}, \qquad (2.6)$$

with $\mathbf{A} = \mathbf{M}^{-1}\mathbf{L}$, where $\mathbf{L} = \alpha\mathbf{K} + \beta\mathbf{M}$ and $\mathbf{K}$ is the finite-element stiffness matrix. Note that $\mathbf{A}$ is $\mathbf{M}$-symmetric; moreover, it follows from the above definition that $\mathbf{\Gamma}_{\text{prior}} = \mathbf{A}^{-2}$. Subsequently, the posterior is a Gaussian $\mathcal{N}(\boldsymbol{m}_{\text{post}}, \mathbf{\Gamma}_{\text{post}})$ with

$$\boldsymbol{m}_{\text{post}} = \mathbf{\Gamma}_{\text{post}}\left(\mathbf{F}^* \mathbf{\Gamma}_{\text{noise}}^{-1} \boldsymbol{d} + \mathbf{\Gamma}_{\text{prior}}^{-1} \boldsymbol{m}_0 \right), \qquad \mathbf{\Gamma}_{\text{post}} = \left(\mathbf{F}^* \mathbf{\Gamma}_{\text{noise}}^{-1} \mathbf{F} + \mathbf{\Gamma}_{\text{prior}}^{-1} \right)^{-1},$$

where $\mathbf{F} : \mathbb{R}_{\mathbf{M}}^n \to \mathbb{R}^q$ is the discretization of the parameter-to-observable map $\mathcal{F}$.

Next, we summarize how to draw samples from and to compute the variance of the discretized posterior measure; here we follow [8], where details of these operations are provided. For a posterior measure $\mathcal{N}(\boldsymbol{m}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$ generating samples requires a decomposition of $\boldsymbol{\Gamma}_{\text{post}}$ in the form, $\boldsymbol{\Gamma}_{\text{post}} = \mathbf{Q}\mathbf{Q}^*$. Then, to generate a sample $\boldsymbol{\nu}$ from $\mathcal{N}(\boldsymbol{m}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$, we draw a realization $\boldsymbol{z}$ from $\mathcal{N}(0, \mathbf{I})$ and compute $\boldsymbol{\nu}$ as

$$\boldsymbol{\nu} = \boldsymbol{m}_{\text{post}} + \mathbf{Q}\mathbf{M}^{-1/2}\boldsymbol{z}. \tag{2.7}$$

The discretized covariance is given by $\mathsf{Cov}\{m_i, m_j\} = \boldsymbol{e}_i^T \boldsymbol{\Gamma}_{\text{post}} \mathbf{M}^{-1} \boldsymbol{e}_j$, for $i, j = 1, \dots, n$. In particular,

$$\mathsf{Var}\{m_i\} = \boldsymbol{e}_i^T \boldsymbol{\Gamma}_{\text{post}} \mathbf{M}^{-1} \boldsymbol{e}_i, \quad i = 1, \dots, n, \tag{2.8}$$

where $\mathsf{Var}\{m_i\}$ denotes the variance of m_i.

2.3. Randomized linear algebra algorithms. One of the major components of our method is a low-rank SVD surrogate for the prior-preconditioned parameter-to-observable map. To compute such low-rank surrogates, we use a randomized SVD [19]. This choice is motivated by the flexibility and the robustness of the method, and by the fact that, as opposed to Krylov subspace methods, it only requires independent matrix-vector products. This aspect is particularly useful for the large-scale problems we target, in which the matrix-vector applications involve expensive PDE solves; see also [7]. Randomized SVD methods can be made very accurate with negligible probability of failure [19].

Computing A-optimal designs requires minimizing the trace of large dense covariance matrices, which, in our target problems, usually have a rapidly decaying spectrum and the eigenvalues are clustered around 0. For such matrices, which are defined implicitly through their applications to vectors, randomized trace estimators provide a reasonably accurate approximation of the trace with a small number of random vectors (see e.g., [31]). These estimators involve only matrix-vector products, which makes them suitable for large scale problems. In particular, randomized trace estimators approximate the trace of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, via Monte-Carlo estimates of the form $\mathsf{tr}(\mathbf{A}) \approx \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} [\boldsymbol{z}^{(i)}]^T \mathbf{A}\boldsymbol{z}^{(i)}$, where the trial vectors $\boldsymbol{z}^{(i)}$ are random n-vectors. A well-known example is the Hutchinson estimator [23], which uses random vectors $\boldsymbol{z}^{(i)}$ with ± 1 entries, each with a probability of $1/2$. Another possibility, used in this paper, is the Gaussian trace estimator, which uses Gaussian random vectors with independent and identically distributed (i.i.d.) standard normal entries. For a description and analysis of different trace estimators, we refer to [3].

3. A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems. In this section, we formulate the A-optimal design problem for infinite-dimensional Bayesian linear inverse problems. The extension of the A-optimal design criterion to the infinite-dimensional setting is described in Section 3.1. In Section 3.2, we specify the mathematical definition of a *design*, and describe how the design is introduced in the Bayesian inverse problem. Finally, in Section 3.3, we formulate the resulting OED optimization problem.

3.1. A-optimal design in Hilbert spaces. In a finite-dimensional inference problem, an A-optimal design minimizes the average posterior variance of the inference parameters [33]. In the linear case, this is accomplished by minimizing the trace of the posterior *covariance matrix*. Since the present work concerns inference problems with

a random field as the inference parameter, we first extend the notion of an A-optimal design to the infinite-dimensional Hilbert space setting.

The average posterior variance of m over the physical domain $\mathcal{D}$ is given by

$$\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} c_{\text{post}}(\boldsymbol{x}, \boldsymbol{x}) \, d\boldsymbol{x},$$

where $|\mathcal{D}|$ denotes the Lebesgue measure of the domain $\mathcal{D}$, and c_{post} is the *covariance function* of m:

$$c_{\text{post}}(\boldsymbol{x}, \boldsymbol{y}) = \mathsf{E}\left\{ \left(m(\boldsymbol{x}) - m_{\text{post}}(\boldsymbol{x}) \right) \left(m(\boldsymbol{y}) - m_{\text{post}}(\boldsymbol{y}) \right) \right\}, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D},$$

where $\mathsf{E}\{\cdot\}$ denotes the expectation operator. Note that c_{post} is related to the *covariance operator* $\mathcal{C}_{\text{post}}$, i.e.,

$$[\mathcal{C}_{\text{post}} u](\boldsymbol{x}) = \int_{\mathcal{D}} c_{\text{post}}(\boldsymbol{x}, \boldsymbol{y}) u(\boldsymbol{y}) \, d\boldsymbol{y}, \quad u \in L^2(\mathcal{D}).$$

The covariance operator $\mathcal{C}_{\text{post}}$ is positive, symmetric, and of trace-class, and thus has real eigenvalues, $\{\lambda_i\}_{i=1}^{\infty}$, and a complete orthonormal set of eigenvectors, $\{e_i\}_{i=1}^{\infty}$. Mercer's Theorem [24, 26] then provides a decomposition of the covariance function c_{post} from the spectral decomposition of $\mathcal{C}_{\text{post}}$ through $c_{\text{post}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_j \lambda_j e_j(\boldsymbol{x}) e_j(\boldsymbol{y})$, where the convergence of the infinite sum is uniform and absolute in $\mathcal{D} \times \mathcal{D}$. From this representation, one obtains,

$$\int_{\mathcal{D}} c_{\text{post}}(\boldsymbol{x}, \boldsymbol{x}) \, d\boldsymbol{x} = \sum_j \lambda_j = \mathsf{tr}(\mathcal{C}_{\text{post}}).$$

Thus, we formulate the A-optimal design problem as that of minimizing the trace of the covariance *operator*, $\mathsf{tr}(\mathcal{C}_{\text{post}})$. This shows that the definition of an A-optimal design in finite dimensions, namely the minimization of the average variance of the estimates, extends naturally to infinite-dimensional Bayesian inverse problems involving a random-field as the inversion parameter. We emphasize that the trace of the posterior covariance operator in the A-optimal design criterion is well defined due to the proper choice of the prior measure for infinite-dimensional inference problems.

3.2. Introducing the design in the Bayesian inverse problem. Let us first specify the notion of a design in the context of our target applications, namely optimal sensor placement. We use a finite-dimensional design space, that is, we fix a set of points $\boldsymbol{x}_i$, $i = 1, \ldots, N_s$ as the set of candidate sensor locations, and associate to each $\boldsymbol{x}_i$ a non-negative weight $w_i \in \mathbb{R}$; see also [16, 18, 33]. The OED problem is then formulated as an optimization problem over the weight vector, $\boldsymbol{w} = (w_1, \ldots, w_{N_s})^T$. The points $\boldsymbol{x}_i$, $i = 1, \ldots, N_s$ can be thought of as a discretization of the *sensor domain*, which is a subset of $\mathcal{D}$.

The design weights can have different interpretations. For example, in classical formulations such as in [33], w_i define a probability mass function, i.e., $w_i \geq 0$ and $\sum w_i = 1$; one may then interpret large weights as ones indicating promising locations for placing sensors. If the inversion is based on a repeatable experiment, weights can also be used to specify the number of experiments performed at each sensor location to control the observation noise; see, e.g., [16]. In many inverse problems, however, experiments cannot be repeated or the mathematical model is not an exact

representation of the physical phenomenon underlying the observations, such that the observation error cannot be controlled. Thus, we prefer weight vectors containing zeros and ones only, indicating absence or presence of sensors over the candidate sensor grid. Unfortunately, solving optimization problems for vectors with binary components is a difficult combinatorial problem. Hence, we employ a relaxation of the problem and consider weights $w_i \in [0, 1]$. In Section 4.5, we devise a method of recovering the desired 0–1 structure using sparsifying penalties combined with a continuation procedure.

Next, we describe the process of introducing the design vector $\boldsymbol{w}$ into the Bayesian inverse problem. Since the design guides the collection of data, the weight vector $\boldsymbol{w}$ enters the inference problem (2.2) through the data likelihood (2.3). The $\boldsymbol{w}$-weighted data-likelihood is given by,

$$\pi_{\text{like}}(\boldsymbol{d}|m; \boldsymbol{w}) \propto \exp\left\{ -\frac{1}{2}(\mathcal{F}m - \boldsymbol{d})^T \mathbf{W}^{1/2} \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{W}^{1/2}(\mathcal{F}m - \boldsymbol{d}) \right\}, \qquad (3.1)$$

where $\mathbf{W} \in \mathbb{R}^{q \times q}$ is a diagonal matrix with weights on its diagonal. For time-dependent problems, with observations collected at sensor locations at discrete points in time, we have $\mathcal{F} : \mathscr{H} \to \mathbb{R}^q$ with $q = N_s N_\tau$, where N_s and N_τ are the number of candidate sensors and observation times, respectively. Here, $\mathbf{W}$ is a block-diagonal matrix having N_τ blocks, where each block is an $N_s \times N_s$ diagonal matrix with $\boldsymbol{w}$ on its diagonal. The posterior covariance operator is thus given by

$$\mathcal{C}_{\text{post}}(\boldsymbol{w}) = (\mathcal{F}^* \mathbf{W}^{1/2} \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{W}^{1/2} \mathcal{F} + \mathcal{C}_0^{-1})^{-1}. \qquad (3.2)$$

From this point on, we work with the discretization of the infinite-dimensional problem, where we follow the discretization strategy described in Section 2.2. In particular, we have the following discretized posterior covariance,

$$\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w}) = (\mathbf{F}^* \mathbf{W}^{1/2} \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{W}^{1/2} \mathbf{F} + \boldsymbol{\Gamma}_{\text{prior}}^{-1})^{-1}. \qquad (3.3)$$

The mean $\boldsymbol{m}_{\text{post}}$ of the discretized posterior measure coincides with the maximum a posteriori probability (MAP) estimate, given as the solution of the minimization problem,

$$\min_{\boldsymbol{m} \in \mathbb{R}^n} \frac{1}{2} \left\langle \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{W}^{1/2}\left(\boldsymbol{d} - \mathbf{F}\boldsymbol{m}\right), \mathbf{W}^{1/2}(\boldsymbol{d} - \mathbf{F}\boldsymbol{m})\right\rangle + \frac{1}{2} \left\langle \boldsymbol{\Gamma}_{\text{prior}}^{-1}\left(\boldsymbol{m} - \boldsymbol{m}_0\right), \boldsymbol{m} - \boldsymbol{m}_0\right\rangle_{\text{M}}.$$

The inverse of the Hessian $\mathbf{H}(\boldsymbol{w})$ of the above functional coincides with $\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w})$ defined in (3.3). In the following, we call the Hessian of the misfit term,

$$\mathbf{H}_{\text{misfit}}(\boldsymbol{w}) = \mathbf{F}^* \mathbf{W}^{1/2} \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{W}^{1/2} \mathbf{F},$$

the *misfit Hessian*; thus, $\mathbf{H}(\boldsymbol{w}) = \mathbf{H}_{\text{misfit}}(\boldsymbol{w}) + \boldsymbol{\Gamma}_{\text{prior}}^{-1}$. In this paper, we consider the case of independent observations, and hence $\boldsymbol{\Gamma}_{\text{noise}}$ is a diagonal matrix. This assumption turns $\mathbf{W}^{1/2} \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{W}^{1/2}$ into a diagonal matrix with diagonal entries w_i/σ_i^2. For simplicity of the presentation, we further assume that $\boldsymbol{\Gamma}_{\text{noise}}$ is a constant multiple of identity, i.e., $\boldsymbol{\Gamma}_{\text{noise}} = \sigma \mathbf{I}$, and set $\sigma = 1$. Thus, the misfit Hessian takes the form $\mathbf{H}_{\text{misfit}} = \mathbf{F}^* \mathbf{W} \mathbf{F}$. The algorithms presented below can be easily modified to accommodate a general diagonal noise covariance matrices $\boldsymbol{\Gamma}_{\text{noise}}$.

3.3. The OED problem. Finally, we can formulate the Bayesian A-optimal experimental design problem as optimization problem for the weight vector $\boldsymbol{w}$. Following the discussions in the previous sections, the OED objective function is the trace of the posterior covariance operator (3.2), which, after discretization is given by (3.3). Additionally, we use a penalization to control the sparsity of the design. Hence, the optimal design vector is the solution to the following optimization problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^{N_s}} \quad \mathsf{tr}\big[\boldsymbol{\Gamma}_{\mathrm{post}}(\boldsymbol{w})\big] + \gamma\Phi(\boldsymbol{w}),$$

$$\text{subject to} \quad 0 \le w_i \le 1, \quad i = 1, \ldots, N_s, \tag{3.4}$$

where $\Phi : \mathbb{R}_+^{N_s} \to [0, \infty)$ is a penalty function and $\gamma \ge 0$ controls the sparsity of the design. We note that the function $\boldsymbol{w} \mapsto \mathsf{tr}\big[\boldsymbol{\Gamma}_{\mathrm{post}}(\boldsymbol{w})\big] = \mathsf{tr}\big[\mathbf{H}(\boldsymbol{w})^{-1}\big]$ is strictly convex due to strict convexity of $\mathbf{X} \mapsto \mathsf{tr}(\mathbf{X}^{-1})$ on the cone of symmetric positive definite matrices (see [28, p. 82]) and the fact that $\boldsymbol{w}$ enters linearly in $\mathbf{H}(\boldsymbol{w})$. Therefore, if the penalty function Φ is convex, (3.4) has a unique solution. An example of a convex penalty is given by $\Phi(\boldsymbol{w}) = \mathbf{1}^T\boldsymbol{w}$, i.e., an ℓ_1-penalty, whose sparsening property has been used extensively in compressive sensing [9, 13] and has also been adapted to OED for inverse problems [16, 18]. In this paper, we will use a continuation approach involving a family $\{\Phi_\varepsilon\}_{\varepsilon>0}$ of penalty functions, which approximate the ℓ_0-"norm". This allows us to find binary optimal design vectors.

We remark that there exists an alternative interpretation of the A-optimal design criterion in the Gaussian linear case considered here. Namely, minimizing the trace of the posterior covariance is equivalent to minimizing the average mean square error (MSE) of the posterior mean, where the average is with respect to the prior measure. This average MSE is also referred to as the Bayes risk of the posterior mean. MSE is a concept in frequentist inference, in which the posterior mean is interpreted as an estimator for the unknown parameter. This frequentist point of view of A-optimal design is used in [16, 18]. For completeness of our presentation, we detail this relation between average MSE and the trace of posterior covariance in Appendix B.

4. Numerical solution of the OED problem (3.4). We begin this section with deriving a decomposition of the misfit Hessian in terms of contributions from different sensors. Then, in Section 4.2, we present an approximation of $\mathsf{tr}(\boldsymbol{\Gamma}_{\mathrm{post}})$ using a randomized trace estimator, and derive expressions for the gradient of the resulting OED objective function. Subsequently, in Sections 4.3 and 4.4, we present algorithmic components that allow efficient realization of these computations. Finally, in Section 4.5, we discuss methods to control the sparsity of the design, i.e., the number of allocated sensors.

4.1. Decomposition of the misfit Hessian. The misfit Hessian plays an important role in the derivative computation of the OED objective function with respect to the sensor location weights. Therefore, we first derive a decomposition of $\mathbf{H}_{\mathrm{misfit}}$ as a weighted sum of terms corresponding to individual sensor locations.

We consider a linear parameter-to-observable map $\mathbf{F}$, which involves a time-dependent PDE. Parameter-to-observable maps with stationary equations are included as special case in the discussion below by considering a single time step. We consider observations at the candidate sensor locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_s}$ in $\mathcal{D}$. For each sensor location, the time evolution of the observation is discretized using Lagrange elements in time (we use piecewise linear elements in this paper) for the nodal time instances $\tau_1, \ldots, \tau_{N_\tau}$ ($\tau_i \in [0, T]$ for $i = 1, \ldots, N_\tau$). These observation times are independent of the time steps used in the integration of the PDE.

The parameter-to-observable map $\mathbf{F}$ takes a parameter vector $\boldsymbol{m} \in \mathbb{R}^n$ and maps it to the space-time observation vector $\bar{\boldsymbol{d}} \in \mathbb{R}^{N_s N_\tau}$:

$$\mathbf{F} : \boldsymbol{m} \overset{\mathbf{S}}{\longmapsto} \bar{\boldsymbol{u}} \overset{\mathbf{B}}{\longmapsto} \bar{\boldsymbol{d}}. \tag{4.1}$$

Here, $\mathbf{S}$ is the discretized PDE solution operator, $\bar{\boldsymbol{u}} \in \mathbb{R}^{n(N_t+1)}$ is the space-time solution vector, and $\mathbf{B}$ is the space-time observation operator. We target designs where the sensor locations coincide for all time observations. Thus, the sensor weight matrix $\mathbf{W}$ can be written as $\mathbf{W} = \sum_{j=1}^{N_s} w_j \mathbf{E}_j$, where $\mathbf{E}_j$ is an $N_s N_\tau \times N_s N_\tau$ block-diagonal matrix, with N_τ blocks, with each block equal to $\boldsymbol{e}_j \otimes \boldsymbol{e}_j = \boldsymbol{e}_j \boldsymbol{e}_j^T$; here $\boldsymbol{e}_j$ denotes the jth coordinate vector in $\mathbb{R}^{N_s}$. Thus, the misfit Hessian for a weight vector $\boldsymbol{w}$ can be decomposed as $\mathbf{H}_{\text{misfit}}(\boldsymbol{w}) = \mathbf{F}^* \mathbf{W} \mathbf{F} = \sum_{j=1}^{N_s} w_j \mathbf{F}^* \mathbf{E}_j \mathbf{F}$, where the matrices $\mathbf{F}^* \mathbf{E}_j \mathbf{F}$, $j = 1, \dots, N_s$ are the *atoms* corresponding to the different sensor locations. This decomposition, which is also used in [33], reveals the identity

$$\frac{\partial \mathbf{H}_{\text{misfit}}(\boldsymbol{w})}{\partial w_j} = \mathbf{F}^* \mathbf{E}_j \mathbf{F}. \tag{4.2}$$

Next, we approximate the OED objective function using trace estimators.

4.2. The OED objective function and its derivative. We consider the problem (3.4) and recall that $\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w}) = \mathbf{H}(\boldsymbol{w})^{-1}$, where the Hessian is an $\mathbf{M}$-symmetric linear mapping on $\mathbb{R}_M^n$. For the numerical solution of the OED problem, we approximate the trace of $\mathbf{H}(\boldsymbol{w})^{-1}$ using a randomized trace estimator (see Section 2.3). The trace estimator-based OED objective functional is

$$\Theta(\boldsymbol{w}) := \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \left\langle \boldsymbol{z}^{(i)}, \mathbf{H}(\boldsymbol{w})^{-1} \boldsymbol{z}^{(i)} \right\rangle_{\text{M}}, \tag{4.3}$$

where $\boldsymbol{z}^{(i)} = \mathbf{M}^{-1/2} \boldsymbol{y}^{(i)}$, $i = 1, \dots, N_{\text{tr}}$, with $\boldsymbol{y}^{(i)}$ appropriately chosen random vectors[1]. Therefore, we consider the following OED optimization problem in our numerical computations,

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{N_s}} \quad & \Theta(\boldsymbol{w}) + \gamma \Phi(\boldsymbol{w}), \\ \text{subject to} \quad & 0 \le w_i \le 1, \quad i = 1, \dots, N_s. \end{aligned} \tag{4.4}$$

Since we will use gradient-based optimization methods to solve (4.4), we need to compute the gradient of $\Theta(\boldsymbol{w})$ with respect to $\boldsymbol{w}$. For $j \in \{1, \dots, N_s\}$, we obtain

$$\begin{aligned} \frac{\partial \Theta(\boldsymbol{w})}{\partial w_j} &= \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \left\langle \boldsymbol{z}^{(i)}, \frac{\partial \mathbf{H}(\boldsymbol{w})^{-1}}{\partial w_j} \boldsymbol{z}^{(i)} \right\rangle_{\text{M}} \\ &= -\frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \left\langle \boldsymbol{z}^{(i)}, \mathbf{H}(\boldsymbol{w})^{-1} \frac{\partial \mathbf{H}(\boldsymbol{w})}{\partial w_j} \mathbf{H}(\boldsymbol{w})^{-1} \boldsymbol{z}^{(i)} \right\rangle_{\text{M}}. \end{aligned}$$

The $\mathbf{M}$-symmetry of $\mathbf{H}(\boldsymbol{w})^{-1}$, the fact that the prior does not depend on $\boldsymbol{w}$, and denoting $\boldsymbol{q}^{(i)} = \mathbf{H}^{-1} \boldsymbol{z}^{(i)}$ yields

$$\frac{\partial \Theta(\boldsymbol{w})}{\partial w_j} = -\frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \left\langle \boldsymbol{q}^{(i)}, \frac{\partial \mathbf{H}_{\text{misfit}}(\boldsymbol{w})}{\partial w_j} \boldsymbol{q}^{(i)} \right\rangle_{\text{M}}, \quad j = 1, \dots, N_s. \tag{4.5}$$

---

[1]In the present work, we rely on Gaussian trace estimators. See Appendix A for a justification of trace estimation in the context of weighted inner products.

Using (4.2) and denoting $\bar{\boldsymbol{d}}^{(i)} = (\boldsymbol{d}^{(i),1}, \ldots, \boldsymbol{d}^{(i),N_\tau})^T := \mathbf{F}\boldsymbol{q}^{(i)}$, with $\boldsymbol{d}^{(i)} \in \mathbb{R}^{N_s}$ the spatial observations corresponding to time τ_i, each of the summands in (4.5) can be computed as

$$
\begin{aligned}
\left\langle \boldsymbol{q}^{(i)}, \frac{\partial \mathbf{H}_{\mathrm{misfit}}(\boldsymbol{w})}{\partial w_j} \boldsymbol{q}^{(i)} \right\rangle_{\mathrm{M}} &= \left\langle \mathbf{F}\boldsymbol{q}^{(i)}, \mathbf{E}_j \mathbf{F}\boldsymbol{q}^{(i)} \right\rangle \\
&= \sum_{\ell=1}^{N_\tau} [\boldsymbol{d}^{(i),\ell}]^T (\boldsymbol{e}_j \otimes \boldsymbol{e}_j)[\boldsymbol{d}^{(i),\ell}] = \sum_{\ell=1}^{N_\tau} \left[d_j^{(i),\ell} \right]^2.
\end{aligned}
\tag{4.6}
$$

To summarize, the evaluation of $\Theta(\boldsymbol{w})$ requires N_{tr} multiplications of vectors with $\mathbf{H}(\boldsymbol{w})^{-1}$. The computation of the derivative of $\Theta(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ additionally requires N_{tr} evaluations of the parameter-to-observable map $\mathbf{F}$. In the Sections 4.3 and 4.4 below, we discuss the efficient realization of these computations.

4.3. Low-rank approximation of the prior-preconditioned parameter-to-observable map. As shown above, the repeated application of $\mathbf{H}(\boldsymbol{w})^{-1}$ is necessary to compute $\Theta(\boldsymbol{w})$ and the gradient of $\Theta(\boldsymbol{w})$ with respect to $\boldsymbol{w}$. In a numerical optimization algorithm to solve (4.4), these computations are required in each iteration. Despite the use of a trace estimator, this is computationally demanding and can render OED for large-scale Bayesian inverse problems infeasible.

As a remedy, we construct a surrogate model for the parameter-to-observable map that can be used to efficiently compute the application of $\mathbf{H}(\boldsymbol{w})^{-1}$ to vectors. The surrogate construction exploits the fact that for a large class of infinite-dimensional inverse problems, $\mathbf{F}$ can be well approximated by a low-rank operator due to properties of the underlying PDE and the limited number of observations; see for instance [8,14]. Thus, we can compute a low-rank SVD surrogate for $\mathbf{F}$ *upfront*, and use this surrogate in all subsequent computations involving $\mathbf{F}$ to find an optimal experimental design.

Since only parameters consistent with the data *and* the prior have a significant posterior probability, only those parameters can influence the OED. Thus, it suffices to compute a surrogate of the *prior-preconditioned* parameter-to-observable map, $\tilde{\mathbf{F}} := \mathbf{F}\boldsymbol{\Gamma}_{\mathrm{prior}}^{1/2}$. The smoothing property of the priors usually employed in infinite-dimensional Bayesian inversion[2] results in $\tilde{\mathbf{F}}$ having faster decaying singular values than $\mathbf{F}$. Thus, the number of forward and adjoint PDE solves required to construct an accurate low-rank SVD surrogate for $\tilde{\mathbf{F}}$ is usually smaller than that for $\mathbf{F}$. For the model problem in Section 6.1.1, this results in a significant speedup for the construction of the low-rank surrogate.

We employ randomized SVD (see Section 2.3) to compute the low-rank surrogate of $\tilde{\mathbf{F}}$. This algorithm only requires the applications of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{F}}^*$ to a set of independent random vectors, which is convenient and can often be implemented efficiently in large-scale computations. In the next section, we show how the low-rank surrogate of $\tilde{\mathbf{F}}$ can be used to compute $\Theta(\boldsymbol{w})$ and its gradient with respect to $\boldsymbol{w}$.

4.4. Efficient computation of $\Theta(\boldsymbol{w})$ and its derivatives. First, we discuss the efficient computation of $\mathbf{H}(\boldsymbol{w})^{-1}\boldsymbol{z}$ for $\boldsymbol{z} \in \mathbb{R}^n$, as required to evaluate $\Theta(\boldsymbol{w})$ and its derivative. Note that the inverse of the Hessian $\mathbf{H}(\boldsymbol{w})$ can be written as [8,25],

$$
\mathbf{H}(\boldsymbol{w})^{-1} = \boldsymbol{\Gamma}_{\mathrm{prior}}^{1/2} \big(\tilde{\mathbf{H}}_{\mathrm{misfit}}(\boldsymbol{w}) + \mathbf{I} \big)^{-1} \boldsymbol{\Gamma}_{\mathrm{prior}}^{1/2},
\tag{4.7}
$$

---

[2]Note that a smoothing property is necessary to render the infinite-dimensional Bayesian inverse problem well-posed, [32].

11

where, for a weight vector $\boldsymbol{w} \in \mathbb{R}^{N_s}$, $\tilde{\mathbf{H}}_{\mathrm{misfit}}(\boldsymbol{w}) = \tilde{\mathbf{F}}^* \mathbf{W} \tilde{\mathbf{F}}$ is the *prior-preconditioned* misfit Hessian. Assuming a rank-r surrogate $\tilde{\mathbf{F}}_r$ for $\tilde{\mathbf{F}}$ is available, $\tilde{\mathbf{H}}_{\mathrm{misfit},r}(\boldsymbol{w}) = \tilde{\mathbf{F}}_r^* \mathbf{W} \tilde{\mathbf{F}}_r$ is the resulting approximation of $\tilde{\mathbf{H}}_{\mathrm{misfit}}(\boldsymbol{w})$ used below. Note that for every weight vector $\boldsymbol{w}$, the rank of $\tilde{\mathbf{H}}_{\mathrm{misfit},r}(\boldsymbol{w})$ is less than or equal to r. One approach to apply $(\tilde{\mathbf{H}}_{\mathrm{misfit},r} + \mathbf{I})^{-1}$ to a vector is to first compute the spectral decomposition $\tilde{\mathbf{H}}_{\mathrm{misfit},r}(\boldsymbol{w}) = \mathbf{V}_r \mathbf{\Lambda}_r \mathbf{V}_r^*$, where $\mathbf{V}_r$ is the matrix containing the eigenvectors of $\tilde{\mathbf{H}}_{\mathrm{misfit},r}(\boldsymbol{w})$ as its columns, and $\mathbf{\Lambda}_r$ is the diagonal matrix with r largest eigenvalues on the diagonal.[3] Then, by the Sherman-Morrison-Woodbury formula [15], we have $(\mathbf{V}_r \mathbf{\Lambda}_r \mathbf{V}_r^* + \mathbf{I})^{-1} = \mathbf{I} - \mathbf{V}_r \mathbf{D}_r \mathbf{V}_r^*$, where $\mathbf{D}_r$ is a diagonal matrix with $\lambda_i / (1 + \lambda_i)$, $i = 1, \ldots, r$, on its diagonal.[4] Defining

$$\hat{\boldsymbol{q}} = \left(\mathbf{I} - \mathbf{V}_r \mathbf{D}_r \mathbf{V}_r^* \right) \mathbf{\Gamma}_{\mathrm{prior}}^{1/2} \boldsymbol{z} \tag{4.8}$$

allows us to compute

$$\boldsymbol{q} = \mathbf{H}(\boldsymbol{w})^{-1} \boldsymbol{z} \approx \mathbf{\Gamma}_{\mathrm{prior}}^{1/2} \hat{\boldsymbol{q}}, \tag{4.9}$$

where the approximation is only due to the use of $\tilde{\mathbf{F}}_r$ instead of $\tilde{\mathbf{F}}$. To compute $\mathbf{F}\boldsymbol{q}$, as needed in (4.6), we use

$$\mathbf{F}\boldsymbol{q} = \mathbf{F} \mathbf{\Gamma}_{\mathrm{prior}}^{1/2} (\tilde{\mathbf{H}}_{\mathrm{misfit}}(\boldsymbol{w}) + \mathbf{I})^{-1} \mathbf{\Gamma}_{\mathrm{prior}}^{1/2} \boldsymbol{z} \approx \tilde{\mathbf{F}}_r \hat{\boldsymbol{q}}, \tag{4.10}$$

where the approximation is again due to the use of the low-rank surrogate for $\tilde{\mathbf{F}}$. We now summarize the procedure for computing $\Theta(\boldsymbol{w})$ and its gradient. Note that once $\tilde{\mathbf{F}}_r$ is available, our method does not require further forward or adjoint PDE solves.

We close this subsection with two remarks concerning the computations involved in Algorithm 1. First, note that while the low-rank SVD surrogate for $\tilde{\mathbf{F}}$ relieves the OED method of repeated PDE solves, we still require the repeated applications of $\mathbf{\Gamma}_{\mathrm{prior}}^{1/2}$. Due to our choice of the prior as the inverse of a squared elliptic operator, this amounts to an elliptic solve, for which optimal complexity solvers (e.g., multigrid) are available.

Next, note that we require the application of $\mathbf{M}^{-1/2}$ to random vectors for the trace estimator in the OED method; see (4.3). The same operation, which transforms vectors from the Euclidean space $\mathbb{R}^n$ to $\mathbb{R}_{\mathbf{M}}^n$ (see Section 2.2) via a mapping that preserves the inner products, is also needed to draw samples from the posterior distribution (see (2.7)). Since the explicit computation of the matrix square root is expensive in high dimensions, we utilize an iterative algorithm to compute the application of the inverse matrix square root to vectors [11]. The method relies on the approximation of the square root via orthogonal polynomials on an interval containing the spectrum of the matrix; since the convergence of this method is fastest for matrices with clustered eigenvalues, we apply it to the mass matrix $\mathbf{M}$, symmetrically preconditioned with the lumped mass matrix, $\mathbf{M}_l$. The eigenvalues of the resulting matrix, $\tilde{\mathbf{M}} = \mathbf{M}_l^{-1/2} \mathbf{M} \mathbf{M}_l^{-1/2}$, are clustered around 1, resulting in fast convergence when applying $\tilde{\mathbf{M}}^{-1/2}$ to vectors as discussed in [11]. Then, instead of $\mathbf{M}^{-1/2}$ we

---

[3]Note that thanks to the low-rank surrogate of the prior-preconditioned parameter-to-observable map $\tilde{\mathbf{F}}_r$, this computation does not require the solution of PDEs.

[4]Note that there exist alternatives to computing a spectral decomposition of $\tilde{\mathbf{H}}_{\mathrm{misfit},r}(\boldsymbol{w})$ and then using the Sherman-Morrison-Woodbury formula. For instance, one can use a Krylov method to compute $\mathbf{H}(\boldsymbol{w})^{-1} \boldsymbol{z}$.

Algorithm 1 Algorithm for computing $\Theta(\boldsymbol{w})$ and its gradient $\boldsymbol{g} = \nabla\Theta(\boldsymbol{w})$ with random vectors $\boldsymbol{z}^{(i)}$, $i = 1,\ldots,N_{\text{tr}}$ for the trace estimator.

Compute $\tilde{\mathbf{H}}_{\text{misfit},r}(\boldsymbol{w}) = \mathbf{V}_r\boldsymbol{\Lambda}_r\mathbf{V}_r^*$ {low-rank approx to $\tilde{\mathbf{H}}_{\text{misfit}}(\boldsymbol{w}) = \tilde{\mathbf{F}}^*\mathbf{W}\tilde{\mathbf{F}}$}

Set $\mathbf{D}_r = \mathsf{diag}\left(\frac{\lambda_1}{1+\lambda_1},\ldots,\frac{\lambda_r}{1+\lambda_r}\right)$

for $i = 1$ to N_{tr} **do**

    Compute $\hat{\boldsymbol{q}}^{(i)} = \left(\mathbf{I} - \mathbf{V}_r\mathbf{D}_r\mathbf{V}_r^*\right)\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{z}^{(i)}$ {as in (4.8)}

    Compute $\boldsymbol{q}^{(i)} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\hat{\boldsymbol{q}}^{(i)}$ {compute $\boldsymbol{q}^{(i)} = \mathbf{H}^{-1}\boldsymbol{z}^{(i)}$ as in (4.9)}

end for

Compute $\Theta(\boldsymbol{w}) = \dfrac{1}{N_{\text{tr}}}\displaystyle\sum_{i=1}^{N_{\text{tr}}}\left\langle\boldsymbol{z}^{(i)},\boldsymbol{q}^{(i)}\right\rangle_{\text{M}}$ {evaluation of $\Theta(\boldsymbol{w})$}

$\boldsymbol{g} = \boldsymbol{0}$ {initialize the gradient vector}

for $i = 1$ to N_{tr} **do**

    $\bar{\boldsymbol{d}} = \tilde{\mathbf{F}}_r\hat{\boldsymbol{q}}^{(i)}$ {compute $\bar{\boldsymbol{d}} = \mathbf{F}\boldsymbol{q}^{(i)}$ as in (4.10)}

    for $j = 1$ to N_s **do**

        $g_{i,j} = \displaystyle\sum_{\ell=1}^{N_\tau}\left(d_j^\ell\right)^2$ $\left\{g_{i,j} = \left\langle\boldsymbol{q}^{(i)},\dfrac{\partial\mathbf{H}_{\text{misfit}}}{\partial w_j}\boldsymbol{q}^{(i)}\right\rangle_{\text{M}}\text{ as in (4.6)}\right\}$

    end for

    $\boldsymbol{g} = \boldsymbol{g} - \frac{1}{N_{\text{tr}}}\boldsymbol{g}_i$

end for

---

use the matrix $\mathbf{L} = \mathbf{M}_l^{-1/2}\tilde{\mathbf{M}}^{-1/2}$ as the isomorphism between $\mathbb{R}^n$ and $\mathbb{R}^n_{\mathbf{M}}$. Note, in particular, that for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$

$$\langle\mathbf{L}\boldsymbol{x},\mathbf{L}\boldsymbol{y}\rangle_{\text{M}} = \boldsymbol{x}^T\tilde{\mathbf{M}}^{-\frac{1}{2}}\mathbf{M}_l^{-\frac{1}{2}}\mathbf{M}\mathbf{M}_l^{-\frac{1}{2}}\tilde{\mathbf{M}}^{-\frac{1}{2}}\boldsymbol{y} = \boldsymbol{x}^T\tilde{\mathbf{M}}^{-\frac{1}{2}}\tilde{\mathbf{M}}\tilde{\mathbf{M}}^{-\frac{1}{2}}\boldsymbol{y} = \langle\boldsymbol{x},\boldsymbol{y}\rangle,$$

which shows that $\mathbf{L}$ is a Hilbert space isomorphism between $\mathbb{R}^n$ and $\mathbb{R}^n_{\mathbf{M}}$.

4.5. Sparsity-enforcing penalty functions. Here, we discuss the penalty term $\gamma\Phi(\cdot)$ in (3.4). If $\Phi(\cdot)$ favors sparse solutions, the degree of sparsity depends on the choice of γ. However, γ only provides an indirect control of the sparsity of the design and in practice adjusting γ and resolving (3.4) might be necessary to obtain a design with a specified (or close to a specified) number of sensors.

Next, we discuss two possible choices for the penalty function $\Phi(\cdot)$. An ℓ_1-penalty function $\Phi(\boldsymbol{w}) = \mathbf{1}^T\boldsymbol{w}$ (see also [16–18]), is known to result in sparse solution vectors $\boldsymbol{w}$. While this choice is convenient due to the convexity of Φ, it results in a weight vector $\boldsymbol{w}$ whose components can take on any value in $[0,1]$. Unfortunately, the direct interpretation of weights $w_i \notin \{0,1\}$ for the placement of sensors is unclear. A practical approach to use such a solution vector is to place sensors in locations where the corresponding weight does not vanish. An alternative to ℓ_1-sparsification is to use non-convex functions for Φ that directly lead to binary weight vectors. This approach is discussed next.

To obtain binary weight vectors, we select penalties that successively approximate the ℓ_0-"norm",[5] which, for $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_0$ is defined as the number of non-zero elements in $\boldsymbol{x}$. One choice for such a family of penalizations is $\Phi_q(\boldsymbol{w}) = \|\boldsymbol{w}\|_q^q$, for $q < 1$. Since

---

[5] Note that $\|\boldsymbol{x}\|_0$ is in fact not a *norm*.

these penalization functions are not Lipschitz continuous at the origin, we use an alternative family of regularizations, namely, for $\varepsilon > 0$,

$$\Phi_\varepsilon(\boldsymbol{w}) := \sum_{i=1}^{N_s} f_\varepsilon(w_i), \quad \text{where} \quad f_\varepsilon(w) = \begin{cases} \dfrac{w}{\varepsilon}, & 0 \le w \le \tfrac{1}{2}\varepsilon, \\ p_\varepsilon(w), & \tfrac{1}{2}\varepsilon < w \le 2\varepsilon, \\ 1, & 2\varepsilon < w \le 1. \end{cases} \tag{4.11}$$

Here, $p_\varepsilon(\cdot)$ is the uniquely defined third order polynomial that makes $f_\varepsilon(x) : [0,1] \to [0,1]$ continuously differentiable. In Figure 4.1, we show these penalty functions for different values of ε; note how $f_\varepsilon(\cdot)$ approximates the ℓ_0-"norm" as $\varepsilon \to 0$.
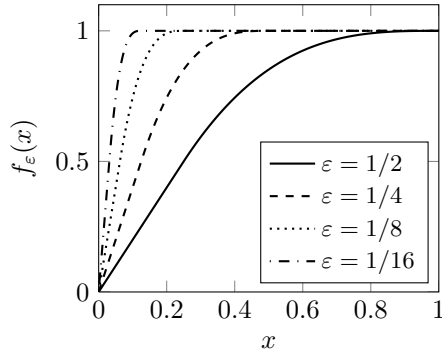

FIG. 4.1. *Graphs of $f_\varepsilon(\cdot)$ as defined in* (4.11) *for $\varepsilon = 1/2^i$, $i = 1, \dots, 4$.*

To cope with potential multiple local minima due to the non-convexity of the penalties Φ_ε, we use a continuation procedure with respect to ε. We fix the regularization parameter $\boldsymbol{\gamma}$ in (3.4) and solve the OED problem with ℓ_1-penalty, resulting in a weight vector $\boldsymbol{w}^0$. We then choose a decreasing sequence of positive real numbers $\{\varepsilon_i\}_{i \ge 1}$, and compute a new weight vector $\boldsymbol{w}^i$ by solving (3.4) with penalty function $\Phi_{\varepsilon_i}(\boldsymbol{w})$ using $\boldsymbol{w}^{i-1}$ as initialization. In practice, once ε_i is sufficiently small, the solution $\boldsymbol{w}^i$ is a 0–1 vector that remains unchanged as ε is further decreased.

A similar continuation strategy is used in topology optimization [4,5], where the target is to design an "optimal structure" by deciding on an optimal distribution of material in a physical domain $\mathcal{D}$. In this application, one seeks a *density* function $\rho : \mathcal{D} \to \{0,1\}$ that characterizes the absence or presence of material in $\mathcal{D}$. One of the main approaches to solving topology optimization problems is to relax the requirement of $\rho \in \{0,1\}$ to $0 \le \rho \le 1$ and to solve a sequence of optimization problems with successively steeper penalty functions to approach a 0–1 solution ρ— similar our Φ_ε-continuation approach outlined above.

In the rest of this paper, we refer to the designs obtained via the above continuation method as Φ_ε-*sparsified designs*. Optimal designs obtained by solving (3.4) with ℓ_1-penalty (followed by thresholding to obtain a 0–1 weight vector) are referred to as ℓ_1-*sparsified designs*. In the numerical experiments presented in Section 6, we compare ℓ_1- and Φ_ε-sparsified designs.

5. Model problem setup. Here, we present the model problem used to study the methods for OED presented in this paper. We consider a time-dependent advection-diffusion equation in which we seek to infer an unknown initial condition from point measurements; see also [1,14,29].
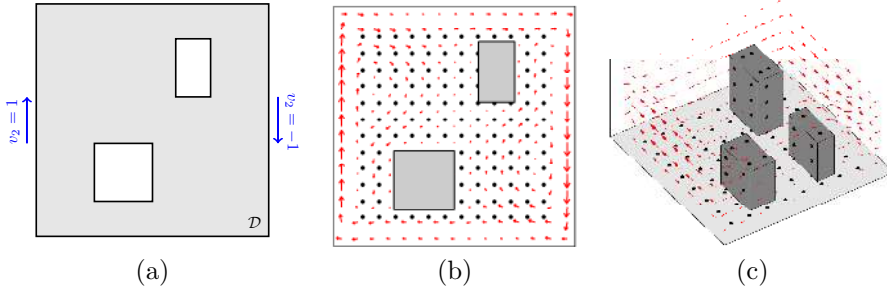
FIG. 5.1. *(a) The computational domain $\mathcal{D}$ for the two-dimensional model problem is $[0,1]^2$ with two rectangular regions (representing buildings) removed. (b) The velocity field $\boldsymbol{v}$ (red arrows) and the candidate sensor locations (black dots) for the two-dimensional problem. (c) The computational domain $\mathcal{D}$ for the three-dimensional model problem is $[0,1]^2 \times [0,0.5]$ with the buildings (gray blocks) removed; shown are also the velocity field (red arrows) and the candidate sensor locations (black dots).*

5.1. The parameter-to-observable map. The PDE in the parameter-to-observable map models diffusive transport in a domain $\mathcal{D} \subset \mathbb{R}^d$, which is depicted in Figure 5.1(a) for $d = 2$ and 5.1(c) for $d = 3$. The domain boundaries $\partial\mathcal{D}$ include the outer edges/faces as well as the internal boundaries of the rectangles/hexahedra, which model buildings. The parameter-to-observable map maps an initial condition $m \in L^2(\mathcal{D})$ to spatial and temporal point observations through the solution of the advection-diffusion equation $u(\boldsymbol{x}, t)$ given by

$$\begin{aligned}
u_t - \kappa\Delta u + \mathbf{v} \cdot \nabla u &= 0 &&\text{in } \mathcal{D} \times (0, T), \\
u(\cdot, 0) &= m &&\text{in } \mathcal{D}, \\
\kappa\nabla u \cdot \boldsymbol{n} &= 0 &&\text{on } \partial\mathcal{D} \times (0, T).
\end{aligned} \tag{5.1}$$

Here, $\kappa > 0$ is the diffusion coefficient and $T > 0$ is the final time. In our numerical experiments, we use $\kappa = 0.001$ for the two-dimensional problem and $\kappa = 0.003$ for the three-dimensional problem. The velocity field $\boldsymbol{v}$, shown in Figures 5.1(b), (c) for the two- and the three-dimensional problems, respectively, is computed by solving the following steady-state Navier-Stokes equation with the side walls driving the flow:

$$\begin{aligned}
-\frac{1}{\mathrm{Re}}\Delta\boldsymbol{v} + \nabla q + \boldsymbol{v} \cdot \nabla\boldsymbol{v} &= 0 &&\text{in } \mathcal{D}, \\
\nabla \cdot \boldsymbol{v} &= 0 &&\text{in } \mathcal{D}, \\
\boldsymbol{v} &= \boldsymbol{g} &&\text{on } \partial\mathcal{D}.
\end{aligned} \tag{5.2}$$

Here, q is pressure, Re is the Reynolds number, set to $\mathrm{Re} = 50$ in the present examples. The Dirichlet boundary data $\boldsymbol{g} \in \mathbb{R}^d$ is given by $\boldsymbol{g} = \boldsymbol{e}_2$ on the left wall of the domain, $\boldsymbol{g} = -\boldsymbol{e}_2$ on the right wall, and $\boldsymbol{g} = \boldsymbol{0}$ everywhere else (see, e.g., Figure 5.1(a) for $d = 2$). To illustrate the physics of the forward problem for a given initial condition, we show snapshots of the time evolution of the state variable u in Figure 5.2; for snapshots of the time evolution for the two-dimensional model problem we refer to [29].

The observation operator $\mathcal{B}$ then extracts the values of u on a set of sensor locations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_s}\} \subset \mathcal{D}$ at times $\{\tau_1, \ldots, \tau_{N_\tau}\} \subset [0, T]$. To summarize, the infinite-dimensional linear parameter-to-observable map $\mathcal{F}$ maps the initial condition m to $q = N_s N_\tau$ observations as follows: First, we solve the time-dependent advection-diffusion equation (5.1), resulting in the space-time solution $u = u(m)$. Then, the
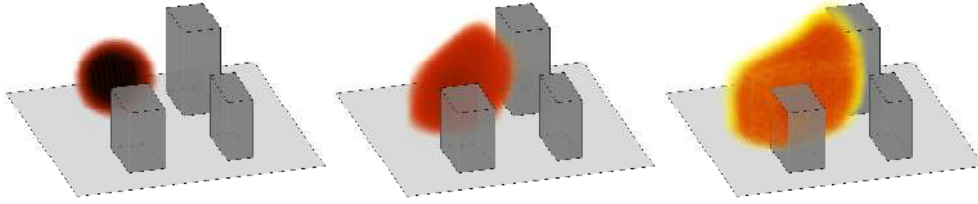
15

FIG. 5.2. *Volume renderings of the time evolution of u for the three-dimensional model problem. The initial condition m is shown in the left, and the middle and right images correspond to snapshots at times $t = 1$ and $t = 2$.*

observation operator extracts point values from u at the measurement locations and times, i.e., evaluates $\mathcal{B}u$. The corresponding discrete parameter-to-observable map $\mathbf{F}$ is obtained by discretizing m and u using, for instance, finite elements.

5.2. The Bayesian inverse problem. Following the Bayesian setup in Section 2.1, we utilize a Gaussian prior measure $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$, with $\mathcal{C}_0 = \mathcal{A}^{-2}$ where $\mathcal{A}$ is an elliptic differential operator as described in Section 2.1, and use an additive Gaussian noise model. Therefore, the solution of the Bayesian inverse problem is the posterior measure, $\mu_{\text{post}} = \mathcal{N}(m_{\text{post}}, \mathcal{C}_{\text{post}})$ with m_{post} and $\mathcal{C}_{\text{post}}$ as given in (2.5). The posterior mean m_{post} is characterized as the minimizer of

$$\mathcal{J}(m) := \frac{1}{2} \|\mathcal{B}u(m) - \boldsymbol{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2} \|\mathcal{A}(m - m_0)\|^2_{L^2(\mathcal{D})},$$

which can also be interpreted as the regularized functional to be minimized in deterministic inversion. Next, we specify the action of the adjoint $\mathcal{F}^*$ of the parameter-to-observable map. Given an observation vector $\bar{\boldsymbol{d}} \in \mathbb{R}^q$ as defined in (4.1), $\mathcal{F}^*\bar{\boldsymbol{d}}$ is computed by solving the *adjoint equation* (see [1, 14, 29]) for the adjoint variable $p = p(\boldsymbol{x}, t)$,

$$
\begin{aligned}
-p_t - \nabla \cdot (p\boldsymbol{v}) - \kappa\Delta p &= -\mathcal{B}^*\bar{\boldsymbol{d}} && \text{in } \mathcal{D} \times (0, T), \\
p(\cdot, T) &= 0 && \text{in } \mathcal{D}, \\
(\boldsymbol{v}p + \kappa\nabla p) \cdot \boldsymbol{n} &= 0 && \text{on } \partial\mathcal{D} \times (0, T),
\end{aligned}
\tag{5.3}
$$

and setting $\mathcal{F}^*\bar{\boldsymbol{d}} = p(\cdot, 0)$. Note that (5.3) is a final value problem, which has to be solved backwards in time.

5.3. Discretization and solvers. The discretization of the forward and adjoint problems uses linear triangular/tetrahedral continuous Galerkin finite elements for two/three space dimensions, and the implicit Euler method in time. The discrete adjoint equation is obtained as the adjoint of the discretized forward equation, i.e., we follow a discretize-then-optimize approach. Due to the large diffusion parameter κ used, no stabilization of the advection-diffusion equation is needed. A factorization of the matrix in each implicit Euler time step is computed upfront, such that every time step (of the forward as well as the adjoint equation) only requires triangular solves.

The OED optimization problems are solved using MATLAB's built-in interior-point solver `fmincon`, to which we supply the OED objective function evaluation and its derivative, which we compute using the methods presented above. The optimization algorithm approximates second derivatives using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [27] method.

6. Numerical results. In this section, we summarize our numerical experiments. In Section 6.1, we provide a detailed numerical study of our method and compare different sparsifications for the two-dimensional model problem. Then, in Section 6.2, we use the three-dimensional model problem to test the scalability and performance of our OED method.

6.1. The two-dimensional model problem. This section contains numerical experiments for the two-dimensional model problem. We use $N_s = 122$ candidate locations for placing sensors as shown in Figure 5.1(b), and use 1864 triangles to discretize the domain. Linear finite elements are used for the parameter m and the state variable u resulting in $n = 1012$ spatial degrees of freedom (and thus inversion parameters). The final time is $T = 4$ and the time interval is discretized using $N_t + 1$ implicit Euler time steps; unless otherwise specified, $N_t = 64$. Observations at the sensor locations are taken at $N_\tau = 19$ equally spaced points in the time interval $[1, 4]$. The prior covariance is specified as described in Section 2.1 with $\alpha = 8 \times 10^{-3}$ and $\beta = 10^{-2}$; these values result in a Gaussian prior that is not overly restrictive or smoothing. Next, we study the low-rank approximations of the forward operator and the misfit Hessian.

6.1.1. Low-rank approximation of F, $\tilde{\mathbf{F}}$ and $\mathbf{H_{misfit}}$. We use a randomized SVD to compute a low-rank surrogate for the parameter-to-observable map $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^q$ and its prior-preconditioned counterpart $\tilde{\mathbf{F}}$, where $q = N_s \times N_\tau$. This directly leads to a low-rank approximation of the prior-preconditioned misfit Hessian, $\tilde{\mathbf{H}}_{\mathrm{misfit}}(\boldsymbol{w}) = \tilde{\mathbf{F}}^* \mathbf{W} \tilde{\mathbf{F}}$, where the diagonal weight matrix $\mathbf{W}$ depends on the design. Since the diagonal entries of $\mathbf{W}$ may vanish, the numerical rank of $\tilde{\mathbf{H}}_{\mathrm{misfit}}(\boldsymbol{w})$ is less than or equal than that of $\tilde{\mathbf{F}}^* \tilde{\mathbf{F}}$, which corresponds to assigning unit weights to all sensors; thus, it suffices to examine the accuracy of the approximation of $\tilde{\mathbf{H}}_{\mathrm{misfit}} = \tilde{\mathbf{F}}^* \tilde{\mathbf{F}}$ only.

First, we study the low-rank approximation of the (prior-preconditioned) parameter-to-observable map. In Figure 6.1 (left), we show the normalized singular values of $\tilde{\mathbf{F}}$ and of $\mathbf{F}$. The eigenvalues decay rapidly and thus both mappings can be accurately approximated by low-rank matrices. Note the faster decay of the singular values of $\tilde{\mathbf{F}}$ compared to those of $\mathbf{F}$, which shows that due to the preconditioning with $\boldsymbol{\Gamma}_{\mathrm{prior}}^{1/2}$, $\tilde{\mathbf{F}}$ can be approximated more efficiently with a low-rank operator than $\mathbf{F}$.

Next, we investigate the influence of spatial and temporal discretization on the prior-preconditioned misfit Hessian; the right plot in Figure 6.1 shows the eigenvalues of $\tilde{\mathbf{H}}_{\mathrm{misfit}}$ for different spatial and temporal discretizations. Note that the spectra lie almost on top of each other. This indicates that the discretizations $\tilde{\mathbf{H}}_{\mathrm{misfit}}$ converge to their infinite-dimensional counterpart, $\tilde{\mathcal{H}}_{\mathrm{misfit}} : L^2(\mathcal{D}) \to L^2(\mathcal{D})$ and that all discretizations studied in Figure 6.1 resolve the dominant physics of the problem.

Finally, we study how the number of candidate sensor locations affects the spectral decay of $\tilde{\mathbf{F}}$. Increasing the number of sensors has the potential to increases the information in the data and thus the numerical rank of $\tilde{\mathbf{F}}$. In Figure 6.2, we plot the singular values of $\tilde{\mathbf{F}}$ for different sensor grids. We observe convergence of the singular value curves as the sensor grids are refined. This is a consequence of the correlation of the information that can be gained from neighboring sensors, which is due to the diffusion term in (5.1).

6.1.2. A-optimal designs with ℓ_1-sparsification. We now apply our method to compute ℓ_1-sparsified A-optimal designs. The weights $\boldsymbol{w}$ found with $\gamma = 6 \times 10^1$ are shown in Figure 6.3 (left). As discussed in Section 4.5, sensors are then placed
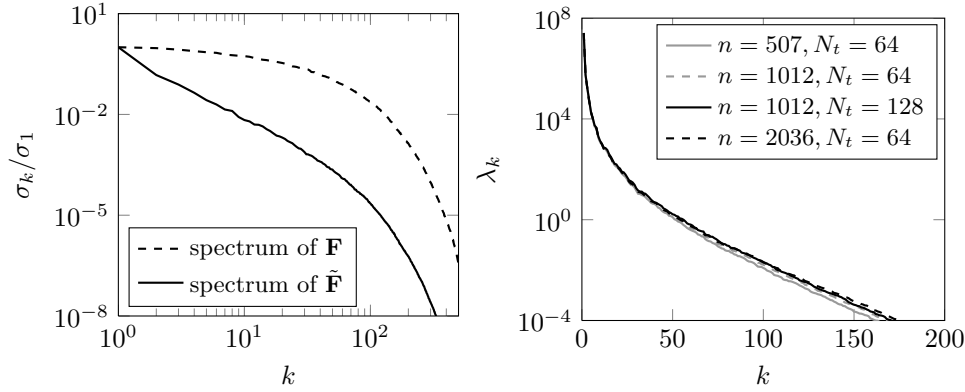
17

FIG. 6.1. *Shown on the left are the normalized singular values of $\tilde{\mathbf{F}}$ and $\mathbf{F}$. The right plot depicts the spectrum of the prior-preconditioned misfit Hessian with unit sensor weights $\tilde{\mathbf{H}}_{misfit} = \tilde{\mathbf{F}}^* \tilde{\mathbf{F}}$ for discretizations with different numbers n of spatial and N_t of temporal unknowns. To compute these spectra, we used a rank $r = 500$ approximation $\tilde{\mathbf{F}}_r$ to diminish the influence of the low-rank approximation for $\tilde{\mathbf{F}}$.*
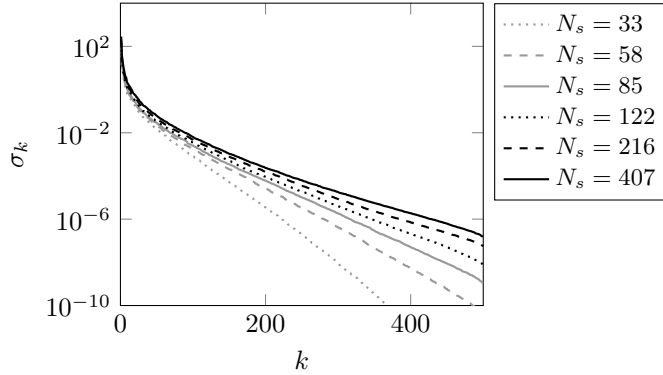

FIG. 6.2. *Singular values of $\tilde{\mathbf{F}}$ for different values of N_s. The results correspond to a discretization of the problem with parameter dimension $n = 1012$ and $N_t = 2^6$ time steps.*

at locations corresponding to non-vanishing weights. To account for numerical errors introduced by the interior point method used to solve (3.4), non-vanishing weights are defined as weights w_i that satisfy $w_i / \sum_j w_j > 4 \times 10^{-3}$; the resulting sensor locations are shown in Figure 6.3 (right). Next, we study the behavior of our method for computing ℓ_1-sparsified designs.

Since we rely on a low-rank surrogate $\tilde{\mathbf{F}}_r$ for the prior-preconditioned parameter-to-observable map $\tilde{\mathbf{F}}$ to avoid repeated PDE solves in the OED method, we first examine the influence of this truncation on the optimal design. For that purpose, we fix $\gamma = 5 \times 10^1$ and solve the A-optimal design problem using approximations $\tilde{\mathbf{F}}_r$ with $r = 10, 15, 20, 30, 40, 60, 80$. In Figure 6.4, we plot the optimal objective value $\Theta(\boldsymbol{w}_{\text{opt}})$ computed using low-rank surrogates $\tilde{\mathbf{F}}_r$ with different r. The convergence of the objective value shows that very similar results are found for low-rank surrogates $\tilde{\mathbf{F}}_r$ with $r \geq 40$. This shows that even a significant compression of the (preconditioned) parameter-to-observable map has little influence on the optimal design.

Next, we study the effect of increasing the parameter dimension (through mesh refinements) on the number of interior-point iterations required to solve the OED
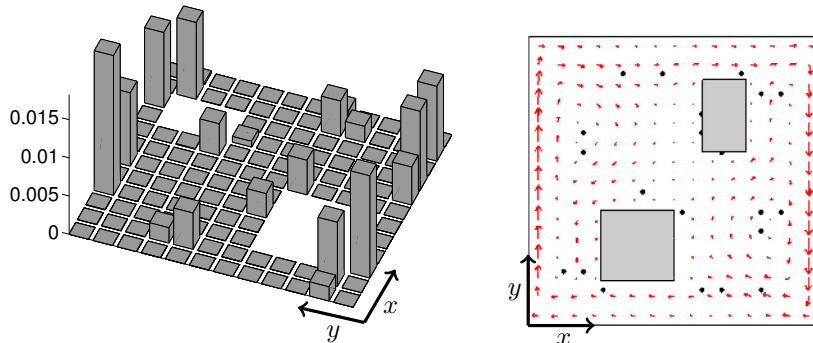
18

FIG. 6.3. *Shown on the left are the weights found by solving the optimization problem* (3.4) *with ℓ_1-sparsification. The right image shows the optimal sensor locations (black dots) obtained by placing sensors at locations with non-vanishing weights. Also shown is the advective flow field $\boldsymbol{v}$ (red arrows).*
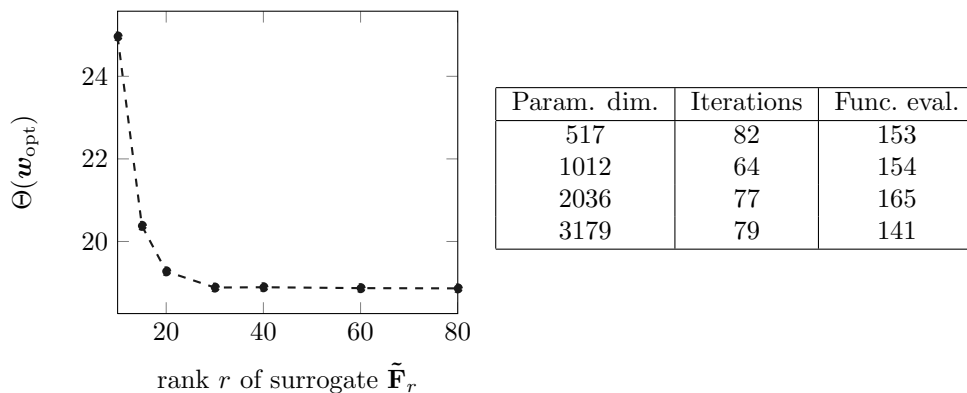

Param. dim.	Iterations	Func. eval.
517	82	153
1012	64	154
2036	77	165
3179	79	141

FIG. 6.4. *Left: Value of $\Theta(\boldsymbol{w}_{opt})$ for an ℓ_1-sparsified A-optimal design obtained with rank-r approximations $\tilde{\mathbf{F}}_r$. Right: Number of interior-point quasi Newton iterations and function evaluations for different discretizations (and thus different parameter dimensions).*

optimization problem (4.4). As observed in Section 6.1.1, increasing the parameter dimension does not increase the numerical rank of $\tilde{\mathbf{F}}$ once the mesh is sufficiently fine. This, and the use of a Newton-type optimization method results in a nearly constant number of interior-point iterations required to solve (4.4), as shown in the table in Figure 6.4.

Finally, we study how the performance of the numerical optimization encountered in our OED method is affected by increasing the number N_s of candidate sensor locations; note that N_s is also the dimension of the weight vector $\boldsymbol{w}$, the unknown in the OED optimization problem (4.4). The results shown in Table 6.1 show that the number of interior-point quasi-Newton iterations and the number of evaluations of $\Theta(\cdot)$ do not increase significantly as the number of sensor candidate locations increases. We again attribute this to the use of a Newton-type method, since Newton's method satisfies—under certain assumptions—a mesh independence property [12].

Regarding the complexity of our method with respect to the number N_s of candidate sensor locations, we draw the following conclusions from Figure 6.2 and Table 6.1: First, the number of PDE solves required to compute the low-rank surrogate $\tilde{\mathbf{F}}_r$ is

TABLE 6.1

Number of interior-point BFGS iterations and evaluations of $\Theta(\cdot)$ required for the solution of (4.4) with ℓ_1-sparsification, for different numbers N_s of sensor candidate locations. The optimization iteration was terminated when the norm of the gradient was decreased by a factor of 10^4.

N_s	33	58	85	122	172	216	264	340
Iterations	58	68	68	81	58	68	75	72
Func. eval.	149	165	141	158	136	137	122	160

bounded with respect to N_s, and second, the number of optimization iterations to solve (4.4) is insensitive to the number of candidate sensor locations, and thus the dimension of the weight vector $\boldsymbol{w}$.

6.1.3. A-optimal designs with Φ_ε-sparsification. Next, we present Φ_ε-sparsified designs obtained by a continuation procedure with respect to ε. Optimal designs for $\gamma = 0.05$ and $\gamma = 0.06$ are shown on the left in Figure 6.5. Note that the optimal sensor locations obtained with $\gamma = 0.06$ is not a subset of the locations obtained with $\gamma = 0.05$, i.e., the designs are not nested.

For the continuation procedure, we use the values $\varepsilon_i = (2/3)^i$, $i = 1, \ldots, 10$. To elucidate the convergence of the weights to the desired binary structure, the weights corresponding to several values of ε_i are shown on the right in Figure 6.5. Note that the behavior of the weights is not monotone as ε is decreased. A possible explanation for this behavior is that, as ε decreases, weights at neighboring locations can get merged into a single weight at a new location.
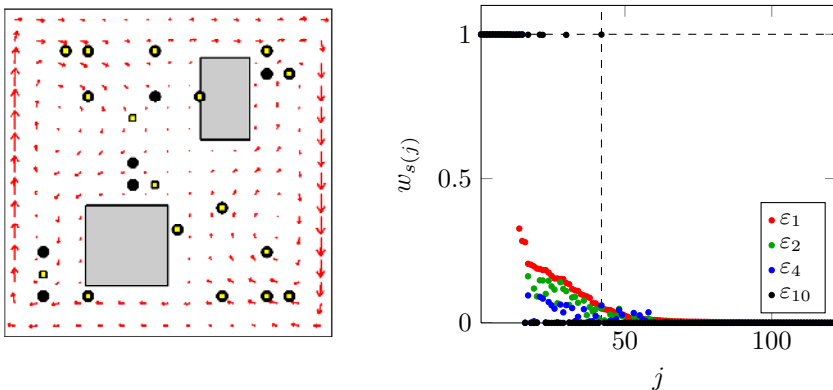

FIG. 6.5. *Left: Sensor locations using Φ_ε-sparsified A-optimal design for $\gamma = 0.05$ (black dots) and for $\gamma = 0.06$ (yellow squares). Right: Convergence of the weights to a 0–1 structure as $\varepsilon \to 0$. The weights are ordered based on the values obtained with ε_1, and shown is the evolution of the weights as ε decreases. Here, $s(j)$ denotes the descending ordering obtained with ε_1.*

To illustrate the decreased variance for A-optimal designs, in Figure 6.6 we compare the pointwise posterior standard deviation obtained with $\gamma = 0.05$ with a uniform and two randomly generated designs employing the same number of sensors. This comparison shows that the A-optimal design provides a notable reduction in the standard deviation (and thus the variance) fields. Next, we compare the performance of sensor placements obtained with ℓ_1- and Φ_ε-sparsification.

6.1.4. Comparing ℓ_1- and Φ_ε-sparsifications. We observe that ℓ_1- and Φ_ε-sparsifications lead to different designs for the same number of sensors. Thus, a
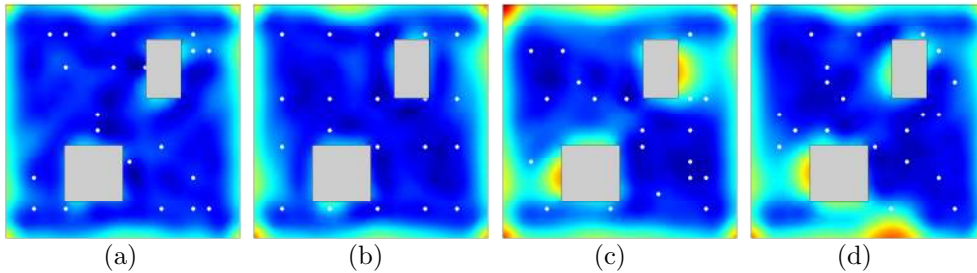
| (a) | (b) | (c) | (d) |

Fig. 6.6. *Shown is the pointwise standard deviation of the posterior for a Φ_ε-sparsified A-optimal design with 20 sensors (a), for a manually chosen uniform design (b), and two randomly generated designs with the same number of sensors (c), (d). The white dots indicate the sensor locations. Red and blue correspond to regions with large and small standard deviation, respectively. We find that compared to the optimal design, the uniform design and the two random designs result in 7%, 36% and 26% increase in the average variance, respectively.*

naturally arising question is (1) *which sparsification results in the better placement of sensors?* Moreover, (2) how much do *optimal* designs improve over *randomly* chosen designs? To answer these questions, we compute optimal designs based on different sparsification strategies and compare with randomly generated designs. We use trace estimators with 100 Gaussian random vectors chosen differently in each OED problem, but report the exact value of $\mathsf{tr}(\mathbf{\Gamma}_{\mathrm{post}}(\boldsymbol{w})) = \mathsf{tr}(\mathbf{H}(\boldsymbol{w})^{-1})$.

Our results are summarized in Figure 6.8. We compute ℓ_1-sparsified designs for various values of γ and report the value of $\mathsf{tr}(\mathbf{H}(\boldsymbol{w})^{-1})$. Similarly, we report $\mathsf{tr}(\mathbf{H}(\boldsymbol{w})^{-1})$ for 0–1 optimal designs computed via Φ_ε-sparsification. Additionally, we compute $\mathsf{tr}(\mathbf{H}(\boldsymbol{w})^{-1})$ for a collection of random sensor configurations. We find that Φ_ε-sparsified designs consistently outperform ℓ_1-sparsified designs and that both improve significantly over randomly chosen designs. Another observation from Figure 6.8 is the diminishing returns as the number of sensors is increased; using more than 20 sensors only results in negligible decrease of the OED objective function value.

6.1.5. Influence of trace estimation. To assess the accuracy of randomized trace estimation for a typical posterior covariance matrix, we compare the exact trace with results from randomized trace estimation with different numbers of random vectors. We find that trace estimators based on 1, 5, 10, 20, 100 vectors estimate the exact trace with an average error of about 15%, 7%, 5%, 2% and 1.5%, respectively. This is consistent with the experiments reported in [3] showing that randomized trace estimation is reasonably accurate with a small number of random vectors, but that many random vectors may be needed to obtain very accurate approximations.

A more relevant question in the context of OED is how the use of trace estimators to compute optimal sensor locations influences the resulting designs. To study this issue, we compute optimal designs with fixed γ, but different trace estimators. The results are shown in Figure 6.7. Note that trace estimation in the OED objective function does have an influence on the designs. More accurate trace estimation decreases the variation in both the number of sensors of the designs and the value of the trace of the posterior covariance. However, as can be seen from Figure 6.8 optimal designs computed using trace estimation (based on 100 random vectors) consistently improve over random designs with respect to reducing the *exact* trace of the posterior covariance. We thus conclude that the use of trace estimation does not have a significant impact on the quality of the optimal designs.
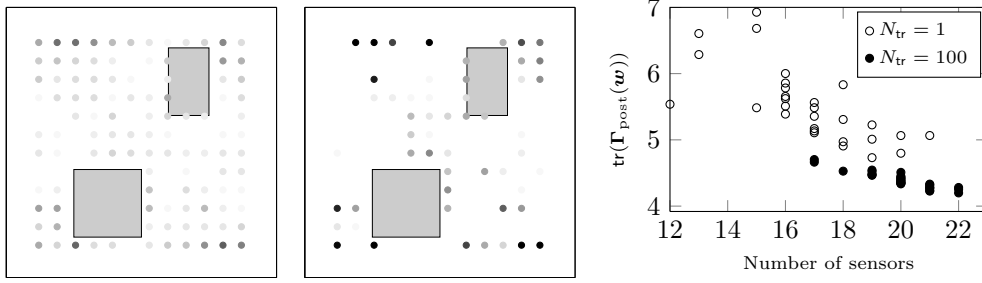
FIG. 6.7. *Sensitivity of Φ_ε-sparsified designs with respect to the trace estimator with fixed $\gamma = 0.05$. The left and the middle image visualize the frequency of a candidate location being part of the optimal design with a trace estimator based on a single random vector (left) and 100 random vectors (middle). The results are based on 30 different realizations of the respective trace estimators. The darker the candidate location, the more often the corresponding sensor was part of the optimal design. Note the decreased variation in the designs in the middle plot, which is due to the increased accuracy of the trace estimator. The right plot depicts the number of sensors and the value of the trace of the posterior covariance for 30 different trace estimators with a single random vector (empty dots) and with 100 random vectors (filled dots). Note that the number of sensors varies for the different realizations and that the more accurate trace estimator results in less scattering with respect to the number of sensors and the value of the trace.*
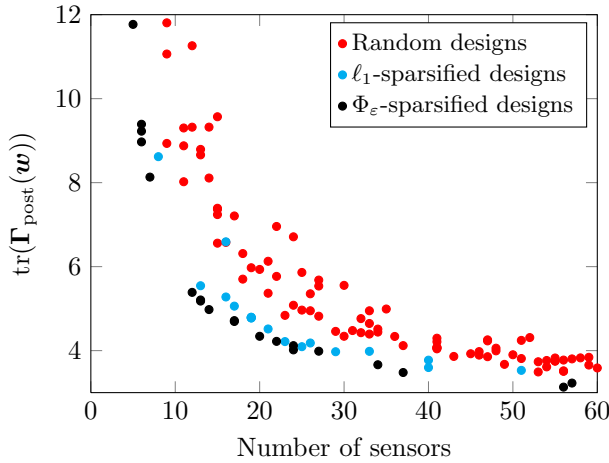

FIG. 6.8. *Comparison of performance for different designs. Shown is the value of $\mathsf{tr}(\mathbf{H}^{-1})$ versus the number of sensors for random designs (red dots), ℓ_1-sparsified designs (cyan dots) and Φ_ε-sparsified designs (black dots).*

6.2. The three-dimensional model problem. The main target of this three dimensional model problem is to study the applicability of our OED method to large-scale inverse problems. The computational domain used is depicted in Figure 5.1(c), where 101 candidate sensor locations are shown as black dots. Note that we allow sensors on the ground and on the sides of the buildings. Observations are collected at $N_\tau = 12$ equally spaced points in the time interval $[1, T]$, where $T = 4$. The parameter and the state variable u in the advection-diffusion equation are discretized using linear finite elements on a tetrahedral mesh with 10,652 spatial degrees of freedom, and 64 implicit Euler time steps are used for the time integration. Thus, the dimension of the inversion parameter is $n = 10,652$.
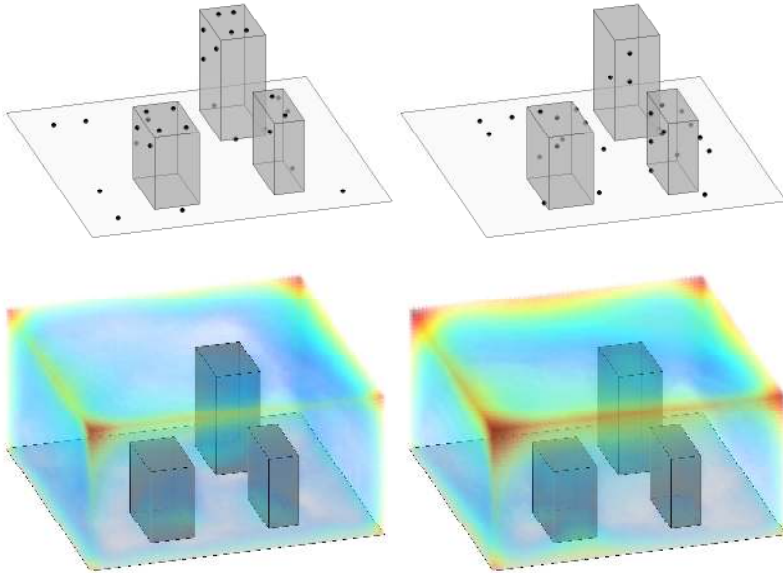
FIG. 6.9. *Top row: Sensor locations (black dots) obtained with A-optimal design with Φ_ε-sparsification (left) and with random design (right). Bottom row: Standard deviation fields of the posterior for an A-optimal design with Φ_ε-sparsification (left) and random design (right). Here, red indicates regions with large standard deviation, and blue corresponds to regions with small standard deviation.*

In our computations, we found a rank-100 SVD surrogate for the prior-preconditioned parameter-to-observable map adequate since the 100th singular value of $\tilde{\mathbf{F}}$ is about four orders of magnitude smaller than the largest singular value. This translates to eight orders of magnitude difference between the first and the 100th eigenvalue of $\tilde{\mathbf{H}}_{\text{misfit}}$. Due to the large parameter space dimension we do not compute $\mathbf{M}^{-1/2}$ explicitly, but use the algorithm for the iterative application of $\mathbf{M}^{-1/2}$ to vectors from [11]; see also Section 4.4. We observe that the algorithm is well converged after 10 applications of $\tilde{\mathbf{M}}$—the relative difference between the iterative application of $\tilde{\mathbf{M}}^{-1/2}$ based on 10 applications of $\tilde{\mathbf{M}}$ compared with that based on 500 applications of $\tilde{\mathbf{M}}$ is approximately 8×10^{-6}. For the solution of the OED optimization problem with $\gamma = 1.2 \times 10^{-1}$ and Φ_ε-sparsification, the continuation procedure presented in Section 6.1.2 is used. To compute the initial ℓ_1-sparsified weight vector $\boldsymbol{w}^0$, 18 interior-point iterations were necessary. This initialization was followed by 10 continuation steps using the penalties Φ_{ε_i}, with $\varepsilon_i = (2/3)^i$; each step required the solution of an auxiliary OED problem, which amounted to a total of 554 interior-point quasi-Newton iterations to arrive at a 0–1 design vector. The iteration for each of these auxiliary OED problems was terminated when the relative residual dropped below 10^{-4} or a maximum of 150 iterations was reached. The number of quasi-Newton iterations in the continuation procedure can likely be decreased by incorporating the continuation procedure into the interior-point algorithm rather than solving an independent interior point problem for each ε_i.

The resulting optimal design is shown in top left image of Figure 6.9; note that several sensors are placed on the top of buildings; these sensors help reduce the variance in the upper parts of the domain. To illustrate the effectiveness of the optimal design to reduce the variance, in Figure 6.9 we show a volume rendering of the point-

23

wise posterior standard deviation obtained with the optimal design, and compare it with a rendering obtained with a randomly generated design that has the same number of sensors.

7. Concluding remarks. We have developed a structure-exploiting efficient numerical method for computing A-optimal experimental designs in infinite-dimensional Bayesian linear inverse problems governed by PDEs. Numerical experiments for the inversion of the initial condition in an advection-diffusion equation indicate that an optimal design can be computed at a cost, measured in forward PDE solves, that is independent of the parameter and candidate sensor dimensions. Moreover, we find that experimental designs obtained with regularized ℓ_0-sparsification are consistently superior to designs obtained with ℓ_1-sparsification, and significantly improve over uniform and random designs.

One limitation of the OED method we have presented is that it relies on linearity of the parameter-to-observable map and assumes Gaussian prior and noise distributions. However, the methods presented here are also applicable in situations where a nonlinear parameter-to-observable map can be well approximated by a linearization over the set of parameters that have significant posterior probability. Moreover, using Gaussian noise and prior distributions is common, particularly in infinite-dimensional inference problems. The computational efficiency of our methods depends on low-rank approximations of the preconditioned parameter-to-observable map, which rely on properties of the associated forward and observation operators. As a consequence of ill-posedness, many parameter-to-observable operators admit such low-rank approximations. Finally, our relaxation of the design problem using a continuous weight vector in combination with sparsification provides just indirect control of the number of sensors through the value of γ. However, this approach makes the combinatorial OED problem of optimal sensor placement computationally tractable.

Possible extensions of the present work include consideration of (1) alternative optimal experimental design criteria that are meaningful in infinite dimensions and (2) nonlinear parameter-to-observable maps. OED with nonlinear parameter-to-observable maps is particularly challenging as, in general, the posterior is non-Gaussian, the (linearization of the) parameter-to-observable map in general depends on the state, parameter, and data variables and thus its low-rank approximation cannot be computed a priori, optimal designs might not be unique even if the sparsifying penalty is convex, and the misfit Hessian depends on (usually unavailable) observations.

Appendix A. The mass-weighted trace estimator. The following basic result justifies the form of the randomized trace estimator for an $\mathbf{M}$-symmetric matrix as defined in section 2.2. The proof given below adapts the arguments given in [3] regarding Gaussian trace estimators for symmetric linear mappings on the standard Euclidean inner product space.

PROPOSITION A.1. *Let $\mathbf{A}$ be an $\mathbf{M}$-symmetric linear mapping on $\mathbb{R}^n_{\mathbf{M}}$, $n \geq 1$. Suppose $\boldsymbol{y}$ is a random n-vector with i.i.d. $\mathcal{N}(0,1)$-entries, and let $\boldsymbol{z} = \mathbf{M}^{-1/2}\boldsymbol{y}$. Then, $T(\mathbf{A}) := \langle \boldsymbol{z}, \mathbf{A}\boldsymbol{z} \rangle_M$ is an unbiased estimator for $\mathsf{tr}(\mathbf{A})$; that is, $\mathsf{E}\{T(\mathbf{A})\} = \mathsf{tr}(\mathbf{A})$.*

Proof. Since $\mathbf{A}$ is $\mathbf{M}$-symmetric, it admits a spectral decomposition, $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$, with $\mathbf{V}$ a matrix with $\mathbf{M}$-orthogonal eigenvectors of $\mathbf{A}$ as its columns (i.e., $\mathbf{V}^T\mathbf{M}\mathbf{V} = \mathbf{I}$) and $\mathbf{\Lambda}$ a diagonal matrix with real eigenvalues $\{\lambda_i\}_{i=1}^n$ of $\mathbf{A}$ on its diagonal. Then,

$$\mathsf{E}\{T(\mathbf{A})\} = \mathsf{E}\left\{\boldsymbol{z}^T\mathbf{M}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^*\boldsymbol{z}\right\} = \mathsf{E}\left\{\boldsymbol{y}^T\mathbf{M}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{M}^{1/2}\boldsymbol{y}\right\} = \mathsf{E}\left\{\boldsymbol{q}^T\mathbf{\Lambda}\boldsymbol{q}\right\},$$

with $\boldsymbol{q} = \mathbf{V}^T\mathbf{M}^{1/2}\boldsymbol{y}$; note that above we also used $\mathbf{V}^* = \mathbf{V}^T\mathbf{M}$. It is straightforward to show that $\boldsymbol{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore,

$$\mathsf{E}\{T(\mathbf{A})\} = \mathsf{E}\left\{\boldsymbol{q}^T\mathbf{\Lambda}\boldsymbol{q}\right\} = \sum_{i=1}^n \lambda_i \mathsf{E}\left\{q_i^2\right\} = \sum_{i=1}^n \lambda_i = \mathsf{tr}(\mathbf{A}),$$

where the penultimate equality follows from the fact that $\mathsf{E}\left\{q_i^2\right\} = 1$, which is the case because q_i^2 is the square of standard normal random variable and is thus χ^2 distributed with one degree of freedom. $\square$

Appendix B. Relation between the trace of posterior covariance and the expected MSE of the posterior mean. Here, we show that

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^q} \left\|\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}} - \boldsymbol{m}\right\|_{\mathbf{M}}^2 \, d\mu_{\boldsymbol{d}|\boldsymbol{m}}(\boldsymbol{d}) \, d\mu_0(\boldsymbol{m}) = \mathsf{tr}(\mathbf{\Gamma}_{\text{post}}), \tag{B.1}$$

where, according to our choice of the noise model, $\mu_{\boldsymbol{d}|\boldsymbol{m}} = \mathcal{N}(\mathbf{F}\boldsymbol{m}, \mathbf{\Gamma}_{\text{noise}})$ and μ_0 is the (discretized) prior measure over $\mathbb{R}_{\mathbf{M}}^n$, $\mu_0 = \mathcal{N}(\boldsymbol{m}_0, \mathbf{\Gamma}_{\text{prior}})$. While (B.1) is known [10] in the context of inference problems in n-dimensional Euclidean space, below we include a proof of this result adjusted to our choice of mass-weighted inner product, which results from a consistent discretization of the infinite-dimensional Bayesian inverse problem. We use the notation $\boldsymbol{m}_{\text{post}} = \boldsymbol{m}_{\text{post}}^{\boldsymbol{d}}$ to make the dependence of the posterior mean to data explicit. For brevity, we write (B.1) as

$$\mathsf{E}_{\mu_0}\left\{\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\left\|\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}} - \boldsymbol{m}\right\|_{\mathbf{M}}^2\right\}\right\} = \mathsf{tr}(\mathbf{\Gamma}_{\text{post}}). \tag{B.2}$$

Suppose μ is a Gaussian measure with mean $\boldsymbol{z}_0$ and covariance $\mathbf{Q}$ defined on a Hilbert space V. In the proof of the following result we use the fact that $\int_V \|\boldsymbol{\xi}\|^2 \, d\mu(\boldsymbol{\xi}) = \mathsf{tr}(\mathbf{Q}) + \|\boldsymbol{z}_0\|^2$. Moreover if $\boldsymbol{z}$ is Gaussian random variable with probability law μ and $\mathbf{A} : V \to W$ is a linear transformation with W a Hilbert space, then $\mathbf{A}\boldsymbol{z} + \boldsymbol{b}$ is also Gaussian with law $\mathcal{N}(\mathbf{A}\boldsymbol{z}_0 + \boldsymbol{b}, \mathbf{A}\mathbf{Q}\mathbf{A}^*)$, where $\mathbf{A}^*$ is the adjoint of $\mathbf{A}$.

PROPOSITION B.1. *Consider a Bayesian linear inverse problem with a Gaussian prior $\mu_0 = \mathcal{N}(\boldsymbol{m}_0, \mathbf{\Gamma}_{\text{prior}})$ on $\mathbb{R}_{\mathbf{M}}^n$ with additive Gaussian noise model as described above. Then, (B.2) holds.*

Proof. An elementary calculation shows that,

$$\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\left\|\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}} - \boldsymbol{m}\right\|_{\mathbf{M}}^2\right\} =$$
$$\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\left\|\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}} - \mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}}\right\}\right\|_{\mathbf{M}}^2\right\} + \left\|\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}}\right\} - \boldsymbol{m}\right\|_{\mathbf{M}}^2, \tag{B.3}$$

Let us consider the first term on the right-hand-side of (B.3); denote $S\boldsymbol{d} = \boldsymbol{m}_{\text{post}}^{\boldsymbol{d}} - \mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\boldsymbol{m}_{\text{post}}^{\boldsymbol{d}}\right\}$. Recalling that $\boldsymbol{d}|\boldsymbol{m} \sim \mathcal{N}(\mathbf{F}\boldsymbol{m}, \mathbf{\Gamma}_{\text{noise}})$, we note that for any fixed $\boldsymbol{m}$, we have $S\boldsymbol{d} = \mathbf{\Gamma}_{\text{post}}\mathbf{F}^*\mathbf{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{d} - \mathbf{F}\boldsymbol{m})$. Therefore, the law of the random variable

$S : (\mathbb{R}^q, \mu_{\boldsymbol{d}|\boldsymbol{m}}) \to \mathbb{R}^n_{\mathbf{M}}$ is $\mu_S = \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{post}}\mathbf{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{post}})$, where $\mathbf{H}_{\text{misfit}} = \mathbf{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{F}$. Therefore,

$$\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\left\|\boldsymbol{m}^{\boldsymbol{d}}_{\text{post}} - \mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\boldsymbol{m}^{\boldsymbol{d}}_{\text{post}}\right\}\right\|^2_{\mathbf{M}}\right\} = \int_{\mathbb{R}^n} \|\boldsymbol{\xi}\|^2_{\mathbf{M}}\, d\mu_S(\boldsymbol{\xi}) = \mathsf{tr}(\boldsymbol{\Gamma}^2_{\text{post}}\mathbf{H}_{\text{misfit}}), \quad \text{(B.4)}$$

where we used that $\mathsf{tr}(\boldsymbol{\Gamma}_{\text{post}}\mathbf{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{post}}) = \mathsf{tr}(\boldsymbol{\Gamma}^2_{\text{post}}\mathbf{H}_{\text{misfit}})$. Next, we consider the expectation over μ_0 of the second term in (B.3). We let $T : (\mathbb{R}^n_{\mathbf{M}}, \mu_0) \to \mathbb{R}^n_{\mathbf{M}}$ be defined by $T\boldsymbol{m} = \boldsymbol{m} - \mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\boldsymbol{m}^{\boldsymbol{d}}_{\text{post}}\right\}$. Then,

$$T\boldsymbol{m} = \boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{\Gamma}^{-1}_{\text{post}} - \mathbf{H}_{\text{misfit}})\boldsymbol{m} - \boldsymbol{\Gamma}_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}}\boldsymbol{m}_0 = \boldsymbol{\Gamma}_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}}(\boldsymbol{m} - \boldsymbol{m}_0). \quad \text{(B.5)}$$

Hence, the law of T is given by $\mu_T = \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} = (\boldsymbol{\Gamma}_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}})\boldsymbol{\Gamma}_{\text{prior}}(\boldsymbol{\Gamma}_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}})^* = \boldsymbol{\Gamma}_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}}\boldsymbol{\Gamma}_{\text{post}}$. Therefore,

$$\mathsf{E}_{\mu_0}\left\{\left\|\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\boldsymbol{m}^{\boldsymbol{d}}_{\text{post}}\right\} - \boldsymbol{m}\right\|^2_{\mathbf{M}}\right\} = \int_{\mathbb{R}^n} \|\boldsymbol{\xi}\|^2_{\mathbf{M}}\, d\mu_T(\boldsymbol{\xi}) = \mathsf{tr}(\boldsymbol{\Gamma}^2_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}}). \quad \text{(B.6)}$$

Hence, using (B.3), (B.4), and (B.6), we have

$$\mathsf{E}_{\boldsymbol{d}|\boldsymbol{m}}\left\{\left\|\boldsymbol{m}^{\boldsymbol{d}}_{\text{post}} - \boldsymbol{m}\right\|^2_{\mathbf{M}}\right\} = \mathsf{tr}(\boldsymbol{\Gamma}^2_{\text{post}}\mathbf{H}_{\text{misfit}}) + \mathsf{tr}(\boldsymbol{\Gamma}^2_{\text{post}}\boldsymbol{\Gamma}^{-1}_{\text{prior}})$$
$$= \mathsf{tr}\big(\boldsymbol{\Gamma}^2_{\text{post}}(\mathbf{H}_{\text{misfit}} + \boldsymbol{\Gamma}^{-1}_{\text{prior}})\big) = \mathsf{tr}(\boldsymbol{\Gamma}_{\text{post}}),$$

where, for the last equality we used $\mathbf{H}_{\text{misfit}} + \boldsymbol{\Gamma}^{-1}_{\text{prior}} = \boldsymbol{\Gamma}^{-1}_{\text{post}}$. $\square$

REFERENCES

[1] Volkan Akçelik, George Biros, Andrei Draganescu, Omar Ghattas, Judith Hill, and Bart van Bloeman Waanders, *Dynamic data-driven inversion for terascale simulations: Real-time identification of airborne contaminants*, in Proceedings of SC2005, Seattle, 2005.

[2] Anthony C. Atkinson and Alexander N. Donev, *Optimum Experimental Designs*, Oxford, 1992.

[3] Haim Avron and Sivan Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, Journal of the ACM (JACM), 58 (2011), p. 17.

[4] Martin P. Bendsøe, *Optimization of structural topology, shape and material*, Springer Verlag, Berlin, Heidelberg, New York, 1995.

[5] Martin P. Bendsøe and Ole Sigmund, *Topology Optimization. Theory, methods and Applicatons*, Springer Verlag, Berlin, Heidelberg, New York, 2003.

[6] Hans Georg Bock, Stefan Körkel, and Johannes P. Schlöder, *Parameter estimation and optimum experimental design for differential equation models*, in Model Based Parameter Estimation, Hans Georg Bock, Thomas Carraro, Willi Jäger, Stefan Körkel, Rolf Rannacher, and Johannes P. Schlöder, eds., vol. 4 of Contributions in Mathematical and Computational Sciences, Springer Berlin Heidelberg, 2013, pp. 1–30.

[7] Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C. Wilcox, *Extreme-scale UQ for Bayesian inverse problems governed by PDEs*, in SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2012.

[8] Tan Bui-Thanh, Omar Ghattas, James Martin, and Georg Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion*, SIAM Journal on Scientific Computing, 35 (2013), pp. A2494–A2523.

[9] Emmanuel J. Candès, Justin Romberg, and Terence Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics, 59 (2006), pp. 1207–1223.

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

[11] Jie Chen, Mihai Anitescu, and Yousef Saad, *Computing $f(A)b$ via least squares polynomial approximations*, SIAM Journal on Scientific Computing, 33 (2011), pp. 195–222.

[12] Peter Deuflhard, *Newton methods for nonlinear problems*, vol. 35 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2004.

[13] David L. Donoho, *Compressed sensing*, IEEE. Transactions on Information Theory, 52 (2006), pp. 1289–1306.

[14] H. Pearl Flath, Lucas C. Wilcox, Volkan Akçelik, Judy Hill, Bart van Bloemen Waanders, and Omar Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM Journal on Scientific Computing, 33 (2011), pp. 407–432.

[15] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, third ed., 1996.

[16] Eldad Haber, Lior Horesh, and Luis Tenorio, *Numerical methods for experimental design of large-scale linear ill-posed inverse problems*, Inverse Problems, 24 (2008), pp. 125–137.

[17] Eldad Haber, Lior Horesh, and Luis Tenorio, *Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems*, Inverse Problems, 26 (2010), p. 025002.

[18] Eldad Haber, Zhuojun Magnant, Christian Lucero, and Luis Tenorio, *Numerical methods for A-optimal designs with a sparsity constraint for ill-posed inverse problems*, Computational Optimization and Applications, (2012), pp. 1–22.

[19] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review, 53 (2011), pp. 217–288.

[20] Lior Horesh, Eldad Haber, and Luis Tenorio, *Optimal experimental design for the large-scale nonlinear ill-posed problem of impedance imaging*, Large-Scale Inverse Problems and Quantification of Uncertainty, (2010), pp. 273–290.

[21] Xun Huan and Youssef M. Marzouk, *Simulation-based optimal Bayesian experimental design for nonlinear systems*, Journal of Computational Physics, 232 (2013), pp. 288–317.

[22] ———, *Gradient-based stochastic optimization methods in Bayesian experimental design*, International Journal for Uncertainty Quantification, (2014). to appear.

[23] Michael F. Hutchinson, *A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines*, Communications in Statistics-Simulation and Computation, 19 (1990), pp. 433–450.

[24] Peter D. Lax, *Functional Analysis*, John Wiley & Sons, New-York, Chicester, Brisbane, Toronto, 2002.

[25] James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM Journal on Scientific Computing, 34 (2012), pp. A1460–A1487.

[26] James Mercer, *Functions of positive and negative type, and their connection with the theory of integral equations*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, (1909), pp. 415–446.

[27] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, Springer Verlag, Berlin, Heidelberg, New York, second ed., 2006.

[28] Andrej Pázman, *Foundations of Optimum Experimental Design*, D. Reidel Publishing Co., 1986.

[29] Noemi Petra and Georg Stadler, *Model variational inverse problems governed by partial differential equations*, Tech. Report 11-05, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, 2011.

[30] Friedrich Pukelsheim, *Optimal Design of Experiments*, John Wiley & Sons, New-York, 1993.

[31] Farbod Roosta-Khorasani and Uri Ascher, *Improved bounds on sample size for implicit matrix trace estimators*. Preprint, 2013.

[32] Andrew M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559.

[33] Dariusz Uciński, *Optimal measurement methods for distributed parameter system identification*, CRC Press, Boca Raton, 2005.

[34] David Williams, *Probability with Martingales*, Cambridge University Press, 1991.