



ESTIMATING AND USING INFORMATION IN INVERSE PROBLEMS

WOLFGANG BANGERTH^{✉1}, CHRIS R. JOHNSON^{✉2}, DENNIS K. NJERU^{✉2}
AND BART VAN BLOEMEN WAANDERS^{✉3}

¹Department of Mathematics, Department of Geosciences,
Colorado State University, Fort Collins, CO 80523-1874, USA

²Scientific Computing and Imaging Institute,
University of Utah, Salt Lake City, UT 84112, USA

³Sandia National Laboratories, MS 0370, PO Box 5800, Albuquerque, NM 87185, USA

(Communicated by Samuli Siltanen)

ABSTRACT. In inverse problems, one attempts to infer spatially variable functions from indirect measurements of a system. To practitioners of inverse problems, the concept of “information” is familiar when discussing key questions such as which parts of the function can be inferred accurately and which cannot. For example, it is generally understood that we can identify system parameters accurately only close to detectors, or along ray paths between sources and detectors, because we have “the most information” for these places.

Although referenced in many publications, the “information” that is invoked in such contexts is not a well understood and clearly defined quantity. Herein, we present a definition of *information density* that is based on the variance of coefficients as derived from a Bayesian reformulation of the inverse problem. We then discuss three areas in which this information density can be useful in practical algorithms for the solution of inverse problems, and illustrate the usefulness in one of these areas – how to choose the discretization mesh for the function to be reconstructed – using numerical experiments.

1. Introduction. Inverse problems – i.e., determining spatially-variable parameters of a system from measurements of its state – are frequently ill-posed. Mathematically, this ill-posedness is often described as the lack of a continuous mapping from the space of measurements to the corresponding parameters reconstructed from a measurement. A consequence of ill-posedness is that a small measurement error can result in a significantly different reconstructed parameter unless the problem is regularized in some way. In order to illustrate this issue, let us consider that we want to identify a spatially varying parameter $q = q(\mathbf{x})$, for instance the density and elastic moduli of the earth in seismology, or the absorption and scattering properties of the human body in biomedical imaging. This identification requires using measurements z of some part of the state of the system under interrogation, e.g., the time-dependent displacement at a seismometer station, or the light intensity at the surface of the body as recorded by the pixels of a camera. If z is corrupted

2020 *Mathematics Subject Classification.* Primary: 65N21, 35R30, 94A17.

Key words and phrases. Information, inverse problems.

*Corresponding author: Wolfgang Bangerth.

by noise of level ε we will get two reconstructions ($q_1 = q_1(\mathbf{x}), q_2 = q_2(\mathbf{x})$) for two measurements (z_1, z_2) that differ only by the realization of the measurement noise. Ideally, we would be able to show that

$$\|q_1 - q_2\| \leq C\|z_1 - z_2\| \quad (1)$$

for some appropriate choice of norms and a constant C of moderate size. The problem is “ill-posed” if such an estimate does not exist. Many inverse problems fall into this category of ill-posedness.

On the other hand, scientists rarely talk about ill-posedness in terms of such estimates; rather, they will say “we don’t know enough about the system, for example because we didn’t measure enough to stably recover all of its features”. In other words, a pragmatic view of inverse problems is that their ill-posedness is at least in part the result of a *lack of information*. Some inverse problems can achieve well-posedness by obtaining different kinds of measurements from the system under consideration, but even if that is not the case, it would already be advantageous to simply *reduce the degree of ill-posedness*. In either case, we may ask whether it is possible to derive estimates of the kind

$$\|J(\mathbf{x})(q_1(\mathbf{x}) - q_2(\mathbf{x}))\| \leq C\|z_1 - z_2\|, \quad (2)$$

again with an appropriate choice of norms. We will call $J(\mathbf{x}) \geq 0$, or a related quantity such as its square root, the *information density*. Equation (2) is motivated by the observation that at places where $J(\mathbf{x})$ is large, we can accurately determine the value of the coefficient $q(\mathbf{x})$ we are looking for (i.e., $|q_1(\mathbf{x}) - q_2(\mathbf{x})|$ must be small to satisfy the inequality). Conversely, the places where $J(\mathbf{x})$ is small coincide with those locations where we have little control over the coefficient, and even small amounts of noise in z may lead to large variations $|q_1(\mathbf{x}) - q_2(\mathbf{x})|$. In the extreme case when the problem is truly ill-posed, $J(\mathbf{x})$ would not be bounded away from zero and consequently $\|J(\mathbf{x})\varphi(\mathbf{x})\|$ would not be a norm of φ .¹

It is unlikely that for practical problems we can find meaningful expressions for J that give rise to provable estimates of the form (2). This is because for many inverse problems, what can and cannot be recovered stably is often not about *where* in space we are, but about which *modes in feature space* (for example low- versus high-frequency components of a function $q(\mathbf{x})$) are identifiable. In our discussions below, we therefore consider estimates such as (2) *aspirational*: We will instead seek statements such as

$$\|j(\mathbf{x})(q_1(\mathbf{x}) - q_2(\mathbf{x}))\| \simeq C\|z_1 - z_2\|, \quad (3)$$

where $j(\mathbf{x})$ takes on the role of the information density, and where \simeq expresses a relationship of the form “behaves conceptually like, but possibly only when spatial discretization and/or regularization is used”. Although proving that any choice of j in (3) implies an estimate of the form (2) is likely impossible (for example because the statement is not in fact true), the conceptual approach of seeking a function $j(\mathbf{x})$ that expresses the idea of an information density of how much we know about q at different points in space will turn out to be useful in practice – as we will demonstrate in Sections 4 and 5.

¹In many inverse problems, for example in imaging, the ill-posedness manifests not by there being *locations* \mathbf{x} at which $q(\mathbf{x})$ is not identifiable. Rather, it is the *high-frequency (Fourier) content* of q that is often not identifiable without regularization. In this case, an equation like (2) with \mathbf{x} replaced by the wave number \mathbf{k} can be considered. Regardless of this obvious difference, we move forward with the derivation as stated.

References to information in inverse problems in the research literature.

The notion of information density is not new, in particular in applications where $q(\mathbf{x})$ is replaced by finite-dimensional parameter vectors $(q_i)_{i=1}^N$. Indeed, similar notions can be found in many areas of inverse and parameter estimation problems in various forms, and among practitioners of inverse problems, there is a degree of “knowledge” that *information* is a key concept. At the same time, practitioners do not appear to have a clear understanding of what information actually means, and uses of this concept in the literature appear to be vaguely defined and disconnected. In practice, references to the term “information” in the literature on inverse problem are almost always qualitative, rather than giving the term a quantitative definition.

Yet, many publications touch on the *concept* of information in inverse problems. The most obvious application of information concepts to inverse problems is in optimal experimental design where the goal of the design of schemes is to measure data about the system to minimize the uncertainty (that is, to maximize the information) in the parameters we wish to recover [7, 13, 32]. This relation to uncertainty is most clearly articulated in the Bayesian setting of optimal experimental design where the *information gain* of the posterior probability distribution over the prior is maximized. However, information, generally defined in less concise terms, is also a topic discussed in other contexts. For example, considering concrete applications, [51] presents Fisher information for a single-particle system and proposes a new uncertainty relationship based on Fisher information. Similarly, [64, 15] discuss the use of a resolution matrix in seismic tomography (see also [50]); related concepts of resolution, resolution length scales, event kernels, sensitivity kernels, Fréchet kernels, or point spread functions also appear in both seismic imaging and a number of other fields, see for example [40, 30, 59, 42, 43, 49, 19, 41, 65, 27, 36]. In many other cases, the literature references the Fisher information matrix that, together with the Cramér-Rao bound, quantifies how accurately we know what the inverse problem seeks to identify [39]; examples include [14], which uses this approach for estimating diffusion in a single particle tracking process; [34], which compares Fisher matrices to the Hessian calculation in boundary value inversion problem using the heat equation; and [46] which presents a preconditioning and regularization scheme based on Fisher information.

These publications generally state that there is a connection between the number, kind, and accuracy of measurements on the one hand, and the uncertainty in the recovered parameters of the inverse problem on the other. But, none of the studies mentioned go on to specifically *identify the role of information in the spatially variable ability to recover parameters in inverse problems* in a systematic way. Let us illustrate this with two publications that use data from seismometer recordings of earthquakes to produce reconstructions of the wave speeds within the Earth. Specifically, [19] presents the GLAD-M15 earth model that is based on 253 earthquake events recorded at seismometer stations around the world; [41] is an improvement over the previous paper and presents the GLAD-M25 model that is based on 1040 events. Because neither earthquakes nor seismometers are uniformly distributed across the earth surface, there are parts of the earth where we have a fairly good understanding of how fast seismic waves travel (namely those parts that are intersected by many “ray paths” from earthquake sources to seismometer stations) whereas there are other parts we do not really know very much about (primarily under the oceans, which are poorly covered by seismometers). In other words, we have “much information” about some parts and “little information” about others;

such statements are certainly common in many inverse problems, but information is not typically a quantitative term in these statements. Indeed, the second paper mentions the term “information” eight times on 21 pages, always in qualitative contexts such as “The ultimate goal is to use every single piece of information in seismograms” and “Figs 2–4 illustrate that much information in seismograms is being assimilated”. Yet, the paper does not present a quantitative definition of “information”. The authors do, however, attempt to quantify the accuracy of their reconstructions: Both [19] and [41] compute “point-spread functions” that assess the size of the smallest features in the earth the method can resolve given the available seismic observations. As the papers illustrate well, the size of the smallest resolvable features varies substantially between different parts of the earth, in accordance with our statements about having more or less information depending on location. Furthermore, the second paper illustrates the improvement that results from using more seismic events by showing that the smallest resolvable features have become smaller in GLAD-M25 compared to GLAD-M15 [41, Section 6.1 and Fig. 7]. However, the evaluations performed there are entirely qualitative: The paper states that “We observe that the 100 km diameter β_V anomaly—which is smaller than what was used in GLAD-M15—at 300 km depth is well recovered”, and “In each case the Gaussian anomaly is well resolved” (end of Section 6.1). It is these kinds of comparisons that we would like to put onto a quantitative basis, and the method we will discuss herein allows us to do exactly this by providing an “information density” that (i) would be able to quantify how much more information is available in some parts of the world than in others, and (ii) that would allow for a quantitative evaluation of the improvement of the second over the first model.

One could similarly analyze papers from many other disciplines that use inverse problems. They may be using different words, but a common feature of the many definitions of resolution, adjoints, sensitivity, and identifiability that can be found in the literature, is that most of these notions originate in, and were developed for *deterministic* inverse problems. On the other hand, “information” is probably best understood as a statistical concept, and a useful definition will therefore be rooted in statistical reformulations of the inverse problems. We will utilize the connection between the Fisher information matrix and the variance of reconstructed parameters, via the Cramér-Rao bound, to derive information densities in Section 2.2 below. If the parameter-to-measurement map is linear, our definition of “information” is related to the sensitivity of this map to changes in parameters; in the nonlinear case, it is related to some average of this sensitivity.

Goals for this paper. The differences in the concepts mentioned above, and the lack of a common language to describe them, presents the motivation for this work. Indeed, while many of the referenced works above at least verbally express the notion that in many inverse problems, we “know more” about the parameters in some parts of the domain than in others, we have not found published works that try to provide quantitative measures of this concept. The current work tries to address this deficit; in concrete terms, our goals are as follows:

- To introduce the notion of an *information density* based on a statistical interpretation of the inverse problem, the Fisher information matrix, and an application of the Cramér-Rao bound.
- To define the information density in a way that respects intuitive notions. In particular, it should satisfy that adding more measurements can only increase the information, not decrease it; that information is inversely proportional to

measurement uncertainties; and that with Gaussian noise and linear models, measuring the same quantity twice increases the information by $\sqrt{2}$.

- To outline a number of applications for which we believe information density can be usefully employed.
- To practically evaluate our concepts in a concrete application, namely the choice of mesh on which to discretize an inverse source-identification problem.

Locating this contribution among the methods for solving inverse problems.

Parameter estimation problems – of which inverse problems are a particular kind – seek information about parameters in a model using measurements (“observations”) of the system’s state. Depending on what kind of, and how much, information one seeks, parameter estimation problems can be formulated in different ways. For instance, *deterministic parameter estimation problems* only seek very limited information about parameters (namely, a “most likely” value) and in return can be solved efficiently even for large numbers of inversion parameters with complicated physics, requiring perhaps a few hundred solutions of the forward model [22, 47, 10, 1]. On the other hand, *Bayesian parameter estimation problems* [57, 58] are formulated to achieve complete statistical characterization of the solution but are much more expensive to solve and therefore limited to relatively small parameter spaces and simpler physics. Deterministic and Bayesian inverse problems are therefore often considered the extremes on a spectrum of formulations. Many methods have been devised that can be located somewhere along this spectrum, providing more information (albeit at a greater computational cost) than just the “best fit” solution that results from the deterministic approach. Our method falls in this category: We seek to augment the solution of the deterministic inverse problem by providing an estimate for the amount of “information” that is associated with each parameter of the solution, but at a cost far smaller than what one would incur by solving the full Bayesian problem.

Overview. Herein, we will address the goals mentioned above by first considering a finite-dimensional, linear model problem in Section 2 that we use to provide a conceptual overview of what we are trying to achieve, followed by the extension of this model problem to the infinite-dimensional case in Section 3. Having so set the stage, in Section 4, we provide “vignettes” for three ways in which we believe information densities can be used in practice. Section 5 then explores one of these – the choice of mesh for discretizing an infinite-dimensional inverse problem – in detail and with numerical and quantitative results. We conclude in Section 6. Two appendices discuss the extension of our work to nonlinear problems (Appendix A) and explain the derivation of the mesh refinement criteria that we compare against the method we propose in Section 5 (Appendix B).

2. A finite-dimensional, linear model problem. In order to explain our ideas, let us first consider a linear, finite-dimensional problem. (In practice, of course, many inverse problems are nonlinear; as we will show in Appendix A, it turns out that nearly everything we say here will carry over to the nonlinear case under common assumptions, namely, if the model is not too nonlinear, or if the noise level in measurements is sufficiently small.) Specifically, let us consider a problem in which a state vector \mathbf{u} is related to source terms via the relationship

$$\mathbf{A}\mathbf{u} = \sum_k q_k \mathbf{s}_k, \tag{4}$$

where \mathbf{s}_k are possible source vectors and q_k are their relative strengths. We assume that the system matrix A is invertible, although it may be ill-conditioned. In the inverse problem, we are then interested in recovering unknown source strengths q_k through a number of (linear, noisy) measurements

$$z_\ell = \mathbf{m}_\ell^T \mathbf{u} + \varepsilon_\ell, \quad (5)$$

where the dot product $\mathbf{m}_\ell \cdot$ corresponds to the ℓ th measurement operator, and ε_ℓ is measurement noise. We assume that we have a guess σ_ℓ for the magnitude of the noise ε_ℓ .

For convenience, let us collect the quantities $q_k, \mathbf{s}_k, z_\ell, \mathbf{m}_\ell$ into vectors and matrices $\mathbf{q}, S, \mathbf{z}, M$, where the \mathbf{s}_k form the columns of S and the \mathbf{m}_ℓ the rows of M . Then, we can state the source strength recovery problem we will consider here as

$$\begin{aligned} \min_{\mathbf{q}, \mathbf{u}} \quad & \frac{1}{2} \|\mathbf{M}\mathbf{u} - \mathbf{z}\|_{\Sigma^{-2}}^2 + \frac{\beta}{2} \|R\mathbf{q}\|^2, \\ \text{such that} \quad & \mathbf{A}\mathbf{u} = S\mathbf{q}. \end{aligned} \quad (6)$$

Here, we have used the weighted norm $\|\mathbf{M}\mathbf{u} - \mathbf{z}\|_{\Sigma^{-2}}^2 = \sum_\ell \frac{1}{\sigma_\ell^2} |\mathbf{m}_\ell^T \mathbf{u} - z_\ell|^2$, where the diagonal matrix $\Sigma_{\ell\ell} = \sigma_\ell$ weighs measurements according to the assumed *certainty* $\frac{1}{\sigma_\ell}$ we have of the ℓ th measurement. We have also added a Tikhonov-type regularization term where β is the regularization parameter and R a matrix that amplifies the undesirable components of \mathbf{q} .

By eliminating the state variable using the state equation (4), we can re-state this problem as an unconstrained, quadratic minimization problem:

$$\min_{\mathbf{q}} \quad \mathcal{J}_{\text{red}}(\mathbf{q}) := \frac{1}{2} \|\mathbf{M}\mathbf{A}^{-1}S\mathbf{q} - \mathbf{z}\|_{\Sigma^{-2}}^2 + \frac{\beta}{2} \|R\mathbf{q}\|^2. \quad (7)$$

It is then not difficult to show that the minimizer \mathbf{q} of this problem satisfies

$$\underbrace{(S^T A^{-T} M^T \Sigma^{-2} M A^{-1} S + \beta R^T R)}_Q \mathbf{q} = S^T A^{-T} M^T \Sigma^{-2} \mathbf{z}. \quad (8)$$

If we consider the noise to be random, we can ask how the solution \mathbf{q} depends on concrete measurements. Specifically, if we have two, presumably nearby, measurements $\mathbf{z}_1, \mathbf{z}_2$, then the following relationship holds for the corresponding solutions $\mathbf{q}_1, \mathbf{q}_2$:

$$Q(\mathbf{q}_1 - \mathbf{q}_2) = S^T A^{-T} M^T \Sigma^{-2} (\mathbf{z}_1 - \mathbf{z}_2). \quad (9)$$

For the following discussions, it is important to point out that the structure of Q guarantees that the matrix is symmetric, positive, and semidefinite; that is, all of its eigenvalues are non-negative. We will assume that the user has chosen β and R in such a way that Q is positive definite, although some of the eigenvalues may be small.

2.1. Defining an “information content” for components of the solution vector. We investigate herein if the relationship (9) between $\mathbf{q}_1 - \mathbf{q}_2$ and $\mathbf{z}_1 - \mathbf{z}_2$ allows us to define stability bounds such as those outlined in (2) or (3) above. In the context of this finite-dimensional situation, such a bound would have the form

$$\|\mathbf{j} \odot (\mathbf{q}_1 - \mathbf{q}_2)\| \simeq C \|\mathbf{z}_1 - \mathbf{z}_2\|, \quad (10)$$

with a constant C of possibly unknown size, and where \odot indicates the Hadamard product that scales each entry of the vector $\mathbf{q}_1 - \mathbf{q}_2$ by the corresponding entry

of the “information vector” \mathbf{j} . (Alternatively, we can interpret $\mathbf{j} \odot (\mathbf{q}_1 - \mathbf{q}_2)$ as $\text{diag}(\mathbf{j})(\mathbf{q}_1 - \mathbf{q}_2)$ where $\text{diag}(\mathbf{j})$ is a diagonal matrix with diagonal entries j_k .)

A *meaningful* statement² such as (10) will not always follow from (8) unless either the action of the matrix $Q = S^T A^{-T} M^T \Sigma^{-2} M A^{-1} S + \beta R^T R$ can somehow be approximated from below by a diagonal matrix, or Q^{-1} be approximated from above by a diagonal matrix. Indeed, we could choose \mathbf{j} to be a vector whose elements are all equal to $j_k := \lambda_{\min}(Q) = [\lambda_{\max}(Q^{-1})]^{-1}$. If, in addition, $C := \|S^T A^{-T} M^T \Sigma^{-2}\|$, then (10) holds true. This approach works as long as the regularization is chosen so that all eigenvalues of Q are reasonably large, i.e., that the problem is well-posed; in practice, however, this choice may over-regularize the problem. Alternatively, we could try to obtain a bound by formally solving (9) to obtain $(\mathbf{q}_1 - \mathbf{q}_2) = Q^{-1} S^T A^{-T} M^T \Sigma^{-2} (\mathbf{z}_1 - \mathbf{z}_2)$, and then trying to find a vector \mathbf{j} obtained from a lower bound of

$$Q^{-1} S^T A^{-T} M^T \Sigma^{-2} = [S^T A^{-T} M^T \Sigma^{-2} M A^{-1} S + \beta R^T R]^{-1} [S^T A^{-T} M^T \Sigma^{-2}].$$

In the end, we have found that choices such as $j_k = \lambda_{\min}(Q)$ obtained by purely algebraic manipulation tend not to be interesting since they do not give us any insight into *which elements of \mathbf{q} can be accurately estimated and which cannot*. Furthermore, if the problem is indeed ill-posed, as the regularization parameter β is reduced, *all* elements j_k will become small, despite the fact that the unregularized problem may only imply that *some* components of \mathbf{q} cannot be stably recovered. As a consequence, we consider (10) as only *aspirational, but not actually realizable* for inverse problems.

In order to define a more useful quantity representing the concept of information, we will appeal to a stochastic (Bayesian) interpretation of the inverse problem [58, 38]. From this perspective, we assume that our measurements \mathbf{z} are stochastic because they are corrupted by noise, and that consequently our recovered coefficients \mathbf{q} are also stochastic variables whose joint probability distribution we would like to infer. If we assume that the components of \mathbf{z} are distributed according to $\mathbf{z} = \hat{\mathbf{z}} + N(0, \Sigma^2)$ with a nominal (but unknown) “exact” measurement value $\hat{\mathbf{z}}$ (that is, we assume that the noise is Gaussian and that our guessed noise levels σ_ℓ are indeed correct), then the desired probability distribution for \mathbf{q} will be of the form

$$p(\mathbf{q}|\mathbf{z}) = \kappa e^{-\mathcal{J}_{\text{red}}(\mathbf{q})} \quad (11)$$

where κ is a normalization constant whose concrete value is not of importance to us.

Given this interpretation, the question of how much we know about the individual components of \mathbf{q} can be related to the *uncertainty* under $p(\mathbf{q}|\mathbf{z})$ – namely, we should choose the information weights j_k as the *inverse of the standard deviation*³ of q_k , that is, equal to $\frac{1}{\sqrt{\text{var}_p(\mathbf{q})_k}}$, where

$$\text{var}_p(\mathbf{q})_k = \int (q_k - \mathbb{E}_p[\mathbf{q}]_k)^2 p(\mathbf{q}|\mathbf{z}) \, d\mathbf{q}. \quad (12)$$

²Equation (8) implies that $\|\mathbf{q}_1 - \mathbf{q}_2\| \leq C \|\mathbf{z}_1 - \mathbf{z}_2\|$ with $C = \|Q^{-1} S^T A^{-T} M^T \Sigma^{-2}\|$, which corresponds to choosing \mathbf{j} in (10) as a vector of ones. At the same time, this estimate reflects no specifics of the problem and we do not consider this choice useful because it does not help us identify *which components of \mathbf{q} can be identified accurately, and which cannot*.

³The inverse of the variance is often called the “precision” with which a parameter is known [31].

Here, $\mathbb{E}_p[\mathbf{q}] = \int \mathbf{q} p(\mathbf{q}|\mathbf{z}) d\mathbf{q}$.⁴ This choice of j_k has the pleasant property of making (10) dimensionally correct also for cases where the components of \mathbf{q} have different physical units. Yet, it turns out to be computationally difficult to obtain the variances $\text{var}_p(\mathbf{q})_k$, and as a consequence we will show in the next section how we can define j_k in a way that allows for an efficient computation.

2.2. Estimating the information content for each component of the solution vector. The definition of an information content j_k based on the inverse of the variance in the stochastic inverse problem makes intuitive sense. The question remains whether these variances can be computed efficiently. As discussed in Remark 2.2 at the end of this section, the answer is no, but we will show next that approximations can be found that allow us to define a quantity j_k that has the right properties while being efficiently computable.

To do so, recall that the variances $\text{var}_p(\mathbf{q})_k = \text{cov}_p(\mathbf{q})_{kk}$ are the diagonal entries of the covariance matrix associated with $p(\mathbf{q}|\mathbf{z})$, where the covariance matrix is defined as

$$\text{cov}_p(\mathbf{q})_{kl} = \int (q_k - \mathbb{E}_p[\mathbf{q}]_k)(q_l - \mathbb{E}_p[\mathbf{q}]_l) p(\mathbf{q}|\mathbf{z}) d\mathbf{q}. \quad (13)$$

As mentioned above, in an ideal world, one would define an information density based on the variances $\text{var}_p(\mathbf{q})_k$, for example by using the inverse of the standard deviation. However, the covariance matrix or a reasonably good approximation can generally not be computed in a computationally tractable way for large problems using sampling techniques such as MCMC unless one utilizes properties of the problem (such as when the forward model is linear and the objective function is quadratic). We do not want to base our approach on using properties that typical applications do not possess. Instead, we use techniques known from statistical analysis. In particular, one can estimate the covariance matrix via the Cramér-Rao bound that states that

$$\text{cov}_p(\mathbf{q}) \geq I_p^{-1}$$

in the sense that $[\text{cov}_p(\mathbf{q}) - I_p^{-1}]$ is a positive semidefinite matrix. Here, I_p is the Fisher information matrix defined by

$$(I_p)_{kl} = -\mathbb{E} \left[\frac{\partial^2}{\partial q_k \partial q_l} \ln p(\mathbf{q}|\mathbf{z}) \right], \quad (14)$$

which for our choice of $p(\mathbf{q}|\mathbf{z})$ and $\mathcal{J}_{\text{red}}(\mathbf{q})$ evaluates to

$$(I_p)_{kl} = \mathbb{E} \left[\frac{\partial^2}{\partial q_k \partial q_l} (-\ln \kappa + \mathcal{J}_{\text{red}}(\mathbf{q})) \right] = \mathbb{E} \left[\frac{\partial^2}{\partial q_k \partial q_l} \mathcal{J}_{\text{red}}(\mathbf{q}) \right] = Q_{kl}.$$

In other words, the Fisher information matrix *can* be computed efficiently, unlike the covariance matrix. At the same time, the estimate above requires us to compute the *inverse* of the Fisher matrix, which for large and ill-conditioned problems is again not computable efficiently and accurately. However, we can use the the following inequality, again known from statistical analysis [25]:

$$\text{var}_p(\mathbf{q})_k = \text{cov}_p(\mathbf{q})_{kk} \geq [I_p^{-1}]_{kk} \geq [(I_p)_{kk}]^{-1}.$$

⁴Because we are considering a linear problem (4) and because the objective function \mathcal{J}_{red} is quadratic, the expectation value $\mathbb{E}_p(q_k)$ in (12) is equal to the solution of the original, deterministic problem (8).

These statements then provide us with an efficient way to estimate $\text{var}_p(\mathbf{q})_k$:

$$\text{var}_p(\mathbf{q})_k \geq [(I_p)_{kk}]^{-1} = [Q_{kk}]^{-1}. \quad (15)$$

In the spirit of the transition from (2) to (3), let us then define the information content of the k th parameter as

$$j_k := \sqrt{Q_{kk}}. \quad (16)$$

As we will discuss below, this definition realizes our goals for what an ‘‘information content’’ should intuitively provide, and we have found this definition to also be *practical*, even though it does not make the aspirational inequality (10) true. Moreover, using different statistical estimates, one could likely have arrived at other ways of defining j_k that still satisfy our goals; the one above has the advantage that it is computable and has all of the properties we ask of it.

Remark 2.1. Based on the definition $Q = S^T A^{-T} M^T \Sigma^{-2} M A^{-1} S + \beta R^T R$, the elements $j_k = \sqrt{Q_{kk}}$ can be computed in different ways by setting parentheses in the defining expression. The first way computes

$$\begin{aligned} Q_{kk} &= \mathbf{e}_k^T Q \mathbf{e}_k \\ &= (\Sigma^{-1} M A^{-1} S \mathbf{e}_k)^T (\Sigma^{-1} M A^{-1} S \mathbf{e}_k) + \beta (R \mathbf{e}_k)^T (R \mathbf{e}_k) \\ &= (\Sigma^{-1} M A^{-1} \mathbf{s}_k)^T (\Sigma^{-1} M A^{-1} \mathbf{s}_k) + \beta (R \mathbf{e}_k)^T (R \mathbf{e}_k) \\ &= (\Sigma^{-1} M \mathbf{h}_k)^T (\Sigma^{-1} M \mathbf{h}_k) + \beta \mathbf{r}_k^T \mathbf{r}_k \\ &= \sum_{\ell} \frac{1}{\sigma_{\ell}^2} (\mathbf{m}_{\ell}^T \mathbf{h}_k)^2 + \beta \mathbf{r}_k^T \mathbf{r}_k, \end{aligned}$$

where \mathbf{e}_k is the k th unit vector and

$$\mathbf{h}_k = A^{-1} S \mathbf{e}_k = A^{-1} \mathbf{s}_k, \quad \mathbf{r}_k = R \mathbf{e}_k.$$

That is, computing the information content vector \mathbf{j} requires the solution of the forward operator A for each of the source terms, plus a few matrix vector products.

An alternative way involves computing $M A^{-1} = (A^{-T} M^T)^T$ first. Because the vectors \mathbf{m}_{ℓ} form the rows of M (and so the columns of M^T), we can compute vectors

$$\mathbf{h}_{\ell}^* = A^{-T} \mathbf{m}_{\ell},$$

and then recognize that

$$Q_{kk} = \sum_{\ell} \frac{1}{\sigma_{\ell}^2} ((\mathbf{h}_{\ell}^*)^T \mathbf{s}_k)^2 + \beta \mathbf{r}_k^T \mathbf{r}_k.$$

This approach requires solving a linear system with A^T for each measurement.

Which of the two ways of computing Q_{kk} is more efficient depends on whether there are more measurements than source terms, or the other way around.⁵

Regardless of the way Q_{kk} (and consequently $j_k = \sqrt{Q_{kk}}$) is computed, it can be interpreted as having contributions from all measurements (through the sum over ℓ) and from regularization. The scalar product $\mathbf{m}_{\ell}^T \mathbf{h}_k$ can be considered as the influence of the forward propagated sources (\mathbf{h}_k) on measurements. On the other hand, the equivalent term $(\mathbf{h}_{\ell}^*)^T \mathbf{s}_k$ corresponds to a view where we first compute an adjoint solution \mathbf{h}_{ℓ}^* that indicates which possible source terms affect a measurement

⁵Clearly, both ways are expensive for real-world cases with many parameters and many measurements. We will come back to this in our conclusions and outlook, Section 6.

functional, and then take the dot product with a concrete source \mathbf{s}_k . Both views represent the sensitivity of measurement functionals to sources.

Interestingly, the formula expresses the intuitive concept that *information is additive*: If there are no measurements and no regularization, then $j = 0$; each measurement in turn adds a non-negative contribution. This is also true for regularization, where the choice of a large value for the regularization parameter β typically indicates that we have great confidence in the correctness of the regularization – in other words, we believe that we *know a lot* about the solution, and so would expect a large value for j_k (which the formula provides).⁶ Finally, because the measurement contributions to Q_{kk} are proportional to $\frac{1}{\sigma_\ell^2}$, and because the information content is the square root of Q_{kk} we have the convenient and reasonable property that *information is inversely proportional to the measurement uncertainty*. As a consequence, the definition of j chosen here matches the goals we have set out at the end of the introduction for an information content or density.

Remark 2.2. The *intent* of the derivations above was to define the information content as j_k as something like the precision, $\frac{1}{\sqrt{\text{var}_p(\mathbf{q})_k}}$, see the end of Section 2.1.

Yet, the concrete form on which we finally landed was to define it as $j_k := \sqrt{Q_{kk}}$, the latter being an approximation of the former via the Cramér-Rao bound and (15). We do not choose the former because in practice computing all diagonal elements Q_{kk} can be done efficiently via the techniques outlined in Remark 2.1, whereas computing all variances $\text{var}_p(\mathbf{q})_k$ (i.e., all diagonal elements of the covariance matrix) is far more expensive, in particular for nonlinear problems (see also Appendix A). In other words, our definition is a pragmatic compromise in that it is not *quite* what we want, but has the right properties and is efficiently computable. At the same time, it is conceivable that the variances can be efficiently *approximated*, for example via randomized linear algebra techniques and sampling methods to compute expectation values. Whether it makes a *practical difference* to use one or the other definition of j_k – and consequently whether it is worth the effort to actually compute the variances – is, of course, a different question; we will leave the answer to this question to future research.

Remark 2.3. In practice, models are rarely linear. For nonlinear models, the Cramér-Rao bound as well as (15) still hold, but perhaps with a larger gap between left and right hand sides. Nonetheless, we can analyze the nonlinear case. Such an analysis is provided in Appendix A in substantial detail, but let us preview the results here already: Using standard statistical and analytical techniques, the definition $j_k := \sqrt{Q_{kk}}$ we have chosen here will still be a good approximation of the inverse of the variance of parameter q_k if (i) the model is not too nonlinear, or (ii) if the noise level is relatively small. Furthermore, the computation of information contents j_k only requires *linearized* solves of the forward or adjoint model, making the relative cost substantially cheaper than in the linear case because the solution of the inverse problem now requires the solution of *nonlinear* forward problems.

⁶Obviously, expressing great confidence in the regularization through the choice of a large value for β does not actually imply that we get a high-fidelity solution of the inverse problems if that original confidence was misplaced – even if it entails that the information content is large.

3. Extension to infinite-dimensional inverse problems. We can extend the reasoning of the previous section to infinite-dimensional inverse problems. Specifically, let us consider the linear source identification problem

$$\mathcal{L}u(\mathbf{x}) = \sum_k q_k s_k(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, \quad (17)$$

where \mathcal{L} is a differential operator acting on functions defined on a domain $\Omega \subset \mathbb{R}^d$, and the equations are augmented by appropriate boundary conditions on $\partial\Omega$ whose details we will skip for the moment. As before, $s_k(\mathbf{x})$ are possible source vectors and q_k are their relative strengths. We again seek to identify source strengths q_k . Importantly, we will assume that the source terms are all of the form

$$s_k(\mathbf{x}) = \chi_{\omega_k}(\mathbf{x}),$$

where χ_{ω_k} is the characteristic function of a subdomain ω_k , and we assume that $\omega_k \cap \omega_l = \emptyset$ for $k \neq l$ and $\bigcup_k \bar{\omega}_k = \bar{\Omega}$. In other words, we seek to identify a source term that is a piecewise constant function defined on a partition of the domain Ω .

We will infer the source strengths q_k through (linear, noisy) measurements

$$z_\ell = \langle m_\ell, u \rangle + \varepsilon_\ell = \int_\Omega m_\ell(\mathbf{x})u(\mathbf{x}) \, d\mathbf{x} + \varepsilon_\ell, \quad (18)$$

where ε_ℓ is again measurement noise assumed to have magnitude σ_ℓ . The formalism we will develop will allow us to assign an *information content* to each q_k . Because the source strengths q_k correspond to characteristic functions s_k of subdomains ω_k , the information content j_k divided by the volume $|\omega_k|$ will define an *information density* $j(\mathbf{x})$, for which we can consider the limit case $|\omega_k| \rightarrow 0$. This limit is not computable, but we can use finite subdivisions into regions ω_k that allow us to approximate it with reasonable accuracy.

To make these concepts concrete, in Section 5 we will consider this model where \mathcal{L} is an advection-diffusion operator, $\mathcal{L} = -D\Delta + \mathbf{b} \cdot \nabla$, and where z_ℓ correspond to point measurements of $u(\mathbf{x})$ (or a well-defined approximation of point measurements if the solution u is not guaranteed to be a continuous function). This example is motivated by a desire to identify sources of air pollution from sparse measurements at a finite number of points.

3.1. Definition of the inverse problem. The inverse problem we have described above then has the following mathematical formulation, where we also include an L_2 regularization term:

$$\min_{\mathbf{q}, u} \mathcal{J}(\mathbf{q}, u) = \frac{1}{2} \sum_\ell \frac{1}{\sigma_\ell^2} |\langle m_\ell, u \rangle - z_\ell|^2 + \frac{\beta}{2} \left\| \sum_k q_k s_k \right\|_{L_2(\Omega)}^2, \quad (19)$$

$$\text{such that } \mathcal{L}u = \sum_k q_k s_k.$$

As in the previous section, we can define a reduced objective function

$$\begin{aligned} \mathcal{J}_{\text{red}}(\mathbf{q}) &= \mathcal{J} \left(\mathbf{q}, \mathcal{L}^{-1} \sum_k q_k s_k \right) \\ &= \frac{1}{2} \sum_\ell \frac{1}{\sigma_\ell^2} \left| \left\langle m_\ell, \mathcal{L}^{-1} \sum_k q_k s_k \right\rangle - z_\ell \right|^2 + \frac{\beta}{2} \left\| \sum_k q_k s_k \right\|_{L_2(\Omega)}^2, \end{aligned} \quad (20)$$

which gives rise to a related stochastic inverse problem with a probability density $p(\mathbf{q}|\mathbf{z})$ defined as in (11).

3.2. Defining the information content. As in Section 2.2, we can again identify the *information content* associated with each parameter q_k via the precision, i.e., inverse of the variance $\text{var}_p(\mathbf{q})_k = \text{cov}_p(\mathbf{q})_{kk}$, and the estimate we have for the variance based on the Fisher information matrix.

In the finite-dimensional case, the Fisher information matrix I_p could be computed by solving one forward problem for each source vector \mathbf{s}_k . The same is true for the current infinite-dimensional situation:

Proposition 3.1. *For the model problem defined above, the Fisher information matrix I_p defined in (14) has the following form:*

$$(I_p)_{kl} = Q_{kl} \quad (21)$$

where

$$Q_{kl} = \sum_{\ell} \frac{1}{\sigma_{\ell}^2} \langle m_{\ell}, h_k \rangle \langle m_{\ell}, h_l \rangle + \beta \int_{\Omega} s_k s_l, \quad (22)$$

and where h_k satisfies the equation

$$\mathcal{L}h_k(\mathbf{x}) = s_k(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega, \quad (23)$$

again augmented by appropriate boundary conditions for h_k .

Proof. Recall that

$$(I_p)_{kl} = -\mathbb{E} \left[\frac{\partial^2}{\partial q_k \partial q_l} \ln p(\mathbf{q}|\mathbf{z}) \right], \quad \text{with} \quad p(\mathbf{q}|\mathbf{z}) = \kappa e^{-\mathcal{J}_{\text{red}}(\mathbf{q})}.$$

Based on the definition of $\mathcal{J}_{\text{red}}(\mathbf{q})$ and the linearity of \mathcal{L} , we then obtain

$$\begin{aligned} (I_p)_{kl} &= \frac{\partial^2}{\partial q_k \partial q_l} \mathcal{J}(\mathbf{q}) \\ &= \frac{\partial^2}{\partial q_k \partial q_l} \left[\frac{1}{2} \sum_{\ell} \frac{1}{\sigma_{\ell}^2} \left\langle m_{\ell}, \mathcal{L}^{-1} \sum_r q_r s_r \right\rangle^2 + \frac{\beta}{2} \left\| \sum_r q_r s_r \right\|_{L_2(\Omega)}^2 \right] \\ &= \sum_{\ell} \frac{1}{\sigma_{\ell}^2} \langle m_{\ell}, \mathcal{L}^{-1} s_k \rangle \langle m_{\ell}, \mathcal{L}^{-1} s_l \rangle + \beta \int_{\Omega} s_k s_l, \end{aligned}$$

as claimed when using $h_k := \mathcal{L}^{-1} s_k$. \square

Proposition 3.2. *The matrix Q can alternatively be expressed through the following formula:*

$$Q_{kl} = \sum_{\ell} \frac{1}{\sigma_{\ell}^2} \langle h_{\ell}^*, s_k \rangle \langle h_{\ell}^*, s_l \rangle + \beta \int_{\Omega} s_k s_l. \quad (24)$$

Here h_{ℓ}^* satisfies the equation

$$\mathcal{L}^* h_{\ell}^*(\mathbf{x}) = m_{\ell}(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega, \quad (25)$$

where \mathcal{L}^* is the adjoint operator to \mathcal{L} , using appropriate boundary conditions for h_{ℓ}^* .

Proof. The proposition follows from the observation that

$$\langle m_\ell, h_k \rangle = \langle m_\ell, \mathcal{L}^{-1} s_k \rangle = \langle \mathcal{L}^{-*} m_\ell, s_k \rangle = \langle h_\ell^*, s_k \rangle.$$

□

Remark 3.3. As in the finite-dimensional case, the Fisher information matrix (21) is easy to compute for problems with either not too many parameters (using (22)) or not too many measurements (then using (24)). In either case, the functions h_k or h_ℓ^* can be computed independently in parallel. Which of the two forms is more efficient depends on whether there are more measurements than source terms or the other way around. That said, in the discussions below, we will want to let $|\omega_k| \rightarrow 0$ and consequently make the number of source terms and parameters infinite, and in that case the adjoint formulation in (24) provides the more useful perspective.

The Fisher information matrix approximates the inverse of the covariance matrix, and the diagonal elements of the Fisher matrix I_p therefore provide a means to estimate the certainty in the corresponding parameters q_k . In the same way as for the finite-dimensional case in (16), we can then define an information content for the parameter k via

$$j_k := \sqrt{Q_{kk}}, \quad (26)$$

where now

$$Q_{kk} = \sum_\ell \frac{1}{\sigma_\ell^2} \langle m_\ell, h_k \rangle^2 + \beta \int_\Omega s_k^2 = \sum_\ell \frac{1}{\sigma_\ell^2} \langle h_\ell^*, s_k \rangle^2 + \beta \int_\Omega s_k^2.$$

3.3. Defining the information density. The discussions in the previous section did not make use of any particular properties of the source basis functions s_k . We now examine the special case of indentifying a piecewise constant source function, i.e., where

$$s_k(\mathbf{x}) = \chi_{\omega_k}(\mathbf{x}).$$

In this case, the information content for the parameter q_k associated with area ω_k is

$$j_k = \sqrt{Q_{kk}} = \sqrt{\sum_\ell \frac{1}{\sigma_\ell^2} \left(\int_{\omega_k} h_\ell^*(\mathbf{x}) dx \right)^2 + \beta |\omega_k|}. \quad (27)$$

Since this quantity scales with the size of the subdomains ω_k , it is reasonable to define a piecewise constant *information density* as

$$\begin{aligned} j(\mathbf{x})|_{\omega_k} &= \frac{1}{|\omega_k|} j_k = \frac{1}{|\omega_k|} \sqrt{Q_{kk}} \\ &= \sqrt{\sum_\ell \frac{1}{\sigma_\ell^2} \left(\frac{1}{|\omega_k|} \int_{\omega_k} h_\ell^*(\mathbf{x}) dx \right)^2 + \beta \frac{1}{|\omega_k|}} \end{aligned} \quad (28)$$

$$\approx \sqrt{\sum_\ell \frac{1}{\sigma_\ell^2} h_\ell^*(\mathbf{x})^2 + \beta \frac{1}{|\omega_k|}}. \quad (29)$$

We can make a number of observations based on these definitions, analogous to the finite-dimensional case of the previous section:

- Because the definition of h_ℓ^* is independent of the choice of ω_k , the formulas shown above can be interpreted as saying that the information density has a component that results from the measurements ℓ (and, in particular, grows

monotonically with the number of measurements), and a component that results from regularization.

- As before, the information density is inversely proportional to the measurement uncertainties σ_ℓ , in the absence of regularization.
- Regularization bounds the amount of information from below: $j(\mathbf{x})|_{\omega_k} \geq \sqrt{\beta/|\omega_k|}$. This dependence on the square root of the regularization parameter is well known [26].
- The information content j_k for a subdomain ω_k decreases under mesh refinement (i.e., as the subdomains become smaller and smaller). This makes sense if we have only finitely many measurements, and the rather weak L_2 regularization term. In order to ensure the continued well-posedness (i.e., an information content j_k that is bounded away from zero) under mesh refinement, we also need to either increase the number of measurements accordingly, or use a stronger regularization term; for the latter, we need to choose regularizers that are “trace-class” – see also the discussion in the Conclusions (Section 6).

Remark 3.4. An important observation is that the definitions of information content and information density above depend only on the forward operator and the measurement functionals, but not on concrete measured values z_ℓ . Consequently, and as anticipated, information quantities can be computed *before* measurements, and they are independent of the specific noise in measurements later obtained. We will come back to this point in Sections 5.2 and 5.3.1, as well as in the discussion of nonlinear problems in Appendix A.

Remark 3.5. We remark that the idea of using variances Q_{kk} as a spatially variable measure of certainty is not new. For example, [49, Section 3.1] illustrates the spatially variable variance for a seismic imaging problem. Yet, the authors’ definition is unclear regarding the role of regularization, misses the square root, and is then discarded as not very useful.

4. Using information densities. Having shown a way to define an information density $j(\mathbf{x})$, the question is whether it is useful. Indeed, there are numerous questions related to the practical solution of inverse problems for which information densities could be useful. In the following subsections, we therefore first outline three vignettes of situations in which the information density could be useful. In Section 5, we then expand on one of these ideas using a concrete numerical example.

In the examples below, we will consider the situation in which we have discretized the source term $q(\mathbf{x}) = \sum_k q_k s_k(\mathbf{x})$ in (17) on a “mesh” \mathbb{T} , as is common in the finite element method. Because there are no differentiability requirements on $q(\mathbf{x})$, it is common to identify the source term as a piecewise constant function on the mesh, and in this case, the source functions s_k are the characteristic functions of the cells K of \mathbb{T} . By identifying the index k with a cell K , (26) and (27) then define an “information content” j_K for each cell K of the mesh.

4.1. Using information densities for regularization. As a first example of where we believe that information densities could be used, we consider the regularization of inverse problems. In a large number of practical applications, one regularizes inverse problems by adding a penalty term to the misfit function for the purpose of penalizing undesirable aspects of the recovered function. For example,

in our definition of the source identification problem in Section 3.1 (see also equation (19)), we have penalized the *magnitude* of the source term to be identified. The strength of this penalization is provided by the factor β .

A practical question is how large this factor β should be. Many criteria have been proposed in the literature [33], but, in practice many studies do not use any of these automatic criteria and instead choose values of β that yield reasonable results based on trial and error.

Moreover, it is clear to many practitioners that regularization may not be necessary to the same degree in all parts of the domain. For example, if measurements are available only in parts of the domain (say, on the boundary), then intuitively more information is available to identify source strengths close to the boundary than deep in the interior of the domain. A particularly obvious example is in seismic imaging: There, we can accurately identify properties of the Earth only in those places that are crossed by ray paths from earthquake sources (predominantly located at plate boundaries) to seismometer stations (predominantly located on land), but not in the rest of the Earth [8, 52, 42, 49, 17]. In the definition of the information density $j(\mathbf{x})$ in (29), this would imply that the first term under the square root would be large along these ray paths, but small elsewhere. In cases such as this, a reasonable approach would be to make the regularization parameter spatially variable: large where little information is available, and small where regularization is not as important, for example so that $j(\mathbf{x}) \geq j_0$. This spatial variability could be achieved by replacing the regularization term in (19) by

$$\frac{1}{2} \left\| \sqrt{\beta(\mathbf{x})} \left(\sum_k q_k s_k(\mathbf{x}) \right) \right\|_{L_2(\Omega)}^2,$$

and defining $\beta(\mathbf{x})$ in some appropriate way.

Using spatially variable regularization is not a new idea (see, for example, [8, 48, 6, 56, 62, 63, 24, 17]), although we are not aware of any references that would provide an overarching, systematic framework for choosing $\beta(\mathbf{x})$. In contrast, the connection between information density $j(\mathbf{x})$ and $\beta(\mathbf{x})$ in (29) has the potential to provide such a systematic approach. A scheme based on this observation also satisfies other considerations that appear reasonable. For example, increasing the number of measurements, or decreasing the measurement error, leads to a larger information density and therefore to a smaller regularization term to satisfy $j(\mathbf{x}) \geq j_0$.

4.2. Using information densities to guide the discretization of an inverse problem. In actual practice, inverse problems are solved by discretization. In our derivation above, we have chosen finitely many source functions s_k that we have assumed are the characteristic functions of “cells” K of some kind of mesh or subdivision of the domain Ω on which $q(\mathbf{x})$ is defined, and then expanded the function we seek as

$$q(\mathbf{x}) = \sum_k q_k s_k(\mathbf{x}).$$

A practical question is how to create this subdivision. Oftentimes, the subdivision is chosen fine enough to resolve the features of interest but coarse enough to keep the computational cost in check. Regularization is frequently used to ensure that an overly fine mesh does not lead to unwanted oscillations in the recovered coefficients – in other words, to keep the problem reasonably well-posed.

Most often in the literature, the mesh for the inverse problem is either uniform or at least chosen a priori through insight into the problem (for approaches in the latter direction, see for example, [18, 8, 42]). On the other hand, discretization is a form of regularization, and it is reasonable to choose the mesh finer where more information is available – say, close to a measurement device – but coarser where our measurements have little information to offer. This idea has been used as a heuristic in the past [8], or at least mentioned (see Section 2.3 of [49] and the references therein), but, as with regularization, no overarching scheme is available to guide this choice of the mesh. [16] is also an example where the mesh is made part of what needs to be estimated in a Bayesian inversion scheme, which in practice appears to lead to meshes that are more refined where information is available. However, [16] is concerned with choosing the mesh used for the solution of the *state equation*, not the discretization of the parameter we seek – although it seems reasonable to assume that the scheme could be adapted to the latter as well. At the same time, the scheme described in [16] requires the solution of a Bayesian inverse problem, which is far more expensive than the deterministic approach we use here.

Moreover, an information density can provide such a guide to determine optimal cell sizes. First, we conjecture that meshes should be graded in such a way that the information content of each cell (i.e., roughly the information density times the measure of a cell) is approximately equal among all cells. We will explore in detail how well this works in practice in Section 5. Second, in mathematical research, mesh refinement cycles are frequently terminated whenever we run out of memory, out of patience, or both, whereas in applications, mesh refinement is stopped whenever an expert deems the solution sufficiently accurate. Either approach is unsatisfactory, and the amount of information available per cell might provide a more rational criterion to stop mesh refinement.

4.3. Using information densities for experimental design. As a final example, we consider optimal experimental design, that is, the question of what, how, or where to measure so as to minimize the uncertainty in recovered parameters given a certain noise level in measurements.

Optimal experimental design for inverse problems is more difficult than for finite-dimensional parameter estimation problems because it is not entirely clear what the objective function should be when minimizing or maximizing by varying the specifics of measuring. For finite-dimensional problems, objective functions include the *A*-, *C*-, *D*-, *E*-, and *T*-optimality criteria, plus many variations [7].

For the infinite-dimensional case (or discretized versions thereof), the choice might be to maximize the information content in all of Ω , or a subset $\omega \subset \Omega$:

$$\phi_{\text{information}}(\{m_\ell\}) = \int_{\omega} j(\mathbf{x}) \, dx,$$

where $\{m_\ell\}$ denotes the set of measurements to be performed and optimization will typically happen over a set of implementable such measurements.

If $\omega = \Omega$, then the integral above reduces to

$$\phi_{\text{information}}(\{m_\ell\}) = \sum_k \sqrt{Q_{kk}},$$

based on the definitions in (27) and (28). Recalling that the matrix Q is the Fisher information matrix, we recognize that the criterion ϕ above is similar to – but distinct from – the generally not very frequently used *T*-optimality criterion that

maximizes the sum of diagonal entries of Q (i.e., the *trace* of Q):

$$\phi_T(\{m_\ell\}) = \sum_k Q_{kk}.$$

5. Numerical examples of using information densities. The previous section provided three vignettes of how we imagine information densities could be used for practical computations. Exploring all of these ideas through numerical examples exceeds the reasonable length of a single publication, and as a consequence we will focus on only mesh refinement.

In the following subsections, we will first lay out the inverse problem we will use as a test case. We will then show numerical results that illustrate the use of information densities as applied to this problem.

All numerical results were obtained with a program that is based on the open-source finite element library `deal.II` [5, 4]. This program is available under an open source license as part of the `deal.II` code gallery at <https://dealii.org/developer/doxygen/deal.II/CodeGallery.html> under the name “Information density-based mesh refinement”.

5.1. The test case. Let us consider the following question: Given an advection-diffusion problem for a concentration $u(\mathbf{x})$, can we identify the sources $q(\mathbf{x})$ of the concentration field from point measurements of u at points ξ_ℓ ? This kind of problem is widely considered in environmental monitoring of pollution sources [53, 44, 61, 37], and also when trying to identify the sources of nuclear radiation.

Mathematically, we assume that the concentration field satisfies the stationary advection-diffusion equation

$$\mathcal{L}u(\mathbf{x}) \equiv \mathbf{b}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) - D\Delta u(\mathbf{x}) = q(\mathbf{x}) \quad \text{in } \Omega, \quad (30)$$

where \mathbf{b} is a (known) wind field and D is the (known) diffusion constant. For simplicity, we will assume homogenous Dirichlet boundary conditions

$$u = 0 \quad \text{on } \partial\Omega. \quad (31)$$

Concretely, for our computations, we will assume that $\Omega = (-1, 1)^2 \subset \mathbb{R}^2$ is a square, $D = 1$, and $\mathbf{b} = (100, 0)^T$. These choices lead to a Péclet number of 200; that is, the problem is advection dominated.

For the inverse problem, we ask whether we can recover the function $q(\mathbf{x})$ (or a discretized version of it) from measurements z_ℓ at a number of points $\xi_\ell \in \Omega$, $\ell = 1, \dots, L$. That is, we consider (18) with $m_\ell(\mathbf{x}) = \delta(\mathbf{x} - \xi_\ell)$ where σ_ℓ is the assumed noise level for the measurement at location ξ_ℓ (see below). We choose these points ξ_ℓ equally distributed around two concentric circles of radius 0.2 and 0.6, centered at the origin, with 50 points on each of the circles, for a total of $L = 100$ measurement points.

For our experiments, we will consider a situation where the data we have, z_ℓ , has been obtained by solving the forward problem with the finite element method, using a known source distribution $q^*(\mathbf{x})$ that is equal to one in a circle of radius 0.2 centered at $(-0.25, 0)^T$. A solution $u^*(\mathbf{x})$ can then be evaluated at the points ξ_ℓ to obtain “synthetic” measurements z_ℓ via

$$z_\ell = \langle m_\ell, u^* \rangle + \varepsilon_\ell = u^*(\xi_\ell) + \varepsilon_\ell. \quad (32)$$

We choose Gaussian noise $\varepsilon_\ell = N(0, \sigma_\ell^2)$ and set $\sigma_\ell = 0.1 \max_{\mathbf{x} \in \Omega} |u^*(\mathbf{x})|$.

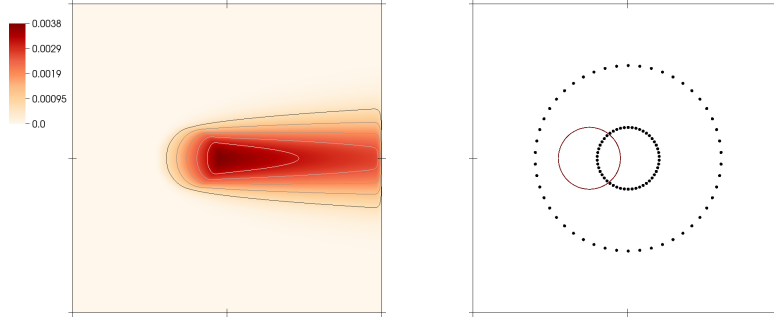


FIGURE 1. *Left: The solution $u^*(\mathbf{x})$ of the forward problem from which we generate “synthetic” measurements z_ℓ via (32). Right: The source term $q^*(\mathbf{x})$ from which we compute synthetic measurements is constant and nonzero only in the solid red circle offset from the center; the detector locations $\xi_\ell, \ell = 1, \dots, L = 100$ are marked by dots.*

To avoid an inverse crime, we solve for u^* on a mesh that is different from the meshes used for all other computations. The solution of this forward problem so computed to obtain synthetic measurements is shown in Fig. 1, along with the locations of the source term and the detector locations.

5.2. The inverse problem. The inverse problem we seek to solve is the identification of the source term $q(\mathbf{x})$ (which we approximate via a finite-dimensional expansion $\sum_k q_k s_k(\mathbf{x})$ in (30), based on the measurements described by (32). We approach this problem by reformulating it in the form of the constrained optimization problem (19), where we set the regularization parameter to $\beta = 10^4$. This problem is then solved by introducing a Lagrangian

$$\mathfrak{L}(u, q, \lambda) = \mathcal{J}(u, q) + \int_{\Omega} \lambda(\mathbf{x}) (\mathcal{L}u(\mathbf{x}) - q(\mathbf{x})) \, dx,$$

and then solving the linear system of partial differential equations that results by setting the derivatives of \mathfrak{L} to zero (that is, the optimality conditions). In strong form, these optimality conditions read

$$\begin{aligned} \mathcal{L}u(\mathbf{x}) &= q(\mathbf{x}), \\ \mathcal{L}^* \lambda(\mathbf{x}) &= - \sum_{\ell} \frac{1}{\sigma_{\ell}^2} (u(\xi_{\ell}) - z_{\ell}) \delta(\mathbf{x} - \xi_{\ell}), \\ \beta q - \lambda &= 0. \end{aligned} \tag{33}$$

The solution of this system of equations is facilitated by discretizing on a finite element mesh. We use continuous, piecewise cubic elements for u and λ , and discontinuous, piecewise constants elements for q .⁷

⁷This choice of higher order finite element spaces for u and λ is akin to solving for the forward and adjoint variables on a finer mesh than the source terms we seek to identify. As a result, we need not worry about satisfying discrete stability properties for the resulting saddle point problem. We can also, in essence, consider the forward and adjoint equation to be solved nearly exactly, with the majority of the discretization error resulting from the discretization of the source term $q(\mathbf{x})$.

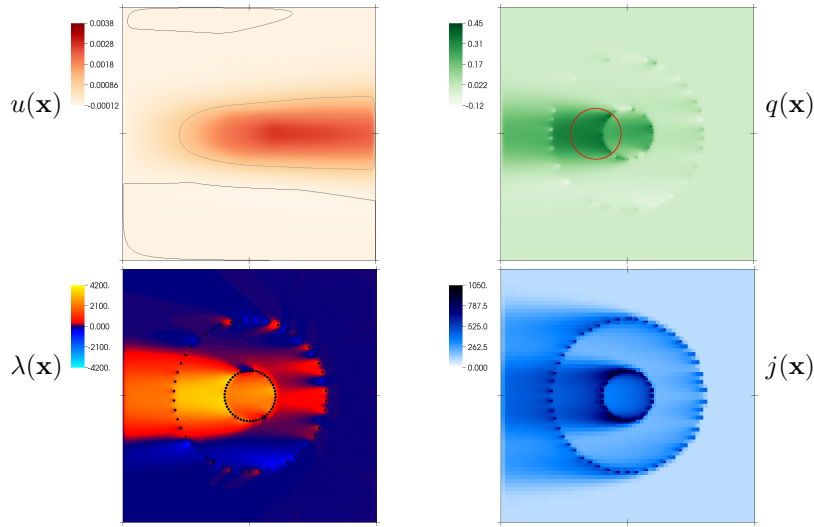


FIGURE 2. The solution of problem (33), computed on a very fine finite element mesh. Top left: The primal variable $u(\mathbf{x})$, shown with the same scale for color and isocontours as in Fig. 1. Top right: The recovered sources $q(\mathbf{x})$, i.e., the solution of the inverse problem. The red circle indicates the location of the source term used in generating the synthetic data. Bottom left: The adjoint variable $\lambda(\mathbf{x})$. Bottom right: The information density $j(\mathbf{x})$ associated with this problem, as defined in (29).

The three components u, q, λ of this solution, computed on a very fine mesh with $256 \times 256 = 65\,536$ cells, for which the coupled problem has 1 248 258 unknowns, are shown in Fig. 2. The maximal value of the recovered source is less than half the maximal size of the “true” source, owing to the effect of the L_2 regularization term. As a consequence, the forward solution u is also too small. Furthermore, the inverse problem places the source in a broader region than where it really is, but this is not surprising: In an advection-dominated problem, it is only possible to say with accuracy that the source is *upstream* of a detector, but not where in the upstream region it actually is unless another detector further upstream indicates that it must be downstream from the latter.

The adjoint variable λ clearly illustrates the effect that the adjoint operator \mathcal{L}^* transports information in the opposite direction $-\mathbf{b}$ of the forward operator \mathcal{L} , and that the sources of the adjoint equation are the residuals $-(u(\xi_\ell) - z_\ell)\delta(\mathbf{x} - \xi_\ell)$; here $u(\xi_\ell) - z_\ell$ reflects the measurement error and, based on our choice of noise above, is Gaussian distributed with both positive and negative values.

The bottom right panel of the figure also shows the information density $j(\mathbf{x})$ that corresponds to this problem, as defined in (29). It illustrates that, given the location of detectors and the nature of the equation, information is primarily available upstream of detector locations. Notably, and as mentioned in Remark 3.4, the information density is based solely on the operator \mathcal{L} and the measurement functionals m_ℓ , but not on the actual measurements z_ℓ (or the noise that is part of z_ℓ).

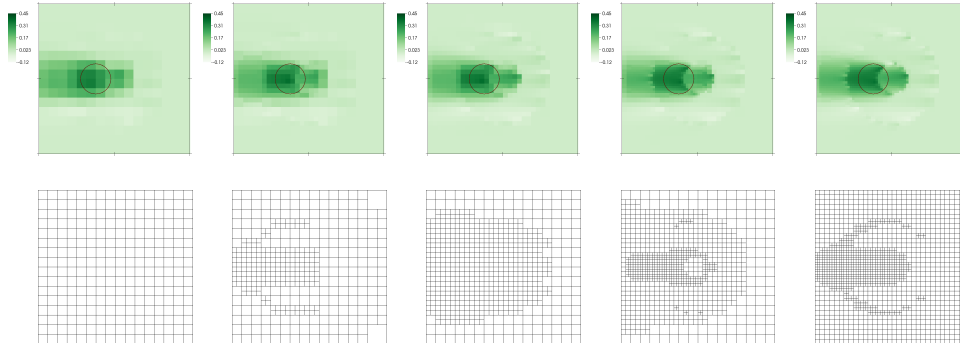


FIGURE 3. Reconstructions (top) on a sequence of meshes (bottom) refined based on the information content of each cell of the mesh.

5.3. Choice of mesh for the inverse problem. The question of interest then is how we can use information densities for mesh refinement. To answer this question, we have repeated the computations discussed above, but instead of using a uniformly refined mesh, we have used a sequence of meshes in which we refine cells hierarchically so as to equilibrate the information content j_k of each cell ω_k , see (26), by always refining those cells that have the largest information content. The reconstructions and the sequence of meshes they are computed on are shown in Fig. 3. For comparison with the computations mentioned above and shown in Fig. 2, the rightmost mesh has 1642 cells and the coupled problem solved on it has 32 582 unknowns.

5.3.1. Comparison with other mesh refinement criteria. The relevant question to ask is whether this mesh is better suited to the task than any other potential mesh. Answering this question is notoriously difficult in inverse problems because, in general, the exact solution of the problem is unknown if only finitely many measurements are available and if regularization is used. As a consequence, it is difficult to answer the question through comparison of convergence rates of different methods, for example.

However, we can sometimes make intuitive comparisons based on experience on “how a good mesh should look”, even though for problems like the one under consideration, it is generally difficult to create such meshes by hand a priori. For comparison with the meshes shown in Fig. 3, we present in Fig. 4 the meshes generated using two different, “traditional” mesh refinement criteria that we will justify in more detail in Appendix B. In both cases, the meshes are obtained by always refining those cells K that have the largest “refinement indicators” η_K . In the top row of Fig. 4, this indicator η_K is the cell-wise norm of the residual of the third equation of (33) and is thus an a posteriori error indicator that can be derived in a way similar to that shown in [9, 11, 12, 35]:

$$\eta_K = \|\beta q - \lambda\|_{L_1(K)}.$$

We will consequently refer to this quantity as the “error estimator”.

In the bottom row of Fig. 4, we show meshes generated by evaluating a finite difference approximation $\nabla_h q(\mathbf{x})$ on each cell by comparing the values of $q|_K$ with the values of q on neighboring cells, and then computing

$$\eta_K = h_K \|\nabla_h q\|_{L_2(K)},$$

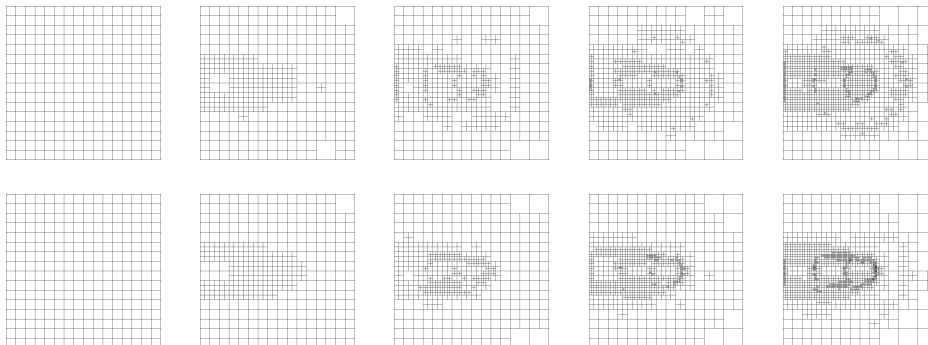


FIGURE 4. Sequences of meshes generated by different mesh refinement criteria. Top: Mesh refinement is driven by an a posteriori error indicator. Bottom: Mesh refinement is driven by a smoothness indicator.

where h_K denotes the diameter of cell K . The choice of η_K is proportional to the interpolation error of a continuous function when approximated by a piecewise constant finite element function (as we do here); the indicator therefore measures where the piecewise constant approximation is likely poor. We will refer to this criterion as the “smoothness indicator”. It is also used in [52], for example.

As outlined in Section 4.2, the literature contains discussions of many other ways to refine meshes for the inverse problem, but we consider the two mentioned above as representative mesh refinement criteria to compare our approach against. We provide a detailed derivation of these two indicators in Appendix B.

The meshes shown in Fig. 4 are structurally similar to those generated based on the information content and shown in Fig. 3. However, they lack the top-bottom symmetry of the ones in Fig. 3 and look generally less organized, owing to the fact that *they are based on the solution of the inverse problem, which is subject to the noise in the measurements*, whereas the information density reflects only how much we *know* about the solution at a specific point in the domain – that is, a quantity that is *independent of the concrete realization of the noise that is part of the measurements*, see also Remark 3.4. Conceptually, the best mesh should be independent of the concrete realization of noise, although dependent on *what* is being measured. The refinement by information content allows us to construct the mesh even before solving the inverse problem because it does not depend on the solution of the inverse problem.

If the individual measurements m_ℓ had had differently sized measurement errors, then this would also have affected the information density-based mesh refinement and led to smaller cells where more accurate measurements are available. In contrast, there is no such direct dependence for the other refinement methods; rather, for those methods, variable noise levels only affect mesh refinement because λ indirectly depends on the error level.

5.3.2. *Quantitative evaluation: Condition numbers of matrices.* A more concrete comparison between meshes would be to measure the degree of ill-posedness of the problem. Of course, we use regularization to make the problem well-posed, but a well-chosen mesh results in a matrix after discretization that has a better condition number than a poorly chosen mesh, and for which the reconstruction is consequently

less sensitive to noise. In practice, the condition number is a poor indicator since it considers only the largest and smallest eigenvalues; we hypothesize that a better criterion would be to ask how many “large” eigenvalues there are, and it is this criterion that we will consider below.

To test this hypothesis, let us consider the discretized version of (33). If we collect the degrees of freedom of a finite element discretization of u into a vector U , and similarly those of λ into a vector Λ and those of q into the vector P (a symbol chosen to avoid confusion with the matrix Q of Section 2), then (33) corresponds to the following system of linear equations after discretization by the finite element method:

$$\begin{aligned} AU &= BP, \\ A^T \Lambda &= -C(U - Z), \\ \beta NP - B^T \Lambda &= 0. \end{aligned} \tag{34}$$

Here, the matrix A corresponds to the discretized operator $\mathcal{L} = \mathbf{b} \cdot \nabla - D\Delta$ acting on the finite element space chosen to discretize the state and adjoint variables, and N is the mass matrix on the finite element space chosen for the source q – that is, on the set of piecewise constant functions $s_k(\mathbf{x})$ associated with the cells of the mesh. The matrix B results from the product $B_{ik} = (\varphi_i, s_k)_\Omega$ between the shape functions for u and q , and C corresponds to terms of the form $C_{ij} = \sum_\ell \frac{1}{\sigma_\ell^2} \varphi_i(\xi_\ell) \varphi_j(\xi_\ell)$. By noting that the matrices A and N are invertible, we can reduce this system of equations to an equation for P by repeated substitution to

$$HP = B^T A^{-1} CZ, \tag{35}$$

where the matrix H is the Schur complement,

$$H = B^T A^{-T} C A^{-1} B + \beta N. \tag{36}$$

The matrix H , which is symmetric and at least positive semidefinite, thus relates the vector of measurements Z to the vector of coefficients P we would like to recover. H can be thought of as the discretized counterpart to the matrix Q in (8). Each eigenvalue of H then corresponds to an eigenvector (“mode”) of the coefficient $q(\mathbf{x})$ we would like to recover. Moreover, large eigenvalues correspond to modes that are insensitive to noise, whereas small eigenvalues correspond to modes that are strongly affected by noise. As a consequence, we would like to aim for discretizations that result in many large and few small eigenvalues.

Fig. 5 provides a numerical evaluation of this perspective. It shows that when refining the mesh using the information content criterion, the eigenvalues of H are further to the top right – in other words, there are more large eigenvalues than when using refinement by the error estimator or the smoothness indicator. This pattern persists after both three refinement cycles (the left part of the figure) and six refinement cycles (the right part).

The stair-step structure of the figure results from the fact that mesh refinement turns one large cell into four small ones. Consequently, in general one large eigenvalue turns into four smaller eigenvalues. By counting the number of derivatives present in the operators that enter into H , we can conjecture that the conditioning of the problem scales with the mesh size h squared; indeed, the levels in the plots confirm that each mesh refinement step reduces the size of the smallest eigenvalues by approximately a factor of $(h_{\text{large}}/h_{\text{small}})^2 = 4$.

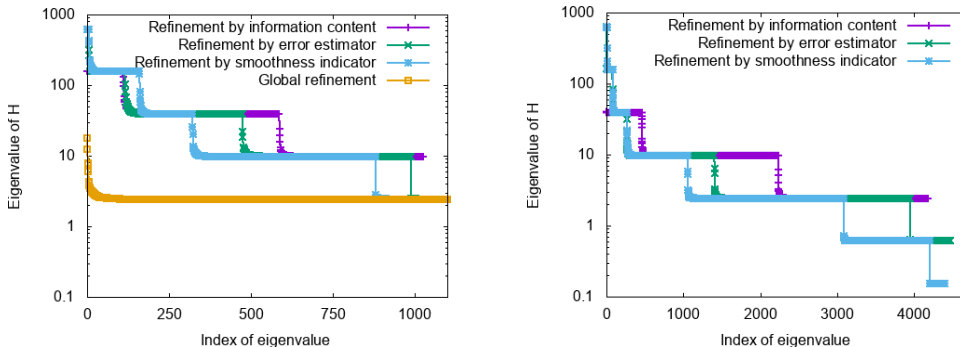


FIGURE 5. Comparison of the eigenvalues of the resolution matrix H for different ways of refining meshes. Left: After three refinement cycles, yielding problems with approximately 1,000 parameters to identify for the adaptive refinement criteria, and 16,384 for global refinement. For the global refinement case, only the first 1,100 eigenvalues are shown; the remaining ones have nearly the same value. Right: After six refinement cycles, resulting in problems with approximately 4,000 parameters for the adaptive refinement criteria. (Global refinement would have resulted in more than one million parameters; the eigenvalues of this matrix could not be computed.)

The left part of the figure also shows the eigenvalues of H for meshes constructed via global refinement, where the globally refined mesh is chosen so that it has the same finest resolution as the adaptively refined ones. Global refinement results in vastly more unknowns than for the adaptively refined meshes, with a large majority of small eigenvalues. These eigenvalues all correspond to small cells to the right of the array of detectors where very little information is available. Given the size of the problem, we were not able to compute the eigenvalues of H following six global refinement steps; the corresponding data are therefore omitted in the right panel.

The comparison shown in the figure confirms the hypothesis we laid out at the beginning of the section: namely, that refining the mesh based on information contents leads to an inverse problem that is better posed than those that result from any of the other refinement criteria we have compared with, in the sense that our approach leads to more large eigenvalues.

6. Conclusions and outlook. In this paper, we have used a statistical approach to define how much information we have about the parameter that is recovered in an inverse problem. More concretely, we have defined a density $j(\mathbf{x})$ that corresponds to *how much we know* about the solution at a given point \mathbf{x} , and derived an explicit expression for it that can be computed. We have then outlined a number of ways in which we believe that this information density can be used, via three vignettes. Finally, we have assessed one of these application areas numerically and showed that basing mesh refinement for inverse problems on information densities indeed leads to meshes that not only visually look more suited to the task than other criteria, but also quantitatively lead to better-posed discrete problems.

At the same time, this paper did not address many areas that would make for very natural next questions, including the following:

- In our work, we have chosen a simple L_2 regularization term $\frac{\beta}{2}\|q\|_{L_2}^2$, see (19). In practice, however, regularization terms would typically be used that penalize oscillations, for example using terms of the form $\frac{\beta}{2}\|\nabla^\alpha q\|_{L_2}^2$, $\alpha \geq 1$. How this would affect the definition of the information density would be interesting to ask in a future study. In order to consider the limit of infinite-dimensional inverse problems, we need to ensure that the covariance operators that result from regularization are of trace class (see, for example, [55, 2, 23]), which is also likely a precondition for the definition of information densities.
- For many inverse problems – such as ultrasound or seismic imaging, or electrical impedance tomography – the quantity we would like to identify is not a right-hand source term, but a coefficient in the operator on the left side of the equation. In these cases, the definition of the information density will have to be linearized around the solution of the inverse problem, which then may make the definition of $j(\mathbf{x})$ dependent on actual noise values. How this affects the usefulness of the information density is a priori unclear, but it would at least require solving the inverse problem before we can compute $j(\mathbf{x})$. We provide a brief outlook at this case in Appendix A.
- The computation of $j(\mathbf{x})$ requires the solution of a number of forward or adjoint problems, which is expensive, especially for three-dimensional inverse problems, even though these problems are all independent of each other and can be computed in parallel. At the same time, although we have shown in Sections 2 and 3 what is necessary to form the complete matrix Q , we need only the diagonal entries of this matrix, see (28). We may be able to compute approximations to the entries Q_{kk} more cheaply, for example by random projections, low-rank approximations, or hierarchical low-rank approximations. Such ideas can, for example, be found in [28, 29, 2, 54], and in [23, Section 5] and [3] for the closely related reduced Hessian matrix.

We leave the exploration of these topics to future work.

Acknowledgments. WB would like to thank Dustin Steinhauer for an introduction to the Fisher information matrix many years ago. WB’s work was partially supported by the National Science Foundation under award DMS-1821210; by award EAR-1925595; and by award OAC-2410847. CRJ’s work was partially supported by the Intel Graphics and Visualization Institutes of XeLLENCE, the National Institutes of Health under award R24 GM136986, the Department of Energy under grant number DE-FE0031880, and the Utah Office of Energy Development.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA-0003525.

Appendix A. On the extension to nonlinear problems. The derivations in both Section 2 (for the finite-dimensional case) and Section 3 (for the infinite-dimensional case) have assumed that the forward model is linear, and that the solution-to-measurement map is as well. This linearity is clearly a severe restriction – many inverse problems are nonlinear: typical examples are optical tomography,

ultrasound imaging, electrical impedance tomography, or seismic imaging. In all of these cases, we seek to identify a coefficient inside a differential operator.

In order to understand what needs to change in the formalism we have outlined in this paper to accommodate the nonlinear case, we stay with the finite-dimensional case for simplicity; the infinite-dimensional case will follow the exact same steps. The argument is essentially one of linearization, predicated on the assumption that measurement errors are small and that the parameter-to-measurement map is sufficiently smooth.

First, let us assume that we replace the forward model (4) by the following model that is (potentially) nonlinear in the parameters q_k we seek:

$$A(S\mathbf{q}; \mathbf{u}) = \mathbf{f}, \quad (37)$$

where $S\mathbf{q}$ is a vector of coefficients that appear in the operator and is expressed in terms of a basis encoded by the columns of the matrix S times the unknown parameters \mathbf{q} we would like to identify. We have also allowed for a right-hand side forcing term \mathbf{f} that is assumed known and that could have been incorporated into A but commonly is not. For a given \mathbf{q} , we then use the following notation to denote the solution of the forward model:

$$\mathbf{u} = A(S\mathbf{q}; \bullet)^{-1}\mathbf{f}. \quad (38)$$

Next, we assume that we obtain measurements using (possibly nonlinear) functionals m_ℓ instead of the linear ones in (5):

$$z_\ell = m_\ell(\mathbf{u}) + \varepsilon_\ell. \quad (39)$$

We can combine the state equation and the measurement process by introducing the parameter-to-measurement map $\mathbf{F}(S\mathbf{q})$:

$$z_\ell = \mathbf{F}(S\mathbf{q})_\ell + \varepsilon_\ell. \quad (40)$$

For comparison with the material in Section 2, note that in the linear case, we have that $A(S\mathbf{q}; \mathbf{u}) = \mathbf{A}\mathbf{u} - S\mathbf{q}$, $\mathbf{f} = 0$, $A(S\mathbf{q}; \bullet)^{-1}\mathbf{f} = A^{-1}S\mathbf{q}$, and $\mathbf{F}(S\mathbf{q}) = MA^{-1}S\mathbf{q}$.

With this parameter-to-measurement operator \mathbf{F} , the deterministic parameter estimation problem (7) then takes on the following form:

$$\min_{\mathbf{q}} \mathcal{J}_{\text{red}}(\mathbf{q}) := \underbrace{\frac{1}{2} \|\mathbf{F}(S\mathbf{q}) - \mathbf{z}\|_{\Sigma^{-2}}^2 + \frac{\beta}{2} \|R\mathbf{q}\|^2}_{=: \mathcal{J}_{\text{red}}(\mathbf{q})}. \quad (41)$$

Let us assume that this problem has a unique minimizer, which we will denote by $\mathbf{q}^* = \mathbf{q}^*(\mathbf{z})$.⁸ Then, under appropriate smoothness assumptions, we can perform a Taylor expansion of the optimality conditions for \mathbf{q}^* . For two random noisy realizations of measurements $\mathbf{z}_1, \mathbf{z}_2$, we obtain that the minimizers $\mathbf{q}_1 = \mathbf{q}(\mathbf{z}_1)$, $\mathbf{q}_2 = \mathbf{q}(\mathbf{z}_2)$ of (41) have to satisfy the following relationship that generalizes (9):

$$Q(\mathbf{q}_1 - \mathbf{q}_2) = S^T F'(S\mathbf{q}^*)^T \Sigma^{-2} (\mathbf{z}_1 - \mathbf{z}_2) + \mathcal{O}(\|\mathbf{z}_1 - \mathbf{z}\|^2) + \mathcal{O}(\|\mathbf{z}_2 - \mathbf{z}\|^2). \quad (42)$$

Here, the matrix $F'(S\mathbf{q}^*)$ is the derivative of the vector function \mathbf{F} with regard to its argument, evaluated at $S\mathbf{q}^*$. In the linear case, $F' = MA^{-1}$ is independent of q^* . The matrix Q above readily generalizes the situation for the linear case and has the form:

$$Q = S^T F'(S\mathbf{q}^*)^T \Sigma^{-2} F'(S\mathbf{q}^*) S + \beta R^T R. \quad (43)$$

⁸There are, of course, problems in which \mathcal{J}_{red} has more than one local minimizer – for example in full waveform seismic wave tomography. In those cases, everything that follows will not be useful because it relies on *local* linearization of the problem.

Ultimately, the information content we seek is based on the precision matrix, which we obtain via the Fisher information matrix from the Bayesian inverse problem (11). Specifically, we will continue to define the information content via the square root of the inverse of the diagonal elements of the covariance matrix (13), which we approximate by way of the Cramér-Rao bound and using (14):

$$j_k := \sqrt{(I_p)_{kk}}, \quad \text{where } (I_p)_{kk} = \mathbb{E} \left[\frac{\partial^2}{\partial q_k^2} \mathcal{J}_{\text{red}}(\mathbf{q}) \right] \quad (44)$$

The differences between the linear and nonlinear cases appear here. First, in the linear case, the Cramér-Rao bound holds with equality, and as a consequence the definition of j_k in the previous equation also satisfies the desired identity (12),

$$j_k := \sqrt{(I_p)_{kk}} = \frac{1}{\sqrt{\text{var}_p(\mathbf{q})_k}}.$$

However, in the nonlinear case, the probability distribution $p(\mathbf{q}|\mathbf{z})$ is no longer Gaussian (because \mathcal{J}_{red} is no longer quadratic in \mathbf{q}), and consequently the last identity of the previous equation may no longer be exact.

Second, in the linear case, we had $I_p = Q$ because $\mathcal{J}_{\text{ref}}(\mathbf{q})$ is a quadratic function, its second derivative is a constant, $\frac{\partial^2}{\partial q_k \partial q_l} \mathcal{J}_{\text{red}}(\mathbf{q}) = Q_{kl}$, and consequently the expectation value in (44) simply evaluates to Q_{kk} . In the nonlinear case, a straightforward Taylor expansion of the definition of \mathcal{J}_{red} around \mathbf{q}^* shows that to first order in $\mathbf{q} - \mathbf{q}^*$,

$$I_p \approx Q + \text{sym}((\mathbf{F}(S\mathbf{q}^*) - \mathbf{z})^T \Sigma^{-2} F''(S\mathbf{q}^*) S S) + \text{higher order terms},$$

with appropriate contractions over the indices of the various matrices and vectors that appear in this expression. Here, $\text{sym}(X) = \frac{1}{2}(X + X^T)$ symmetrizes the matrix X .

The first correction term beyond Q in the previous expression corresponds to the difference between the Gauss-Newton and Newton matrices in least-squares minimization [45]. The correction is small if either the problem is nearly linear (F'' is small), or if the residual $\mathbf{F}(S\mathbf{q}^*) - \mathbf{z}$ is small – essentially a condition on whether the model is able to accurately predict the data and whether the noise level is small.

The next order term of the higher order corrections in the expression above is of the form $\mathbb{E}[\text{sym}(P(\mathbf{q} - \mathbf{q}^*))] = \text{sym}(P \mathbb{E}[\mathbf{q} - \mathbf{q}^*])$, where P is a rank-3 tensor composed of a lengthy list of terms proportional to at least second derivatives of \mathbf{F} , often multiplied by the residual $\mathbf{F}(S\mathbf{q}^*) - \mathbf{z}$. As above, for problems that either have a small residual or are nearly linear, P by itself is already small. Furthermore, if the noise level is small, we should expect the posterior probability to be localized near \mathbf{q}^* , nearly symmetric around \mathbf{q}^* , and therefore for the expectation value $\mathbb{E}[\mathbf{q} - \mathbf{q}^*]$ to be small.

Whether or not these correction terms will be small for a concrete application depends, of course, on both how nonlinear the operator \mathbf{F} is (i.e., how large \mathbf{F}'' is) and how large we expect the measurement errors (and consequently the residuals) to be. Assuming that one or the other is indeed small for a particular case, it seems reasonable to expect that our choice (16) for the information content, namely

$$j_k := \sqrt{Q_{kk}},$$

may still be a useful one for nonlinear problems.

Having settled the question of how to define the information content j_k in the nonlinear case, there remains the question of how to compute it. Given (43), we

find that it can be computed in much the same way as outlined in Section 2.2, Remark 2.1. In analogy to the derivations there, note that now

$$\begin{aligned} Q_{kk} &= \mathbf{e}_k^T S^T F'(S\mathbf{q}^*)^T \Sigma^{-2} F'(S\mathbf{q}^*) S \mathbf{e}_k + \beta \mathbf{r}_k^T \mathbf{r}_k \\ &= (\Sigma^{-1} F'(S\mathbf{q}^*) S \mathbf{e}_k)^T (\Sigma^{-1} F'(S\mathbf{q}^*) S \mathbf{e}_k) + \beta \mathbf{r}_k^T \mathbf{r}_k \\ &= \sum_{\ell} \frac{1}{\sigma_{\ell}^2} \eta_{\ell k}^2 + \beta \mathbf{r}_k^T \mathbf{r}_k, \end{aligned}$$

where

$$\eta_{\ell k} = F'_{\ell}(S\mathbf{q}^*) S \mathbf{e}_k.$$

These quantities require a solve with the linearized forward model, linearized around the solution \mathbf{q}^* of the deterministic problem. Similar considerations apply when computing Q_{kk} via adjoint solutions (the ‘‘alternative way’’ in Remark 2.1).

The cost of these forward solves is the same as for the linear case, but two other considerations come into play. First, the solution of the nonlinear inverse problem (in which we are of course primarily interested) is generally more expensive than for linear problems, and so the cost of computing information contents j_k may or may not be a major cost any more, unlike in the linear case. On the other hand, in the nonlinear case, the definition of Q_{kk} and consequently j_k depends on the solution \mathbf{q}^* of the deterministic problem and, thus, on the noisy data \mathbf{z} from which \mathbf{q}^* was computed. This is unfortunate: One of the major benefits of using j_k touted in Remark 3.4 was that it did *not* depend on the noisy data, unlike all other criteria we know of. Whether this dependence is strong enough in practice to be important will remain an open question for now; as in the discussions about the difference between Q and I_p , we can speculate that this is a higher order concern, unlike for example in the mesh refinement criteria in Fig. 4, where the noise in the data was clearly visible in the structure of the generated mesh.

Appendix B. On the derivation of traditional mesh refinement criteria.

Section 5.3.1 contains a comparison of our information-density based mesh refinement criterion against two other mesh refinement criteria that we have called ‘‘error estimator’’ and ‘‘smoothness indicator’’. Both of these can be derived in ways that follow common finite element theory. In the following, we will attempt to provide an intuitive derivation of their form, without attempting to be precise and complete.

The starting point for the derivation of both forms is the optimality condition (33) – a system of partial differential equations – for the deterministic problem we are trying to solve. For simplicity, let us repeat this set of equations here:

$$\begin{aligned} \mathcal{L}u(\mathbf{x}) &= q(\mathbf{x}), \\ \mathcal{L}^* \lambda(\mathbf{x}) &= - \sum_{\ell} \frac{1}{\sigma_{\ell}^2} (u(\xi_{\ell}) - z_{\ell}) \delta(\mathbf{x} - \xi_{\ell}), \\ \beta q - \lambda &= 0. \end{aligned}$$

In order to compute a numerical solution, we seek finite-dimensional approximations $u_h(\mathbf{x})$, $\lambda_h(\mathbf{x})$, $q_h(\mathbf{x})$ of the exact solution $u(\mathbf{x})$, $\lambda(\mathbf{x})$, $q(\mathbf{x})$ of this problem. We will use the finite element method; specifically, we will use piecewise polynomials of degree p_u, p_{λ}, p_q for this approximation. The appropriate choice in the current context is to use continuous shape functions for u_h and λ_h (because we apply a second-order differential operator, for which the weak form of the equations above requires H^1

solutions of which the continuous functions are a subset), and discontinuous shape functions for q_h (because no differential operator is ever applied to this variable).

To assess the quality of the mesh we have chosen, we would like to quantify the “error”, i.e., the difference between u and u_h , and similarly for the other variables. For the second order operator \mathcal{L} we consider, the appropriate norm to measure this error is

$$e := \left(\|\nabla(u - u_h)\|_{L^2(\Omega)}^2 + \|\nabla(\lambda - \lambda_h)\|_{L^2(\Omega)}^2 + \|q - q_h\|_{L^2(\Omega)}^2 \right)^{1/2}. \quad (45)$$

There are now two conceptually different ways in which we can approach estimating this error: *a priori* and *a posteriori* error estimates. These two approaches will yield the two criteria we used in Section 5.3.1.

A priori error estimates are based on the premise that for many equations and under certain conditions,⁹ one can show that on every cell K of the mesh, we have [20, 21]

$$\begin{aligned} \|\nabla(u - u_h)\|_{L^2(K)} &\leq c_u \|\nabla(u - I_{h,p_u}u)\|_{L^2(K)} &&\leq C_u h_K^{p_u} \|\nabla^{p_u+1}u\|_{L^2(K)}, \\ \|\nabla(\lambda - \lambda_h)\|_{L^2(K)} &\leq c_\lambda \|\nabla(\lambda - I_{h,p_\lambda}\lambda)\|_{L^2(K)} &&\leq C_\lambda h_K^{p_\lambda} \|\nabla^{p_\lambda+1}\lambda\|_{L^2(K)}, \\ \|q - q_h\|_{L^2(K)} &\leq c_q \|q - I_{h,q}q\|_{L^2(K)} &&\leq C_q h_K^{p_q} \|\nabla^{p_q+1}q\|_{L^2(K)}, \end{aligned}$$

where $c_\bullet, C_\bullet < \infty$ are constants we know exist but generally do not know the value of, and $I_{h,p}$ denotes the interpolation operator from a function to a discrete, piecewise polynomial function represented on the mesh with polynomial degree p . The first inequality in each line above represents the “best-approximation” property that holds for many equations and finite element schemes, whereas the latter is the Bramble-Hilbert inequality for polynomial approximation. h_K denotes the diameter of cell K .

Because we were primarily interested in the mesh used to discretize the coefficient q in Section 5.3, we chose $p_u = p_\lambda = 3, p_q = 0$. In other words, we used cubic elements to represent u, λ accurately, but approximated q only with piecewise constant functions. With the considerations above, the *a priori* error estimate (45) would then have the form

$$\begin{aligned} e &= \left(\sum_K \left[\|\nabla(u - u_h)\|_{L^2(K)}^2 + \|\nabla(\lambda - \lambda_h)\|_{L^2(K)}^2 + \|q - q_h\|_{L^2(K)}^2 \right] \right)^{1/2} \\ &\leq C \left(\sum_K \left[h_K^6 \|\nabla^4 u\|_{L^2(K)}^2 + h_K^6 \|\nabla^4 \lambda\|_{L^2(K)}^2 + h_K^2 \|\nabla q\|_{L^2(K)}^2 \right] \right)^{1/2} \\ &\approx C \left(\sum_K h_K^2 \|\nabla q\|_{L^2(K)}^2 \right)^{1/2}. \end{aligned}$$

The last approximation is because for sufficiently small h_K , the first two terms in square brackets are much smaller than the third because their exponents are much

⁹Whether the conditions that allow us to derive such error estimates are actually satisfied for (33) is perhaps not actually very important here. This is because we do not seek to strictly bound the error; instead, our goal is to derive a criterion that can tell us where the error is small or large, and that can consequently be used to *refine the mesh*. Experience in the finite element community has long shown that for this, less rigorous goal, *heuristic* derivations of refinement criteria “by analogy” are sufficient and often surprisingly effective.

larger. In practice, the exact solution q is of course unknown, but we can derive a computable quantity for mesh refinement purposes by defining

$$\eta_K = h_K \|\nabla_h q_h\|_{L_2(K)}$$

for every cell, where q_h is the computed approximate solution, and $\nabla_h q_h$ is a finite difference approximation of the gradient ∇q based on the piecewise constant function q_h . We have omitted the (unknown) constant C because we are only interested in comparing which cells have large and which have small errors, rather than actually computing a reasonable approximation for e . This derivation of η_K therefore leads to the “smoothness indicator” used in Section 5.3.1. The term “smoothness” indicates that we are measuring the size of derivatives of the function (\mathbf{x}).

The second approach toward deriving estimates for the error e , called “a posteriori error estimation”, is premised on first computing a numerical solution u_h, λ_h, q_h and then deriving an estimate for e that does not require knowledge of the true solution u, λ, q . Omitting many details (but see [12, 60]), we typically end up with an estimate that contains norms of the residual (i.e., the degree to which the left and right hand sides of (33) are not actually equal), times a constant of again unknown size, times powers of the mesh size h_K in the same way as above. That is, these error estimators will have the form

$$\begin{aligned} e &\leq C \left\{ \sum_K \left[h_K^6 \left(\|\mathcal{L}u_h - q_h\|_{L^2(K)}^2 + \text{jump}(u_h) \right) \right. \right. \\ &\quad \left. \left. + h_K^6 \left(\left\| \mathcal{L}^* \lambda_h + \sum_{\ell} \frac{1}{\sigma_{\ell}^2} (u(\xi_{\ell}) - z_{\ell}) \delta(\mathbf{x} - \xi_{\ell}) \right\|_{L^2(K)}^2 + \text{jump}(\lambda_h) \right) \right. \right. \\ &\quad \left. \left. + h_K^2 \|\beta q_h - \lambda_h\|_{L^2(K)}^2 \right] \right\}^{1/2} \\ &\approx C \left(\sum_K h_K^2 \|\beta q_h - \lambda_h\|_{L^2(K)}^2 \right)^{1/2}. \end{aligned}$$

In these formulas, $\text{jump}(u_h)$ and $\text{jump}(\lambda_h)$ are “jump terms” that arise due to integration by parts and whose specific form is not important in the current context. Crucially, again for reasons of high powers of h_K , we can estimate e in the last line by omitting everything but the term that contains the residual of the last line of the optimality conditions (33). Again dropping the unknown constant C for the same reasons as above, we can define an indicator for the size of the error on cell K by

$$\eta_K = h_K \|\beta q_h - \lambda_h\|_{L_2(K)}.$$

This is nearly the “error estimator” form of η_K presented in Section 5.3.1 – the difference being that we use the L_1 -norm instead of the L_2 -norm above.¹⁰ We note that this derivation is essentially what all publications do that base their mesh refinement decisions on a posteriori error estimates, perhaps up to the choice of norm or using weighting factors derived from dual solutions (see, for example, [9, 12, 10, 35], along with many others). The estimator η_K we use here is then really

¹⁰The choice of norm is immaterial for the derivation: At some point in the development, one has to apply Hölders inequality and can choose exponents, which can produce either of the two norms mentioned.

just a “redux” of what other papers use, adapted to the fact that we have chosen $p_u, p_\lambda \gg p_q$.

The two choices for the refinement criterion η_K we have derived in these ways thus represent the two most widely used ways to derive error estimators for finite element discretizations. There, of course, exists a large collection of publications that derive variations of such estimators, perhaps also including ways of assessing the sizes of the constants that appear in the estimates, but experience has shown that – at least for the purposes of mesh refinement – the meshes obtained from different estimators vary little and look conceptually very similar within each of the two broad categories of deriving these estimators. As a consequence, we consider the two choices we have presented here as representative for these categories, and we do not expect that any variation on the approach would produce meshes that are fundamentally better or worse. This, in particular, includes the methods presented in the papers mentioned in the previous paragraph. In contrast, the refinement indicator based on the information density is an entirely different approach that has no resemblance to either of the two methods outlined above – in fact, it makes no reference to any numerical solution of the optimality conditions at all!

REFERENCES

- [1] V. Akçelik, G. Biros, O. Ghattas, J. Hill, D. Keyes and B. van Bloemen Waanders, [Parallel algorithms for PDE-constrained optimization](#), chapter 16, SIAM, 2006, 291-322.
- [2] A. Alexanderian, N. Petra, G. Stadler and O. Ghattas, [A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized \$\ell_0\$ -sparsification](#), *SIAM Journal on Scientific Computing*, **36** (2014), A2122-A2148.
- [3] I. Ambartsumyan, W. Boukaram, T. Bui-Thanh, O. Ghattas, D. Keyes, G. Stadler, G. Turkiyyah and S. Zampini, [Hierarchical matrix approximations of Hessians arising in inverse problems governed by PDEs](#), *SIAM J. Sci. Comput.*, **42** (2020), A3397-A3426.
- [4] D. Arndt, W. Bangerth, B. Blais, M. Fehling, R. Gassmüller, T. Heister, L. Heltai, U. Köcher, M. Kronbichler, M. Maier, P. Munch, J.-P. Pelteret, S. Proell, K. Simon, B. Turcksin, D. Wells and J. Zhang, [The deal.II library, version 9.3](#), *Journal of Numerical Mathematics*, **29** (2021), 171-186.
- [5] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin and D. Wells, [The deal.II finite element library: Design, features, and insights](#), *Computers & Mathematics with Applications*, **81** (2021), 407-422, <https://arxiv.org/abs/1910.13247>.
- [6] S. R. Arridge and M. Schweiger, [Inverse methods for optical tomography](#), in *Information Processing in Medical Imaging* (eds. H. H. Barrett and A. F. Gmitro), Springer Berlin Heidelberg, Berlin, Heidelberg, 1993, 259-277.
- [7] A. C. Atkinson and A. N. Donev, *Optimum Experimental Design*, Clarendon Press, Oxford, 1992.
- [8] L. Auer, L. Boschi, T. W. Becker, T. Nissen-Meyer and D. Giardini, [Savani: A variable resolution whole-mantle model of anisotropic shear velocity variations based on multiple data sets](#), *J. Geophys. Res. Solid Earth*, **119** (2014), 3006-3034.
- [9] W. Bangerth, *Adaptive Finite Element Methods for the Identification of Distributed Parameters in Partial Differential Equations*, PhD thesis, University of Heidelberg, 2002.
- [10] W. Bangerth, [A framework for the adaptive finite element solution of large inverse problems](#), *SIAM J. Sc. Comput.*, **30** (2008), 2965-2989.
- [11] W. Bangerth and A. Joshi, [Adaptive finite element methods for the solution of inverse problems in optical tomography](#), *Inverse Problems*, **24** (2008), 034011, 22 pp.
- [12] W. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Birkhäuser Verlag, 2003.
- [13] H. T. Banks, K. Holm and F. Kappel, [Comparison of optimal design methods in inverse problems](#), *Inverse Problems*, **27** (2011), 075002, 31 pp.
- [14] A. J. Berglund, [Statistics of camera-based single-particle tracking](#), *Physical Review E*, **82** (2010).

- [15] J. G. Berryman, [Analysis of approximate inverses in tomography in tomography I. Resolution analysis of common inverses](#), *Optimization and Engineering*, **1** (2000), 87-115
- [16] D. Bigoni, Y. Chen, N. Garcia Trillos, Y. Marzouk and D. Sanz-Alonso, [Data-driven forward discretizations for Bayesian inversion](#), *Inverse Problems*, **36** (2020), 105008, 27 pp.
- [17] R. Bonadio, S. Lebedev, T. Meier, P. Arroucau, A. J. Schaeffer, A. Licciardi, M. R. Agius, C. Horan, L. Collins, B. M. O'Reilly, P. W. Readman and Ireland Array Working Group, [Optimal resolution tomography with error tracking and the structure of the crust and upper mantle beneath Ireland and Britain](#), *Geophysical Journal International*, **226** (2021), 2158-2188.
- [18] L. Borcea, V. Druskin and A. V. Mamonov, [Circular resistor networks for electrical impedance tomography with partial boundary measurements](#), *Inverse Problems*, **26** (2010), 045010, 30 pp.
- [19] E. Bozdağ, D. Peter, M. Lefebvre, D. Komatitsch, J. Tromp, J. Hill, N. Podhorszki and D. Pugmire, [Global adjoint tomography: First-generation model](#), *Geophysical Journal International*, **207** (2016), 1739-1766.
- [20] D. Braess, *Finite Elements*, Cambridge University Press, 1997.
- [21] S. C. Brenner and R. L. Scott, *The Mathematical Theory of Finite Elements*, 2nd edition, Springer, Berlin-Heidelberg-New York, 2002.
- [22] T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler and L. C. Wilcox, [Extreme-scale UQ for Bayesian inverse problems governed by PDEs](#), in *Proceedings of the Supercomputing 2012 conference*, ACM/IEEE SC Conference Series, 2012.
- [23] T. Bui-Thanh, O. Ghattas, J. Martin and G. Stadler, [A computational framework for infinite-dimensional bayesian inverse problems part i: The linearized case, with application to global seismic inversion](#), *SIAM J. Sci. Comput.*, **35** (2013), A2494-A2523.
- [24] C. Cohen-Bacrie, Y. Goussard and R. Guardo, [Regularized reconstruction in electrical impedance tomography using a variance uniformization constraint](#), *IEEE Trans. Med. Imag.*, **16** (1997), 562-571.
- [25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 2006.
- [26] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman, 1974.
- [27] F. A. Dahlen, S.-H. Hung and G. Nolet, [Fréchet kernels for finite-frequency traveltimes—I. Theory](#), *Geophys. J. Int.*, **141** (2000), 157-174.
- [28] S. Das, [Efficient Calculation of Fisher Information Matrix: Monte Carlo Approach Using Prior Information](#), PhD thesis, Johns Hopkins University, 2007.
- [29] S. Das, J. C. Spall and R. Ghanem, [Efficient monte carlo computation of fisher information matrix using prior information](#), *Computational Statistics & Data Analysis*, **54** (2010), 272-289.
- [30] A. Datta-Gupta, J. Xie, N. Gupta, M. J. King and W. J. Lee, [Radius of investigation and its generalization to unconventional reservoirs](#), *J. Petroleum Techn.*, **63** (2011), 52-55.
- [31] M. H. DeGroot, *Optimal Statistical Decisions*, Wiley-Interscience, 1970.
- [32] A. F. Emery and A. V. Nenarokomov, [Optimal experiment design](#), *Measurement Science and Technology*, **9** (1998), 864-876.
- [33] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, 1996.
- [34] Y. Favennec, [Hessian and Fisher matrices for error analysis in inverse heat conduction problems](#), *Numerical Heat Transfer, Part B: Fundamentals*, **52** (2007), 323-340.
- [35] A. Griesbaum, B. Kaltenbacher and B. Vexler, [Efficient computation of the Tikhonov regularization parameter by goal-oriented adaptive discretization](#), *Inverse Problems*, **24** (2008), 025025, 20 pp.
- [36] S.-H. Hung, F. A. Dahlen and G. Nolet, [Fréchet kernels for finite-frequency traveltimes—II. Examples](#), *Geophys. J. Int.*, **141** (2000), 175-203.
- [37] A. Jamshidi, J. M. V. Samani, H. M. V. Samani, A. Zanini, M. G. Tanda and M. Mazaheri, [Solving inverse problems of unknown contaminant source in groundwater-river integrated systems using a surrogate transport model based optimization](#), *Water*, **12** (2020), 2415.
- [38] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2006.
- [39] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall, 1993.
- [40] F. J. Kuchuk, [Radius of investigation for reserve estimation from pressure transient well tests](#), *SPE*, 120515, 1-22.

- [41] W. Lei, Y. Ruan, E. Bozdağ, D. Peter, M. Lefebvre, D. Komatitsch, J. Tromp, J. Hill, N. Podhorszki and D. Pugmire, [Global adjoint tomography—model GLAD-M25](#), *Geophysical Journal International*, **223** (2020), 1-21.
- [42] Q. Liu and Y. J. Gu, [Seismic imaging: From classical to adjoint tomography](#), *Tectonophysics*, **566-567** (2012), 31-66.
- [43] R. Montelli, G. Nolet, F. A. Dahlen and G. Masters, [A catalogue of deep mantle plumes: New results from finite-frequency tomography](#), *Geochemistry, Geophysics, Geosystems*, **7** (2006).
- [44] R. M. Neupauer, B. Borchers and J. L. Wilson, [Comparison of inverse methods for reconstructing the release history of a groundwater contamination source](#), *Water Resources Research*, **36** (2000), 2469-2475.
- [45] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, New York, 1999.
- [46] S. Nordebo and M. Gustavsson, [A priori modeling for gradient based inverse scattering algorithms](#), *Progress in Electromagnetics Research B*, **16** (2009), 407-432.
- [47] N. Petra, J. Martin, G. Stadler and O. Ghattas, [A computational framework for infinite-dimensional bayesian inverse problems, part ii: Stochastic newton mcmc with application to ice sheet flow inverse problems](#), *SIAM J. Sci. Comput.*, **36** (2014), A1525–A1555.
- [48] B. W. Pogue, T. O. McBride, J. Prewitt, U. L. Österberg and K. D. Paulsen, [Spatially variant regularization improves optical tomography](#), *Appl. Optics*, **38** (1999), 2950-2961.
- [49] N. Rawlinson, A. Fichtner, M. Sambridge and M. K. Young, [Seismic tomography and the assessment of uncertainty](#), in *Advances in Geophysics*, Elsevier, **55** (2014), 1-76.
- [50] J. Ritsema, A. Deuss, H. J. van Heijst and J. H. Woodhouse, [S40RTS: A degree-40 shear-velocity model for the mantle from new rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements](#), *Geophysical Journal International*, **184** (2010), 1223-1236.
- [51] E. Romera, P. Sánchez-Moreno and J. S. Dehesa, [The Fisher information of single-particle systems with a central potential](#), *Chemical Physics Letters*, **414** (2005), 468-472.
- [52] M. Sambridge and N. Rawlinson, [Seismic tomography with irregular meshes](#), in *Seismic Earth: Array Analysis of Broadband Seismograms* (eds. A. Levander and G. Nolet), American Geophysical Union, **157** (2005), 49-65.
- [53] T. H. Skaggs and Z. J. Kabala, [Recovering the release history of a groundwater contaminant](#), *Water Resources Research*, **30** (1994), 71-79.
- [54] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio and Y. Marzouk, [Optimal low-rank approximations of Bayesian linear inverse problems](#), *SIAM J. Sci. Comput.*, **37** (2015), A2451-A2487.
- [55] A. M. Stuart, [Inverse problems: A Bayesian perspective](#), *Acta Numerica*, **19** (2010), 451-559.
- [56] L. Tang, G. Hamarneh and R. Abugharbieh, [Reliability-driven, spatially-adaptive regularization for deformable registration](#), in *Biomedical Image Registration, 4th International Workshop, WBIR 2010, Lübeck, Germany, July 11-13, 2010. Proceedings* (eds. B. Fischer, B. M. Dawant and C. Lorenz), vol. 6204 of Lecture Notes in Computer Science, Springer, 2010, 173-185.
- [57] A. Tarantola, *Inverse Problem Theory*, Elsevier, Amsterdam, New York, 1987.
- [58] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, 2004.
- [59] D. W. Vasco, H. Keers and K. Karasaki, [Estimation of reservoir properties using transient pressure data: An asymptotic approach](#), *Water Resources Res.*, **36** (2000), 3447-3465.
- [60] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Wiley/Teubner, New York, Stuttgart, 1996.
- [61] J. A. Vrugt, P. H. Stauffer, T. Wöhling, B. A. Robinson and V. V. Vesselinov, [Inverse modeling of subsurface flow and transport properties: A review with new developments](#), *Vadose Zone Journal*, **7** (2008), 843-864.
- [62] M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann and S. M. Smith, [Mixture models with adaptive spatial regularization for segmentation with an application to MRI data](#), *IEEE Trans. Med. Imag.*, **24** (2005), 1-11.
- [63] Y. Yin, E. A. Hoffman and C.-L. Lin, [Lung lobar slippage assessed with the aid of image registration](#), *Med. Image. Comput. Comput. Assist. Interv.*, **13** (2010), 578-585.
- [64] H. Zhang and C. H. Thurber, [Estimating the model resolution matrix for large seismic tomography problems based on lanczos bidiagonalization with partial reorthogonalization](#), *Geophysical Journal International*, **170** (2007), 337-345.

- [65] L. Zhao, T. H. Jordan, K. B. Olsen and P. Chen, [Fréchet kernels for imaging regional earth structure based on three-dimensional reference models](#), *Bulletin of the Seismological Society of America*, **95** (2005), 2066-2080.

Received February 2025; revised September 2025; early access November 2025.