Numerical Algebra, Control and Optimization

doi:10.3934/naco.2022036

EFFICIENT LEARNING METHODS FOR LARGE-SCALE OPTIMAL INVERSION DESIGN

Julianne Chung[✉1], Matthias Chung[✉*1]

Silvia Gazzola[✉2], Mirjeta Pasha[✉3]

[1]Department of Mathematics
Emory University, Atlanta, GA 30322, USA

[2]Department of Mathematical Sciences
University of Bath, Bath BA2 7AY, UK

[3]Department of Mathematics
Tufts University, Medford, MA, USA

Abstract. In this work, we investigate various approaches that use learning from training data to solve inverse problems, following a bi-level learning approach. We consider a general framework for *optimal inversion design*, where training data can be used to learn optimal regularization parameters, data fidelity terms, and regularizers, thereby resulting in superior variational regularization methods. In particular, we describe methods to learn optimal p and q norms for $\mathrm{L}^p - \mathrm{L}^q$ regularization and methods to learn optimal parameters for regularization matrices defined by covariance kernels. We exploit efficient algorithms based on Krylov projection methods for solving the regularized problems, both at training and validation stages, making these methods well-suited for large-scale problems. Our experiments show that the learned regularization methods perform well even when there is some inexactness in the forward operator, resulting in a mixture of model and measurement error.

1. **Introduction.** Inverse problems arise in many important science and engineering applications such as biomedical and astronomical imaging, satellite surveillance, and seismic monitoring [42, 18]. Two of the main challenges to solving large-scale inverse problems are (i) ill-posedness of the problem, whereby small noise or errors in the data can and often do lead to large errors in the solution, and (ii) the large size of the problem, which for some applications is on the order of millions of observations and billions of unknown parameters. A standard way to solve inverse problems is to follow a variational approach, where solutions are computed by minimizing a pre-determined energy functional that depends upon assumptions regarding the statistical distribution of the observational noise, the forward model, and any prior knowledge about the properties of the unknown solution. Although a significant

2020 *Mathematics Subject Classification.* Primary: 65F22, 65K10; Secondary: 62F15.

Key words and phrases. Bi-level learning, learning priors, variational regularization, Krylov projection methods, inverse problems.

The first author is supported by NSF grants DMS-1654175 and DMS-1723005. The second author is supported by NSF grant DMS-1723005 and DMS-215266. The third author is supported by EPSRC grant EP/T001593/1. The fourth author is supported by NSF grant DMS-1502640.

* Corresponding author: Matthias Chung, matthias.chung@emory.edu.

This paper is handled by Andreas Mang as the guest editor.

amount of research has gone into developing efficient optimization methods to solve variational problems, the formulation of the optimization problem relies on standard assumptions that may not hold in general and that, moreover, may further rely on additional unknown (hyper)parameters.

In this work, we describe a general *optimal inversion design* (OID) framework for solving inverse problems, where the goal is to use available training data to design an optimal energy functional for variational inversion. In order to introduce the OID learning problem, we begin with a discrete linear inverse problem of the form,

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{true}} + \mathbf{e}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ represents a given forward model that is also known as the parameter-to-observation map, $\mathbf{b} \in \mathbb{R}^m$ stores available observations corrupted by some unknown additive noise $\mathbf{e} \in \mathbb{R}^m$, and $\mathbf{x}_{\text{true}}$ contains unknowns that should be recovered. We assume that the inverse problem is ill-posed, and therefore regularization is needed to compute stable, reasonable approximations of $\mathbf{x}_{\text{true}}$. The aim of regularization is to incorporate prior knowledge about the solution. There are many forms of regularization ranging from spectral filtering methods to variational regularization methods to iterative regularization, and many combinations and variants of these [42, 33]. In its general form, we consider approaches where the regularized solution can be computed as

$$\widehat{\mathbf{x}}(\boldsymbol{\theta}) \in \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\ \mathcal{J}(\mathbf{x}, \mathbf{A}, \mathbf{b}; \boldsymbol{\theta}) + \mathcal{R}(\mathbf{x}; \boldsymbol{\theta}), \tag{2}$$

where the overall loss is composed of a data fitting term $\mathcal{J}$, which incorporates the forward process $\mathbf{A}$ and information about the measurement process, such as the noise distribution in the observations $\mathbf{b}$, and a regularization functional $\mathcal{R}$ that integrates prior knowledge of $\mathbf{x}_{\text{true}}$. While determined by the underlying statistics, the selection of $\mathcal{J}$ and $\mathcal{R}$ is problem dependent and remains a crucial yet heuristic choice for the inversion process [18]. Here we assume that such *design* choices may be represented by some *design parameters* $\boldsymbol{\theta} \in \mathbb{R}^\ell$, often also referred to as *hyperparameters* [32].

Within this work we focus on a particular form of (2) which is given by

$$\widehat{\mathbf{x}}(\boldsymbol{\theta}) \in \underset{\mathbf{x}}{\arg\min}\ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda \|\mathbf{L}(\boldsymbol{\beta})\mathbf{x}\|_q^q, \tag{3}$$

with design parameters $\boldsymbol{\theta} = [\lambda; p; q; \boldsymbol{\beta}]$, where $\lambda, p, q \in \mathbb{R}^+$ and $\boldsymbol{\beta} \in \mathbb{R}^{\ell_{\boldsymbol{\beta}}}$. Note, with utilizing $\in$ instead of $=$ in (3) we merely emphasize the potential non-uniqueness of the minimizer. Here, $\|\cdot\|_s$ denotes the L^s-norm for $s \geq 1$ and a homogeneous function without all norm properties for $0 < s < 1$. This formulation encompasses many popular variational regularization methods. For instance:

1. For fixed $\mathbf{L}(\boldsymbol{\beta}) = \mathbf{L} \in \mathbb{R}^{r\times n}$ and $\boldsymbol{\theta} = [\lambda, p, q]^\top$, problem (3) is an $\mathrm{L}^p - \mathrm{L}^q$ type regularized problem,
$$\min_{\mathbf{x}}\ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda \|\mathbf{L}\mathbf{x}\|_q^q. \tag{4}$$
2. For fixed $p = q = 2$ and $\boldsymbol{\theta} = [\lambda, \boldsymbol{\beta}]^\top$, problem (3) may include a design-dependent operator $\mathbf{L} : \mathbb{R}^{\ell_{\boldsymbol{\beta}}} \to \mathbb{R}^{r\times n}$ in the regularization term, i.e.,
$$\min_{\mathbf{x}}\ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{L}(\boldsymbol{\beta})\mathbf{x}\|_2^2. \tag{5}$$
Within a Bayesian approach $\mathbf{L}(\boldsymbol{\beta})$ may be regarded as an inverse square root of a positive definite parameteric prior covariance matrix. Consequently, a minimizer of (5) may then constitute a maximum a posteriori estimate [51, 50].

Both problems (4) and (5) depend on the particular choice of the design parameters $\boldsymbol{\theta}$, and the main question is how to *optimally* select $\boldsymbol{\theta}$?

Assume that we are given the distribution of $\mathbf{x}_{\text{true}}$ and $\mathbf{e}$. Then optimal design parameters $\boldsymbol{\theta}$ may be selected by minimizing the Bayes risk, i.e.,

$$\min_{\boldsymbol{\theta}\in\Omega} \ \tfrac{1}{2}\mathbb{E}\,\|\widehat{\mathbf{x}}(\boldsymbol{\theta}) - \mathbf{x}_{\text{true}}\|_2^2 \tag{6a}$$

$$\text{subject to } \widehat{\mathbf{x}}(\boldsymbol{\theta}) \text{ solving (3)}, \tag{6b}$$

where Ω is a set of feasible design choices and $\mathbb{E}$ is the expected value. By minimizing the expected mean squared error (6a), the optimal design parameters are expected to perform well on average, leading to reconstructions $\widehat{\mathbf{x}}(\boldsymbol{\theta})$ that minimize the Bayes risk. While other design criteria are available, we focus on this design criterion, which is referred to as *A-design* in the field of optimal experimental design [62, 7].

For problems where the distribution of $\mathbf{x}_{\text{true}}$ is unknown or not obtainable, but training data are readily available, we consider empirical Bayes risk design problems, where the training data are used to approximate the expected value in (6a). Assume that we are given a set of training data consisting of J true models $\mathbf{x}_{\text{true}}^1, \dots, \mathbf{x}_{\text{true}}^J \in \mathbb{R}^n$ and simulated observations $\mathbf{b}^1, \dots, \mathbf{b}^J \in \mathbb{R}^m$, e.g., by data simulation through (1). Then we consider the empirical Bayes risk OID problem,

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Omega} \ \tfrac{1}{2J}\sum_{j=1}^{J} \left\|\widehat{\mathbf{x}}^j(\boldsymbol{\theta}) - \mathbf{x}_{\text{true}}^j\right\|_2^2 \tag{7a}$$

$$\text{subject to } \widehat{\mathbf{x}}^j(\boldsymbol{\theta}) \text{ solving (3) for data } \mathbf{b} = \mathbf{b}^j. \tag{7b}$$

In other words, the design problem is a bi-level optimization problem where the goal is to find the parameters $\boldsymbol{\theta}$ that minimize the sample average of reconstruction errors for some training set [37, 38, 22, 16, 4]. The *outer* optimization problem (7a) is referred to as the design problem, while the variational regularization problem (7b) is referred to as the *inner* problem.

Overview of main contributions. In this work, we describe efficient learning techniques to solve the overall design problem (7). Although this framework can incorporate various variational regularization techniques, we focus on the two scenarios described above. Learning the regularization parameter λ has been previously considered in various contexts, but to the best of our knowledge, learning optimal values of p and q (i.e., $\boldsymbol{\theta} = [\lambda, p, q]$) for the $\mathrm{L}^p - \mathrm{L}^q$ regularized problem and learning optimal parameters for covariance kernel matrices (i.e., $\boldsymbol{\theta} = [\lambda, \boldsymbol{\beta}]$) have not been considered in an OID framework before. We will show that these methods can handle various uncertainties in the problem, from mitigating errors in the forward model to resolving unknown parameters in the prior and noise assumptions. Furthermore, to efficiently handle the inner problem, we exploit recent developments in Krylov projection methods such as [25] and [47] and incorporate these methods into our bi-level optimization framework.

An outline of the paper is as follows. In Section 2 we provide a brief overview on previous research on learning methods for solving inverse problems. Section 3 is devoted to computational approaches for learning design parameters in an OID framework including details on iterative projection methods for solving the inner problems (7b). In Section 4, we provide numerical results for various image deblurring and tomography applications that demonstrate the effectiveness and benefits of our approaches. Conclusions are provided in Section 5.

2. **Previous works on learning for inverse problems.** Supervised learning techniques have gained significant interest in the inverse problems community as a way to combine model-driven and data-driven approaches for solving inverse problems. A comprehensive overview can be found in [6]. Two predominant classes of supervised learning approaches have emerged for solving inverse problems: empirical Bayes risk minimization approaches that are related to optimal experimental design techniques [38] and techniques based on deep learning tools such as neural networks and variational autoencoders [55]. Supervised training approaches for solving inverse problems were first formally introduced in Haber and Tenorio [37], where a bi-level optimization problem of the form (7) was considered for learning optimal parameters for the regularization functional. One of the many advantages of these learning approaches is that the learned (parameterized) regularization functional is tailored to a specific forward operator and noise level of the data. There have been various extensions of this idea (e.g., to learn optimal spectral filters [22, 23], optimal weighted and multi-parameter Tikhonov parameters [39, 45], and optimal weighted TV parameters [44]). An additional advantage is that empirical Bayes risk minimization approaches can exploit existing computationally efficient optimization techniques and incorporate a wide range of state constraints [65]. Furthermore, they are general in that different design objective functions can be incorporated [38], they can be used to learn critical information such as optimal sampling patterns (e.g., for MRI [66]) and design setup for experiments (e.g., for tomography [65]), and they have rich theoretical connections to Bayesian experimental design [3, 46]. The main concerns of this approach include the need to solve an expensive bi-level optimization problem [29, 28, 27, 16] and bias towards the training set, since reconstructed parameters are only good on average.

The other major class of supervised learning techniques to take root in the inverse problems community consists of methods that exploit deep learning techniques (e.g., neural networks and variational autoencoders), see e.g., review papers [6, 59, 58, 55]. Initially, deep learning was used mainly for postprocessing of solutions to improve solution quality (e.g., image denoising) or for performing tasks such as classification. However, deep neural networks are now being considered for solving inverse problems by learning the mapping from observation-to-reconstruction [41, 72] or by learning an appropriate auto-encoder network (e.g., a generative adversarial network) to serve as a proxy for the regularizer [57, 40, 54, 61]. However, a major disadvantage for many of these network learning approaches is that due to the large number of network parameters, a massive amount of training data is required, which may not be readily available. Furthermore, it is important to have a well-tuned network and a good choice of parameters (e.g., batch size, epochs, learning rate) for the stochastic optimization algorithms, prior to training the networks.

Although there have been significant developments in both classes of supervised learning approaches for inverse problems, there are still open problems. In particular, as mentioned in Section 1, we are interested in learning the appropriate L^p and L^q-norms along with the regularization parameter λ in (4). This problem is most related to the work by De los Reyes and Schönlieb [28], where parameter learning methods were considered for learning the noise model in variational image denoising by estimating weights for different noise models. However, rather than consider a weighted version of pre-determined noise models, our approach seeks an appropriate L^p-norm to resolve any errors or uncertainties in the data-fit term and combines it with an optimally-selected L^q-norm for the regularization term. For the problem

of learning optimal parameters for a regularization operator (5), special cases have been considered in supervised learning frameworks (e.g., [37] considered different regularization terms for different regions of the solution and [4] considered a bilevel optimization learning framework for learning the fractional Laplacian parameter). However, learning the kernel parameters in an OID framework remains a challenge, especially when the prior precision matrix (i.e., the inverse of the covariance matrix) or its square root $\mathbf{L}(\boldsymbol{\beta})$ is not readily available for all design parameters $\boldsymbol{\beta}$. In general, significant computational challenges may arise within large-scale bilevel optimization problem (7), which we address next.

3. Computational OID for variational inverse problems.

In this section, we describe computational approaches for the OID problem,

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Omega}{\arg\min}\ \mathcal{P}(\boldsymbol{\theta}) = \tfrac{1}{2J}\sum_{j=1}^{J}\left\|\widehat{\mathbf{x}}^j(\boldsymbol{\theta}) - \mathbf{x}^j_{\text{true}}\right\|_2^2 \tag{8a}$$

$$\text{s.t.}\ \ \widehat{\mathbf{x}}^j(\boldsymbol{\theta}) = \underset{\mathbf{x}}{\arg\min}\ \left\|\mathbf{A}\mathbf{x} - \mathbf{b}^j\right\|_p^p + \lambda\left\|\mathbf{L}(\boldsymbol{\beta})\mathbf{x}\right\|_q^q \quad j = 1,\dots,J, \tag{8b}$$

where $\boldsymbol{\theta} = [\lambda, p, q, \boldsymbol{\beta}]$ with $\lambda, p, q > 0$.

Bi-level optimization problems such as (8) are notoriously difficult to solve. For instance, simple non-convexity in the inner problem (such as those encountered when $p, q < 1$ in (8b)) may lead to discontinuities in the outer design problem, [30, 67]. As an example, consider the toy problem $\min_\theta\ \widehat{x}(\theta)$ where the inner problem $\widehat{x}(\theta) = \arg\min_x\ (x-1)^2(x+1)^2 + \theta x$ is non-convex in x. For $\theta = 0$ two global minima $\widehat{x}(0) = \pm 1$ exist and for any $\theta < 0$ and $\theta > 0$ we have $\widehat{x}(\theta) > 1$ and $\widehat{x}(\theta) < -1$, respectively. Hence the outer design function is discontinuous at $\theta = 0$. Many of the existing theoretical results rely on strict assumptions of differentiability and convexity, we refer the interested reader to various monographs on bilevel optimization problems, for instance [30, 8, 31].

Various approaches exist to address bi-level optimization problems. One approach is to cast the inner problem as a constraint and utilize "off-the-shelf" constrained optimization methods, such as augmented Lagrangians or interior-point methods. Computational challenges arise in this approach, since the inner problem results in non-standard equality and inequality constraints [16, 27, 4, 45]. Another approach commonly used in the PDE constrained optimization literature is to eliminate the constraints by approximately solving for $\widehat{\mathbf{x}}^j$, yielding a reduced problem. Such approaches were used to compute optimal error filters in [22, 38, 39]; however, for regularized solutions that have a nontrivial dependence on $\boldsymbol{\theta}$ (e.g., for general variational regularization methods (2) where the inner problem does not admit a closed form solution) such methods do not apply. In a third approach, the potentially discontinuous outer design problem is treated with non-gradient based global optimization methods such as evolutionary methods [70, 67]. Note that global optimization methods tend to be significantly more expensive than methods from convex optimization. Hence, special care must be taken in reducing the overall computational cost.

We approach the computational bottleneck from two directions. First, since the inner problem consists of a variational linear inverse problem, we take advantage of recently developed and highly efficient iterative solvers. Specifically, for the $\mathrm{L}^p - \mathrm{L}^q$ type problems, we use a majorization-minimization (MM) approach together with

generalized Krylov subspaces (GKS), dubbed MM-GKS [12]; for the parametric kernel learning problem, we use a generalized Golub-Kahan (genGK)-based method, sometimes in its hybrid version, dubbed genHyBR [25]. We refer to the discussions in Sections 3.1 and 3.2, respectively, for more details. Second, we utilize surrogate optimization techniques, also referred to as Bayesian optimization, for the outer problem (8a); see [36, 60]. One major advantage of surrogate optimization methods is that they construct a surrogate objective function and evaluate the surrogate instead of the true objective to find global minimizers[1], thereby reducing the overall number of inner solves (8b) in our design problem. Further, since the inner problems (8b) are solved using iterative methods, the solutions $\widehat{\mathbf{x}}^j(\boldsymbol{\theta})$ are affected by stopping criteria and any further hyper-parameters of the iterative method. As a consequence, the objective function $\mathcal{P}(\boldsymbol{\theta})$ in (8a) may be non-convex and may even have discontinuities. Surrogate optimization methods are versatile probabilistic methods and therefore suitable in our settings. Since surrogate optimization methods can be seen as sophisticated importance sampling methods, these methods suffer from the curse of dimensionality. Our problem setup avoids this major disadvantage of surrogate optimization methods by restricting the number of parameters for the outer optimization problem.

More precisely, a surrogate optimization method takes samples of the objective function given as $\mathcal{S}_K = \{\boldsymbol{\theta}_k, \mathcal{P}(\boldsymbol{\theta}_k)\}_{k=1}^K$ on a predetermined box $\Omega \subset \mathbb{R}^\ell$ and builds a surrogate model $s_K : \Omega \to \mathbb{R}$ by extrapolating the objective function (8a) beyond the sample points $\mathcal{S}_K$. For instance, surrogate models may be constructed using radial basis functions [69] or Gaussian processes [36]. Typically the surrogate model matches $\mathcal{P}$ exactly at points $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$, hence interpolating the true objective function $\mathcal{P}$ at $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$. From the surrogate model s_K, a merit or so called *acquisition function* $m_K : \Omega \to \mathbb{R}$ is constructed that balances the trade-off between exploitation and exploration [5]. A commonly used acquisition function is the expected improvement function, where the surrogate model predicts low objective function values by means of known sample locations and values, as well as taking into account uncertainty of unexplored regions.

In this work we utilize standard Matlab libraries for surrogate optimization provided by the global optimization toolbox [1] for outer problem (8a). Next we describe two computational OID problems: learning optimal p and q norms and learning optimal hyperparameters for the prior.

3.1. Learning optimal p and q norms. OID with $\boldsymbol{\theta} = [\lambda, p, q]$.

Consider OID problem (8) where $\mathbf{L}(\boldsymbol{\beta}) = \mathbf{L}$ is fixed and $\boldsymbol{\theta} = [\lambda, p, q]$, i.e., learning the optimal regularization parameter, data fidelity norm, and regularization norm.

There are various reasons why one would want to learn an optimal p. While it is well-known that $p = 2$ should be considered when the noise follows an i.i.d. Gaussian distribution and that $0 < p < 2$ should be considered when the available data are perturbed by non-Gaussian noise, it is unclear what to use when there is a mixture of noise corrupting the data. For specific statistical models of noise, e.g., mixed Gaussian and Poisson noise that arise from Charge Coupled Device detectors, a reformulation to a weighted least-squares problem has been considered, see e.g., [9, 17, 52, 12] and references therein. However, the reformulation relies on an approximation using knowledge about the noise statistics, which is not necessarily available in practice.

[1] In Bayesian optimization, it is common to consider equivalent maximization problems.

Moreover, learning p can be relevant when the forward operator $\mathbf{A}$ used to solve the inner problem (8b) is inexact. That is, the adopted forward model is (slightly) different from the one used to generate the training data; e.g., deblurring problems using erroneous point spread functions or tomographic reconstruction problems with slightly mismatched projection angles. Estimating and correcting for model errors represent important yet challenging tasks when solving inverse problems. For problems where the user has strong knowledge about the parameterization of the forward model, there are sophisticated ways of accounting for inexactness in the forward operator, see e.g., [24, 63]. In a learning context, recent approaches to learn implicit and explicit corrections to the operator using neural networks was considered in [56] and learning non-Gaussian models was considered in [68]. However, for many scenarios where a good parameterization does not exist or the goal is not necessarily to determine the model correction itself, we show that it is possible to mitigate inexactness in the forward model as well as resolve any faults in the noise assumptions by determining a better norm for the data-fit term. That is, we use the OID framework to determine a proper choice of the norm in the data-fidelity function, which is purely informed by the availability of training data, to mitigate any effects of inexactness in the forward operator or faults in our noise assumptions.

Learning an optimal value for q in the regularization term is important as well, as this encodes prior knowledge about the solution. The most common choice is Tikhonov regularization ($q = 2$), but for promoting sparsity in the solution, $q = 1$ provides a numerically appealing approximation to the computationally NP-hard $q = 0$ case. More recently, regularization techniques that allow a generic choice of $q > 0$ have been developed [47, 21]. Nevertheless, for such techniques, the choice of a suitable q that accommodates the properties of the desired solution is not always obvious, hence learning q becomes crucial in many applications where its choice can be informed by training data.

Thus, we consider OID problem (8) with $\boldsymbol{\theta} = [\lambda, p, q]$, where the efficiency of the approach relies on the ability to quickly and accurately compute $\mathrm{L}^p - \mathrm{L}^q$ regularized solutions (i.e., solving (8b) with $\mathbf{L}(\boldsymbol{\beta}) = \mathbf{L}$ fixed). This can be challenging, especially in a large-scale setting. Although various optimization methods such as primal-dual gradient descent methods [73, 34, 19] could be used, we consider iterative projection methods, which approximate $\widehat{\mathbf{x}}(\boldsymbol{\theta})$ by solving (8b) in a reduced-dimensional (projected) subspace; such methods can be considered as special instances of MM strategies [49, 53], as summarized below.

Rewriting the s-norm as $\|\mathbf{x}\|_s = \left(\sum_{j=1}^n |x_j|^{s-2} x_j^2\right)^{1/s}$ and by approximating $|x| \approx (x^2 + \varepsilon^2)^{1/2} =: \phi_\varepsilon(x)$ with $\varepsilon > 0$ for $0 < s \leq 1$ and $\varepsilon = 0$ for $s > 1$, we obtain (an approximation of) $\|\mathbf{x}\|_s^s$ by

$$\|\mathbf{x}\|_s^s \approx \sum_{j=1}^n (x_j^2 + \varepsilon^2)^{(s-2)/2}\, x_j^2 = \sum_{j=1}^n \phi_\varepsilon(x_j)^{s-2}\, x_j^2. \tag{9}$$

By defining $\mathbf{S}_{s,\varepsilon}(\mathbf{x})$ as a diagonal matrix dependent on $\mathbf{x}$, with

$$\mathbf{S}_{s,\varepsilon}(\mathbf{x}) = \mathrm{diag}\big(\big[\phi_\varepsilon(x_1)^{s-2}, \ldots, \phi_\varepsilon(x_n)^{s-2}\big]\big)\ , \tag{10}$$

we get

$$\|\mathbf{x}\|_s^s \approx \left\| (\mathbf{S}_{s,\varepsilon}(\mathbf{x}))^{1/2}\, \mathbf{x} \right\|_2^2, \tag{11}$$

where we define the square root elementwise. Hence,

$$\left\|(\mathbf{S}_{p,\varepsilon}(\mathbf{A}\mathbf{x}-\mathbf{b}))^{1/2}\ (\mathbf{A}\mathbf{x}-\mathbf{b})\right\|_2^2+\lambda\left\|(\mathbf{S}_{q,\varepsilon}(\mathbf{L}\mathbf{x}))^{1/2}\ \mathbf{L}\mathbf{x}\right\|_2^2 \tag{12}$$

is a sufficiently smooth approximation of the objective function in (4). Assuming an approximation $\mathbf{x}_k$ to $\mathbf{x}_{\text{true}}$ is available, we consider the *quadratic tangent majorant* of (12) at $\mathbf{x}_k$ (omitting a constant term), i.e.,

$$\mathcal{M}\left(\mathbf{x},\mathbf{x}_k\right)=\left\|\left(\mathbf{S}_{p,\varepsilon}^k\right)^{1/2}(\mathbf{A}\mathbf{x}-\mathbf{b})\right\|_2^2+\lambda\left\|\left(\mathbf{S}_{q,\varepsilon}^k\right)^{1/2}\mathbf{L}\mathbf{x}\right\|_2^2, \tag{13}$$

where we defined

$$\mathbf{S}_{p,\varepsilon}^k=\mathbf{S}_{p,\varepsilon}(\mathbf{A}\mathbf{x}_k-\mathbf{b})\qquad\text{and}\qquad\mathbf{S}_{q,\varepsilon}^k=\mathbf{S}_{q,\varepsilon}(\mathbf{L}\mathbf{x}_k)\,; \tag{14}$$

for details see [47]. Given a point $\mathbf{x}_k$, we compute $\mathbf{x}_{k+1}$ as an approximate solution minimizing (13). This process is iterated to approximate a solution of (4), and it is referred to as MM.

Classical methods for MM (which, in this particular instance, coincide with IRLS methods [10]) involve minimizing (13) by, e.g., applying CGLS, and result in time-consuming inner-outer iterative strategies. Although adaptive tolerances for solving the inner problem can be used to accelerate these methods without significant impact on the solution, see e.g., [64], some hand tuning may be required. Instead, we consider recently developed strategies that bypass classical IRLS schemes and approximate a solution of minimizing (12) by simultaneously computing a new approximation $\mathbf{x}_{k+1}$ and updating the weights $\mathbf{S}_{p,\varepsilon}^{k+1},\mathbf{S}_{q,\varepsilon}^{k+1}$ in (14). These methods involve projections on generalized Krylov subspaces (GKS) [47, 12], and we refer to them as MM-GKS.

Specifically, the GKS-based solver considered here computes $\widehat{\mathbf{x}}(\boldsymbol{\theta})$ starting from an initial approximate solution $\mathbf{x}_0$ belonging to an initial approximation subspace $\text{ran}(\mathbf{V}_0^{\text{GKS}})=\text{ran}(\mathbf{V}_h)$ generated by, e.g., performing $1\le h\ll\min\{m,n\}$ steps of Golub–Kahan bidiagonalization applied to $\mathbf{A}$ with initial vector $\mathbf{b}$. Then, at the $(k+1)$st iteration, one computes the (skinny) QR factorizations,

$$\mathbf{S}_{p,\varepsilon}^k\mathbf{A}\mathbf{V}_{k+1}^{\text{GKS}}=\mathbf{Q}_p\mathbf{R}_p,\quad\mathbf{S}_{q,\varepsilon}^k\mathbf{L}\mathbf{V}_{k+1}^{\text{GKS}}=\mathbf{Q}_q\mathbf{R}_q. \tag{15}$$

where $\mathbf{V}_{k+1}^{\text{GKS}}=[\mathbf{V}_k^{\text{GKS}},\mathbf{v}_{\text{new}}]$ and $\mathbf{v}_{\text{new}}$ is the normalized residual vector $\mathbf{A}^\top(\mathbf{A}\mathbf{x}_k-\mathbf{b})+\lambda\mathbf{L}^\top\mathbf{L}\mathbf{x}_k$. The $(k+1)$st approximate solution reads $\mathbf{x}_{k+1}=\mathbf{V}_{k+1}^{\text{GKS}}\mathbf{y}_{k+1}\in\text{ran}(\mathbf{V}_{k+1}^{\text{GKS}})$, where

$$\mathbf{y}_{k+1}=\underset{\mathbf{y}\in\mathbb{R}^{k+1}}{\arg\min}\ \left\|\mathbf{R}_p\mathbf{y}-\mathbf{Q}_p^\top\left(\mathbf{S}_{p,\varepsilon}^k\right)^{1/2}\mathbf{b}\right\|_2^2+\lambda\|\mathbf{R}_q\mathbf{y}\|_2^2, \tag{16}$$

and where the projected problem is obtained by plugging in the factorizations in (15) into the functional (13). GKS-based solvers can be applied to many instances of (4), provided that matrix-vector products with $\mathbf{L}$ are cheap to compute, and $k\ll\min\{m,n\}$.

3.2. **Learning design-dependent operators. OID with $\boldsymbol{\theta}=[\lambda;\boldsymbol{\beta}]$.** Learning approaches can also be used to estimate hyperparameters for regularization functionals that belong to a parametric family of regularizers (e.g., those defined from a kernel function). We consider OID problem (8) where $p=q=2$ and $\mathbf{L}(\boldsymbol{\beta})$ and its inverse are not readily available, but matrix vector multiplications with $\mathbf{Q}(\boldsymbol{\beta})=(\mathbf{L}(\boldsymbol{\beta})^\top\mathbf{L}(\boldsymbol{\beta}))^{-1}$ can be done efficiently. For example, with Gaussian random fields, the entries of the prior covariance matrix are computed directly as

$\mathbf{Q}_{ij}(\boldsymbol{\beta}) = \kappa(r_{ij};\boldsymbol{\beta})$ where $\kappa(\,\cdot\,;\boldsymbol{\beta})$ is a covariance kernel function that depends on some parameters in $\boldsymbol{\beta}$ and $r_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$, with $\mathbf{z}_i$ corresponding to spatial points in the domain. Although the matrix $\mathbf{Q}(\boldsymbol{\beta})$ may be dense and the inverse or symmetric factorization is not available, matrix-vector multiplications with $\mathbf{Q}(\boldsymbol{\beta})$ can often be done efficiently.

We consider two families of covariance matrices that are built from parameterized kernels: the squared exponential covariance matrix and the Matérn covariance matrix [71]. Given a hyperparameter β that plays the role of the characteristic length-scale, the *squared exponential kernel* is defined as

$$\kappa(r;\beta) = \exp\left(-\frac{r^2}{2\beta^2}\right). \tag{17}$$

Given two hyperparameters β_1 and β_2 that define the smoothness and length scale respectively, the *Matérn kernel* is defined as

$$\kappa(r;\beta_1,\beta_2) = \frac{1}{2^{\beta_1-1}\Gamma(\beta_1)}\left(\frac{\sqrt{2\beta_1}r}{\beta_2}\right)^{\beta_1} K_{\beta_1}\left(\frac{\sqrt{2\beta_1}r}{\beta_2}\right), \tag{18}$$

where $\Gamma(\,\cdot\,)$ is the Gamma function and $K_{\beta_1}(\,\cdot\,)$ is the modified Bessel function of the second kind of order β_1. Note that, oftentimes in the literature, the Matérn parameters are denoted as $\nu = \beta_1$ and $\ell = \beta_2$, where simplifications of the kernel function can be made for half integers $\nu = p + 1/2, p \in \mathbb{N}^+$. We do not impose this constraint here.

In most inverse problems settings, the kernel parameters must be selected *prior* to solving the inverse problem and oftentimes appropriate choices come from expert knowledge. There exist various approaches in Bayesian statistics for estimating hyperparameters for covariance functions (e.g., cross-validation and maximum likelihood) [71]. The process, which is referred to as model selection, seeks to estimate the hyperparameters directly from the data, but these methods can be computationally infeasible, especially for large-scale problems. For inverse problems in imaging, semivariogram methods were considered in [11] for estimating Matérn parameters, but this approach only works for problems where the observation grid and the solution grid are the same (e.g., in deblurring and denoising). We remark that learning approaches that use training data to estimate parameters defining the regularizer have been considered in [37, 4, 22]; however, contrary to existing methods that work with the precision matrix directly, here we consider regularizers that arise in Bayesian approaches and that correspond to prior covariance matrices defined using parametric kernel functions.

We exploit genGK approaches for efficient inner solves (8b) requiring only matrix-vector products with the prior covariance matrix. More specifically, we are interested in solving (1) where $\mathbf{e} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\lambda^{-1}\mathbf{Q}(\boldsymbol{\beta}))$ where $\mathbf{Q}(\boldsymbol{\beta})$ is defined above. By Bayes' formula,

$$\pi(\mathbf{x}|\mathbf{b}) \propto \pi(\mathbf{b}|\mathbf{x})\pi(\mathbf{x}) \propto \exp\left(-\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 - \lambda\mathbf{x}^\top\mathbf{Q}(\boldsymbol{\beta})^{-1}\mathbf{x}\right).$$

The maximum a posteriori approximation of $\mathbf{x}$ can be found by minimizing the negative log-likelihood of $\pi(\mathbf{x}|\mathbf{b})$, i.e.,

$$\widehat{\mathbf{x}}(\boldsymbol{\theta}) = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|^2_{\mathbf{Q}(\boldsymbol{\beta})^{-1}} \tag{19}$$

with $\boldsymbol{\theta} = [\lambda;\boldsymbol{\beta}]$ which, since $\mathbf{Q}(\boldsymbol{\beta}) = (\mathbf{L}(\boldsymbol{\beta})^\top\mathbf{L}(\boldsymbol{\beta}))^{-1}$, is equivalent to (5). In this setting, an iterative projection method based on the genGK bidiagonalization can

be used to approximate (19). After performing a change of variables (to avoid computations with $\mathbf{Q}(\boldsymbol{\beta})^{-1}$), the k-th iteration of the genGK method is given by

$$\widehat{\mathbf{x}}_k(\boldsymbol{\theta}) = \mathbf{Q}(\boldsymbol{\beta})\mathbf{V}_k^{\text{genGK}}\widehat{\mathbf{y}}_k(\boldsymbol{\theta}),$$

$$\text{where } \widehat{\mathbf{y}}_k(\boldsymbol{\theta}) = \underset{\mathbf{y}\in\mathbb{R}^k}{\arg\min} \left\| \mathbf{B}_k^{\text{genGK}}\mathbf{y} - \|\mathbf{b}\|\mathbf{e}_1 \right\|_2^2 + \lambda\|\mathbf{y}\|_2^2 .$$

The matrices above satisfy the partial genGK matrix factorization, i.e.,

$$\mathbf{A}\mathbf{Q}(\boldsymbol{\beta})\mathbf{V}_k^{\text{genGK}} = \mathbf{U}_{k+1}^{\text{genGK}}\mathbf{B}_k^{\text{genGK}},$$

$$\text{with } \mathbf{V}_k^{\text{genGK}} \in \mathbb{R}^{n\times k}, \quad \text{and} \quad \mathbf{B}_k^{\text{genGK}} \in \mathbb{R}^{(k+1)\times k},$$

together with another similar factorization involving $\mathbf{A}^\top$. We refer to [25] for the original derivation.

We conclude this section by mentioning that an upside of all the solvers for (8b) described so far is that λ can be adaptively set during the iterations, i.e., they can be reformulated as so-called "hybrid methods". When solving (8) where λ is a design parameter that is fixed for each instance of the inner problem (3), we will not take advantage of this feature of hybrid methods. However, we may still be able to exploit this feature of hybrid methods, for OID where $\boldsymbol{\theta} = \boldsymbol{\beta}$, and λ is selected automatically. Numerical comparisons will be presented in Section 4.2.

4. Numerical experiments.

In this section, we provide OID examples to show that learned regularization methods perform well for various inverse problems. In Section 4.1, we consider an example from image deblurring, where we learn optimal norms for both the data fit and the regularization term, in addition to an optimal regularization parameter, in order to handle different noise types and to mitigate impacts from an imprecise forward operator. Then, in Section 4.2, we consider an example from tomographic reconstruction, where optimal parameters are found for parametric prior covariance matrices. For all of the experiments, we assess the quality of a reconstructed solution using the Relative Reconstruction Error (RRE) norm defined by $\text{RRE}(\mathbf{x}) = \frac{\|\mathbf{x}-\mathbf{x}_{\text{true}}\|_2}{\|\mathbf{x}_{\text{true}}\|_2}$, for some reconstruction $\mathbf{x}$.

4.1. OID with $\boldsymbol{\theta} = [\lambda, p, q]^\top$.

The goal of this section is to investigate the performance of OID for learning optimal parameters $\lambda, p,\ q$, with $\mathbf{L} = \mathbf{I}$, for image deblurring. For the training and validation datasets, we consider satellite images obtained from the NASA website [2], where each image contains 256×256 pixels. We use 10 images of satellites with 8 random affine transformations, giving a total of 80 training images, and 5 images of satellites with 6 random affine transformations, giving 30 validation images. Samples of the training and validation images are provided in Figure 1.

For the forward model, we consider a blurring process defined by an isotropic Gaussian blur centered at location (χ_1, χ_2), where the point spread function $\mathbf{P}$ has entries

$$[P]_{i,j} = c(\sigma_1, \sigma_2) \exp\left(-\frac{(i-\chi_1)^2}{2\sigma_1^2} - \frac{(j-\chi_2)^2}{2\sigma_2^2} \right), \tag{20}$$

where $c(\sigma_1, \sigma_2)$ is a scaling factor. In the following we use the notation $\mathbf{A} = \mathbf{A}(\sigma_1, \sigma_2)$ to highlight the dependence of the matrix $\mathbf{A}$ on the blurring parameters, and we consider reflective boundary conditions. We mention that other boundary conditions can be used, provided matrix-vector multiplications can be performed efficiently.

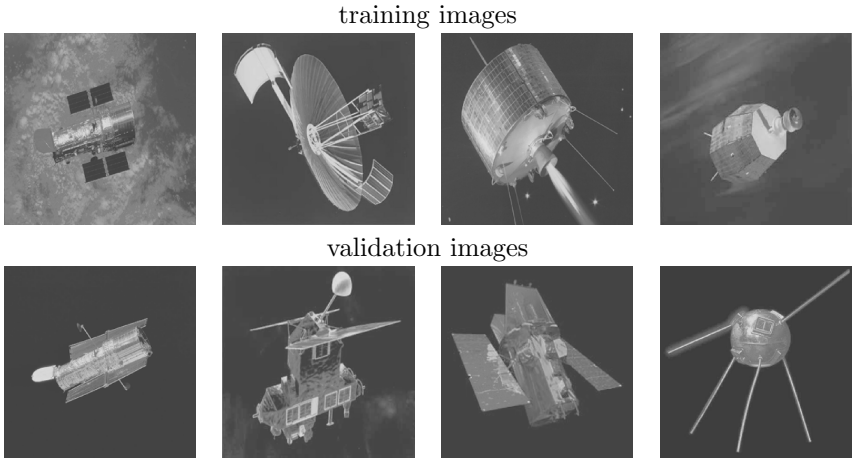

FIGURE 1. Four prototype true images used for generating the training set (top row) and validation set (bottom row) in the OID experiment with $\boldsymbol{\theta} = [\lambda, p, q]^\top$.

The observed image is obtained as in (1), with $\mathbf{A}(2.5, 2.5)$ and $\mathbf{e}$ being impulse noise, with noise level selected uniformly at random between 10% and 50%. More specifically impulse noise is obtained when the entries of the vector $\mathbf{b}$ are constructed as follows

$$\mathbf{b}_i = \begin{cases} (\mathbf{A}\mathbf{x}_{\text{true}})_i & \text{with probability } 1-\eta, \\ u_i & \text{with probability } \eta, \end{cases}$$

where $0 \leq \eta < 1$ denotes the (relative) noise level and u_i is a number chosen randomly in the range of values of $\mathbf{A}\mathbf{x}_{\text{true}}$. Although the images were generated using $\mathbf{A}(2.5, 2.5)$, we consider reconstruction methods that use a different model matrix $\mathbf{A}(2.3, 3.1)$, i.e., we introduce errors in the Gaussian blur parameters to model the realistic situation where the forward operator contains uncertainty and does not match data from the actual model. We emphasize that our approach can incorporate general model errors (not just parameterized blurs), e.g., where the true forward model is a matrix perturbation of the forward model matrix.

Using the OID approaches described in Section 3.1 with MM-GKS solvers for the inner problem, we compute the following optimal parameters.

- First, for fixed p and q we learn λ only. For example, we denote '$\text{OID}_{\lambda,2,2}$' as OID with $\theta = \lambda$ and $p = q = 2$, we denote '$\text{OID}_{\lambda,1,2}$' as OID with $\theta = \lambda$, $p = 1$, and $q = 2$, and we denote '$\text{OID}_{\lambda,2,1}$' as OID with $\theta = \lambda$, $p = 2$, and $q = 1$. The value of the regularization parameters so obtained are 0.4168, 0.0796, and 2.1376, respectively.
- Then we learn the regularization parameter, fit-to-data norm, and regularization norm triplet. We refer to this approach as '$\text{OID}_{\lambda,p,q}$' and we obtained the values of $\widehat{\boldsymbol{\theta}} = [\ \widehat{\lambda},\ \widehat{p},\ \widehat{q}\ ]^\top = [0.0916, 1.0154, 1.9194]^\top$.

All OID approaches use surrogate optimization with a maximum of 200 iterations and with lower and upper bounds of $10^{-8} \leq \lambda \leq 10$ for $\text{OID}_{\lambda,2,2}$ and $\text{OID}_{\lambda,1,2}$, and lower and upper bounds of $10^{-3} \leq \lambda \leq 10^{-1}$, $0.1 \leq p, q \leq 2.15$. For the inner problem, we prescribed 50 iterations of MM-GKS and 100 iterations of CGLS (for

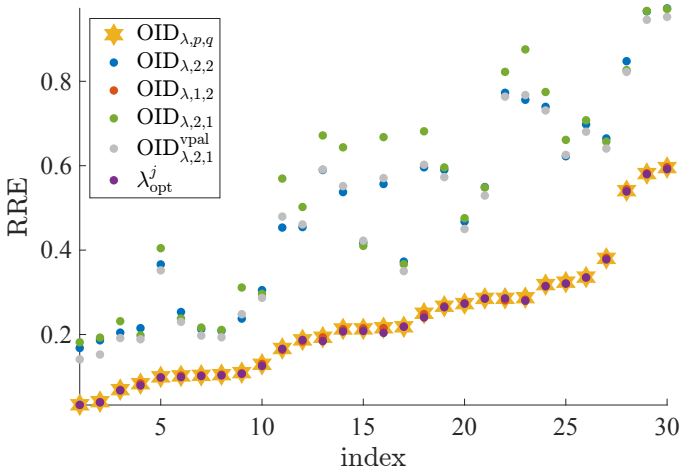

FIGURE 2. For the image deblurring problem with model error and impulse noise, we provide scatter plots of RRE norms for OID$_{\lambda,p,q}$, OID$_{\lambda,2,2}$, OID$_{\lambda,1,2}$, and OID$_{\lambda,2,1}$ (where the inner problem is solving using MM-GKS). Results for OID$^{\text{vpal}}_{\lambda,2,1}$ correspond to using a variable projection augmented Lagrangian method to solve the inner problem. Each column of dots corresponds to one sample from the validation set, where the indices have been sorted based on the RRE norms for OID$_{\lambda,p,q}$. As a further comparison, λ^j_{opt} corresponds to RRE norms for (21), where the optimal regularization parameter is selected for each image using the learned $\widehat{p}$ and $\widehat{q}$.

$p = q = 2$) at both training and validation stages. For all of the results in this section, we use the sample mean to center the data, prior to learning.

Using OID computed parameters, we obtain reconstructions for each of the validation images. In Figure 2 we provide the RRE norms for OID$_{\lambda,p,q}$ (marked by yellow stars) for each validation image, where the index for the validation set has been sorted based on the RRE norms for OID$_{\lambda,p,q}$. RRE norms for OID$_{\lambda,2,2}$ (blue dots), OID$_{\lambda,1,2}$ (red dots), and OID$_{\lambda,2,1}$ (green dots) are also provided for each validation image. Since most of the blue and green dots lie above the yellow stars, we conclude that OID$_{\lambda,p,q}$ consistently performs better than OID$_{\lambda,2,2}$ and OID$_{\lambda,2,1}$, as expected. We also observe that RREs for OID$_{\lambda,1,2}$ are similar to the RREs for OID$_{\lambda,p,q}$, since a good p-norm for the data misfit for impulse noise is expected to be close to the 1-norm.

To verify the use of MM-GKS for solving L^1-regularized problems, we also compare to a method that directly handles an L^1-norm regularizer through a splitting approach. More specifically, we utilize a variable projection augmented Lagrangian (vpal) method [26] that exploits variable projection methods for solving the inner lasso problem directly utilizing shrinkage on the L^1-norm regularization. These results are denoted as OID$^{\text{vpal}}_{\lambda,2,1}$ and correspond to the gray dots in Figure 2. The value of the computed OID regularization parameter was 0.5964.

To investigate the impact of the regularization parameter λ, we also provide results for the OID$_{\lambda,\widehat{p},\widehat{q}}$ method, i.e., OID where the previously computed values for p and q are fixed. Namely, the optimal regularization parameter λ^j_{opt} is computed

for each validation image as

$$\begin{aligned} \lambda^j_{\text{opt}} &= \arg\min_{\lambda} \ \tfrac{1}{2}\left\|\widehat{\mathbf{x}}^j(\lambda) - \mathbf{x}^j_{\text{true}}\right\|_2^2 , \\ \text{where } \ \widehat{\mathbf{x}}^j(\lambda) &= \arg\min_{\mathbf{x}} \ \left\|\mathbf{A}\mathbf{x} - \mathbf{b}^j\right\|_{\widehat{p}}^{\widehat{p}} + \lambda \left\|\mathbf{x}\right\|_{\widehat{q}}^{\widehat{q}} , \end{aligned} \tag{21}$$

with the $\text{OID}_{\lambda,p,q}$ computed values $\widehat{p} = 1.0154$, $\widehat{q} = 1.9194$. RRE values for each validation image are provided as purple dots in Figure 2. As expected, the images reconstructed using the optimal regularization parameter for each validation image have smaller RRE values than the images reconstructed using $\text{OID}_{\lambda,p,q}$ parameters. However, we stress that this approach is not feasible in practice and that the OID results are not far off. Reconstructed images along with RRE values for one validation image are provided in Figure 3. We observe that the $\text{OID}_{\lambda,p,q}$ reconstruction does not contain artifacts that are present in the $\text{OID}_{\lambda,2,2}$ reconstruction, and the reconstruction with the optimal regularization parameter is only slightly better and nearly indistinguishable from the $\text{OID}_{\lambda,p,q}$ reconstruction.

Next, we investigate the impact of model error on the noise and hence the data-fit term, for which it is well-known that the choice of p is directly related to the statistics of the observation error. For one satellite image $\mathbf{x}_{\text{true}}$, we consider the observed image that is generated using $\mathbf{A}(2.5, 2.5)$ and we consider observation errors coming from two sources: 1% additive Gaussian noise and model error by using $\mathbf{A}(5.6, 5.6)$ for reconstructions instead of $\mathbf{A}(2.5, 2.5)$. Images of the additive Gaussian noise, the model error, i.e., $\mathbf{A}(5.6, 5.6)\mathbf{x}_{\text{true}} - \mathbf{A}(2.5, 2.5)\mathbf{x}_{\text{true}}$, and the sum of the two sources of errors are provided in the top row of Figure 4 from left to right. These represent pixel-wise quantities. In the lower frame, we provide a density plot of the combined error, along with the density functions corresponding to $p = 2$ and $p = 1.3835$ (the best p-norm density fit for this image).

We observe that the combined model and measurement error, which resembles a heavy-tailed distribution, results in a noise distribution that is no longer Gaussian (i.e., ignoring the model error and using $p = 2$ may not be appropriate). Indeed, even with additive Gaussian noise, there exists a value for p that better resembles the noise statistics when model error is present. Thus, without additional prior knowledge, changing the norm for the data-fit term (in effect learning the statistics of the combined additive and model error from training data) is a reasonable approach to handle model error.

4.2. **OID for learning kernel parameter(s) and regularization parameter.**

We consider a seismic imaging problem (namely, PRseismic from [43]) that can be modeled as (1), with $\mathbf{x}_{\text{true}}$ containing a smooth image; see Figure 5. $\mathbf{A}$ represents 2D seismic travel-time tomography, using $n_{\text{s}} = 256$ sources located on the right boundary and $n_{\text{r}} = 512$ receivers (seismographs) scattered along the left and top boundaries. The rays are transmitted from each source to each receiver. The noisy observations are provided in Figure 5; here $n = 65{,}536 = 256^2$.

We generated a set of 30 training images, some samples are provided in Figure 6. These were obtained by randomizing the parameters used to define the "smooth" image in IRTools [35]. Then for each training image, we simulated noisy observations using realizations of a Gaussian random vector with $\mathbf{0}$ mean and noise level $\eta = \frac{\|\mathbf{e}\|_2}{\|\mathbf{A}\mathbf{x}_{\text{true}}\|_2}$ uniformly chosen between 10^{-2} and 10^{-1}.

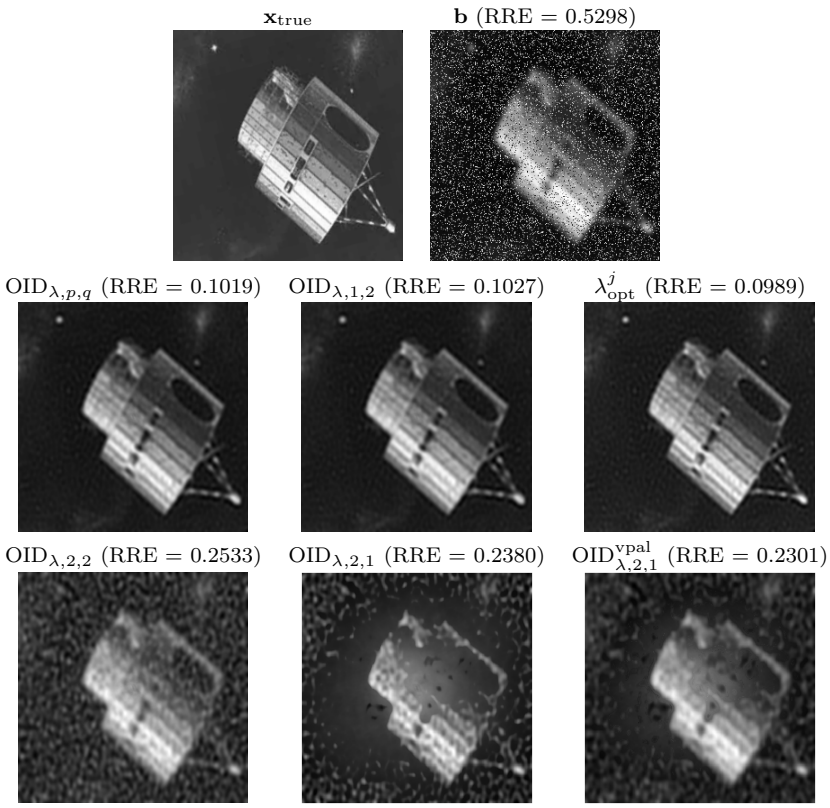

FIGURE 3. For one sample of the validation data set, we provide in the top row the true image and the observed image. In the second row, we provide the $\text{OID}_{\lambda,p,q}$ reconstruction, the $\text{OID}_{\lambda,1,2}$ reconstruction, and the reconstruction computed using the optimal regularization parameter for this image, which is provided for comparison purposes only. In the bottom row are reconstructions for $\text{OID}_{\lambda,2,2}$ and $\text{OID}_{\lambda,2,1}$, and $\text{OID}^{\text{vpal}}_{\lambda,2,1}$. RRE values are provided in the titles.

Using the training data, we consider two kernel functions, the squared exponential kernel function (17) and the Matérn kernel function (18). For each kernel function, we provide OID results for the following two scenarios:

- 'OID' corresponds to solving (8) with $\boldsymbol{\theta} = [\lambda, \boldsymbol{\beta}]$, where the genGK-based iterative projection method described in Section 3.2 is used to solve the inner problem (8b). Here we note that, since λ is being learned in the OID problem, the only stopping criteria used for the inner problem are based on tolerances on the residual norm.
- OID with $\boldsymbol{\theta} = \boldsymbol{\beta}$, where genHyBR is used for selecting λ according to WGCV (weighted generalized cross validation) and the full suite of stopping criteria are used within genHyBR. We refer to this approach as 'OID-wgcv'.

Both OID approaches use surrogate optimization with a maximum of 20 iterations and with lower and upper bounds of $10^{-1} \leq \lambda \leq 100$, $0.5 \leq \beta_1 \leq 15$, and $5 \cdot 10^{-2} \leq$

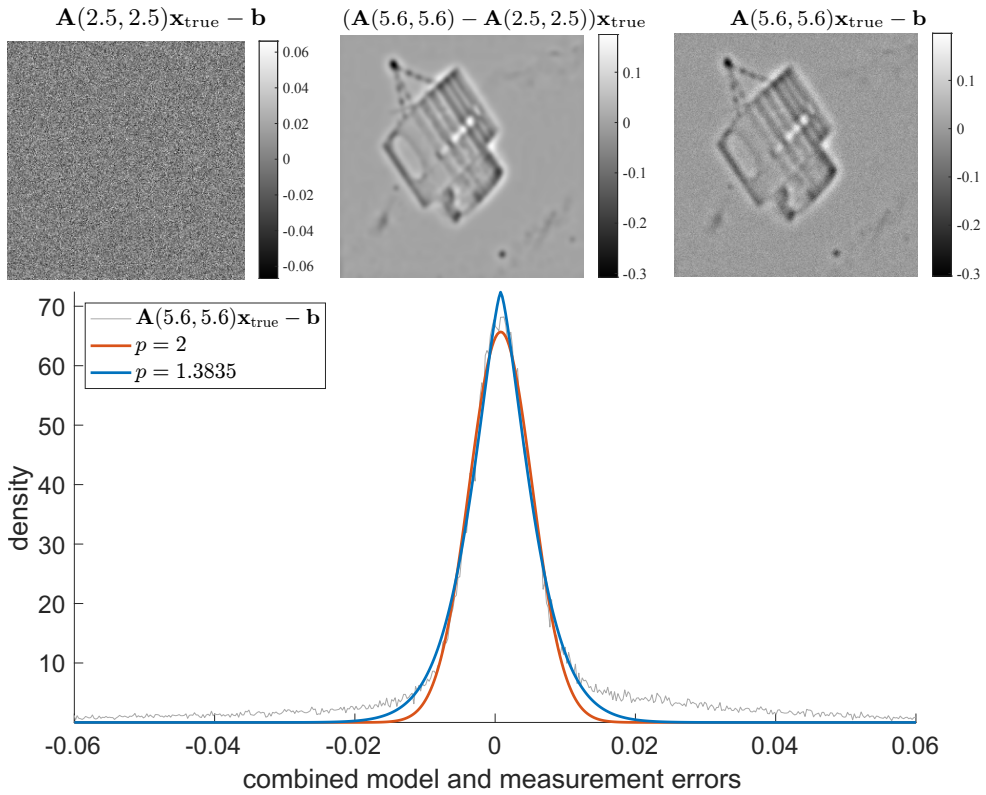

FIGURE 4. Investigating the impact of model error on the overall noise statistics. From left to right in the top row, we provide pixel-wise values of the additive Gaussian noise (of level 1%), the model error, and the sum of these two errors. The density plot for the combined error is provided, along with the density function for $p = 2$ (corresponding to Gaussian noise) and the best density fit to the true errors $\mathbf{A}(5.6, 5.6)\mathbf{x}_{\text{true}} - \mathbf{b}$, i.e., $p = 1.3835$.

$\beta_2 \leq 0.7$ for the Matérn kernel and $10^{-1} \leq \lambda \leq 100$, and $0.01 \leq \beta \leq 0.5$ for the squared exponential kernel. Computed OID parameters are given in Table 1.

Then similar to how we generated the training images, we generated 100 validation images and their corresponding observations. Using the OID parameters in Table 1, we obtained reconstructions for the validation set. The overall mean squared reconstruction error for the validation images is provided in the last column of Table 1. Individually computed RRE norms for OID and OID-wgcv for each validation image are provided as yellow and red dots in Figure 8 respectively, where the indices have been sorted based on the RRE norms for OID. Notice that most of the red dots lie above the yellow dots, and thus OID tends to perform better than OID-wgcv for this example.

For comparison, we use the approach described in [20] that estimates hyper-parameters directly from the the sample covariance matrix constructed from the training data, followed by genHyBR with WGCV for selecting λ. Following [20], let $\widehat{\mathbf{Q}}$ be the sample covariance matrix constructed from the training dataset, then

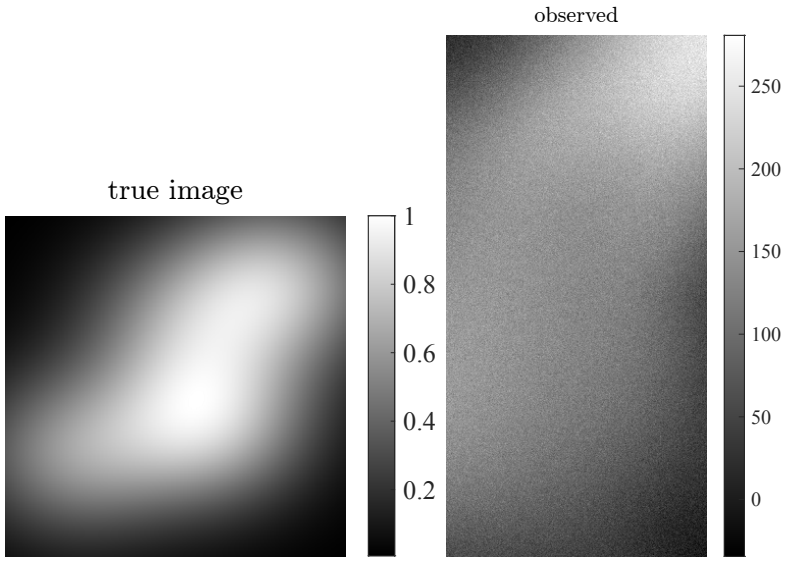

FIGURE 5. Seismic image reconstruction example. The true image (left) contains 256×256 pixels and represents a smooth medium. The noisy sinogram image (right) represents projection data from a setup with 256 sources and 512 receivers.

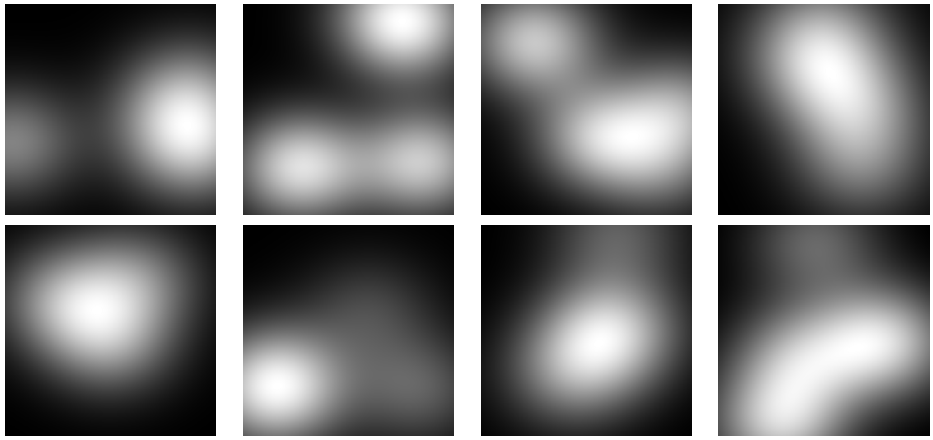

FIGURE 6. Seismic example - random samples from the training set

Matérn parameters were estimated by solving

$$(\hat{\nu}, \hat{\ell}) = \underset{\nu>0,\ell>0}{\arg\min} \left\| \mathbf{Q}(\nu, \ell) - \widehat{\mathbf{Q}} \right\|_{\mathrm{F}}^2, \tag{22}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Once the parameters are computed, they can be used to define $\mathbf{Q} = \mathbf{Q}(\hat{\nu}, \hat{\ell})$, which can be used directly in generalized hybrid methods. For computational feasibility, a stochastic approximation is used for the

objective function in (22), i.e.,

$$\left\|\mathbf{Q}(\nu,\ell)-\widehat{\mathbf{Q}}\right\|_{\mathrm{F}}^2=\mathbb{E}\left\|(\mathbf{Q}(\nu,\ell)-\widehat{\mathbf{Q}})\boldsymbol{\xi}\right\|_2^2, \tag{23}$$

where $\boldsymbol{\xi}$ is a random variable such that $\mathbb{E}\boldsymbol{\xi}=\mathbf{0}$ and $\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}^\top)=\mathbf{I}$. Using a Hutchinson trace estimator, we let $\boldsymbol{\xi}^{(i)}\in\mathbb{R}^n$ for $i=1,2,\ldots,M$ be realizations of a Rademacher distribution (i.e., $\boldsymbol{\xi}$ consists of ±1 with equal probability), and we consider the approximate optimization problem,

$$(\check{\nu},\check{\ell})=\underset{\nu>0,\ell>0}{\arg\min}\,\frac{1}{M}\sum_{i=1}^{M}\|(\mathbf{Q}(\nu,\ell)-\widehat{\mathbf{Q}})\boldsymbol{\xi}^{(i)}\|_2^2\,. \tag{24}$$

Similar to the approach described in [20], we used an interior-point method (`fmincon.m` in MATLAB) to minimize (24) with $M=100$, and refer to this approach as sample covariance (SC). We extend this approach to be used for estimating the squared exponential kernel parameter β and provide the computed hyperparameters in the row labeled 'SC' in Table 1. We remark that this approach uses only the training data and not the model, noise or observations for learning the kernel parameter. On the contrary, OID incorporates this information. Also, for comparison, we provide results for standard Tikhonov regularization ($\mathbf{Q}=\mathbf{I}$) with the optimal λ selected for each sample. The goal of this comparison is to show that including the prior is critical for this example. Scatter plots of RRE values for both approaches are provided in Figure 7. Density graphs of the reconstruction errors are provided in Figure 8, and one reconstructed image is provided in Figure 9.

Matérn	λ	$\boldsymbol{\beta}$	$\mathcal{P}$, validation
OID	18.8313	5.0312, 0.3344	**6.0833**
OID	wgcv	12.2812, 0.3344	29.9412
SC	wgcv	123.1735, 0.2011	49.3447
sq. exp.	λ	β	$\mathcal{P}$, validation
OID	50.0500	0.2550	**3.4468**
OID	wgcv	0.3163	24.7547
SC	wgcv	0.2010	47.6977

TABLE 1. Computed values of the hyperparameters for OID, along with the mean reconstruction errors for the validation set. OID with λ computed using WGCV corresponds to using OID-wgcv for estimating $\boldsymbol{\beta}$ only. 'SC' corresponds to estimating $\boldsymbol{\beta}$ directly from the sample covariance matrix as described in [20] and then using genHyBR with WGCV.

Finally, we investigate the properties of the design objective function (8a) for OID with the squared exponential kernel. In Figure 10, we provide a contour plot of the design objective function, where the white point corresponds to the OID computed values. Notice that there is wide region of values for λ and β that result in small and similar design objective values, with a smaller range of good choices for β.

5. **Conclusions and extensions.** In this work, we have presented a unified framework for optimal inversion design for large-scale inverse problems. We have described learning approaches for computing hyperparameters from training data that

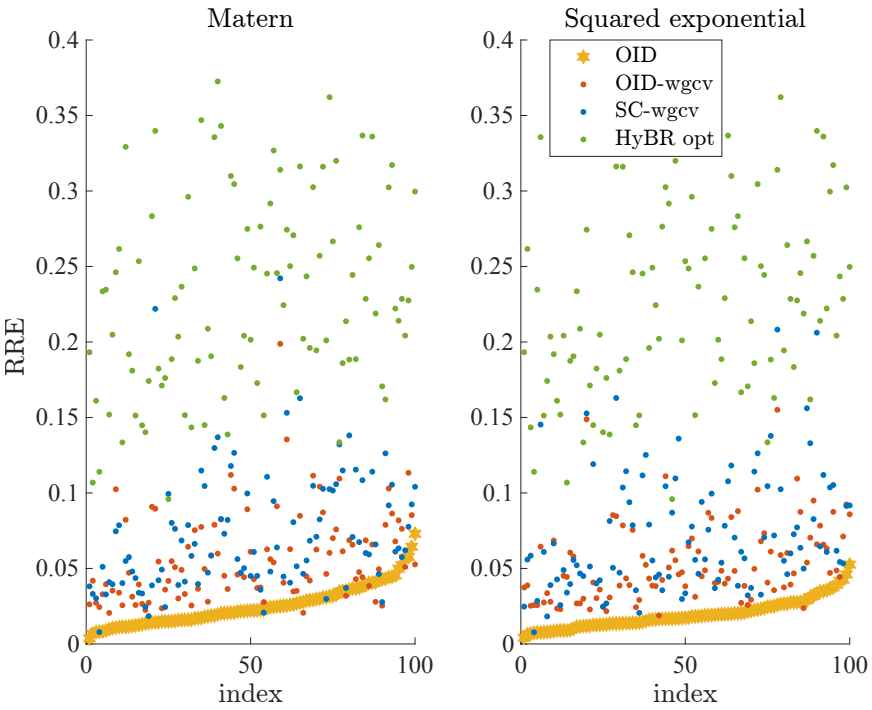

FIGURE 7. For the seismic example, we provide scatter plots of RRE norms for OID, OID-wgcv, SC-wgcv, and HyBR opt. Each column of dots corresponds to one sample from the validation set, where the indices have been sorted based on the RRE norms for OID.

exploit Krylov subspace methods for efficiently solving regularized inner problems within bi-level schema. In particular, we considered OID for learning the norm exponent in the data-fit and regularization term, as well as for learning the regularization parameter. Furthermore, we considered OID for learning parameters for kernel functions used to define prior covariance matrices. Numerical experiments showed that OID methods can compute hyperparameters that deliver quality reconstructions, even in especially relevant scenarios where there is a mixture of noise and model error in the data (e.g., due to the presence of inaccuracies in the forward operator).

We remark that there are other cases where data-driven, optimal inverse frameworks can be used. The focus of our applications is image processing, and more particularly, image deblurring and computerized tomography; nevertheless, the learning approaches that we propose here can be applied to broad applications outside the field of image processing. Moreover, general regularization matrices can be used when considering OID with $\boldsymbol{\theta} = [p, q, \lambda]^\top$, including: discretizations of the derivative operators when solutions with edge preserving properties are desired, wavelet and framelet transformations like in [12, 13, 14, 15, 48] when the solution is sparse in a transformed domain, or fractional Laplacian regularizers where smoothness is determined by a fractional exponent [4]. Furthermore, a wide variety of design criteria can be easily incorporated in this framework, although it is not necessarily the

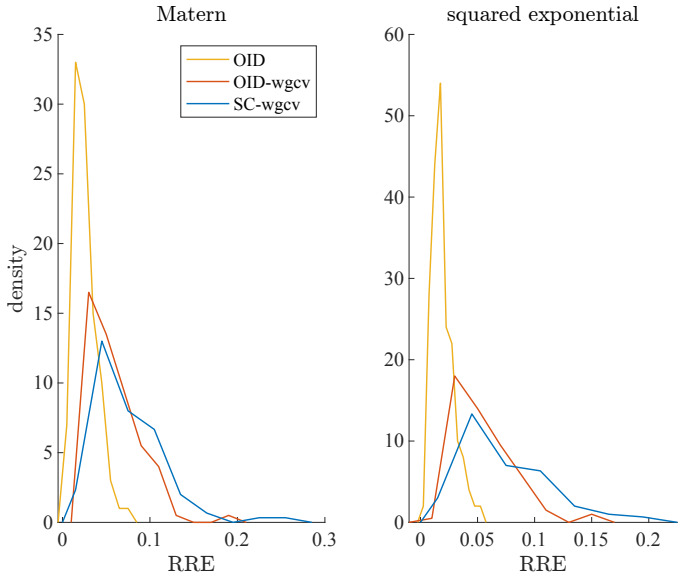

FIGURE 8. For the validation set of the seismic example, we provide histograms of the RRE norms for OID, OID-wgcv, and SC-wgcv.

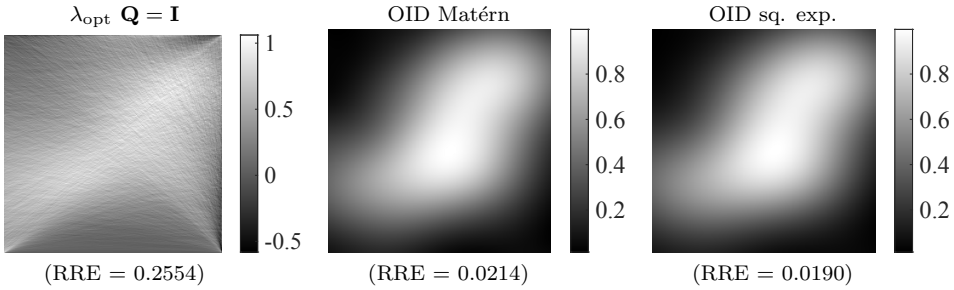

FIGURE 9. For the validation image in Figure 5, we provide reconstructions obtained with HyBR opt and OID reconstructions for the Matérn and squared exponential kernels.

case that changing the design objective results in significant changes to the design parameters. For both numerical examples, we experimented with various design functions, including minimizing the average norm of the errors, $\|\widehat{\mathbf{x}}^j(\boldsymbol{\theta}) - \mathbf{x}^j_{\text{true}}\|_s^2$, (i.e., changing the norm with $s = 2$ in (8a) to $s = 1$ or $s = \infty$), and we observed minor changes to the computed design parameters. Other design objectives such as those leading to min-max optimization problems could also be considered. Other extensions include stochastic approximation methods for problems where the training set is very large and hence empirical Bayes risk minimization problems become computationally intractable. Furthermore, although adaptive schemes have previously been used to accelerate bilevel optimization schemes, additional investigations are needed in the context of surrogate optimization where surrogate models build on all previous function evaluations. These are topics of future investigations.

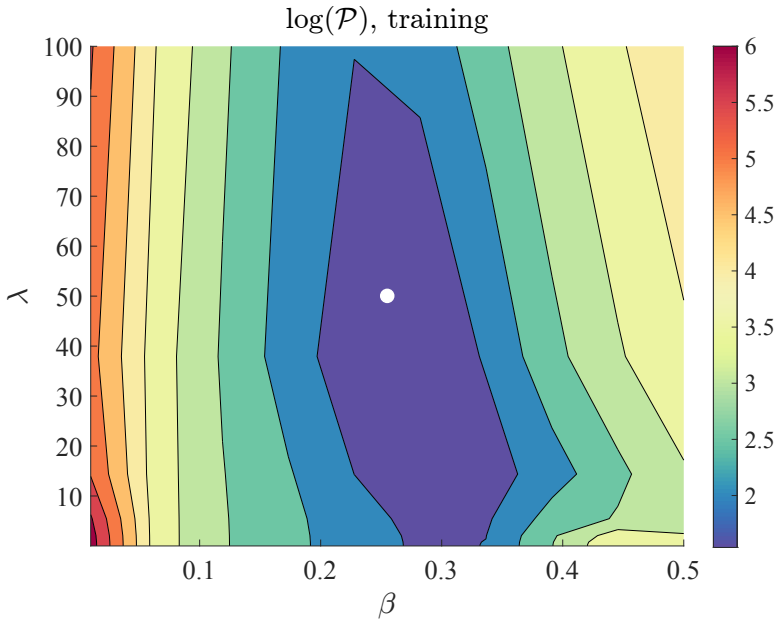

FIGURE 10. Design objective for OID with the squared exponential kernel for the seismic example. The filled contour corresponds to OID, and the white point denotes the OID computed values.

Acknowledgments. This work was initiated as a part of the Statistical and Applied Mathematical Sciences Institute (SAMSI) Program on Numerical Analysis in Data Science in 2020. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF). MP gratefully acknowledges the support from ASU Research Computing facilities for the computing resources used for testing purposes.

REFERENCES

[1] Global optimization toolbox, https://www.mathworks.com/help/gads/index.html?s_tid=CRUX_lftnav, Accessed: 2021-09-20.
[2] Nasa, http:www.nasa.gov.
[3] A. Alexanderian, P. J. Gloor, O. Ghattas et al., On Bayesian A-and D-optimal experimental designs in infinite dimensions, *Bayesian Analysis*, **11** (2016), 671-695.
[4] H. Antil, Z. W. Di and R. Khatri, Bilevel optimization, deep learning and fractional Laplacian regularization with applications in tomography, *Inverse Problems*, **36** (2020), 064001.
[5] F. Archetti and A. Candelieri, The acquisition function, in *Bayesian Optimization and Data Science*, Springer, (2019), 57-72.
[6] S. Arridge, P. Maass, O. Öktem and C.-B. Schönlieb, Solving inverse problems using data-driven models, *Acta Numerica*, **28** (2019), 1-174.
[7] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*, 1992.
[8] J. F. Bard, *Practical Bilevel Optimization: Algorithms and Applications*, Vol. 30, Springer, Berlin, 2013.
[9] J. M. Bardsley and J. G. Nagy, Covariance-preconditioned iterative methods for nonnegatively constrained astronomical imaging, *SIAM Journal on Matrix Analysis and Applications*, **27** (2006), 1184-1197.
[10] A. Björk, *Numerical Methods for Least Squares Problems*, SIAM, 1996.

[11] R. D. Brown, J. M. Bardsley and T. Cui, Semivariogram methods for modeling Whittle–Matérn priors in Bayesian inverse problems, *Inverse Problems*, **36** (2020), 055006.
[12] A. Buccini, M. Pasha and L. Reichel, Modulus-based iterative methods for constrained $\ell_p - \ell_q$ minimization, *Inverse Problems*, **36** (2020), 084001.
[13] A. Buccini, Y. Park and L. Reichel, Numerical aspects of the nonstationary modified linearized Bregman algorithm, *Applied Mathematics and Computation*, **337** (2018), 386-398.
[14] A. Buccini, M. Pasha and L. Reichel, Projected Bregman in Krylov subspaces., *Mathematics of Computation*, **78** (2009), 1515-1536.
[15] J.-F. Cai, S. Osher and Z. Shen, Linearized Bregman iterations for frame-based image deblurring, *SIAM Journal on Imaging Sciences*, **2** (2009), 226-252.
[16] L. Calatroni, C. Cao, J. C. De Los Reyes, C.-B. Schönlieb and T. Valkonen, Bilevel approaches for learning of variational imaging models, *Variational Methods: In Imaging and Geometric Control*, **18** (2017), 2.
[17] L. Calatroni, J. C. De Los Reyes and C.-B. Schönlieb, Infimal convolution of data discrepancies for mixed noise removal, *SIAM Journal on Imaging Sciences*, **10** (2017), 1196-1233.
[18] D. Calvetti and E. Somersalo, *An Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*, Vol. 2, Springer Science & Business Media, 2007.
[19] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision*, **40** (2011), 120-145.
[20] T. Cho, J. Chung and J. Jiang, Hybrid projection methods for large-scale inverse problems with mixed Gaussian priors, *Inverse Problems*, **37** (2021), 044002.
[21] J. Chung and S. Gazzola, Flexible Krylov methods for ℓ_p regularization, *SIAM Journal on Scientific Computing*, **41** (2019), S149-S171.
[22] J. Chung, M. Chung and D. P. O'Leary, Designing optimal spectral filters for inverse problems, *SIAM Journal on Scientific Computing*, **33** (2011), 3132-3152.
[23] J. Chung and M. I. Español, Learning regularization parameters for general-form Tikhonov, *Inverse Problems*, **33** (2017), 074004.
[24] J. Chung and J. G. Nagy, An efficient iterative approach for large-scale separable nonlinear inverse problems, *SIAM Journal on Scientific Computing*, **31** (2010), 4654-4674.
[25] J. Chung and A. K. Saibaba, Generalized hybrid iterative methods for large-scale Bayesian inverse problems, *SIAM Journal on Scientific Computing*, **39** (2017), S24-S46.
[26] M. Chung and R. Renaut, The variable projected augmented Lagrangian method, `arXiv:2207.08216`.
[27] J. C. De los Reyes, C.-B. Schönlieb and T. Valkonen, Bilevel parameter learning for higher-order total variation regularisation models, *Journal of Mathematical Imaging and Vision*, **57** (2017), 1-25.
[28] J. C. De los Reyes and C.-B. Schönlieb, Image denoising: learning the noise model via nonsmooth PDE-constrained optimization, *Inverse Problems & Imaging*, **7** (2013), 1183-1214.
[29] J. C. De los Reyes and D. Villacís, *Bilevel Optimization Methods in Imaging*, Springer International Publishing, 2022.
[30] S. Dempe, *Foundations of Bilevel Programming*, Kluwer Academic Publishers, New York, 2002.
[31] S. Dempe, V. Kalashnikov, G. A. Pérez-Valdés and N. Kalashnykova, *Bilevel Programming Problems*, Springer, Berlin, 2015.
[32] M. M. Dunlop, T. Helin and A. M. Stuart, Hyperparameter estimation in Bayesian MAP estimation: Parameterizations and consistency, `arXiv:1905.04365`.
[33] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Vol. 375, Springer Science & Business Media, 1996.
[34] E. Esser, X. Zhang and T. F. Chan, A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science, *SIAM Journal on Imaging Sciences*, **3** (2010), 1015-1046.
[35] S. Gazzola, P. C. Hansen and J. G. Nagy, Ir tools: a matlab package of iterative regularization methods and large-scale test problems, *Numerical Algorithms*, **81** (2019), 773-811.
[36] R. B. Gramacy, *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Chapman and Hall/CRC, 2020.
[37] E. Haber and L. Tenorio, Learning regularization functionals-a supervised training approach, *Inverse Problems*, **19** (2003), 611.
[38] E. Haber, L. Horesh and L. Tenorio, Numerical methods for experimental design of large-scale linear ill-posed inverse problems, *Inverse Problems*, **24** (2008), 055012.

[39] E. Haber, L. Horesh and L. Tenorio, Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems, *Inverse Problems*, **26** (2010), 025002.
[40] M. Haltmeier and L. V. Nguyen, Regularization of inverse problems by neural networks, arXiv:2006.03972.
[41] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock and F. Knoll, Learning a variational network for reconstruction of accelerated MRI data, *Magnetic Resonance in Medicine*, **79** (2018), 3055-3071.
[42] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*, SIAM, 2010.
[43] P. C. Hansen and J. S. Jørgensen, Air tools ii: algebraic iterative reconstruction methods, improved implementation, *Numerical Algorithms*, **79** (2018), 107-137.
[44] M. Hintermüller, C. N. Rautenberg, T. Wu and A. Langer, Optimal selection of the regularization function in a weighted total variation model. part II: Algorithm, its analysis and numerical tests, *Journal of Mathematical Imaging and Vision*, **59** (2017), 515-533.
[45] G. Holler, K. Kunisch and R. C. Barnard, A bilevel approach for parameter learning in inverse problems, *Inverse Problems*, **34** (2018), 115012.
[46] X. Huan and Y. Marzouk, Gradient-based stochastic optimization methods in bayesian experimental design, *International Journal for Uncertainty Quantification*, **4** (2014), 479-510.
[47] G. Huang, A. Lanza, S. Morigi, L. Reichel and F. Sgallari, Majorization–minimization generalized krylov subspace methods for $\ell_p - \ell_q$ optimization applied to image restoration, *BIT Numerical Mathematics*, **57** (2017), 351-378.
[48] J. Huang, M. Donatelli and R. H. Chan, Nonstationary iterated thresholding algorithms for image deblurring, *Inverse Problems & Imaging*, **7** (2013), 717-736.
[49] D. R. Hunter and K. Lange, A tutorial on MM algorithms, *The American Statistician*, **58** (2004), 30-37.
[50] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Vol. 160, Springer Science & Business Media, 2006.
[51] J. P. Kaipio, V. Kolehmainen, M. Vauhkonen and E. Somersalo, Inverse problems with structural prior information, *Inverse Problems*, **15** (1999), 713.
[52] M. Kubínová and J. G. Nagy, Robust regression for mixed Poisson–Gaussian model, *Numerical Algorithms*, **79** (2018), 825-851.
[53] K. Lange, *MM Optimization Algorithms*, SIAM, 2016.
[54] H. Li, J. Schwab, S. Antholzer and M. Haltmeier, NETT: Solving inverse problems with deep neural networks, *Inverse Problems*, **36** (2020), 065005.
[55] A. Lucas, M. Iliadis, R. Molina and A. K. Katsaggelos, Using deep neural networks for inverse problems in imaging: beyond analytical methods, *IEEE Signal Processing Magazine*, **35** (2018), 20-36.
[56] S. Lunz, A. Hauptmann, T. Tarvainen, C.-B. Schönlieb and S. Arridge, On learned operator correction in inverse problems, *SIAM Journal on Imaging Sciences*, **14** (2021), 92-127.
[57] S. Lunz, O. Öktem and C.-B. Schönlieb, Adversarial regularizers in inverse problems, *Advances in Neural Information Processing Systems*, **31** (2018), 8507-8516.
[58] M. T. McCann, K. H. Jin and M. Unser, Convolutional neural networks for inverse problems in imaging: A review, *IEEE Signal Processing Magazine*, **34** (2017), 85-95.
[59] G. Ongie, A. Jalal, C. A. M. R. G. Baraniuk, A. G. Dimakis and R. Willett, Deep learning techniques for inverse problems in imaging, *IEEE Journal on Selected Areas in Information Theory*, **1** (2020), 39-56.
[60] M. A. Osborne, R. Garnett and S. J. Roberts, Gaussian processes for global optimization, in 3^{rd} *International Conference on Learning and Intelligent Optimization (LION3)*, (2009), 1-15.
[61] J. Prost, A. Houdard, A. Almansa and N. Papadakis, Learning local regularization for variational image restoration, arXiv:2102.06155.
[62] F. Pukelsheim, *Optimal Design of Experiments*, SIAM, 2006.
[63] N. A. B. Riis, Y. Dong and P. C. Hansen, Computed tomography with view angle estimation using uncertainty quantification, *Inverse Problems*, **37** (2021), 065007.
[64] P. Rodríguez and B. Wohlberg, Efficient minimization method for a generalized total variation functional, *IEEE Transactions on Image Processing*, **18** (2008), 322-332.
[65] L. Ruthotto, J. Chung and M. Chung, Optimal experimental design for inverse problems with state constraints, *SIAM Journal on Scientific Computing*, **40** (2018), B1080-B1100.

[66] F. Sherry, M. Benning, J. C. De los Reyes, M. J. Graves, G. Maierhofer, G. Williams, C.-B. Schönlieb and M. J. Ehrhardt, Learning the sampling pattern for MRI, *IEEE Transactions on Medical Imaging*, **39** (2020), 4310-4321.
[67] A. Sinha, P. Malo and K. Deb, A review on bilevel optimization: from classical to evolutionary approaches and applications, *IEEE Transactions on Evolutionary Computation*, **22** (2017), 276-295.
[68] D. Smyl, T. N. Tallman, J. A. Black, A. Hauptmann and D. Liu, Learning and correcting non-Gaussian model errors, *Journal of Computational Physics*, **432** (2021), 110152.
[69] M. Urquhart, E. Ljungskog and S. Sebben, Surrogate-based optimisation using adaptively scaled radial basis functions, *Applied Soft Computing*, **88** (2020), 106050.
[70] T. Weise, Global optimization algorithms-theory and application, Self-Published Thomas Weise.
[71] C. K. Williams, *Gaussian Processes for Machine Learning*, Vol. 2, MIT Press, 2006.
[72] K. Zhang, W. Zuo, S. Gu and L. Zhang, Learning deep CNN denoiser prior for image restoration, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 3929-3938.
[73] M. Zhu and T. Chan, An efficient primal-dual hybrid gradient algorithm for total variation image restoration, UCLA CAM Report, **34** (2008), 8-34.

Received October 2021; 1^{st} revision August 2022; Final revision October 2022; Early access December 2022.