SIAM J. IMAGING SCIENCES
Vol. 14, No. 1, pp. 92–127

On Learned Operator Correction in Inverse Problems*

Sebastian Lunz†, Andreas Hauptmann‡, Tanja Tarvainen§, Carola-Bibiane Schönlieb†, and Simon Arridge¶

Abstract. We discuss the possibility of learning a data-driven explicit model correction for inverse problems and whether such a model correction can be used within a variational framework to obtain regularized reconstructions. This paper discusses the conceptual difficulty of learning such a forward model correction and proceeds to present a possible solution as a forward-adjoint correction that explicitly corrects in both data and solution spaces. We then derive conditions under which solutions to the variational problem with a learned correction converge to solutions obtained with the correct operator. The proposed approach is evaluated on an application to limited view photoacoustic tomography and compared to the established framework of the Bayesian approximation error method.

Key words. model correction, inverse problems, operator learning, deep learning, variational methods, photoacoustic tomography

AMS subject classifications. 65K10, 65F22, 94A08, 47A52

DOI. 10.1137/20M1338460

1. Introduction.

In inverse problems it is usually considered imperative to have an accurate forward model of the underlying physics. Nevertheless, such accurate models can be computationally highly expensive due to possible nonlinearities, large spatial and temporal dimensions, as well as stochasticity. Thus, in many applications approximate models are used in order to speed up reconstruction times and to comply with hardware and cost restrictions. As a consequence the introduced approximation errors need to be taken into account

*Received by the editors May 15, 2020; accepted for publication (in revised form) October 22, 2020; published electronically January 26, 2021.
https://doi.org/10.1137/20M1338460

Funding: The work of the authors was partially supported by the Academy of Finland projects 312123, 312342 (Finnish Centre of Excellence in Inverse Modeling and Imaging, 2018-2025), 334817, 314411, the Jane and Aatos Erkko Foundation, the British Heart Foundation grant NH/18/1/33511, the CMIC-EPSRC platform grant (EP/M020533/1), and the EPSRC-Wellcome grant WT101957. The work of the first author was supported by the EPSRC grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis, and the Cantab Capital Institute for the Mathematics of Information. The work of the fourth author was supported by the Leverhulme Trust project "Breaking the non-convexity barrier," the Philip Leverhulme Prize, the EPSRC grants EP/S026045/1, EP/T003553/1, the EPSRC Centre grant EP/N014588/1, the Wellcome Innovator Award RG98755, the RISE projects CHiPS and NoMADS, the Cantab Capital Institute for the Mathematics of Information, and the Alan Turing Institute. The work of the fifth author was supported by the EPSRC grants EP/N022750/1, EP/T000864/1.

†University of Cambridge, Department of Applied Mathematics and Theoretical Physics, Cambridge, (lunz@math.cam.ac.uk, cbs31@cam.ac.uk).
‡University of Oulu, Research Unit of Mathematical Sciences; University College London, Department of Computer Science, London (Andreas.Hauptmann@oulu.fi).
§University of Eastern Finland, Department of Applied Physics, Kuopio; University College London, Department of Computer Science, London (tanja.tarvainen@uef.fi).
¶University College London, Department of Computer Science, London (s.arridge@ucl.ac.uk).

when solving ill-posed inverse problems or a degradation of the reconstruction quality can be expected.

For instance, in classical computerized tomography with a relatively high dose, models based on ray transforms are sufficiently accurate for the reconstruction task, whereas the full physical model would incorporate stochastic X-ray scattering events. Nevertheless, in some cone beam computerized tomography applications the dose is typically relatively low with a large field of view and hence scattering becomes more prevalent [38] and simple models based on the ray transform are not enough to guarantee sufficient image quality. However, as these scattering events are stochastic, accurate models would be too expensive for practical image reconstruction. Therefore, the basic model is used as an approximation with an appropriate correction that accounts for the full physical phenomena [47].

In applications where the forward model is given by the solution of a partial differential equation, model reduction techniques are often used to reduce computational costs [8, 14, 39]. Such reductions lead to known approximation errors in the model and can be corrected for by explicit modeling [4, 23]. Recently, with the possibility of combining deep learning techniques with classical variational methods, approximate models are now also used in the framework of learned image reconstruction [20]. In this case, the approximate model is embedded in an iterative scheme and updates are performed by a convolutional neural network (CNN). Here, model correction is performed *implicitly* by the network while computing the iterative updates.

In this paper we investigate the possibility of correcting such approximation errors *explicitly* with data-driven methods, in particular, using a CNN. In what follows, we restrict ourselves to linear inverse problems, with both theory and experiments considering the linear case only. However, we expect many of the challenges and approaches discussed here to be relevant and to give insight into the nonlinear case as well. Let $x \in X$ be the unknown quantity of interest we aim to reconstruct from measurements $y \in Y$, where X, Y are Hilbert spaces and x and y fulfil the relation

$$Ax = y, \tag{1.1}$$

where $A : X \to Y$ is the accurate forward operator modeling the underlying physics sufficiently accurately for any systematic error to be well below the noise level of the acquisition. We assume that the evaluation of accurately operator A is computationally expensive and we rather want to use an approximate model $\widetilde{A} : X \to Y$ to compute x from y. In doing so, we introduce an inherent approximation error in (1.1) and have

$$\widetilde{A}x = \widetilde{y} \tag{1.2}$$

leading to a systematic model error

$$\delta y = y - \widetilde{y}. \tag{1.3}$$

Remark 1.1. In general, the range and domain of $\widetilde{A}$ might be different than those of A. To simplify the remainder of this paper we assume, unless otherwise stated, that appropriate projections between the range and domain of the approximate operator $\widetilde{A}$ as well as the range and domain of the accurate operator A are included in the implementation of $\widetilde{A}$, so that expressions such as (1.3) are well-defined.

In this work, we consider corrections for this approximation error via a parameterizable, possibly nonlinear, mapping $F_\Theta : Y \to Y$, applied as a correction to $\widetilde{A}$. This leads to a corrected operator A_Θ of the form

$$A_\Theta = F_\Theta \circ \widetilde{A}. \tag{1.4}$$

We aim to choose the correction F_Θ such that ideally $A_\Theta(x) \approx Ax$ for some $x \in X$ of interest. Restricting the corrected operator A_Θ to be a composition of the approximate operator $\widetilde{A}$ and a parameterizable correction yields various advantages compared to fully parameterizing the corrected operator $A_\Theta : X \to Y$, without utilizing the knowledge of $\widetilde{A}$. It avoids having to model the typically global dependencies of A in the learned correction and allows us to employ generic network architectures for F_Θ, such as the popular U-Net [34].

The primary question that we aim to answer is, whether such corrected models (1.4) can be subsequently used in variational regularization approaches that find a reconstruction x^* as

$$x^* = \arg\min_{x \in X} \frac{1}{2} \|A_\Theta(x) - y\|_Y^2 + \lambda R(x) \tag{1.5}$$

with regularization functional R and associated hyperparameter λ. Apart from investigating the practical performance of (1.5), we will discuss conditions on the model correction that need to be satisfied to guarantee convergence of solutions to (1.5) to the accurate solution as the corrected operator A_Θ approaches the accurate operator A. We provide theoretical results, which show that variational regularization strategies can be applied under certain conditions. In particular, as we will discuss in this study, while it is fairly easy to learn a model correction that fulfils (1.4), it cannot be readily guaranteed to yield high-quality reconstructions when used within the variational problem (1.5). This is a conceptual difficulty caused by a possible discrepancy in the range of the adjoints of A and $\widetilde{A}$ that can be an inherent part of the approximate model and hence first order methods to solve (1.5) yield nondesirable results.

To overcome this restriction, we introduce a forward-adjoint correction that combines an explicit forward model correction with an explicit correction of the adjoint. We will show that such a forward-adjoint correction—if trained sufficiently well—provides a descent direction for a gradient scheme to solve (1.5) for which we can guarantee convergence to a neighborhood of the solution obtained with the accurate operator A.

This work fits into the wider field of learned image reconstruction techniques that have sparked large interest in recent years [5, 22, 25]. In particular, we are motivated by model-based learned iterative reconstruction techniques that have shown to be highly successful in a variety of application areas [1, 2, 17, 21, 36]. These methods generally mimic iterative gradient descent schemes and demonstrate impressive reconstruction results with often considerable speedups [18], but are mostly empirically motivated and lack convergence guarantees. In contrast, this paper follows a recent development in understanding how deep learning methods can be combined with classical reconstruction algorithms, such as variational techniques, to retain established theoretical results on convergence. Whereas most studies are concentrated on learning a regularizer [27, 31, 33, 37], we concentrate here on the operator only and keep

Table 1

Symbol	Description	Definition
X	Reconstruction space	Hilbert space, norm $\|\cdot\|_X$, product $\langle\cdot,\cdot\rangle_X$
Y	Measurement space	Hilbert space, norm $\|\cdot\|_Y$, product $\langle\cdot,\cdot\rangle_Y$
A	Exact forward operator	$A : X \to Y$
$\widetilde{A}$	Approximate forward operator	$\widetilde{A} : X \to Y$
F_Θ	Parameterizable correction in Y	$F_\Theta : Y \to Y$
G_Φ	Parameterizable correction in X	$G_\Phi : X \to X$
A_Θ	Corrected forward operator	$A_\Theta : X \to Y$, $A_\Theta = F_\Theta \circ \widetilde{A}$
A^*_Φ	Corrected adjoint	$A^*_\Phi : Y \to X$, $A^*_\Phi = G_\Phi \circ \widetilde{A}^*$
$Df(t)$	Fréchet derivative of f at t	$Df(t) : \mathbf{dom}(f) \to \mathbf{rng}(f)$
		$f(t+\delta t) = f(t) + Df(t)\delta t + \mathcal{O}(\delta t^2)$
R	Regularization functional	$R : X \to \mathbb{R}_+$
$\mathcal{L}$	Variational functional with A	$\mathcal{L}(x) = \frac{1}{2}\|Ax - y\|^2_Y + \lambda R(x)$
$\mathcal{L}_\Theta$	Variational functional with A_Θ	$\mathcal{L}_\Theta(x) = \frac{1}{2}\|A_\Theta(x) - y\|^2_Y + \lambda R(x)$

a fixed, analytical form for the regularizer. Further, related works that consider learned corrections by utilising explicit knowledge of the operator range are [7, 9, 37]. Another line of research examines the incorporation of imperfectly known forward operators into a fully variational model [10, 29] as well as perturbations in [13, 32]. We note also the connection to the concept of calibration in a Bayesian setting [26].

This paper is organized as follows. In section 2, we introduce the concept of model correction and compare it to previous work in the field. In section 3, we discuss forward corrections and demonstrate their limitations. To overcome these limitations, we introduce the forward-adjoint corrections in section 4, where we also present convergence results for this correction. This is followed by a discussion of computational challenges and the experimental setup in section 5. Finally, in section 6, we demonstrate the performance of the discussed approaches on two data sets for limited view photoacoustic tomography.

Glossary. To improve readability throughout the paper we provide a glossary (see Table 1) with the definition of frequently used notation.

2. Learning a model correction.

As we have motivated above, we only consider an explicit model correction (1.4) in this study and leave the regularization term untouched. Therefore, we will discuss in the following how a model correction using data-driven methods is possible and what the main challenges are.

Before we turn to the discussion of an *explicit* correction, it is important to make the distinction from an *implicit* correction in the framework of learned iterative reconstructions. In particular, we concentrate here on learned gradient schemes [1], which can be formulated by a network Λ_Θ, that is designed to mimic a gradient descent step. In particular, we train the networks to perform an iterative update, such that

$$x_{k+1} = \Lambda_\Theta\left(\nabla_x \frac{1}{2}\|Ax_k - y\|^2_Y, x_k\right), \tag{2.1}$$

where $\nabla_x \frac{1}{2}\|Ax_k - y\|^2_Y = A^*(Ax_k - y)$. Now, one could use an approximate model instead of the accurate model and compute an approximate gradient given by $\widetilde{A}^*(\widetilde{A}x_k - y)$ for the update in (2.1), as proposed in [20]. The network Λ_Θ then implicitly corrects the model error to produce the new iterate. That means, the correction and a prior are, hence, trained simultaneously with the update in (2.1). Such approaches are typically trained by using a loss function, like the L^2-loss, to measure the distance between reconstruction and a ground-truth phantom.

On the other hand, in the explicit approach that we pursue here, we aim to learn a correction A_Θ that is independent of the regularization use. It can hence be trained using knowledge of the accurate and approximate operator alongside training data in either X or Y, without requiring pairs of measurements and their corresponding ground-truth phantoms. In a scenario where the operators cannot been accessed directly, samples of pairs from the two operators can even be sufficient to fit an explicit operator correction. While implicit methods have been shown to perform well in practice [20], our approach will yield an explicit correction and as such can be used in combination with any regularization functional and builds on the established variational framework. Furthermore, we note that the study of explicit methods also allows one to uncover and investigate some of the fundamental challenges of model correction that might easily be left ignored in implicit approaches.

Thus, we will concentrate our discussion in the following on how an explicit data correction can be achieved, how the correction of the model $\widetilde{A}$ can be parameterized by a neural network, and how this can be incorporated into a variational framework.

2.1. Approximation error method (AEM). A well-established approach to incorporate model correction into a reconstruction framework, such as (1.5), is given by Bayesian approximation error modeling [23, 24]. Let us shortly recall, that in Bayesian inversion we want to determine the posterior distribution of the unknown x given y, and by Bayes' formula we obtain

$$p(x|y) = p(y|x)\frac{p(x)}{p(y)}. \tag{2.2}$$

Thus, the posterior distribution is characterized by the likelihood $p(y|x)$ and the chosen prior $p(x)$ on the unknown. Typically, the likelihood $p(y|x)$ is modeled using accurate knowledge of the forward operator $A : X \to Y$ as well as the noise model. In the AEM, the purpose is now to adjust the likelihood by examining the difference between the (accurate) forward operator A and its approximation $\widetilde{A}$ of the model (1.1)–(1.2) as

$$\varepsilon = \delta y = Ax - \widetilde{A}x. \tag{2.3}$$

Including an additive model for the measurement noise e, this leads to an observation model

$$y = \widetilde{A}x + \varepsilon + e. \tag{2.4}$$

We model the noise e independently of x as Gaussian $e \sim \mathcal{N}(\eta_e, \Gamma_e)$, where η_e and Γ_e are the mean and covariance of the noise. Further, the model error ε is approximated as Gaussian $\varepsilon \sim \mathcal{N}(\eta_\varepsilon, \Gamma_\varepsilon)$ and is modeled independently of noise e and unknown parameters x leading to a Gaussian distributed total error $n = \varepsilon + e$, $n \sim \mathcal{N}(\eta_n, \Gamma_n)$, where η_ε and η_n are means and

Γ_ε and Γ_n are the covariance matrices of model error and total errors, respectively. This leads to a so-called enhanced error model [23] with a likelihood distribution of the form

$$p(y|x) \sim \exp\left(-\frac{1}{2}\|L_n(\widetilde{A}x - y + \eta_n)\|_Y^2\right),$$

where $L_n^{\mathrm{T}} L_n = \Gamma_n^{-1}$ is a matrix square root such as the Cholesky decomposition of the inverse covariance matrix of the total error. In the case of Gaussian white noise with a zero mean and a constant standard deviation σ, this can be written as

$$p(y|x) \sim \exp\left(-\frac{1}{2\sigma}\|L_\varepsilon(\widetilde{A}x - y + \eta_\varepsilon)\|_Y^2\right),$$

where $L_\varepsilon^{\mathrm{T}} L_\varepsilon = \Gamma_\varepsilon^{-1}$. This could be used to motivate writing the variational problem (1.5) in the form

$$x^* = \arg\min_{x \in X} \frac{1}{2}\|L_\varepsilon(\widetilde{A}x - y + \eta_\varepsilon)\|_Y^2 + \lambda R(x). \tag{2.5}$$

In order to utilize the approach, the unknown distribution of the model error needs to be approximated. That can be obtained, for example, by simulations [4, 41] as follows. Let $\{x^i, i = 1, \ldots, N\}$ be a set of samples drawn from a training set. The corresponding samples of the model error are then

$$\varepsilon^i = Ax^i - \widetilde{A}x^i \tag{2.6}$$

and the mean and covariance of the model error can be estimated from the samples as

$$\eta_\varepsilon = \frac{1}{N}\sum_{i=1}^{N} \varepsilon^i, \tag{2.7}$$

$$\Gamma_\varepsilon = \frac{1}{N-1}\sum_{i=1}^{N} \varepsilon^i(\varepsilon^i)^{\mathrm{T}} - \eta_\varepsilon \eta_\varepsilon^{\mathrm{T}}. \tag{2.8}$$

2.2. Learning a general model correction. The classical Bayesian AEM provides an affine linear correction of the likelihood in (2.5) and by construction is limited to cases where the error between accurate and approximate models (2.3) can be approximated as normally distributed. As this can be too restrictive in certain cases to describe more complicated errors, we will now address a more general concept of learning a nonlinear explicit model correction.

That is, given an accurate underlying forward model A, we aim to find a (partially) learned operator A_Θ which we consider as an explicitly corrected approximate model of the form (1.4). To do so, we need to set a notion of distance between A and A_Θ in order to assess the quality of the approximation. A seemingly natural notion of distance between two operators would be the supremum norm over elements in X, that is, we consider here

$$\|A - A_\Theta\|_{X \to Y} := \sup_{x \in X : \|x\| = 1} \|Ax - A_\Theta(x)\|_Y. \tag{2.9}$$

However, in many relevant applications it is impossible to find a correction of the form $A_\Theta = F_\Theta \circ \widetilde{A}$ that achieves low uniform approximation error, making this notion of distance too restrictive. For instance, if we consider the case of a learned a posteriori correction of some approximate model $\widetilde{A}$ with a parameterizable mapping $F_\Theta : Y \to Y$ that fulfils (1.4), then the approximate model $\widetilde{A}$ can exhibit a nullspace $\mathbf{kern}(\widetilde{A})$ that is different from that of the accurate operator and, in particular, is potentially much larger. Thus, there may exist a (or several) $v \in \mathbf{kern}(\widetilde{A})$ with $Av \neq 0$. Any corrected operator $A_\Theta = F_\Theta \circ \widetilde{A}$ then exhibits an error in the sense of (2.9) of at least $\|Av\|_Y$, as

$$\begin{aligned}\|A - A_\Theta\|_{X\to Y} &\geq \max\{\|Av - F_\Theta(0)\|_Y, \|A(-v) - F_\Theta(0)\|_Y\} \\ &\geq \min_{y\in Y} \max\{\|Av - y\|_Y, \| - Av - y\|_Y\} \\ &= \|Av\|_Y,\end{aligned}$$

where in the last equality we have used that the point minimizing the maximum of the distance to two other points is the center of the line through those points. In our case, the center of the line between Av and $-Av$ is always the origin of the coordinate system 0, independently of the choice of A and v. In other words, the information in direction v is lost in the approximate model and would need to be recovered subsequently by the correction F_Θ. If there are several such nontrivial $v \in \mathbf{kern}(\widetilde{A})$, a uniform correction becomes increasingly difficult in the form of (2.9). We will illustrate this difficulty in the following section 2.2.1.

While aiming for a uniform correction is impractical, it can nevertheless be possible to correct the operator $\widetilde{A}$ using an a posteriori correction as in (1.4), provided a weaker notion of operator distance is employed. Here, we propose an empirical, learned notion of operator correction, that is optimized for a training set of points $\{x^i, i = 1, \dots, N\}$, similar to section 2.1. More precisely, we examine the average deviation of A_Θ from A as

$$\frac{1}{N}\sum_i^N \|A_\Theta(x^i) - Ax^i\|_Y \tag{2.10}$$

in a suitable norm $\|\cdot\|$. In this notion, it is sufficient for the operators to be close in the mean for a given training set and hence we call this a *statistical* or *learned* correction with respect to the chosen training set. For instance, if the kernel direction $v \in \mathbf{kern}(\widetilde{A})$ is orthogonal to the sample x^i, the information lost in direction v is not crucial for representing the data of interest. Alternatively, the kernel direction v might be highly correlated with another direction $w \notin \mathbf{kern}(\widetilde{A})$ in the sense that $\langle x^i, v\rangle \approx \langle x^i, w\rangle$ for all i. Then the result of Av can be inferred from $\widetilde{A}w$, even though $\widetilde{A}v = 0$.

To conclude this section, we note that in many cases we cannot hope to find a uniform model correction, but that correcting the model error can be still attempted using the notion of learned correction, quantified by (2.10). This is possible even if the operators A and $\widetilde{A}$ are exhibiting different kernel spaces, as long as the training set $\{x^i, i = 1, \dots, N\}$ exhibits sufficient structure to compensate for the loss of information in the approximate model.

Remark 2.1. We consider nonlinear corrections $A_\Theta = F_\Theta \circ \widetilde{A}$ in this paper even when correcting a linear operator A from a linear approximation $\widetilde{A}$, as in our computational examples.

We have three main motivations to do so. First, there are well-established nonlinear network architectures, such as U-Net [34], that are highly powerful and in fact have considerably fewer parameters than a fully parameterized linear map when the method is applied to applications in 3 dimensions, making the nonlinear approach scalable. Second, when considering nonlinear corrections, a generalization to the context of nonlinear operators will be easier. Finally and most importantly, while the operators A and $\widetilde{A}$ might be linear, the region of interest in image and data space where we need a good correction is highly nonlinear, in the sense that the samples x^i in (2.10) are drawn from a distribution with nonlinear support. This makes nonlinear corrections considerably more powerful in correcting model errors than their linear counterparts.

2.2.1. A toy case: Downsampling. In order to illustrate the challenge of a learned operator correction, we consider a toy case. Here, the accurate forward model A is given by a downsampling operator with an averaging filter, while the approximate model $\widetilde{A}$ simply skips every other sample. Concretely, we consider $x \in \mathbb{R}^n$, $y \in \mathbb{R}^{n/2}$ and $\widetilde{A}, A \in \mathbb{R}^{n/2\times n}$, given by

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & & & & & \\ & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix} \quad \text{and } \widetilde{A} = \begin{pmatrix} 1 & 0 & & & & & \\ & 0 & 1 & 0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & 0 & 1 & 0 \end{pmatrix}. \tag{2.11}$$

Clearly, both operators have very different kernel spaces, with A vanishing on inputs of even magnitude with alternating sign, whereas $\widetilde{A}$ vanishes for every v with $v[j] = 0$, with index j even, and any value for j *odd*. In other words, the null space is spanned by the unit vectors with odd index, $\mathbf{kern}(\widetilde{A}) = \{\mathbf{e}_j \mid 0 < j \le n,\ j \text{ even}\}$. In fact, by the same argument as above, these $v \in \mathbf{kern}(\widetilde{A})$ with $\|v\|_\infty = 1$ are such that the uniform approximation error for any correction will be $\|Av - F_\Theta(\widetilde{A}v)\|_\infty \ge \|Av\|_\infty \ge 0.25$ for all $v \in \mathbf{kern}(\widetilde{A})$.

This example exhibits the two features described in the previous section: First, a uniform correction in the sense of (2.9) is impossible due to different kernel spaces. However, a learned correction in the mean (2.10) is possible on some data $\{x^i, i = 1,\dots,N\}$ consisting of piecewise constant functions: On these samples the two operators $\widetilde{A}$ and A already coincide everywhere except near jumps, where a weighted average can be employed to correct the approximation error.

2.3. Solving the variational problem. We now aim to solve an inverse problem given the corrected model A_Θ by solving the associated variational problem (1.5). In this context it is natural to require that the solutions of the two minimization problems, involving the operator correction A_Θ and A, are close, that is,

$$\arg\min_{x\in X} \frac{1}{2}\|A_\Theta(x) - y\|_Y^2 + \lambda R(x) \approx \arg\min_{x\in X} \frac{1}{2}\|Ax - y\|_Y^2 + \lambda R(x). \tag{2.12}$$

Note that this formulation is different than the AEM (2.5), where the data fidelity term is given by $\|L_\varepsilon(\widetilde{A}x - y + \eta_\varepsilon)\|_Y^2$. Solutions to (1.5) are then usually computed by an iterative algorithm. Here we consider first order methods to draw connections to learned iterative schemes [1, 2, 17]. In particular, we consider a classic gradient descent scheme, assuming differentiable R. Then, given an initial guess x_0, we can compute a solution by the following iterative process:

$$x_{k+1} = x_k - \gamma_k \nabla_x \left(\frac{1}{2} \|Ax_k - y\|_X^2 + \lambda R(x_k) \right) \tag{2.13}$$

with appropriately chosen step size $\gamma_k > 0$. When using (2.13) for the corrected operator it seems natural to ask for a *gradient consistency* of the approximate gradient

$$\nabla_x \|A_\Theta(x) - y\|_X^2 \approx \nabla_x \|Ax - y\|_X^2 \tag{2.14}$$

and hence we can identify

$$\sum_{i=1}^{N} \left\| \nabla_x \left\| A_\Theta(x^i) - y^i \right\|_X^2 - \nabla_x \left\| Ax^i - y^i \right\|_X^2 \right\| \tag{2.15}$$

as another relevant measure of quality for model corrections within the variational framework, if gradient schemes are used to solve (1.5). In the following we will discuss the possibilities of obtaining a correction, such that we can guarantee a closeness of solutions in the sense of (2.12).

3. Forward model correction.

We will now present the possibility of correcting the forward model only and discuss resulting shortcomings of this approach. More precisely, in a forward model correction, the approximate operator $\widetilde{A} : X \to Y$ is corrected using a neural network $F_\Theta : Y \to Y$ that is trained to remove artefacts in data space for a given training set. This leads to a corrected operator of the form $A_\Theta = F_\Theta \circ \widetilde{A}$.

3.1. The adjoint problem.

To solve the minimization problem (1.5) with the learned forward operator with an iterative scheme such as (2.13), we need to compute the gradient of the data fidelity. We recall that the corrected operator $A_\Theta = F_\Theta \circ \widetilde{A}$, where the correction F_Θ is given by a nonlinear neural network. Following the chain rule we obtain the following gradient:

$$\frac{1}{2} \nabla_x \|A_\Theta(x) - y\|_2^2 = \widetilde{A}^* \left[DF_\Theta(\widetilde{A}x) \right]^* \left(F_\Theta(\widetilde{A}x) - y \right). \tag{3.1}$$

Here, we denote by $DF_\Theta(y)$ the Fréchet derivative of F_Θ at y, which is a linear operator $Y \to Y$, whereas the gradient for the correct data fidelity term is simply given by

$$\frac{1}{2} \nabla_x \|Ax - y\|_Y^2 = A^*(Ax - y).$$

That means, to satisfy the gradient consistency condition (2.14), we would need

$$\widetilde{A}^* \left[DF_\Theta(\widetilde{A}x) \right]^* \left(F_\Theta(\widetilde{A}x) - y \right) \approx A^*(Ax - y). \tag{3.2}$$

On the other hand, if we train the forward model correction, only requiring consistency in data space by minimizing (2.10), we will only ensure consistency of the residuals $F_\Theta(\widetilde{A}x) - y \approx$

$Ax - y$, but not full gradient consistency as in (2.14). In order to enforce gradient consistency we need to control the derivative of the network $DF_\Theta(\widetilde{A}x)$ and consequently also need to take the adjoint into consideration when training the forward correction. This could be done by adding an additional penalty term to (2.10) that penalizes the network for exhibiting an adjoint different from A^*. For that purpose, let us examine the adjoint of the linearization of the correction operator A_Θ around a point x:

$$(DA_\Theta(x))^* [y] = \widetilde{A}^* \left(DF_\Theta(\widetilde{A}x)\right)^* [y].$$

With this we can consider the following additional penalty term in the training:

$$\left\| \left(A^* - \widetilde{A}^* \circ \left[DF_\Theta(\widetilde{A}x)\right]^*\right)(r)\right\|_X, \quad \text{where } r = F_\Theta(\widetilde{A}x) - y \tag{3.3}$$

and choose r to be the residual in data space $F_\Theta(\widetilde{A}x) - y$ that arises when minimizing the data fidelity term as in (3.1).

However this solution comes with its own drawback. As we can see in (3.1), the range of the corrected fidelity term's gradient (3.1) is limited by the range of the approximate adjoint, $\mathbf{rng}(\widetilde{A}^*)$. Thus, we identify the key difficulty here in the differences of the range of the accurate and the approximate adjoints rather than the differences in the forward operators themselves, which links back to the discussion in 2.2.

Indeed, a correction of the forward operator via composition with a parameterized model F_Θ in measurement space is not able to yield gradients close to the gradients of the accurate data term if $\mathbf{rng}(\widetilde{A}^*)$ and $\mathbf{rng}(A^*)$ are too different. This problem is exacerbated if the dimensions of these two spaces differ and we cannot expect to find a correction that satisfies the gradient consistency (3.2) and, related to Remark 1.1, even suitable projections in $\widetilde{A}$ would not be sufficient to compensate for this. This observation can be made precise in the following theorem.

Theorem 3.1 (unlearnability of a gradient consistent forward model correction). *Let A and $\widetilde{A}$ be compact linear operators from X to Y and the solutions*

$$\hat{x} \in \arg\min_x \frac{1}{2}\|Ax - y\|_Y^2, \tag{3.4}$$

$$\hat{x}_\Theta \text{ critical point of } \frac{1}{2}\|A_\Theta(x) - y\|_Y^2 \tag{3.5}$$

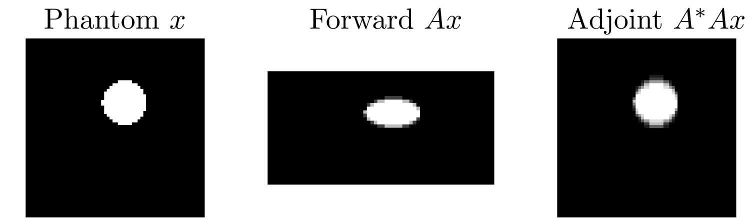
be given. If $\widetilde{x}_0 \in \mathbf{rng}(\widetilde{A}^*)$ and $\hat{x} \notin \overline{\mathbf{rng}(\widetilde{A}^*)}$, then a gradient descent algorithm for the functional in (3.5), initialized with $\widetilde{x}_0$, yields a solution such that $\hat{x}_\Theta \neq \hat{x}$ for any $\hat{x}$ solving (3.4).

Proof. This follows directly from the update equations for solving (3.5) by

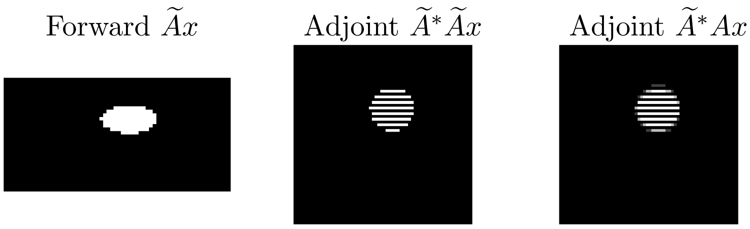
$$\widetilde{x}_{k+1} = \widetilde{x}_k - \lambda_k \Delta\widetilde{x}_k$$

with

$$\Delta\widetilde{x}_k := \frac{1}{2}\nabla_{\widetilde{x}_k}\|A_\Theta(\widetilde{x}_k) - y\|_Y^2 = \widetilde{A}^* \left[DF_\Theta(\widetilde{A}\widetilde{x}_k)\right]^* \left(F_\Theta(\widetilde{A}\widetilde{x}_k) - y\right). \tag{3.6}$$



(a) Application of the accurate forward operator and its adjoint



(b) Application of the approximate forward operator and its adjoint

Figure 1. *Illustration of mapping properties for the toy case. As we can see, the range of the adjoint and approximate adjoint are essentially different. Even if the approximate adjoint $\widetilde{A}^*$ is applied to the ideal data Ax (bottom right), representing a perfect fit of the forward model, the range of the approximate adjoint* $\mathbf{rng}(\widetilde{A}^*)$ *makes it impossible to compute a consistent gradient in* (2.14) *without further modifications.*

If $\widetilde{x}_0 \in \mathbf{rng}(\widetilde{A}^*)$ then $\Delta\widetilde{x}_0 \in \mathbf{rng}(\widetilde{A}^*)$, and hence $\widetilde{x}_1 \in \mathbf{rng}(\widetilde{A}^*)$. By induction this is true for all $k > 0$, i.e., $\widetilde{x}_k \in \mathbf{rng}(\widetilde{A}^*)$, $\forall k$ and thus any limit point $\hat{x}_\Theta \in \overline{\mathbf{rng}(\widetilde{A}^*)}$ lies in the closure of the range of $\widetilde{A}^*$. Since $\hat{x} \notin \mathbf{rng}(\widetilde{A}^*)$ it follows that $\hat{x} \neq \hat{x}_\Theta$ for any limit point of a gradient descent algorithm for solving (3.5). ■

Thus, a correction of the forward model by requiring only consistency in data space does not in fact ensure consistency of the data term, when solving a variational problem. Additionally, according to Theorem 3.1 even including an additional penalty term in the form of (3.3) does not solve this problem.

3.1.1. Illustration with the toy case. Going back to the toy case from section 2.2.1, where we considered a downsampling operation, the approximate operator was chosen such that the null space is spanned by the unit vectors with even index. The range of the adjoint can then be characterized by the identity $\mathbf{rng}(\widetilde{A}^*) = (\mathbf{kern}(\widetilde{A}))^\perp$ and hence we have $\mathbf{rng}(\widetilde{A}^*) = \{\mathbf{e}_j \mid 0 \leq j \leq n,\ j \text{ odd}\}$. It is now clear, that we cannot compute any solution $x^* \notin \mathbf{rng}(\widetilde{A}^*)$ by the updates in (3.6), if we initialize them with $\widetilde{x}_0 \in \mathbf{rng}(\widetilde{A}^*)$, since all updates are restricted to the range of the adjoint of the approximate operator. This problem is illustrated in Figure 1, where we consider an imaging problem for illustrative purposes and x is vectorized before the operators in (2.11) are applied. Whereas the difference in the forward operator is minimal for this example, the range of the approximate adjoint makes it impossible to recover the phantom without further adjustments after application of the adjoint, which will be addressed in the next section.

4. Forward-adjoint correction.

As is evident from the last section, a forward model correction that is computed to minimize (2.10) in data space alone is not sufficient to compute the actual reconstruction in a variational framework. We additionally require consistency in the gradients of the data fidelity term (2.15) which in turn boils down to a condition for a correction on the adjoint of the corrected forward operator in image space, motivated by (3.3). We will refer to such a correction in data and image space as a *forward-adjoint correction*, as we will learn a correction of the forward operator, as well as a correction of the adjoint (backward).

4.1. Obtaining a forward-adjoint correction.

The goal is now to obtain a gradient consistent model correction. To achieve this we propose to learn two networks. That is, we learn a network F_Θ that corrects the forward model and another network G_Φ that corrects the adjoint, such that we have

$$A_\Theta := F_\Theta \circ \widetilde{A}, \quad A^*_\Phi := G_\Phi \circ \widetilde{A}^*.$$

These corrections are obtained as follows. Given a set of training samples (x^i, Ax^i), we train the forward correction F_Θ acting in measurement space Y with the loss

$$\min_\Theta \sum_i \|F_\Theta(\widetilde{A}x^i) - Ax^i\|_Y. \tag{4.1}$$

In an analogous way, we correct the adjoint with the network G_Φ acting on image space X with the loss

$$\min_\phi \sum_i \|G_\Phi(\widetilde{A}^* r^i) - A^* r^i\|_X. \tag{4.2}$$

Here, we can choose the direction $r^i = F_\Theta(\widetilde{A}x^i) - y^i$ as in (3.3) for the adjoint loss. This ensures that the adjoint correction is in fact trained in directions relevant when solving the variational problem.

At evaluation time, the corrected operators can then be used to compute approximate gradients of the data fidelity term $\|Ax - y\|_Y^2$. The gradient then takes the form

$$A^*(Ax - y) \approx \left(G_\Phi \circ \widetilde{A}^*\right)\left(F_\Theta(\widetilde{A}x) - y\right). \tag{4.3}$$

Let us note that the separate correction of the adjoint and the forward operator comes with a change of philosophy compared to existing methods for forward operator correction as presented in section 2.1. Instead of trying to fit a single corrected operator A_Θ that is already parameterized according to its use within the data fidelity term of a variational problem, we fit a nonlinear corrected operator A_Θ whose use within the variational problem requires to fit the *gradient* of the data term directly. This gradient fit takes the form as in (4.3). We use the gradient of the data fidelity term to directly obtain the gradient of the variational functional for our corrected operator, allowing us to perform minimization techniques like gradient descent. We take the obtained critical point of these dynamics as the reconstruction. Note that the approximate gradient cannot be associated with a variational functional for the forward-adjoint method anymore. Instead, the gradient is parameterized directly, without parametrizing the variational functional first.

Remark 4.1. We note that such a separate correction in image and data space can be related to learned primal dual (LPD) methods [2], where the correction is performed implicitly as described in section 2. This explains in part why LPD approaches might be especially suitable for applications with an imperfectly known operator; see also [44].

In the following section we will discuss how these dynamics relate to the original variational problem and we will see that they can in fact take us close to the original reconstruction if both the forward and adjoint are fit sufficiently well.

4.2. Convergence analysis. The purpose of this section is to show that sufficiently small training losses can ensure that gradient descent over (1.5) converges to a neighborhood of the reconstruction $\hat{x}$, obtained with the accurate operator A. The section relates to the forward-adjoint correction (4.3) and uses the notation of this approach. In the case of forward-adjoint correction, these loss functions are given by

$$\|Ax - A_\Theta(x)\|_Y \text{ and } \|\left(A^* - A^*_\Phi\right)(A_\Theta(x) - y)\|_X. \tag{4.4}$$

Let us now consider for any $y \in Y$ the two functionals

$$\mathcal{L}(x) := \frac{1}{2}\|Ax - y\|_Y^2 + \lambda R(x),$$
$$\mathcal{L}_\Theta(x) := \frac{1}{2}\|A_\Theta(x) - y\|_Y^2 + \lambda R(x)$$

associated with the variational problem for the reconstruction x from the measurement y. We will show connections between the reconstruction $\hat{x} := \arg\min_x \mathcal{L}(x)$ using the accurate operator A and the solutions $\hat{x}_\Theta \in \arg\min_x \mathcal{L}_\Theta(x)$ obtained with our corrected operator A_Θ.

When considering the gradient descent dynamics over $\mathcal{L}_\Theta$, we do not refer to the actual gradient over $\mathcal{L}_\Theta$ but instead consider the direct fit to the gradient of the form $A^*_\Phi(A_\Theta(x) - y) + \lambda\nabla R(x)$ as discussed in the last section. In a slight abuse of notation we will nevertheless denote this gradient as $\nabla^\dagger\mathcal{L}_\Theta := A^*_\Phi(A_\Theta(x) - y) + \lambda\nabla R(x)$ to keep the notation easy to read in the remainder of this section. If R is merely subdifferentiable, then $\nabla R(x)$ denotes an element in the subgradient of R.

For the remainder of this chapter, we make the following assumption on the regularization functional R.

Assumption 4.2 (strong convexity). We assume that the regularization functional R is strongly convex and denote the strong convexity constant by m.

Remark 4.3. Assumption 4.2 in particular holds for R being the Tikhonov regularization functional $R(x) = \|x\|_X^2$ and for the pseudo-Huber loss $R(x) = \int_{[0,1]^2} \delta \left[\sqrt{1 + \frac{1}{\delta^2}\|\nabla_t x(t)\|^2} - 1\right]$ for a bounded function $x : [0,1]^2 \mapsto \mathbb{R}$ and $\delta > 0$ which we use in the experimental section. For operators A with bounded inverse it is sufficient for the regularization functional to be convex to ensure strong convexity of the resulting variational functional $\mathcal{L}$. In this case, strong convexity of the regularization functional is not required.

This allows us to use the following two fundamental lemmas on the behavior of $\mathcal{L}$ near the minimum of the variational functional. As a direct consequence of 4.2 and the convexity of the data term for linear forward operators we will from now on assume $\mathcal{L}$ to be strongly convex.

Lemma 4.4 (proximity to minimizer). *Let $\mathcal{L}$ be strongly convex. Then for every ϵ there is a $\delta > 0$ such that for any x*

$$\mathcal{L}(x) - \mathcal{L}(\hat{x}) \leq \delta \implies \|x - \hat{x}\|_X \leq \epsilon, \tag{4.5}$$

where $\hat{x} := \arg\min_x \mathcal{L}(x)$.

Proof. By the definition of strong convexity we have

$$\mathcal{L}(x) \geq \mathcal{L}(\hat{x}) + \langle s_{\hat{x}}, x - \hat{x}\rangle_X + \frac{m}{2}\|x - \hat{x}\|_X^2,$$

where $s_{\hat{x}} \in \partial\mathcal{L}(\hat{x})$ is in the subdifferential of $\mathcal{L}$ at $\hat{x}$. Using $0 \in \partial\mathcal{L}(\hat{x})$ yields

$$\delta \geq \mathcal{L}(x) - \mathcal{L}(\hat{x}) \geq \frac{m}{2}\|x - \hat{x}\|_X^2$$

which proves the claim by setting $\delta = \frac{\epsilon^2}{2m}$. ■

Lemma 4.5 (lower gradient norm bound). *Let $\mathcal{L}$ be strongly convex. For every ϵ there is a $\delta > 0$ such that for any x*

$$\|x - \hat{x}\|_X > \epsilon \implies \forall s \in \partial\mathcal{L}(x) : \|s\|_X > \delta, \tag{4.6}$$

where $\partial\mathcal{L}(x)$ denotes the subdifferential of $\mathcal{L}$ at x and $\hat{x} := \arg\min_x \mathcal{L}(x)$.

Proof. By the definition of strong convexity

$$\mathcal{L}(\hat{x}) \geq \mathcal{L}(x) + \langle s_x, \hat{x} - x\rangle_X + \frac{m}{2}\|x - \hat{x}\|_X^2,$$

where again s_x denotes an element in the subdifferential of $\mathcal{L}$ around x. Then by the Cauchy–Schwarz inequality

$$\mathcal{L}(\hat{x}) - \mathcal{L}(x) - \frac{m}{2}\|x - \hat{x}\|_X^2 \geq -\|s_x\|_X\|\hat{x} - x\|_X.$$

Using $\mathcal{L}(\hat{x}) - \mathcal{L}(x) < 0$ by assumption shows

$$\frac{m}{2}\|x - \hat{x}\|_X^2 \leq \|s_x\|_X\|\hat{x} - x\|_X$$

and, hence, $\|s_x\|_X \geq \frac{m}{2}\|x - \hat{x}\|_X$, which proves the result. ■

Remark 4.6. The assumption of strong convexity is used in the following results via Lemmas 4.4 and 4.5 only. While it is a sufficient condition for these to hold, it is not necessary. In particular, if the variational functional is not strongly convex but such that 4.4 and 4.5 hold true, the following results still apply.

We now turn to show that a minimizer $\hat{x}_\Theta$ of the approximate functional can in fact be computed with a gradient descent scheme and that this leads minimizer in fact to the accurate reconstruction $\hat{x}$. We begin by extending Lemma 4.5 to include the regularization term. For this purpose, we consider the alignment of the variational gradients including the regularization term

$$\cos \Phi_v(x) := \frac{\langle \nabla \mathcal{L}(x), \nabla^\dagger \mathcal{L}_\Theta(x)\rangle}{\|\nabla \mathcal{L}(x)\|^2}. \tag{4.7}$$

We show how the alignment can be used as the key quantity to guarantee convergence of the approximate dynamics to a a neighborhood of the accurate solution. We remark again the abuse of notation $\nabla^\dagger \mathcal{L}_\Theta(x) := A^*_\Phi(A_\Theta(x) - y) + \lambda \nabla R(x)$.

Proposition 4.7 (convergence under alignment constraints). *Assume that outside a neighborhood U of the minimizer $\hat{x}$ of the exact functional $\mathcal{L}$ we have*

$$\cos \Phi(x) > \delta_1 > 0$$

for some $\delta_1 > 0$. Then eventually the gradient descent dynamics over $\mathcal{L}_\Theta$ will reach the neighborhood U.

Proof. Denote by $x_\Theta(t)$ the trajectory of the reconstruction under the gradient flow

$$\partial_t x_\Theta(t) = -\nabla^\dagger \mathcal{L}_\Theta(x_\Theta(t)).$$

Consider the evaluation of the variational loss $\mathcal{L}$ that invokes the correct forward operator A. Using the bound of the alignment as in Lemma 4.8, we can bound

$$\begin{aligned}\partial_t \mathcal{L}(x_\Theta(t)) &= \langle \nabla \mathcal{L}(x_\Theta(t)), \partial_t x_\Theta(t)\rangle_X = -\langle \nabla^\dagger \mathcal{L}(x_\Theta(t)), \nabla \mathcal{L}_\Theta(x_\Theta(t))\rangle_X \\ &\leq -\delta_1 \cdot \|\nabla \mathcal{L}(x)\|_X^2.\end{aligned}$$

As long as $\Theta(t)$ has not reached the neighborhood U, by (4.6), we have $\|\nabla \mathcal{L}(x)\|_X > \delta_2$ for some δ_2 and hence

$$\partial_t \mathcal{L}(x_\Theta(t)) \leq -\delta_1 \cdot \|\nabla \mathcal{L}(x)\|_X^2 \leq -\frac{1}{2}\delta_1\delta_2 =: -c < 0.$$

The gradient flow dynamics induced by $\nabla^\dagger \mathcal{L}_\Theta$ hence induce a decrease of $\mathcal{L}$ at a rate that is globally bounded by c outside a neighborhood U around $\hat{x}$, concluding the proof by Lemma 4.4. ∎

We have shown that even though the corrected operator A_Θ is potentially nonlinear, the gradient dynamics induced by $\nabla^\dagger \mathcal{L}_\Theta$ can in fact minimize the variational problem with the accurate operator A, effectively minimizing the associated variational functional $\mathcal{L}$ and leading us close to the accurate solution $\hat{x}$. The proposition is based on an assumption about the alignment $\cos \Theta$. We will directly track this quantity in our experimental section, making sure the convergence results can be applied to our experimental findings. The training loss, however, is not based on the alignment directly, but rather minimizes a combination of forward and adjoint loss. We have in fact found that this combination of loss functionals is both more interpretable and more stable than directly minimizing the alignment. The following lemma and theorem show that these loss functions in fact minimize a lower bound on the alignment and hence a sufficiently well-trained correction can also be guaranteed to yield results close to the minimizer $\hat{x}$ of the variational functional involving the exact operator A. In this context, a well-trained correction is such that it achieves sufficiently low training errors.

Lemma 4.8 (complete gradient alignment bound). *Let $\mathcal{L}$ and $\mathcal{L}_\Theta$ be defined as above. We have the lower bound*

$$\cos \Phi_v \geq 1 - \frac{\|A\|_{X\to Y}\|(A - A_\Theta)(x)\|_Y + \|(A^* - A^*_\Phi)(A_\Theta(x) - y)\|_X}{\|\nabla \mathcal{L}(x)\|_X},$$

where $\cos \Phi_v$ is defined as in (4.7).

Proof. A straightforward calculation shows

$$\begin{aligned} &\frac{\langle \nabla \mathcal{L}(x), \nabla^\dagger \mathcal{L}_\Theta(x)\rangle_X}{\|\nabla \mathcal{L}(x)\|_X^2} \\ &= \frac{\langle \nabla \mathcal{L}(x), \nabla \mathcal{L}(x)\rangle_X}{\|\nabla \mathcal{L}(x)\|_X^2} + \frac{\langle \nabla^\dagger \mathcal{L}_\Theta(x) - \nabla \mathcal{L}(x), \nabla \mathcal{L}(x)\rangle_X}{\|\nabla \mathcal{L}(x)\|_X^2} \\ &\geq 1 - \frac{\|\nabla^\dagger \mathcal{L}_\Theta(x) - \nabla \mathcal{L}(x)\|_X}{\|\nabla \mathcal{L}(x)\|_X}. \end{aligned}$$

The result follows by using the bound

$$\begin{aligned} &\|A^*(Ax - y) - A^*_\Phi(A_\Theta(x) - y)\|_X \\ &\leq \|A\|_{X\to Y}\|(A - A_\Theta)(x)\|_Y + \|(A^* - A^*_\Phi)(A_\Theta(x) - y)\|_X, \end{aligned}$$

which itself emerges directly from the triangular inequality applied to the identity

$$A^*(Ax - y) - A^*_\Phi(A_\Theta(x) - y) = A^*(A - A_\Theta)(x) + (A^* - A^*_\Phi)(A_\Theta(x) - y). \qquad \blacksquare$$

Theorem 4.9 (convergence to a neighborhood of $\hat{x}$). *Let $\epsilon > 0$ and pick δ as in (4.6). Assume both the adjoint and forward operator are fit up to a $\delta/4$-margin, i.e.,*

$$\|A\|_{X\to Y}\|(A - A_\Theta)(x_n)\|_Y < \delta/4, \quad \|(A^* - A^*_\Phi)(A_\Theta(x_n) - y)\|_X < \delta/4 \tag{4.8}$$

for all y and x_n obtained during gradient descent over $\mathcal{L}_\Theta$. Then eventually the gradient descent dynamics over $\mathcal{L}_\Theta$ will reach an ϵ neighborhood of the accurate solution $\hat{x}$.

Proof. We apply 4.7, with the neighborhood U chosen as the ϵ ball around $\hat{x}$. Using Lemma 4.8, we can bound

$$\cos \Phi \geq 1 - \frac{\|A\|_{X\to Y}\|(A - A_\Theta)(x)\|_Y + \|(A^* - A^*_\Phi)(A_\Theta(x) - y)\|_X}{\|\nabla \mathcal{L}(x)\|_X} \geq 1 - \frac{\delta/4 + \delta/4}{\|\nabla \mathcal{L}(x)\|_X}.$$

As long as $\|x_\Theta(t) - \hat{x}\|_X \geq \epsilon$, by (4.6), we have $\|\nabla \mathcal{L}(x)\|_X > \delta$ and hence

$$\cos \Phi \geq 1 - \frac{\delta/2}{\delta} > 0.$$

We can hence apply 4.7 to conclude the proof. $\blacksquare$

Overall, we have thus shown that a sufficiently well-trained nonlinear corrected operator A_Θ induces gradient dynamics $\nabla^\dagger \mathcal{L}_\Theta$ that lead close to the accurate solution $\hat{x}$.

We note that the main assumption in Theorem 4.9 is that the learned operator A_Θ has to be sufficiently close to the accurate operator A *throughout* the minimization trajectory, in the sense of (4.8). While this corresponds directly to the quantities of the loss functions that the approximations A_Θ and A^*_Φ were trained on, it includes any x_n occurring during the gradient descent dynamics. Thus, we will discuss the concept of adding exactly these samples x_n to the training set in the next chapter, effectively making our training loss function minimize exactly the relevant quantities $\|(A - A_\Theta)(x_n)\|_Y$ and $\|(A^* - A^*_\Phi)(A_\Theta(x_n) - y)\|_X$.

Remark 4.10. The above Theorem 4.9 makes use of both proximity of the forward operator as well as of the adjoints. While this is necessary to guarantee convergence of the gradient descent dynamics to a neighborhood of the accurate solution, it is not strictly necessary to guarantee proximity of the minimizers of $\mathcal{L}_\Theta$ and of $\mathcal{L}$. In fact, in Appendix B we show that under certain assumptions a good forward approximation quality is sufficient to ensure closeness of minimizers, without considering a specific optimization scheme. While this result is interesting from a theoretical viewpoint, Theorem 4.9 is essential for supporting and explaining the experimental results in this study.

5. Computational considerations.

In the following we will first address some details on the training procedures and then continue to present the design of experiments to evaluate the performance of the discussed approaches. In particular, as we mentioned above, in order to ensure the convergence in Theorem 4.9, we need to make sure that the forward fit as well as the backward fit in (4.8) are satisfied throughout the minimization process, which makes a special recursive training of the corrections necessary.

5.1. Recursive training.

Let us now address how to ideally choose the training sets for the forward-adjoint correction to ensure a good fit of the forward correction F_Θ by minimizing (4.1) and the adjoint correction G_Φ with (4.2). To create the training set, there are two possibilities. Either we are given a set of measurements $\{y^i, i = 1, \dots, N\}$ or, alternatively, if we are given a set of samples in image space $\{x^i, i = 1, \dots, N\}$, then we need to create a corresponding set of measurements by applying the accurate model $y^i = Ax^i + e^i$ with the addition of noise e^i. Either way, given the set of measurements y^i we need to train F_Θ and G_Φ on a meaningful starting point for the gradient descent to solve the variational problem; a natural candidate would be to choose the backprojection $x^i_0 = \widetilde{A}^* y^i$.

Training the corrected operators A_Θ and A^*_Φ with the samples $\{(x^i_0, Ax^i_0)\}$ only yields operator corrections that approximate A and A^* well for samples x that are close to backprojections of measurements. However, the purpose of this paper is to learn a correction of $\widetilde{A}$ that can be used *within the variational problem* to obtain a solution close to the one obtained using the accurate operator A. We observe that training A_Θ on the backprojections $x^i_0 = \widetilde{A}^* y^i$ only is not sufficient to achieve this goal. While this leads to A_Θ being a good approximation to A for the first iterates in the gradient descent scheme, the approximation quality tends to deteriorate for later iterates, making A_Θ not a good appproximation to A anymore. Such a behavior is in fact what one would heuristically expect, as A_Θ has never been trained on later iterates to match the accurate operator.

This connects to the assumptions made in the convergence Theorem 4.9, where we assume low approximation errors for both the forward and the adjoint at *all* iterates of the gradient descent scheme. We hence need to ensure a uniformly low approximation error at any iterate to be able to guarantee convergence and it is in particular not sufficient to ensure a low approximation error at the initial point of the minimization of the variational problem only.

A natural solution to mitigate this problem is to include later iterates of the variational problem into the training samples for the corrected operator. More precisely, given some weights Θ of the correction operator, denote by $\{x_n^i\}$ the iterates obtained following the dynamics

$$x_{n+1}^i = x_n^i - \mu \left[A_\Phi^*(A_\Theta(x_n^i) - y_n^i) + \lambda \nabla R(x) \right], \tag{5.1}$$

where μ denotes the step size. We add these samples to the original training set $\{(x^i, Ax^i)\}$, i.e., we also train on $\{(x_n^i, Ax_n^i)\}$ for all $n < N_{\text{iter}}$ and i. Here N_{iter} is the maximal number of gradient descent steps we take. This allows us to ensure that the corrections A_Θ, as well as A_Φ^* for the forward-adjoint method, are fit consistently well at any iterate x_n^i of the gradient descent dynamics.

A major drawback of this approach is the additional computational burden it incurs during training. Obtaining the iterates of the minimization to solve the variational problem requires performing the minimization at training time. To reduce the additional computational burden one can make use of the fact that the gradient of the data term for the learned operator correction A_Θ has to be computed for two different purposes. First, it is used to perform minimization over the variational functional and, second to further train the A_Θ to better match the accurate operator. One can hence perform this computation only once, using it for both purposes. This reduces computational costs particularly when training on every iterate of the minimization over the variational functional, in which case little overhead cost compared to regular training is inflicted.

Additionally, the trajectory (5.1) depends on the network weights Θ. The training samples can hence change during training and convergence is not clear a priori. Empirically, we find that training on the full trajectory (x_n^i, Ax_n^i) for $n < N_{\text{iter}}$ from the beginning tends to be unstable, as this will lead to most training samples differing greatly from both the original training distribution as well as the accurate trajectory we are finally interested in. There are, however, two effective solutions to this problem: First, one could alternatively train on the trajectory obtained when using the accurate operator A, avoiding instabilities in the beginning of training. This, however, could lead to errors accumulating during training. We found that the most effective solution is to have N_{iter} increase from 1 to some $N_{\max}$ during training. With this approach, we start off by training on the original samples x_0^i only and then add in more samples from the trajectory as training proceeds. We have noticed that once trained on backprojections, adding later iterates to the training set does not change the behavior of the learned correction on backprojections by much. In this sense, one can interpret the latter approach to recursive training as gradually extending the domain the correction is valid on, without considerably changing the behavior of the correction on the part of the image domain that it is already valid on. This heuristically explains why recursive training can be performed very stably when gradually increasing N_{iter}.

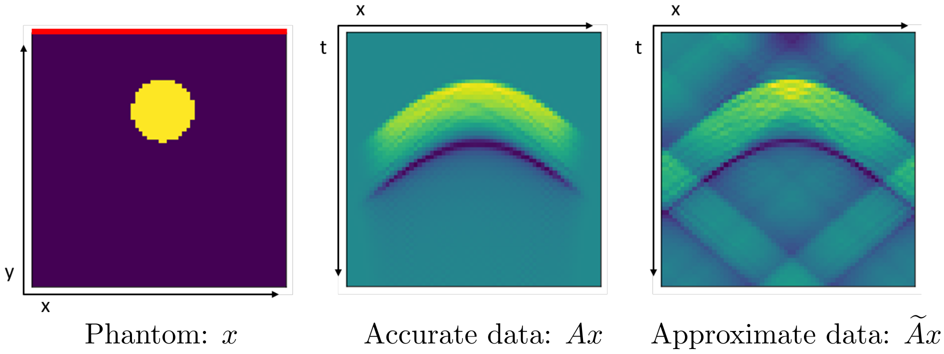


Figure 2. *Illustration of the limited view imaging scenario under consideration. Left: numerical phantom with a line detector (red line). Middle: ideal data from the accurate forward model. Right: data obtained with an approximate model with clearly visible aliasing artefacts.*

5.2. Experimental design. For a practical application we consider photoacoustic tomography (PAT) in two dimensions; for more details on PAT see [6] and the discussion in Appendix A. Here, the measurement data are given as a set of time series in a limited view geometry measured with a line detector at the surface, which we visualize as a space-time image in Figure 2. In this limited view scenario, the reconstruction task is already a very challenging inverse problem in itself even with the accurate operator available; we refer to [30, 45] for details. Here, the accurate model A is given by a pseudospectral time-stepping model [42, 43], whereas the approximate model $\widetilde{A}$ is given by a regriding and fast Fourier transform which neglects the effect of singularities and introduces systematic errors in the forward mapping [11, 28]. In particular, to avoid singularities in the approximate model we threshold incident waves with an angle up to $\theta_{\max} = 60°$ from normal incidence, which means that this part of the data is inevitably lost. Nevertheless, the approximate forward model still exhibits strong aliasing artefacts, as can be clearly seen in Figure 2 indicating that this application is an ideal candidate for this study. For more details on the models, we refer to the discussion in Appendix A. We developed the majority of code in Python using the TensorFlow package and using the k-Wave MATLAB (R2018b) toolbox [42] for some calculations concerning the accurate operator. We used a single Quadro P6000 to conduct the experiments.

Model corrections under consideration. We evaluate the forward only method with a gradient penalty term as described in section 3 as well as the forward-adjoint approach as outlined in section 4.[1] For both of these methods, we conduct experiments with a model trained on back-projected measurements only and with a model that has been trained using recursive training (section 5.1). As a baseline method, we compare this to the widely used AEM approach as outlined in section 2.1, a linear approach to model correction. We finally compare this to reconstructions obtained with the uncorrected operator as well as to the reconstruction the accurate operator yields. This allows us to assess how well various correction approaches are able to correct the shortcomings of the uncorrected operator.

[1]Code is available at https://github.com/lunz-s/ModelCorrection

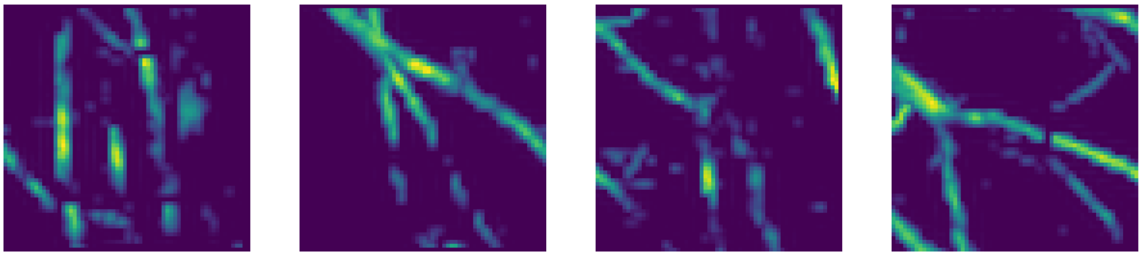


Figure 3. *Examples from the vessel set used for training of the model correction. The phantoms were obtained from segmented CT scans to provide a realistic ground-truth image for photoacoustic imaging of vessel structures.*

Measurement setup. We consider a limited view problem in this study, where measurements are only taken on top of the target with a line detector, as indicated in Figure 2. In particular, we consider an image size of 64×64, the measurements are taken with a line detector of the same width as the target, and $t = 64$ time points, resulting in a measurement space of the same size, i.e., 64×64. The detector is modeled as a Fabry–Pérot sensor [46] with wide bandwidth and no directivity. Since both image and data space can be represented as a two-dimensional image, it is reasonable to use the same network architecture for both spaces.

Training samples. For the evaluation of the various model correction methods, we utilize two different sets of samples. First, a simple synthetic set of "ball" images, consisting of circles of varying intensity in $[0.75, 1]$, with fixed radius, but random location on an empty, zero intensity background. We employ a total of 4096 ball samples for fitting the correction and an additional 64 for evaluation. An example of a ball image and the corresponding data are illustrated in Figure 2. Second, a realistic vessel set that has been obtained by segmenting vessels from three-dimensional (3D) CT scans to provide realistic phantoms, see [21] for details. For this study, the 3D volumes have been projected to two dimensions by a maximum intensity projection and subsequently cropped to the intended target size; we note that all samples are normalized between $[0, 1]$. Examples of the obtained vessel phantoms are displayed in Figure 3. We use 2760 unique vessel phantoms for training, augmented by a rotation by 90° for a training set of 5520 samples in total. We evaluate on a separate test set containing 64 samples. All phantoms had a resolution of 64^2 and resolution in data space is the same for both, correct and approximate model. The phantoms are used to generate synthetic measurements $y^i := Ax^i + e^i$ by applying the accurate operator A and adding Gaussian white noise at 1% of the maximum value in measurement space.

Training scheme. For every measurement y^i, we compute $x_0^i := 4 \cdot \widetilde{A}^* y$ as an initial reconstruction. We choose to rescale the adjoint $\widetilde{A}^* y$ by a factor of 4 as in our measurement setup we typically have $\|Ax\|_Y \approx \frac{1}{2}\|x\|_X$ and $\|A^* y\|_X \approx \frac{1}{2}\|y\|_Y$. This is due to the fact that we measure along a line on one side of the object only, hence recording only half the energy emitted on the measurement device. This ensures that the average intensity of the backprojection roughly matches the one of both the ground truth and the minimizer of the variational functional. It allows us to keep the norm of the reconstruction approximately stable throughout solving the variational problem (5.5) and hence makes operator approximations more robust throughout the trajectory of minimizing (5.5).

Given a set of training samples y^i, we then train the forward approximation with the loss term

$$\sum_i \underbrace{\left\| F_\Theta(\widetilde{A}x_0^i) - Ax_0^i \right\|_Y}_{\text{Forward Loss}} + \underbrace{\left\| \left(A^* - \widetilde{A}^* \left[DF_\Theta(\widetilde{A}x_o^i) \right]^* \right) \left(F_\Theta(\widetilde{A}x_0^i) - y^i \right) \right\|_X}_{\text{Adjoint Loss}}, \tag{5.2}$$

weighting the forward and adjoint loss equally. In the case of a forward-adjoint correction, the forward approximation is trained using the loss

$$\sum_i \left\| F_\Theta(\widetilde{A}x_0^i) - Ax_0^i \right\|_Y, \tag{5.3}$$

while the adjoint is trained with the loss

$$\sum_i \left\| \left(G_\Phi \circ \widetilde{A}^* - A^* \right) \left(F_\Theta(\widetilde{A}x_0^i) - y^i \right) \right\|_X. \tag{5.4}$$

Note that the quasi-adjoint of the approximate operator $A^*_\Phi := G_\Phi \circ \widetilde{A}^*$ as well as the adjoint of the forward approximation in (5.2) is evaluated in direction $r := F_\Theta(\widetilde{A}x_0^i) - y^i$. This loss is chosen to be consistent with the terms arising during a gradient descent based optimization of (5.5), as shown in the previous chapters.

If recursive training is applied, we additionally compute the iterates of a gradient descent scheme on the penalty functional

$$\arg\min_x \|A_\Theta(x) - y^i\| + \lambda R(x). \tag{5.5}$$

All losses are summed over the later iterates x_n^i with $n \geq 0$, instead of taking the initial point x_0^i only. To make recursive training stable, the number of recursive steps considered during training is gradually increased to the maximal value, instead of beginning by training on the full trajectory from the start as outlined in section 5.1.

Network details. The networks F_Θ and G_Φ are built with a U-Net [34] architecture, that has been particularly popular in the image reconstruction community including applications to PAT [3, 12, 15] and other modalities [16, 19, 22]. We follow the standard architecture with 4 downsampling and the same amount of upsampling blocks, each containing two convolutional layers with filters of size 5×5. We employed average pooling for downsampling and transpose convolutions for upsampling layers. We note, that the proposed framework is agnostic to the employed architecture; we expect similar results with other sufficiently expressive network architectures.

Solving the variational problem. We employ gradient descent with a fixed step size of 0.2 for all experiments to solve the variational problem (5.5), which we have seen can lead to a near-optimal reconstruction given sufficient approximation quality in section 4.2. We additionally add a positivity constraint $x_n \geq 0$ everywhere to the minimization that we incorporate using projected gradient descent. This means we cut the negative part of every iterate to 0 everywhere, as negative values are nonphysical.

As regularization functional R we choose the pseudo-Huber varation functional

$$R(x) := \sum_{i,j} \delta \left[\sqrt{1 + \frac{1}{\delta^2}[(x[i+1,j] - x[i,j])^2 + (x[i,j+1] - x[i,j])^2]} - 1 \right] \tag{5.6}$$

to reconstruct $x \in \mathbb{R}^{64\times 64}$. Here $x[i,j]$ denotes the pixel of x at location i along the vertical and j along the horizontal axis. This functional approximates the L^2-norm of the gradient of the reconstruction for small values and the L^1-norm for large values of the gradient, coinciding with total variation (TV) in the limit $\delta \to 0$. The parameter δ specifies the characeristic length at which the behavior of the regularization functional changes from approximating L^2 to L^1. We chose $\delta = 0.01$ for all experiments. We remark that this functional is strongly convex on all bounded domains for all $\delta > 0$, with the strong convexity constant depending on δ and the diameter of the imaging domain. The latter is in our case specified by the constraint $x[i,j] \in [0,1]$.

The regularization parameter λ is tuned for every experiment and baseline individually via a grid search over a logarithmically evenly spaced grid with grid points being a factor of $log(10)$ apart. The best parameter was chosen in terms of L^2 distance to the ground-truth image.

6. Computational results.

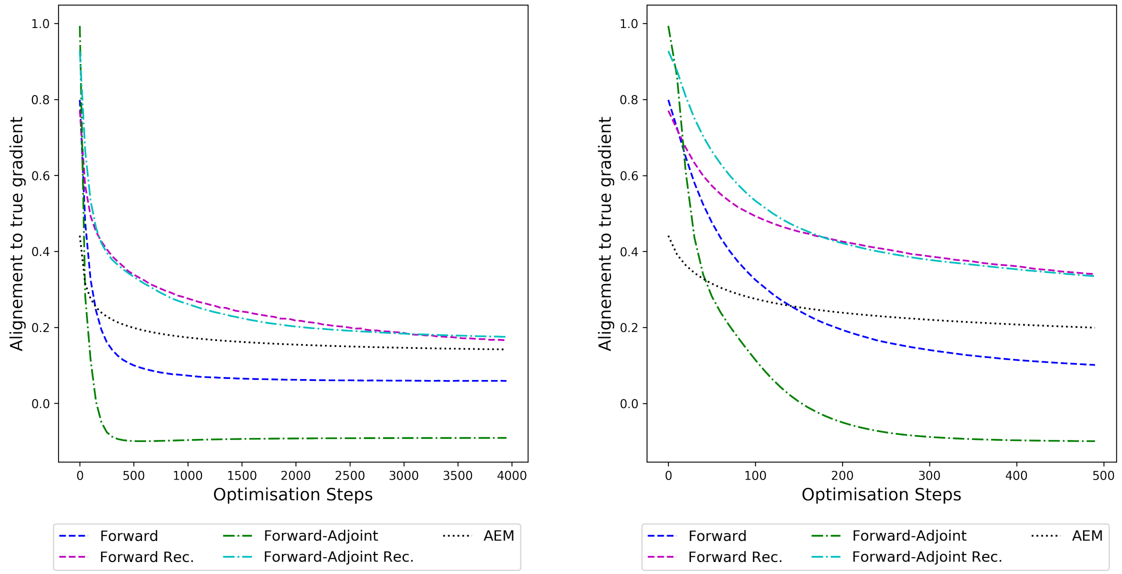
Synthetic ball phantoms. To evaluate the proposed approaches we solve the variational problem employing the various approaches for model correction for a set of samples generated from a test set that is different from the samples used for fitting the correction. We use the same Huber regularization functional and regularization parameter as discussed in the last paragraph.

First, we investigate the correction accuracy in terms of the alignment of the gradient of the data fidelity term with the accurate gradient $A^*(Ax_n - y)$ throughout the minimization of the variational functional in Figure 4. As a notion of alignment we consider

$$\cos \Phi_v(x) = \frac{\left\langle A^*\left(Ax_n - y\right), \left(G_\Phi \circ \widetilde{A}^*\right)\left(F_\Theta(\widetilde{A}x) - y\right)\right\rangle_X}{\left\| A^*\left(Ax_n - y\right)\right\|_X \left\| \left(G_\Phi \circ \widetilde{A}^*\right)\left(F_\Theta(\widetilde{A}x) - y\right)\right\|_X} \tag{6.1}$$

in the case of the forward-adjoint method. For the forward only and AEM methods, the expression $\left(G_\Phi \circ \widetilde{A}^*\right)\left(F_\Theta(\widetilde{A}x) - y\right)$ is replaced by the appropriate gradient of the corrected data fidelity term. Equation (6.1) is a slight deviation from (4.7) used in the theory section. This is to ensure good comparability with the baseline AEM and better interpretability by rescaling the alignment with the norm of the approximate gradient. This also makes different choices of regularization parameters more comparable. In the theory section we instead rescale with the norm of the accurate gradient only, making the proofs more straightforward.

We note that all correction methods apart from the AEM approach start at a high alignment of > 0.8 at the first iterate. However, only the forward-adjoint based methods are able to achieve an alignment of > 0.95 at the first iterate. Forward only approaches that rely on fitting a correction in measurement space only are limited by the range of the adjoint $\widetilde{A}^*$ as discussed in section 4.



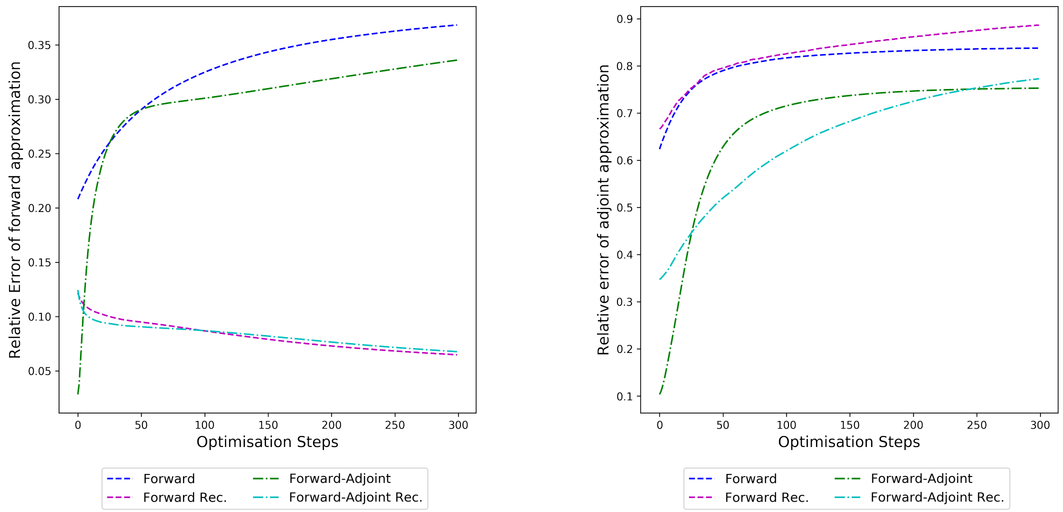
(a) Full trajectory (b) First 500 steps

Figure 4. *Alignment* (6.1) *of approximate gradient to the gradient of the accurate data term* $A^*(Ax_n - y)$ *for each approach on the ball test set of* 64 *samples. The alignment is recorded over all minimization steps for solving the associated variational problem. On the left* (a) *for the full trajectory and on the right* (b) *for the first* 500 *steps.*

However, the alignment starts decreasing rapidly over the minimization of the variational problem, dropping below 0 for the forward-adjoint method before the 200th iterate. The recursive versions of the forward and forward-adjoint methods, as discussed in section 5.1, are able to mitigate some of this shortcoming. While the alignment between accurate gradient and the correction also declines throughout the minimization of the variational problem when employing recursive training, the decline is significantly less steep and occurs at a later stage of the minimization. We also note that the alignment never drops under 0.2 for recursively trained corrections.

The benchmark AEM method is not able to correct the gradient as accurately as any of the methods we discussed for the first iterates of the variational problem. However, it does not exhibit a decline of the alignment as drastic as any of the other methods throughout the minimization process. This can be explained by the lower expressive power of AEM compared to the corrections based on neural networks that does not allow the method to fit the accurate gradient as well for early iterates but prevents overfitting on later iterates, leading to the method being stable throughout the minimization of the variational functional.

The different behaviors of forward and forward-adjoint methods as well as their recursive counterparts is investigated in Figure 5. We note that in terms of the forward approximation error, applying recursive training makes the key difference in terms of keeping a low error throughout gradient descent. For the adjoint approximation error we note that methods based on the forward scheme that fit a single operator are not able to achieve low error,



(a) Relative approximation error of forward operator

(b) Relative approximation error of adjoint operator

Figure 5. *Approximation error of the model correction compared to the accurate operator on the ball test set of* 64 *samples, tracked throughout the first* 300 *steps of the gradient descent scheme. Left* (a): *relative error of the forward approximation as defined in* (5.3). *Right* (b): *relative error for the adjoint, as defined for the forward only in* (5.2) *and for the forward-adjoint method in* (5.4).

even at the first iterate due to the fundamental limitations of the method. Forward-adjoint methods on the other hand are able to fit the accurate adjoint well at the first iterates, but also suffer from deteriorated approximation quality for later steps.

In Figure 6, we see evolution of the data term $\|Ax_n - y\|_Y$ evaluated using the accurate operator A in order to test if the corrections minimize the original variational problem. We note that both recursive methods are able to effectively minimize the data term quickly, with both converging stably to their respective minimal value. This empirical observation shows that the learned reconstructions in fact lead to a variational energy that satisfies Lemma 4.4 to ensure closeness of minimizer. We note that forward-adjoint recursive is able to achieve a lower data loss than its forward only counterpart, which is consistent with the behavior observed in Figure 4. It is interesting to note, that both methods are able to minimize the accurate data term significantly better than the baseline AEM. When omitting recursive training both the forward only and the forward-adjoint algorithm are not able to minimize the accurate data term well.

Finally, we evaluate the model correction in terms of the distance of the reconstruction to the ground-truth image, measured by the relative L^2 error shown in Figure 7. We note that all approximation approaches outperform the uncorrected operator in this metric. Both corrections, forward and forward-adjoint, without recursive training lead to a decrease in reconstruction error reconstruction quality for the first 300 optimization steps, stagnating or even deteriorating afterwards. This is again consistent with the findings in Figure 4, which show that the gradient generated by these methods does not align with the accurate

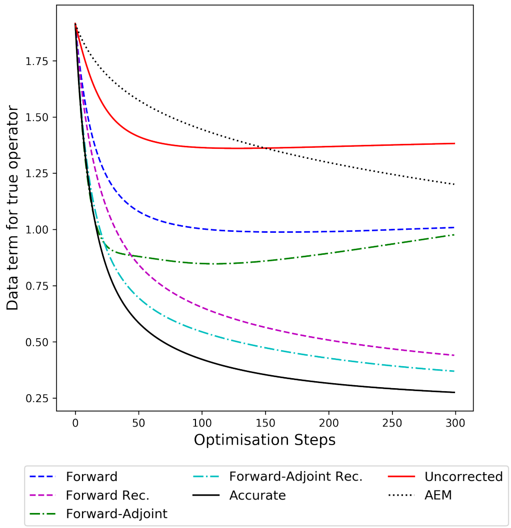


Figure 6. *True data term $\|Ax_n - y\|_Y$ evaluated for all methods on the ball test set of 64 samples, tracked throughout the gradient descent scheme.*

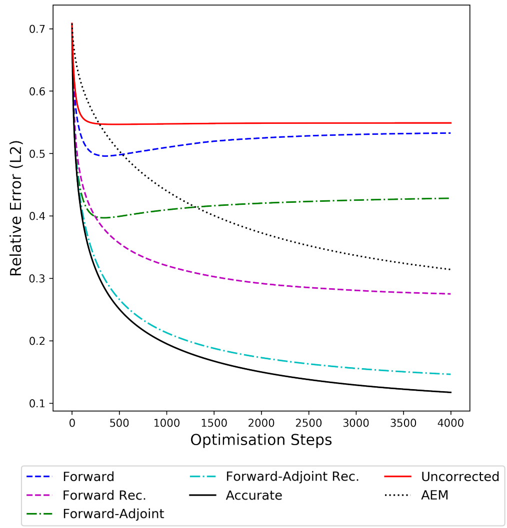
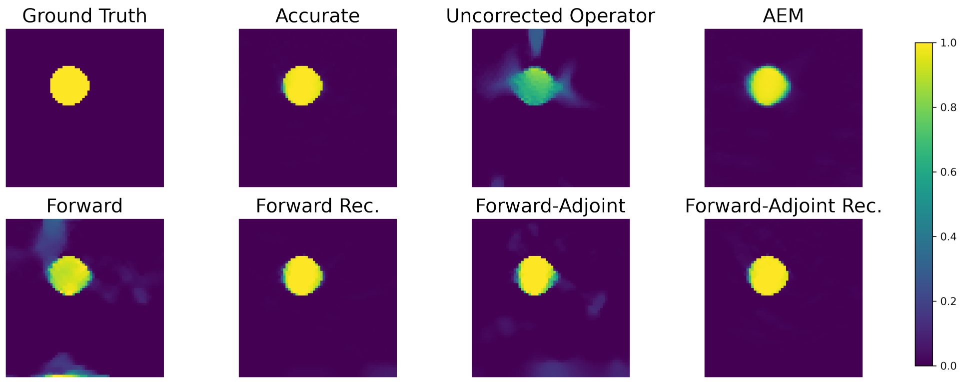


Figure 7. *Relative reconstruction error (L2) for all methods on the ball test set of 64 samples, tracked throughout the gradient descent scheme.*

gradient any more at this point of the minimization. The recursive counterparts of the forward and forward-adjoint method produce considerably better results, with the recursive forward-adjoint method generating reconstructions that are nearly of the same quality as the ones generated by the accurate operator. The baseline with AEM is converging more slowly than any of the other methods but is able to produce high-quality results after 4000 gradient descent steps that are on par with the forward recursive method, but are significantly outperformed by the recursive forward-adjoint method.

For a qualitative evaluation, we show obtained reconstructions in Figure 8 for all methods discussed and two samples with different behavior. In the first example, where the ball is close to the line detector, we note that all methods are able to correct the errors introduced by the approximate operator to some extent. However, both the forward and forward-adjoint method introduce background artefacts when not trained recursively. These artefacts disappear when recursive training is applied, leading to near perfect reconstructions. Compared to AEM as baseline, which is able to correct the approximate operator without introducing background artefacts, the correction by AEM introduces blurred edges of the ball that are not observed by any of the neural network based corrections we are investigating. The second sample is particularly more challenging, with the ball being far from the detector exhibiting stronger limited view artefacts and consequently the approximate operator introduces severe artefacts if uncorrected. For the corrections without recursive training we see again that both approaches, forward and forward-adjoint, introduce background artefacts. For the forward method, these artefacts cannot be suppressed by applying recursive training, leaving a severe artefact at the boundary of the domain. Only the recursive forward-adjoint is able to produce a reconstruction that is nearly on par with the reconstruction obtained with the accurate operator and that

(a) Reconstructions for phantom close to the line detector



(b) Reconstructions for phantom far from the line detector

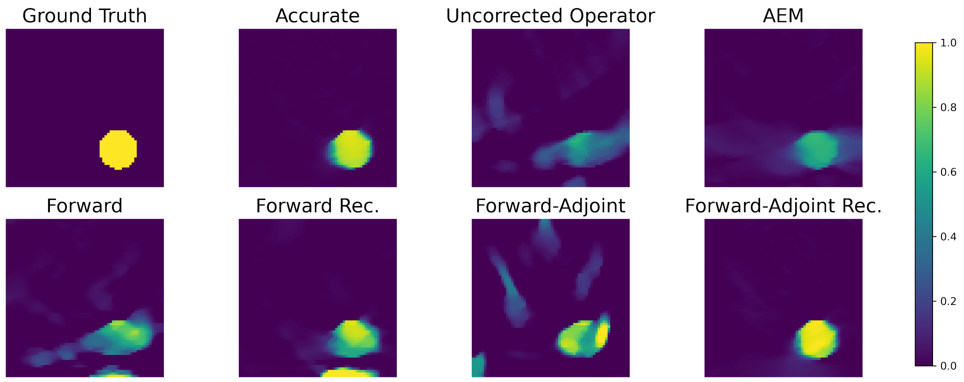
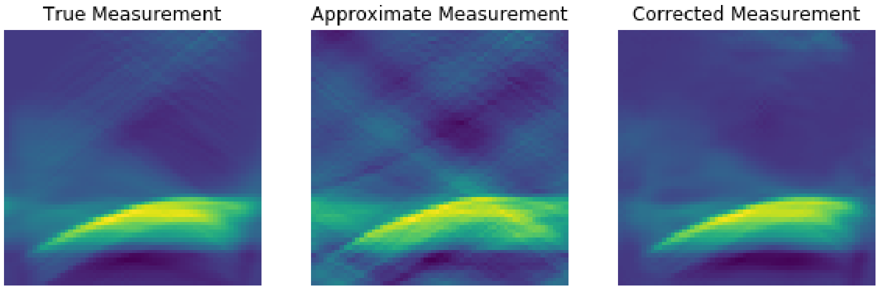


Figure 8. *Reconstructions for the various model correction algorithms for two samples from the ball set. We show the results after* 4000 *steps of gradient descent. Huber regularization is used. Top* (a)*: Phantom close to the detector, which corresponds to an easy setting for limited view PAT. Bottom* (b)*: Phantom far from the the detector, which corresponds to a very challenging setting.*

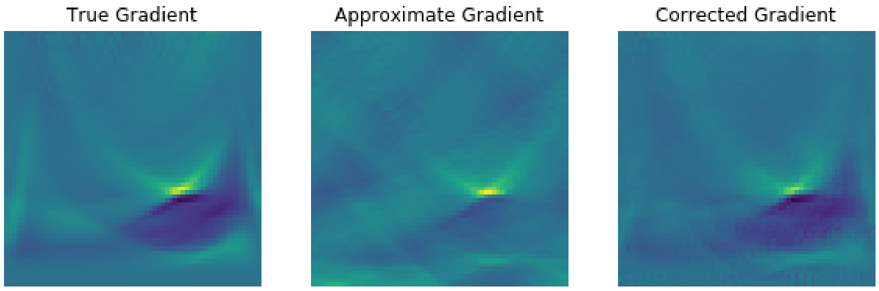
does not exhibit any obvious artefacts. The baseline with AEM also introduces background artefacts leaking from the ball, but those are more structured and less severe than those of all other methods apart from the forward-adjoint recursive approach which gives the best visual results in this setting as well. The visual quality of the reconstructions hence coincides with the quantitative results discussed in Figure 7.

Figure 9 visualizes the effect of the forward-adjoint recursive approach on the ball images, showing Ax_0 $\widetilde{A}x_0$, and $A_\Theta(x_0)$ as well as the gradients of the data term for each of the operators A, $\widetilde{A}$, and A_Θ. The visualizations are computed for sample (b) in Figure 8 on the ball samples. We see that the forward-adjoint approach is in fact able to correct for approximation artefacts *both* in the forward operator as well as in its adjoint, leading to a good approximation of the accurate gradient of the data term.

Vessel phantoms. The results on the vessel phantoms quantitatively match the overall behavior observed on the ball set. The alignment, as shown in Figure 10, is again initially higher



(a) Measurements computed at the initial reconstruction for the exact, approximate, and corrected operators. From left to right Ax_0, $\widetilde{A}x_0$, $A_\Theta(x_0)$.



(b) Gradients of the data term, computed at the initial reconstruction for the exact, approximate, and corrected operators.

Figure 9. *Estimated measurements and gradients at initialization of the gradient descent scheme for a sample from the ball images.*

with forward-adjoint methods achieving higher values as forward only methods. If no recursive training is applied, alignment declines very quickly. AEM is again generating gradients of comparatively low initial alignment, that however stays relatively steady throughout solving the variational problem. We note that the overall alignment is significantly lower than in the case of the ball samples, reflecting the additional difficulty of the vessel set.

The relative error of the reconstructions compared to the ground truth can be seen in Figure 11. We again see both the forward and forward-adjoint methods fail to improve reconstruction quality further early into the minimization process if recursive training is omitted. In case recursive training is applied, both methods lead to a clear improvement over the uncorrected operator, with the forward-adjoint approach again performing considerably better than the forward only. On the vessel samples we however note a considerably larger gap between the forward-adjoint correction and the accurate operator that is caused by the extremely challenging nature of the vessel set. The AEM baseline converges slowly on the vessels, an indication that the estimated covariance matrix is fairly ill-conditioned. We hence additionally report the reconstruction quality at convergence, which we observed after 20000 steps of gradient descent. While this is a competitive reconstruction, it is still outperformed slightly

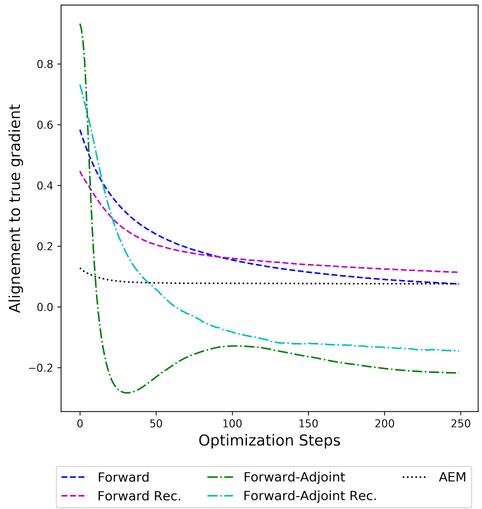


Figure 10. *Alignment* (6.1) *of approximate gradient to the gradient of the accurate data term* $A^*(Ax_n - y)$ *for each method on the vessel test set with* 64 *samples, recorded over the* 250 *steps of solving the associated variational problem.*

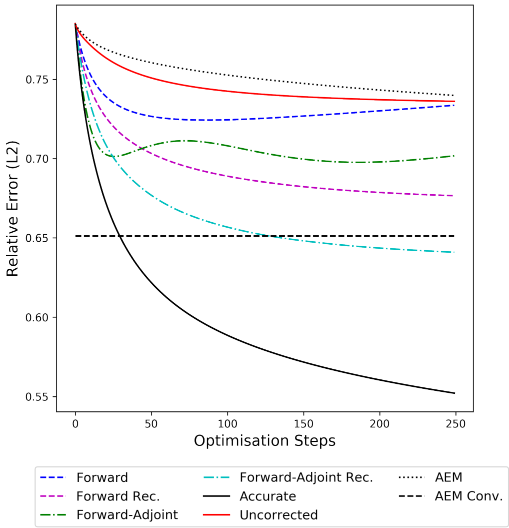


Figure 11. *Relative reconstruction error (L2) for all methods on the vessel test set with* 64 *samples, tracked throughout the gradient descent scheme.* 250 *steps of gradient descent were performed for all methods but AEM, where* 20000 *steps were taken.*

by the recursive forward-adjoint method. We remark that we have applied early stopping for all other methods on the vessel samples.

We present reconstructions for all discussed methods for two samples in Figure 12. We note for the first sample that the vessel structure at the right of the image completely disappears when using the uncorrected approximation. In fact, the corresponding measurement is severely reduced due to the thresholding of incident waves in the approximate model. Hence, no correction method is able to fully recover the vessel structure at the right of the first sample, with AEM, forward method, and forward-adjoint method coming closest. For all correction methods we observe a deterioration in reconstruction quality compared to the accurate operator. We note that the recursive forward method seems to lead to striping artefacts. Consistent with the quantitative results in Figure 11 the forward-adjoint recursive reconstructions are of the highest visual quality compared to the other reconstructions using a model correction, leading to sharper results than the AEM baseline and to fewer artefacts than methods based on the forward only approach or those omitting recursive training. We remark that, up to some extent, perceived differences in smoothness can also be caused as the regularization parameter has been optimized for all methods individually and hence might differ slightly between reconstructions.

To this end, we note that the training set with a total of 2760 samples (5520 with rotations) is fairly small when taking into account the complexity of the vessel structures; see, for instance, the discussion with respect to AEM in [35]. It is hence possible that the remaining gap in reconstruction quality to the accurate operator could be closed further by using a

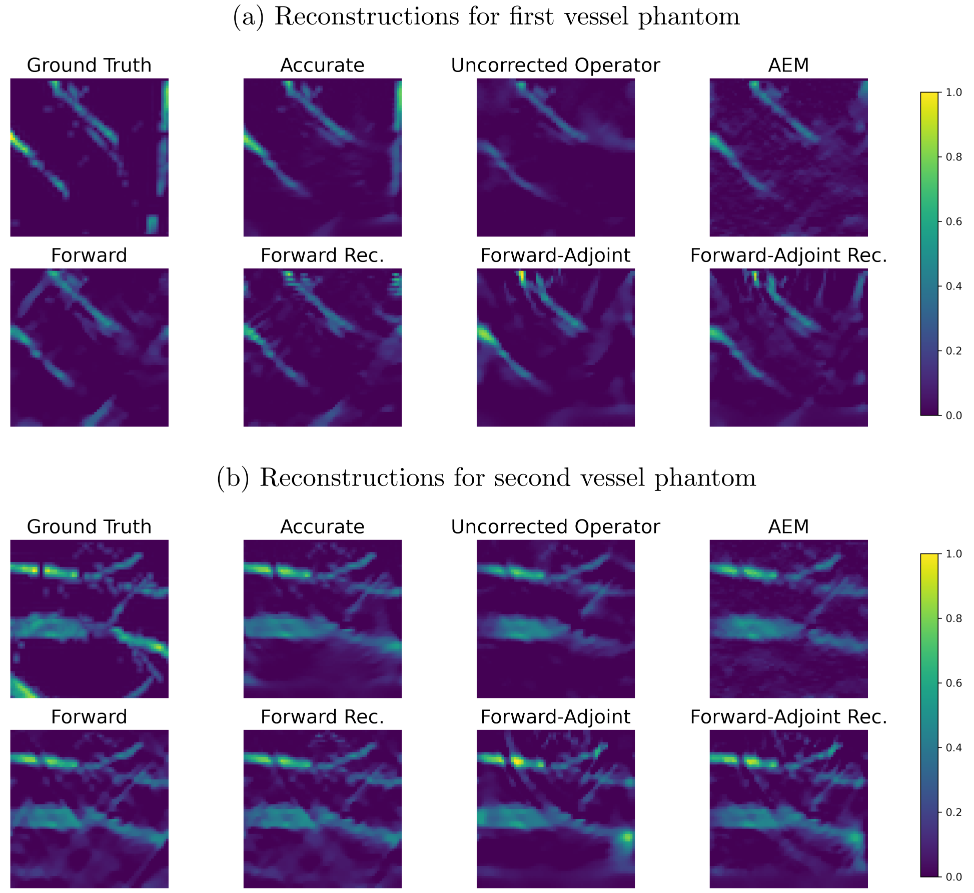


Figure 12. *Reconstructions on vessels using the various operator corrections. We show the results after* 250 *iterations of gradient descent for all methods but AEM, for which* 20000 *iteration steps were taken. Huber regularization is used.*

more extensive training set. However, we expect that the gap cannot be closed completely on samples with a complexity comparable to the vessel phantoms as too much information might be lost in the thresholding step of the approximate operator that cannot be recovered even when taking into account the structure of the samples with highly parameterized learned corrections. This underlines the necessity of a statistical correction as discussed throughout section 2 to compensate for lost kernel directions in the approximate operator.

Model transfer between vessel and ball phantoms. In this paragraph we investigate how well the operator corrections trained on either the ball or the vessel samples generalize to the other of the two data sets. In particular, we discuss using models trained on balls to reconstruct vessels and vice versa. The aim of these experiments is to obtain a first understanding on how well-trained model corrections generalize to new data sets in general, especially if the new set is very different from the training data in terms of image characteristics.

When using models trained on the ball samples and tested on vessel images, we notice that the model gives reasonable corrections at the initialization of the variational scheme for the

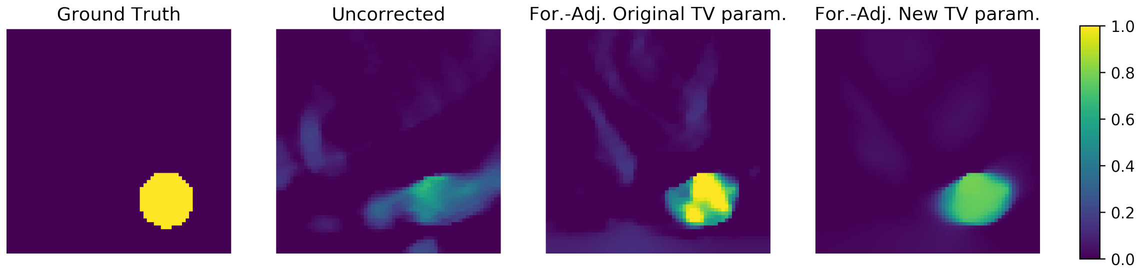


Figure 13. *Models trained on vessel samples, evaluated on ball samples. From left to right: Ground-truth image, reconstruction using the uncorrected operator, reconstruction using a recursive forward-adjoint correction with the same TV parameter as used on vessel data, reconstruction using a recursive forward-adjoint correction with new optimal TV parameter.*

vessel samples, yielding corrected gradients. Nevertheless, the correction quality deteriorated rapidly during the gradient descent steps and the final reconstruction was not satisfactory compared to reconstructions obtained with the uncorrected approximate operator $\widetilde{A}$. We hypothesize that the ball data were too distinct from the vessel samples and that the structure of the ball data were too simple for the learned model to perform reasonably on the much more complicated vessel data. In particular, the learned corrections were potentially fit very tightly to data and measurements induced by the ball phantoms that do not contain the same level of complexity as the vessel phantoms. Heuristically speaking, the data manifold of the ball samples seems to be too low dimensional to generalize to other data.

On the other hand, when using the forward-adjoint recursive model trained on the vessel samples on the ball samples, we obtained results that are clear improvements over reconstructions obtained with the uncorrected operator and are even comparable to the nonrecursively trained methods on the ball data. We do, however, not match the performance of the forward-adjoint recursive model trained on the ball samples themselves. Figure 13 shows reconstructions on a ball sample for various methods trained on the vessel samples. The reconstructions show a well-localized ball reconstruction with fairly sharp edges even in the challenging case of the ball sample located far from the detector plate. The results can be compared to results obtained with methods trained on the ball samples, as shown in Figure 8. The visual assessment of reconstruction quality matches the quantitative results in terms of L^2 error as shown in Table 2.

Finally, we note in both Figure 13 and Table 2 that adopting the regularization parameter λ of the forward-adjoint correction trained on vessel samples to a new optimal value for the ball data yields considerable improvements in performance. This demonstrates one of the main advantages of explicit corrections over their implicit counterparts, as separating between model correction and regularization allows for an adaption of the regularization parameter to the task, independently of the model correction learned.

7. Conclusion. In this paper, we have introduced various approaches to learn a data-driven explicit model correction for inverse problems to be employed within a variational reconstruction framework. We have investigated several strategies to learn such a correction, starting with a simple forward correction for which we pointed out some fundamental limita-

Table 2

Performance of the recursive forward-adjoint correction on ball samples. We evaluate the performance of models trained on vessel samples and compare to models trained on ball samples. Results are reported in terms of the L^2 error compared to the ground-truth image.

	Training data	L^2 error
Accurate operator	-	0.11
Approximate operator	-	0.55
Forward-adjoint	balls	0.15
For.-adj. (old TV param.)	vessels	0.40
For.-adj. (new TV param.)	vessels	0.35

tions. In particular, we observed that this approach is limited by the range of the adjoint of the approximate operator when employed in a gradient descent scheme and is therefore unable to fully correct all modeling errors. To mitigate this, we have proposed a forward-adjoint correction as an alternative approach, overcoming these limitations by fitting an independent adjoint correction.

To ensure a model correction that can be employed throughout the optimization process and avoid overfitting the initial reconstruction, we proposed to augment all methods with a recursive training scheme. For the recursive forward-adjoint correction we provided a theoretical convergence analysis to show that the method approximates the accurate solution when trained to a sufficiently low loss. Finally, we have shown the potential of our approach on the task of limited view PAT, demonstrating our theoretical considerations in practice and showing improved results compared to the commonly used AEM.

For the data chosen, the algorithm can be trained very quickly, requiring 12h for non-recursive experiments and around 16h for their recursive counterparts. For images larger than the 64×64 format used in the paper, the number of operations scales linearly with the number of pixels and hence quadratically with resolution in 2 dimensions and cubically in 3 dimensions. The actual increase in computational time might scale lower than the increase in operators as a larger number of operations per layer increases the potential for parallelization. The number of network parameters, however, does not necessarily change with resolution. Higher resolutions might make a deeper architecture appropriate, but the increase in weights caused by this would typically be strongly sublinear.

This work is orthogonal to previous attempts at using neural networks to learn operator corrections that were exclusively focused on the idea of implicit model corrections, learning the correction operator, and a reconstruction prior simultaneously in an end-to-end trained reconstruction network. While this approach comes with advantages in terms of performance, our explicit model correction allows us to flexibly use any prior model alongside the corrected operator and can be integrated in the well-established framework of variational regularization. Furthermore, our work unveils some of the challenges in model correction that are hidden in implicit schemes. Our findings can be used to inspire the design of novel implicit algorithms and allows for an analysis of implicit correction in future studies. In particular, our observations on the limitations of the range of the adjoint of the approximation motivates the use of corrections in both reconstruction and data space for implicit model correction, motivating the use of algorithms such as LPD [2].

In future work one could apply the proposed method to different fields of application, such as CT. In this application, the accurate model can be obtained by expensive photon-level Monte Carlo simulations, whereas a computationally efficient approximation is given by the widely used ray transform. In general, applications to inverse problems involving nonlinear operators are an interesting direction deserving further study; we refer to a related study exploring first ideas in this direction [40]. A class of very challenging applications are settings where we do not have explicit access to the accurate forward operator, but instead have access to empirical measurements only. Examples of such problems are tomography with slightly wrong estimated angles or deconvolution problems with errors in the point-spread function. These problems differ from the setting considered in this paper, where explicit access to the accurate operator was given and the approximation was performed to overcome computational constraints. In particular, the concept of recursive training, as presented here, requires explicit access to the accurate operator and is thus not readily applicable for problems where we have access to empirical measurements only, making them particularly challenging. We believe that in such settings, alternate training regimes that are not fully supervised and make use of secondary measures will be needed, estimating the approximation error from the data itself.

Finally, we mention a possible combination of the proposed approach with AEM techniques. Since the latter, after training, yields a multivariate normal distribution as an estimate of the distribution of model errors it becomes increasingly unreliable as the non-Gaussianity of the accurate distribution increases. However, after an initial nonlinear correction of the form A_Θ described here, the AEM could be reestimated using such a model. Commensurately, the estimated statistics of the model error from the AEM could be used in place of the simple L^2-loss used in the training in (5.2) and (5.3) for example (i.e., the norm implied in the space Y). A possible future research direction could therefore be to iterate these approaches with a view to obtaining a more accurate probabilistic estimate of the eventual remaining model errors.

Appendix A. An approximate model for photoacoustic tomography.

Here we discuss the accurate and approximate model as previously used in [20]. In PAT a short pulse of near-infrared light is absorbed by chromophores in biological tissue. For a sufficiently short pulse, the quantity of interest will result as a spatially varying pressure increase x, which will initiate an ultrasound pulse (*photoacoustic effect*), that then propagates to the tissue surface. The measurement consists of the detected waves in space-time at the boundary of the tissue; this set of pressure time series constitutes the measured photoacoustic data y.

For the forward model, this acoustic propagation is commonly modeled by an initial value problem for the wave equation [11],

$$(\partial_{tt} - c^2\Delta)p(\mathbf{x},t) = 0, \quad p(\mathbf{x},t=0) = x(\mathbf{x}), \quad \partial_t p(\mathbf{x},t=0) = 0, \text{ with } \mathbf{x} \in \mathbb{R}^2. \tag{A.1}$$

The measurement is then modeled as a linear operator $\mathcal{M}$ acting on the pressure field $p(\mathbf{x},t)$ restricted to the boundary of the computational domain Ω and a finite time window

$$y = \mathcal{M}\, p_{|\partial\Omega\times(0,T)}. \tag{A.2}$$

Together, (A.1) and (A.2) define the linear forward model that we consider in this study:

$$Ax = y \tag{A.3}$$

from initial pressure x to the measured time series y. This accurate forward model can be simulated by a pseudospectral time-stepping model as outlined in [42, 43].

For the approximate model, we can exploit the fact that in our case the measurement points lie on a line ($\mathbf{x}_2 = 0$) outside the support of x; the pressure there can be related to x by [11, 28]

$$p(\mathbf{x}_1, t) = \frac{1}{c^2}\mathcal{F}_{k_1}\left\{\mathcal{C}_\omega\left\{B(k_1,\omega)\tilde{x}(k_1,\omega)\right\}\right\}, \tag{A.4}$$

where $\tilde{x}(k_1,\omega)$ is obtained from $\hat{x}(k)$ via the dispersion relation $(\omega/c)^2 = k_1^2 + k_2^2$ and $\hat{x}(k) = \mathcal{F}_{\mathbf{x}}\{x(\mathbf{x})\}$ is the 2-dimensional Fourier transform of $x(\mathbf{x})$. $\mathcal{C}_\omega$ is a cosine transform from ω to t, $\mathcal{F}_{k_1}$ is the 1-dimensional inverse Fourier transform from k_1 to $\mathbf{x}_1$ on the detector. The weighting factor,

$$B(k_1,\omega) = \omega/\left(\operatorname{sgn}(\omega)\sqrt{(\omega/c)^2 - k_1^2}\right), \tag{A.5}$$

contains an integrable singularity which means that if (A.4) is evaluated by discretization on a rectangular grid (and thus enabling the application of FFT for efficient calculation), then aliasing in the measured data $p(\mathbf{x}_1, t)$ results. Consequently, evaluating (A.4) using FFT leads to a *fast but approximate* forward model. In fact, we can control the degree of aliasing, by avoiding the singularity, that means in practice all components of B for which $k_1^2 > (\omega/c)^2 \sin^2\theta_{\max}$ are set to zero. This is equivalent to assuming only waves arriving at angles up to $\theta_{\max}$ from normal incidence are detected. We note, that there is a trade-off: the greater the range of angles included, the greater the aliasing. Finally, this results in a thresholded weighting factor $\widetilde{B}$ and hence the relation (A.4) using $\widetilde{B}$ defines the approximate model for this study: $\widetilde{A}x = y$.

Appendix B. Addition to theoretical results. In this section, we only investigate the question of closeness of minimizers, without investigating if the minimizers of $\mathcal{L}_\Theta$—that involves a nonlinear operator in the data term—can be identified efficiently using a gradient descent based algorithm. To answer this question, we will assume that the learned corrected operator A_Θ approximates the ground-truth operator A sufficiently well, uniformly on some manifold $\mathcal{D}$ that contains the minimizer of $\mathcal{L}$. These assumptions represent the situation of a well-fit forward approximation on the data manifold $\mathcal{D}$ that we assume all relevant reconstructions to lie on.

While it is difficult to check these assumptions in practice, the purpose of this discussion is to give a more complete theoretical view of the problem at hand, demonstrating that under sufficient assumptions closeness of forward operators is sufficient to deduce closeness of minimizers. However, this does not guarantee that the minimum can be found with a gradient descent algorithm or that a gradient descent algorithm even stays on the manifold $\mathcal{D}$ of good approximation quality. As a theoretical underpinning for the experiments conducted in this paper, Theorem 4.9 should hence instead be considered as a the main theorem.

Proposition B.1 (proximity of minimizers). *Denote by $\mathcal{D} \subset X$ the manifold of possible reconstructions that the operator approximation was trained on using empirical risk mini-*

mization (4.1)*. Let* $\mathcal{L}$ *be strongly convex. Assume further that* $\mathcal{D}$ *and the measurement noise is bounded. Hence for any* $y = Ax_1 + \epsilon$ *for some* $x_1 \in \mathcal{D}$ *and for any* $x_2 \in \mathcal{D}$ *we have* $\|Ax_2 - y\|_Y \leq C$ *(boundedness). Let* $\epsilon > 0$*. Denote by* δ *the corresponding quantity as in Lemma* 4.4*; without loss of generality let* $\delta \leq 32C^2$*. Assume further that* A_Θ *has been trained such that* $\sup_{x\in\mathcal{D}} \|A_\Theta(x) - Ax\|_Y \leq \delta/4C$*. Denote by*

$$\hat{x} := \arg\min_x \mathcal{L}(x), \quad \hat{x}_\Theta \in \arg\min_x \mathcal{L}_\Theta(x),$$

the reconstructions computed via the variational problem using either the accurate operator A *or the corrected operator* A_Θ*, respectively. Note that the minimizer is unique for the functional* $\mathcal{L}$ *by strong convexity, but not necessarily for the functional* $\mathcal{L}_\Theta$*. Then for any* $y \in B_r(0)$ *that is such that both* $\hat{x}, \hat{x}_\Theta \in \mathcal{D}$*, we have*

$$\|\hat{x} - \hat{x}_\Theta\|_X < \delta.$$

Proof. First note that $|\mathcal{L}_\Theta(x) - \mathcal{L}(x)| \leq \delta/2$ for any $x \in \mathcal{D}$, as

$$\begin{aligned}|\mathcal{L}_\Theta(x) - \mathcal{L}(x)| &= \frac{1}{2}\left|\|A_\Theta(x) - y\|_Y^2 - \|Ax - y\|_Y^2\right| \\ &\leq \|Ax - y\|_Y \|Ax - A_\Theta(x)\|_Y + \frac{1}{2}\|Ax - A_\Theta(x)\|_Y^2 \\ &\leq C \cdot \frac{\delta}{4C} + \delta\frac{1}{2}\frac{32C^2}{(8C)^2} = \delta/2\end{aligned}$$

By taking the minimum, this in particular implies that

$$|\mathcal{L}_\Theta(\hat{x}_\Theta) - \mathcal{L}(\hat{x})| \leq \delta/2.$$

We conclude via

$$|\mathcal{L}(\hat{x}_\Theta) - \mathcal{L}(\hat{x})| \leq \ |\mathcal{L}(\hat{x}_\Theta) - \mathcal{L}_\Theta(\hat{x}_\Theta)| + |\mathcal{L}_\Theta(\hat{x}_\Theta) - \mathcal{L}(\hat{x})| \leq \ \delta/2 + \delta/2 = \delta,$$

which finishes the proof using (4.5). ■

Remark B.2. The assumption that y is such that $\hat{x}, \hat{x}_\Theta \in \mathcal{D}$, can be interpreted as a necessity for y to have emerged from an underlying image that is close to the manifold of reconstructions $\mathcal{D}$ that the correction A_Θ has been trained on. Put differently, we require y to be an actual realistic measurement, similar to those used to train the model correction A_Θ.

Acknowledgments. The authors acknowledge helpful discussions with Jonas Adler, Jari Kapio, Yury Korolev, Ozan Öktem, and Peter Maass amongst others.

REFERENCES

[1] J. Adler and O. Öktem, *Solving ill-posed inverse problems using iterative deep neural networks*, Inverse Problems, 33 (2017), 124007.

[2] J. Adler and O. Öktem, *Learned primal-dual reconstruction*, IEEE Trans. Med. Imaging, 37 (2018), pp. 1322–1332.

[3] S. Antholzer, M. Haltmeier, and J. Schwab, *Deep learning for photoacoustic tomography from sparse data*, Inverse Prob. Sci. Eng., 27 (2019), pp. 987–1005.
[4] S. Arridge, J. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen, *Approximation errors and model reduction with an application in optical diffusion tomography*, Inverse Problems, 22 (2006), pp. 175–195.
[5] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, *Solving inverse problems using data-driven models*, Acta Numer., 28 (2019), pp. 1–174.
[6] P. Beard, *Biomedical photoacoustic imaging*, Interface Focus, 1 (2011), pp. 602–631.
[7] Y. E. Boink and C. Brune, *Learned SVD: Solving Inverse Problems via Hybrid Autoencoding*, preprint, https://arxiv.org/abs/1912.10840 (2019).
[8] L. Borcea, V. Druskin, A. Mamanov, S. Moskow, and M. Zaslvsky, *Reduced order models for spectral domain inversion: Embedding into the continuous problem and generation of internal data*, Inverse Problems, 36 (2020), 055010.
[9] T. A. Bubba, G. Kutyniok, M. Lassas, M. Maerz, W. Samek, S. Siltanen, and V. Srinivasan, *Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography*, Inverse Problems, 35 (2019), 064002.
[10] M. Burger, Y. Korolev, and J. Rasch, *Convergence rates and structure of solutions of inverse problems with imperfect forward models*, Inverse Problems, 35 (2019), 024006.
[11] B. Cox and P. Beard, *Fast calculation of pulsed photoacoustic fields in fluids using k-space methods*, J. Acoust. Soc. Amer., 117 (2005), pp. 3616–3627.
[12] N. Davoudi, X. L. Deán-Ben, and D. Razansky, *Deep learning optoacoustic tomography with sparse data*, Nature Mach. Intell., 1 (2019), pp. 453–460.
[13] H. Egger, J.-F. Pietschmann, and M. Schlottbom, *Identification of chemotaxis models with volume-filling*, SIAM J. Appl. Math., 75 (2015), pp. 275–288.
[14] R. W. Freund, *Model reduction methods based on Krylov subspaces*, Acta Numer., 12 (2003), pp. 267–319.
[15] S. Guan, A. Khan, S. Sikdar, and P. Chitnis, *Fully dense UNet for 2D sparse photoacoustic tomography artifact removal*, IEEE J. Biomed. Health Inform., 24 (2020), pp. 568–576.
[16] S. J. Hamilton and A. Hauptmann, *Deep D-bar: Real-time electrical impedance tomography imaging with deep neural networks*, IEEE Trans. Med. Imaging, 37 (2018), pp. 2367–2377.
[17] 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*, Magn. Reson. Med., 79 (2018), pp. 3055–3071.
[18] A. Hauptmann, J. Adler, S. Arridge, and O. Öktem, *Multi-scale learned iterative reconstruction*, IEEE Trans. Comput. Imaging, to appear.
[19] A. Hauptmann, S. Arridge, F. Lucka, V. Muthurangu, and J. A. Steeden, *Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease*, Magn. Reson. Med., 81 (2019), pp. 1143–1156.
[20] A. Hauptmann, B. Cox, F. Lucka, N. Huynh, M. Betcke, P. Beard, and S. Arridge, *Approximate k-space models and deep learning for fast photoacoustic reconstruction*, in International Workshop on Machine Learning for Medical Image Reconstruction, Springer, Cham, Switzerland, 2018, pp. 103–111.
[21] A. Hauptmann, F. Lucka, M. Betcke, N. Huynh, J. Adler, B. Cox, P. Beard, S. Ourselin, and S. Arridge, *Model based learning for accelerated, limited-view 3d photoacoustic tomography*, IEEE Trans. Med. Imaging, 39 (2018), pp. 1382–1393.
[22] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, *Deep convolutional neural network for inverse problems in imaging*, IEEE Trans. Image Process., 26 (2017), pp. 4509–4522.
[23] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer, New York, 2005.
[24] J. Kaipio and E. Somersalo, *Statistical inverse problems: Discretization, model reduction and inverse crimes*, J. Comput. Appl. Math., 198 (2007), pp. 493–504.
[25] E. Kang, J. Min, and J. C. Ye, *A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction*, Med. Phys., 44 (2017), pp. e360–e375.
[26] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. R. Stat. Soc. Ser. B Stat. Methodol., 63 (2001), pp. 425–464.

[27] E. Kobler, A. Effland, K. Kunisch, and T. Pock, *Total deep variation for linear inverse problems*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Piscataway, NJ, 2020, pp. 7549–7558.
[28] K. P. Koestli, M. Frenz, H. Bebie, and H. P. Weber, *Temporal backward projection of optoacoustic pressure transients using Fourier transform methods*, Phys. Med. Biol., 46 (2001), pp. 1863–1872.
[29] Y. Korolev and J. Lellmann, *Image reconstruction with imperfect forward models and applications in deblurring*, SIAM J. Imaging Sci., 11 (2018), pp. 197–218.
[30] P. Kuchment and L. Kunyansky, *Mathematics of thermoacoustic tomography*, European J. Appl. Math., 19 (2008), pp. 191–224.
[31] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier, *Nett: Solving inverse problems with deep neural networks*, Inverse Problems, 36 (2020), 065005.
[32] A. Lorz, J.-F. Pietschmann, and M. Schlottbom, *Parameter identification in a structured population model*, Inverse problems, 35 (2019), 095008.
[33] S. Lunz, O. Öktem, and C.-B. Schönlieb, *Adversarial regularizers in inverse problems*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2018, pp. 8507–8516.
[34] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, Switzerland, 2015, pp. 234–241.
[35] T. Sahlström, A. Pulkkinen, J. Tick, J. Leskinen, and T. Tarvainen, *Modeling of errors due to uncertainties in ultrasound sensor locations in photoacoustic tomography*, IEEE Trans. Med. Imaging, 39 (2020), pp. 2140–2150.
[36] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, *A deep cascade of convolutional neural networks for dynamic MR image reconstruction*, IEEE Trans. Med. Imaging, 37 (2017), pp. 491–503.
[37] J. Schwab, S. Antholzer, and M. Haltmeier, *Deep null space learning for inverse problems: Convergence analysis and rates*, Inverse Problems, 35 (2019), 025008.
[38] J. H. Siewerdsen and D. A. Jaffray, *Cone-beam computed tomography with a flat-panel imager: Magnitude and effects of X-ray scatter*, Med. Phys., 28 (2001), pp. 220–231.
[39] D. Smyl and D. Liu, *Less is often more: Applied inverse problems using hp-forward models*, J. Comput. Phys., 399 (2019), 108949.
[40] D. Smyl, T. N. Tallman, J. A. Black, A. Hauptmann, and D. Liu, *Learning and Correcting Non-Gaussian Model Errors*, preprint, https://arxiv.org/abs/2005.14592 (2020).
[41] T. Tarvainen, A. Pulkkinen, B. T. Cox, J. P. Kaipio, and S. R. Arridge, *Bayesian image reconstruction in quantitative photoacoustic tomography*, IEEE Trans. Med. Imaging, 32 (2013), pp. 2287–2298.
[42] B. E. Treeby and B. T. Cox, *k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields*, J. Biomed. Opt., 15 (2010), 021314.
[43] B. E. Treeby, J. Jaros, A. P. Rendell, and B. T. Cox, *Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method.*, J. Acoust. Soc. Amer., 131 (2012), pp. 4324–4336, https://doi.org/10.1121/1.4712021.
[44] V. Vishnevskiy, R. Rau, and O. Goksel, *Deep variational networks with exponential weighting for learning computed tomography*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, Switzerland, 2019, pp. 310–318.
[45] L. W. Y Xu, G. Ambartsoumian, and P. Kuchment, *Reconstructions in limited-view thermoacoustic tomography*, Med. Phys., 31 (2004), pp. 724–733.
[46] E. Zhang, J. Laufer, and P. Beard, *Backward-mode multiwavelength photoacoustic scanner using a planar Fabry-Perot polymer film ultrasound sensor for high-resolution three-dimensional imaging of biological tissues*, Appl. Opt., 47 (2008), pp. 561–577.
[47] L. Zhu, Y. Xie, J. Wang, and L. Xing, *Scatter correction for cone-beam CT in radiation therapy*, Med. Phys., 36 (2009), pp. 2258–2268.