Iterated preconditioned LSQR method for inverse problems on unstructured grids

2014 Inverse Problems 30 075009

(http://iopscience.iop.org/0266-5611/30/7/075009)

OPEN ACCESS
IOP Publishing

Made open access 16 July 2014
Inverse Problems

Inverse Problems **30** (2014) 075009 (27pp)
doi:10.1088/0266-5611/30/7/075009

Iterated preconditioned LSQR method for inverse problems on unstructured grids

S R Arridge[1], **M M Betcke**[1] **and L Harhanen**[2]

[1] Department of Computer Science, University College London, London, UK
[2] Department of Engineering Design and Production, Aalto University, PO Box 14200, FI-00076 Aalto, Finland

E-mail: Simon.Arridge@cs.ucl.ac.uk, M.Betcke@ucl.ac.uk and lauri.harhanen@aalto.fi

Received 19 December 2013, revised 14 April 2014
Accepted for publication 12 May 2014
Published 25 June 2014

Abstract

This article presents a method for solving large-scale linear inverse imaging problems regularized with a nonlinear, edge-preserving penalty term such as total variation or the Perona–Malik technique. Our method is aimed at problems defined on unstructured meshes, where such regularizers naturally arise in unfactorized form as a stiffness matrix of an anisotropic diffusion operator and factorization is prohibitively expensive. In the proposed scheme, the nonlinearity is handled with lagged diffusivity fixed point iteration, which involves solving a large-scale linear least squares problem in each iteration. Because the convergence of Krylov methods for problems with discontinuities is notoriously slow, we propose to accelerate it by means of priorconditioning (Bayesian preconditioning). Priorconditioning is a technique that, through transformation to the standard form, embeds the information contained in the prior (Bayesian interpretation of a regularizer) directly into the forward operator and thence into the solution space. We derive a factorization-free preconditioned LSQR algorithm (MLSQR), allowing implicit application of the preconditioner through efficient schemes such as multigrid. The resulting method is also matrix-free i.e. the forward map can be defined through its action on a vector. We illustrate the performance of the method on two numerical examples. Simple 1D-deblurring problem serves to visualize the discussion throughout the paper. The effectiveness of the proposed numerical scheme is demonstrated on a three-dimensional problem in fluorescence diffuse optical tomography with total variation regularization derived algebraic multigrid preconditioner, which is the type of large scale, unstructured mesh problem, requiring matrix-free and factorization-free approaches that motivated the work here.

Keywords: ill-posed inverse problems, inhomogeneous diffusion, priorconditioning, Krylov methods, LSQR

(Some figures may appear in colour only in the online journal)

1. Introduction

Inverse problems arise in almost all fields of science when details of a model have to be determined from a set of observed data. Formally, we consider a mapping $A: X \to Y$ as the *forward problem* and the inversion of this mapping as the *inverse problem*. A defining characteristic of such problems is that they are *ill-posed* whenever the forward mapping is a compact operator which, for infinite-dimensional linear operators, is equivalent to a gradual decay to zero for the singular values of the mapping A. As a consequence of the ill-posedness of A, its inversion is unstable and requires *regularization*, see e.g. [1–3].

In this work, we are interested in linear problems after appropriate discretization, where $X = \mathbb{R}^{n_X}$, $n_X \in \mathbb{N}$ is the space of images (*large-scale* for two or three spatial dimensions) and $Y = \mathbb{R}^{n_Y}$, $n_Y \in \mathbb{N}$ is the data space. Broad class of such problems includes image deblurring, denoising and inpainting, tomography based on the radon transform (x-ray CT) or attenuated radon transform (PET, SPECT), or fluorescence diffuse optical tomography (fDOT).

For many of those problems, the matrix representation of A becomes too large to form explicitly, even for moderately sized images. Therefore in this paper we focus on methods using an implicit representation of A, which is sometimes referred to as the *matrix-free* approach. Examples are projection operators in x-ray CT, PET and SPECT, and the solution of direct and adjoint partial differential equations in fDOT.

1.1. Linear problem with nonlinear edge-preserving regularizer

The problem we aim to solve is of the form

$$\min_{f \in X} \left[\Phi(f) := \frac{1}{2} \left\| g - Af \right\|^2 + \tau R(f) \right], \tag{1}$$

where A is a discrete forward operator (possibly defined only as its action on a vector), g denotes the data, f the unknown solution and $R(f)$ is an appropriate discretization of a nonlinear regularizer of the type

$$\mathcal{R}(f) := \int_\Omega r\left(\left| \nabla f(x) \right| \right) \mathrm{d}x, \tag{2}$$

with $r \in C^1\left(\left[0, \infty\right) \right)$ and $\tau > 0$ controls the relative weight of the data fidelity and the regularization term in (1). The function r acting on $|\nabla f(x)|$ makes such regularizers of immediate interest to imaging problems, where preservation of the edges is desirable. Two prominent examples are

- total variation [4] (sparsity prior for edges) and its smoothed approximation

$$r(t) = t\,, \quad \text{or} \quad r(t) = T\sqrt{1 + \left(t/T\right)^2}\,, \tag{3}$$

- Perona–Malik [5] (edge-preserving prior)

$$r(t) = \frac{1}{2}T^2 \log\left(1 + \left(t/T\right)^2\right), \quad \text{or} \quad r(t) = \frac{1}{2}T^2\left(1 - \exp\left(-t^2/T^2\right)\right). \tag{4}$$

In (4), T is a threshold parameter indicating a level of image structure below which edges are considered as noise; we apply the same interpretation to this parameter in (3) to illustrate the generic approach, even though in the total variation literature this factor is usually stated as a purely numerical correction.

1.2. Lagged diffusivity fixed point iteration

In this paper we solve (1) by searching for a critical point, $\frac{\mathrm{d}\Phi}{\mathrm{d}f} = 0$, which amounts to solution of the nonlinear normal equation

$$\frac{\mathrm{d}\Phi}{\mathrm{d}f} = A^{\mathrm{T}}\left(Af - g\right) + \tau R'\left(f\right) = 0. \tag{5}$$

In the continuous setting, the Fréchet derivative of $\mathcal{R}$ at f in the direction h is defined as

$$\mathcal{R}'\left(f\right)h := \int_{\Omega} \frac{r'\left(\left|\nabla f\left(x\right)\right|\right)}{\left|\nabla f\left(x\right)\right|} \nabla f\left(x\right) \cdot \nabla h\left(x\right) \mathrm{d}x, \tag{6}$$

which is the weak form of the inhomogeneous diffusion operator

$$\mathcal{M}_f := -\nabla \cdot c_f\left(x\right) \nabla \tag{7}$$

acting on f with $c_f\left(x\right) = \frac{r'\left(\left|\nabla f\left(x\right)\right|\right)}{\left|\nabla f\left(x\right)\right|}$. With M_f representing a discretization of $\mathcal{M}_f$ we define the discrete derivative as $R'\left(f\right) := M_f f$.

As the nonlinearity is restricted to the regularizer, we solve (5) using the lagged diffusivity fixed point iteration

$$A^{\mathrm{T}}\left(Af_k - g\right) + \tau M_{f_{k-1}} f_k = 0, \tag{8}$$

introduced in [6] originally for total variation image denoising. This approach amounts to using successive linearizations where at kth nonlinear iteration step, we use the currently available approximation f_{k-1} to form the diffusion operator $M_{f_{k-1}}$. An alternative derivation of (8) can be found in [7].

Using symmetric factorization $M_{f_{k-1}} = L^{\mathrm{T}}L$, we can formally write the corresponding linear regularizer as $\frac{1}{2}\|Lf\|^2$ and hence the corresponding linear least squares problem reads

$$\min_{f \in X}\left[\Phi_k\left(f\right) := \frac{1}{2} \parallel g - Af \parallel^2 + \tau \frac{1}{2} \parallel Lf \parallel^2\right]. \tag{9}$$

1.3. Bayesian interpretation

It is well known that the minimizer of the Tikhonov regularized least squares problem (1) can be interpreted as the maximizer of the *posterior* distribution [8]

$$\pi\left(f|g\right) \propto \pi_Y\left(g|f\right)\pi_X\left(f\right), \tag{10}$$

where $\pi_Y\left(g|f\right)$ denotes the *likelihood* of observing data g for a known value of the parameter f and $\pi_X\left(f\right)$ is the measurement independent *prior*.

Under the assumption of zero-mean white Gaussian noise (i.e. noise covariance has already been incorporated into A and g), the unnormalized likelihood becomes

$$\pi_Y\left(g|f\right) \propto \exp\left\{-\frac{1}{2}\|g - Af\|^2\right\}.$$

For image reconstruction problems, choosing a prior π_X modelling the distribution of images with weak or vanishing correlation across edges

$$\pi_X\left(f\right) \propto \exp\left\{-\tau R\left(f\right)\right\},$$

we arrive at (1). The linearized problem (9) can be interpreted as choosing a zero-mean Gaussian prior with a (possibly improper) covariance C_X defined via $C_X^{-1} = \tau M_{f_{k-1}} = \tau L^{\mathrm{T}} L$.

The linear problem with nonlinear edge-preserving regularizer admits a Bayesian interpretation in terms of hyperpriors [9]. Hyperprior models the amplitude of the edges as a random variable, which estimation becomes a part of the inverse problem. A method alternatingly updating the parameters of primary interest and hyperparameters results in a lagged diffusivity-type method. In particular, choosing Gamma or inverse Gamma distribution as the hyperprior results in a lagged diffusivity iteration with TV or Perona–Malik regularizer, respectively.

1.4. Overview of the contribution

We present a *matrix and factorization-free* algorithm for *large-scale* linear inverse problems with *nonlinear regularizers*. The nonlinearity is limited to the regularization term, and is handled with lagged diffusivity fixed point iteration [6], which amounts to repeated solution of the problem linearized at the previous iterate.

We derive a new accelerated LSQR algorithm, MLSQR, for solution of linearized least squares problems. MLSQR is analytically equivalent to LSQR applied to the preconditioned least square problem with the preconditioner L^{-1}, $M_{f_{k-1}} = L^{\mathrm{T}} L$, but in MLSQR the preconditioner is applied implicitly avoiding the costly factorization of $M_{f_{k-1}}$ and at the same time allowing for using efficient algorithms such as multigrid methods for applying the preconditioner.

Preconditioning with L^{-1} is equivalent to transformation to the standard form and recently has also been termed priorconditioning [8–12] due to its interpretation as whitening of the prior in Bayesian framework. Using GSVD decomposition, we show that the Krylov subspace corresponding to the priorconditioned problem (problem in standard form) indeed amplifies the directions spanning the covariance of the prior, in contrast to the Krylov subspace for the original regularized problem. This explains the observed accelerated convergence for problems with discontinuities when priorconditioning with regularizers of the form (2).

The remainder of this paper is organized as follows. The following two sections deal with the solution of the linear least square problems arising from the linearization of the normal equations in each step of the lagged diffusivity iteration. In section 2, we compare the Krylov space solvers applied to the linearized problem in the original form (unpriorconditioned problem) and standard form (priorconditioned problem). We discuss various benefits of priorconditioning and we demonstrate the superiority of the Krylov space spanned by the solver for the priorconditioned problem. Section 3 derives the MLSQR algorithm, for matrix and factorization-free solution of the preconditioned least squares problem (14). Section 4 discusses the nonlinear solver based on successive linearizations and subspace

priorconditioning for solution of the nonlinear regularized problem (1). The performance of the method is demonstrated on a three-dimensional fDOT problem in section 5. We conclude with a summary of the results and a discussion of prospective research directions.

2. Priorconditioning of Krylov spaces

In this section we discuss solution of the linearized regularized least squares problem (9) using Krylov solvers. In particular we compare the effect of Krylov solvers applied to the original (unpriorconditioned) problem (9) and the problem in standard form (priorconditioned problem) (14).

2.1. Krylov methods for linear inverse problems

The regularized least squares problem (9) can be rewritten as

$$\min_{f\in X} \left\| \begin{bmatrix} A \\ \sqrt{\tau}L \end{bmatrix} f - \begin{bmatrix} g \\ 0 \end{bmatrix} \right\|, \tag{11}$$

which can be solved with the LSQR algorithm [13, 14]. LSQR is analytically equivalent to CG method applied to the corresponding *regularized (but unpriorconditioned)* normal equation

$$\left(A^{\mathrm{T}}A + \tau M\right) f = A^{\mathrm{T}} g, \tag{12}$$

and solves (11) (and equivalently (9)) over the Krylov space

$$\mathcal{K}^{A^{\mathrm{T}}A+\tau M} = \mathrm{span}\left\{ A^{\mathrm{T}}g, \left(A^{\mathrm{T}}A + \tau M\right)A^{\mathrm{T}}g, \ldots, \left(A^{\mathrm{T}}A + \tau M\right)^{i_{\max}-1} A^{\mathrm{T}}g \right\}. \tag{13}$$

For readability in this and the next section which deal with the linearized problem we drop the subscript indicating the connection to the nonlinear problem i.e $M := M_{f_{k-1}}$.

We note the following difficulties with methods based on the space $\mathcal{K}^{A^{\mathrm{T}}A+\tau M}$:

- If the regularization functional is designed to promote edges, solutions exhibit slow convergence due to the slow build up of high frequencies. As an example the operator considered in section 2.4 is a convolution which eigenvectors are Fourier-type modes with the oscillation frequency growing as the corresponding eigenvalue decreases. Such a modal basis provides only a poor approximation for discontinuities, which is the reason why Krylov methods converge slowly for the deconvolution of functions with jumps.
- Regularization can be controlled either through the parameter τ or the limit on the dimension $i_{\max}$ of the solution subspace in (13). While the truncation index $i_{\max}$ is usually determined implicitly within an iterative algorithm by the choice of a stopping rule, change of τ requires a recomputation of the subspace (13) from scratch.

2.2. Solution space priorconditioning

Assuming L is invertible, the change of basis

$$\hat{f} = Lf, \quad \hat{A} = AL^{-1},$$

transforms (9) into the *standard form* [15–17]

$$\min_{\hat{f}\in LX}\left[\hat{\Phi}\left(\hat{f}\right) := \frac{1}{2}\,\|\,g - \hat{A}\hat{f}\,\|^2 + \tau\frac{1}{2}\,\|\,\hat{f}\,\|^2\right]. \tag{14}$$

Since the transformed variable $\hat{f}$ has now standard multivariate normal distribution, it is sometimes referred to as *whitened* in the signal processing literature. In Bayesian context, transformation to standard form is also referred to as *priorconditioning* [8–12].

The normal equation corresponding to the *priorconditioned* problem (14) is exactly the normal equation (12) with symmetrically (split) preconditioning

$$\left(L^{-\mathrm{T}}A^{\mathrm{T}}AL^{-1} + \tau I\right)\hat{f} = L^{-\mathrm{T}}A^{\mathrm{T}}g, \qquad f = L^{-1}\hat{f}\,. \tag{15}$$

Hence, the solution space spanned by LSQR applied to (14) is

$$\mathcal{K}^{M^{-1}A^{\mathrm{T}}A} = \mathrm{span}\left\{M^{-1}A^{\mathrm{T}}g,\, M^{-1}A^{\mathrm{T}}AM^{-1}A^{\mathrm{T}}g,\, \ldots,\, \left(M^{-1}A^{\mathrm{T}}A\right)^{i_{\max}-1}M^{-1}A^{\mathrm{T}}g\right\}.$$

Note that for the priorconditioned formulation, the solution f is contained in $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ while the actual subspace built by the Krylov solver for the transformed solution $\hat{f}$ is $L\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$. Due to the shift invariance of the Krylov spaces, $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A} = \mathcal{K}^{M^{-1}A^{\mathrm{T}}A+\tau I}$ for any choice of τ. Hence, in the preconditioned normal equation (15), τ acts solely as a damping parameter.

Application of LSQR to (14) instead of (9) has a number of benefits:

- Local structure of the prior is embedded directly in the transformed operator $\hat{A}$ allowing quick convergence, see discussion in section 2.3.
- The corresponding Krylov space is the same for all values of τ due to translation invariance.
- As the prior is embedded into the transformed operator, it is possible to exploit the prior even without an explicit regularization term, i.e. setting $\tau = 0$.
- The transformed variables are *dimensionless*, which removes issues with physical units if further transformations, e.g. exponentiation, are used.

2.3. Accelerating convergence of LSQR

For well-conditioned problems preconditioning is an established technique to accelerate the convergence of Krylov subspace methods through clustering the eigenvalues of the transformed problem. For ill-posed inverse problems, the mapping A is a discretization of a compact operator which singular values accumulate at zero, rendering such clustering impossible. Therefore, some different form of preconditioning is necessary, cf [18–23] for some earlier work.

In this article we advocate acceleration by priorconditioning [8, 10–12] which *embeds the prior information like discontinuities directly into the forward operator*, which would otherwise require a large number of iterations to be built up in the Krylov subspace. This information is embedded in all the eigenvectors even those corresponding to the large eigenvalues and hence into the Krylov subspace from early on. As a result the priorconditioned solver requires only a few iterations to converge. Technically, priorconditioning can be achieved through preconditioning with $M^{-1} = C_X$, a preconditioner based on the properties of the expected solution rather than those of the forward mapping. In contrast to preconditioning, no clustering of eigenvalues takes place. On the contrary, the condition number of the priorconditioned matrix is usually larger than that of the original matrix.

In an attempt to shed light on the phenomena we express the corresponding Krylov spaces, $\mathcal{K}^{A^{\mathrm{T}}A+\tau M}$ and $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ in terms of the general singular value decomposition (GSVD) of the pair (A,L). The GSVD of (A, L) for L square and of the same column dimension as A can be written in the following simplified form

$$A = U\Gamma X^{-1}, \quad L = VX^{-1} \tag{16}$$

with orthogonal matrices $V^{\mathrm{T}}V = I$, $U^{\mathrm{T}}U = I$. Here we use the following, nonstandard scaling of the GSVD vectors X:

$$X^{\mathrm{T}}A^{\mathrm{T}}AX = \Gamma^2, \quad X^{\mathrm{T}}L^{\mathrm{T}}LX = I,$$

with Γ a matrix with the generalized singular values on the diagonal, which up to the ordering are the singular values of the matrix AL^{-1}. Note, that by construction X simultaneously diagonalize $A^{\mathrm{T}}A$ and $L^{\mathrm{T}}L$.

With the above defined GSVD matrices we have

$$A^{\mathrm{T}}A + \tau L^{\mathrm{T}}L = X^{-\mathrm{T}}\left(\Gamma^2 + \tau I\right)X^{-1}, \quad L^{-\mathrm{T}}A^{\mathrm{T}}AL^{-1} = V\Gamma^2V^{\mathrm{T}},$$

and the corresponding Krylov spaces can now be written as

$$\mathcal{K}^{A^{\mathrm{T}}A+\tau M} = \operatorname{span}\left\{A^{\mathrm{T}}g, \left(X^{-\mathrm{T}}\left(\Gamma^2 + \tau I\right)X^{-1}\right)A^{\mathrm{T}}g, \ldots, \left(X^{-\mathrm{T}}\left(\Gamma^2 + \tau I\right)X^{-1}\right)^{i_{\max}-1}A^{\mathrm{T}}g\right\} \tag{17}$$

and

$$\begin{aligned}\mathcal{K}^{M^{-1}A^{\mathrm{T}}A} &= L^{-1}\operatorname{span}\left\{L^{-\mathrm{T}}A^{\mathrm{T}}g, \left(V\Gamma^2\ V^{\mathrm{T}}\right)L^{-\mathrm{T}}A^{\mathrm{T}}g, \ldots, \left(V\Gamma^{2(i_{\max}-1)}V^{\mathrm{T}}\right)L^{-\mathrm{T}}A^{\mathrm{T}}g\right\} \\ &= \operatorname{span}\left\{XX^{\mathrm{T}}A^{\mathrm{T}}g, \left(X\Gamma^2X^{\mathrm{T}}\right)A^{\mathrm{T}}g, \ldots, \left(X\Gamma^{2(i_{\max}-1)}X^{\mathrm{T}}\right)A^{\mathrm{T}}g\right\}.\end{aligned} \tag{18}$$

Here we made use of $M^{-1} = L^{-1}L^{-\mathrm{T}} = XV^{\mathrm{T}}VX^{\mathrm{T}} = XX^{\mathrm{T}}$.

Recall that the covariance of the prior $C_X = \tau^{-1}M^{-1} = \tau^{-1}XX^{\mathrm{T}}$. Hence the generalized singular vectors X span the covariance matrix C_X of the prior (even though X are not orthogonal). Inspecting Krylov vectors in (18) reveals that the subspace $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ is dominated by the features of the prior spanned by those x_i corresponding to the large generalized singular values while those spanned by x_i corresponding to the small generalized singular values get damped.

On the other hand, we observe that $X^{-\mathrm{T}}$ spans the range of the anisotropic diffusion matrix $M = \tau^{-1}C_X^{-1} = X^{-\mathrm{T}}X^{-1}$. The GSVD form of (17) reveals that each Krylov iteration boosts the directions in the range of the regularizer M spanned by $X^{-\mathrm{T}}$ corresponding to the large singular values while damping those corresponding to the small ones with respect to the last Krylov vector; however there is no direct relation to the initial vector as in case of (18).

The GSVD has been previously used to describe the Krylov spaces above in [24] but no connection was made to the covariance of the prior. In [24], a projection method is derived for solving the linear problem (9) in case L is not invertible, where the projection subspace comes from joint bidiagonalization of A and L. Interestingly, this method yields a solution which also lies in the subspace $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$.

The preconditioning interpretation allows for an important departure from the prior-conditioning and the standard form framework, namely for *approximate inversion of M*. In our fDOT example we use such a symmetric approximate inverse obtained with algebraic

multi-grid without observing deterioration of the quality of the solution. Formal analysis of the effect of the approximate inversion is out of scope of the present paper.

While each priorconditioned iteration is more expensive than its unpriorconditioned counterpart due to the application of M^{-1}, for many large-scale problems the additional cost is insignificant and outweighted by the benefit of faster convergence; see the discussion in section 6.

2.4. Example problem: deconvolution in 1D

Throughout the paper we use the following 1D deconvolution problem to illustrate properties of the method. Assume that the matrix A in the observation model

$$g = Af + n \tag{19}$$

is a discretization of the stationary convolution operator

$$\mathcal{A}f\,(x) := \int_{-\infty}^{\infty} \mathcal{K}_{\sigma_f}\big(|x - x'|\big) f\,(x')\,\mathrm{d}x',$$
$$\mathcal{K}_{\sigma_f}(x) = \sqrt{\frac{2}{\pi\sigma_f^2}}\,\exp\left[-\frac{x^2}{2\sigma_f^2}\right],$$

and $n \sim \mathcal{N}(0, \sigma_n^2 I)$ represents isotropic Gaussian white noise with zero mean. Figure 1(a) shows the target function f along with its convoluted, noisy version g with $\sigma_f = 0.03$, $\sigma_n = 0.01$. The functions were evaluated on a regular grid with 512 samples over the interval [0, 1].

To examine the effect of priorconditioning we solve (19) regularized with M an ideal Perona–Malik regularizer (4) (left equation) based on the *true* solution $f = f_{\text{true}}$ and threshold T=0.005, discretized with finite difference method. The first four basis vectors of the respective Krylov spaces $\mathcal{K}^{A^{\mathrm{T}}A}$, $\mathcal{K}^{A^{\mathrm{T}}A+\tau M}$ (with $\tau = 1$) and $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ are shown in figures 1(b)–(d). The vectors in $\mathcal{K}^{A^{\mathrm{T}}A}$ do not contain any prior information on the discontinuities, and so they are completely smooth. The third and fourth basis vectors in $\mathcal{K}^{A^{\mathrm{T}}A+\tau M}$ start to show *local* oscillations around the jumps of the true solution superposed with smooth components, which is consistent with the local nature of the anisotropic differential operator underlying M in $A^{\mathrm{T}}A + \tau M$. The amplitude of the local oscillations is governed by the parameter τ. On the other hand the vectors in the Krylov subspace $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$, strongly resemble the structure of the prior covariance $C_X = \tau^{-1}M^{-1}$, and with high probability contain the edges of f_{true} used for construction of M.

3. Implicitly preconditioned LSQR

On structured grids, it is natural to define the prior by constructing the matrix L directly using for instance finite difference discretization. As such L are typically nonsquare and hence noninvertible, numerical methods based on the transformation to standard form revert to A-weighted pseudoinverse cf [15, 24]. On unstructured meshes, in contrast, the prior is usually given through a scaled inverse of the prior covariance, $M := (\tau C_X)^{-1} = L^{\mathrm{T}}L$, rather than its factor L. M is typically a second order differential operator cf (7) which can be cheaply stored due to its sparse representation. Taking a naïve approach, the construction of the

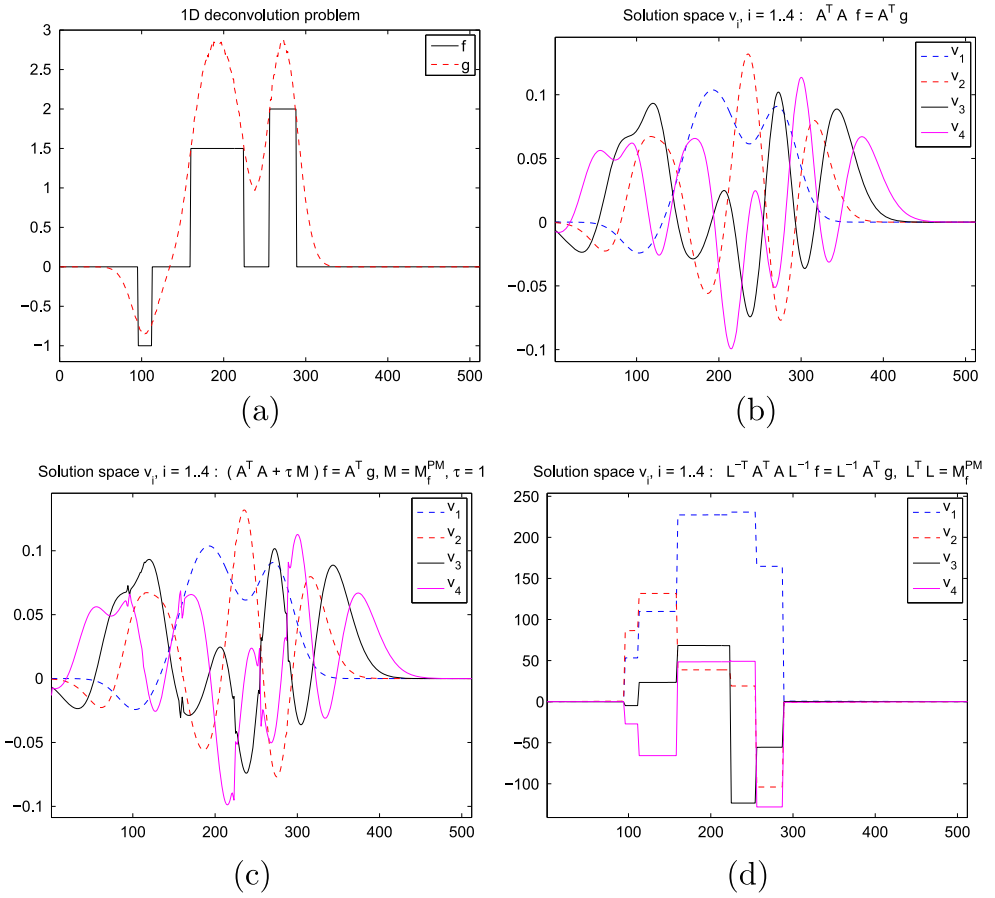

Figure 1. (a) True solution f and data g. First four basis vectors of (b) $\mathcal{K}^{A^{\mathrm{T}}A}$ ($\tau = 0$), (c) $\mathcal{K}^{A^{\mathrm{T}}A+\tau M}$ ($\tau = 1$, applied to (11)), (d) $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ (applied to (14)).

priorconditioned problem (14) requires an explicit symmetric factorization $L^{\mathrm{T}}L = M$ with an invertible L, e.g. Cholesky decomposition. For large-scale problems such a factorization may be too expensive to compute, store and apply: although M is sparse, L will have considerable fill-in. Furthermore, for nonlinear problems the factorization would have to be recomputed at each linearization step of a nonlinear iteration. In this section we propose a factorization-free preconditioned LSQR algorithm (MLSQR), which solves the priorconditioned formulation (14) without actually factorizing M.

The split preconditioned symmetric system, (15) with $\tau = 0$,

$$L^{-\mathrm{T}}A^{\mathrm{T}}AL^{-1}\hat{f} = L^{-\mathrm{T}}A^{\mathrm{T}}g, \qquad f = L^{-1}\hat{f}$$

can be solved by preconditioned conjugate gradient without the need to provide a factorization $L^{\mathrm{T}}L$ for M (MCG) [25]. The key is to use suitable scalar products. For the left preconditioned normal equation

$$M^{-1}A^{\mathrm{T}}Af = M^{-1}A^{\mathrm{T}}g, \tag{20}$$

MCG utilizes the M-weighted scalar product in which $M^{-1}A^{\mathrm{T}}A$ is self-adjoint:

$$\begin{aligned}\left(M^{-1}A^{\mathrm{T}}Af,\, g\right)_M &= \left(M^{-1}A^{\mathrm{T}}Af,\, Mg\right) = \left(A^{\mathrm{T}}Af,\, g\right)\\ &= \left(f,\, A^{\mathrm{T}}Ag\right) = \left(f,\, MM^{-1}A^{\mathrm{T}}Ag\right) = \left(f,\, M^{-1}A^{\mathrm{T}}Ag\right)_M.\end{aligned}$$

Alternatively, the right preconditioned normal equation

$$A^{\mathrm{T}}AM^{-1}\bar{f} = A^{\mathrm{T}}g, \qquad M^{-1}\bar{f} = f, \tag{21}$$

can be solved by preserving symmetry with the M^{-1}-weighted scalar product, in which $A^{\mathrm{T}}AM^{-1}$ is self-adjoint. It is easy to see that both the left and right preconditioned variants produce the same sequence of computations and hence they are equivalent up to the round-off errors. In the following in this context *equivalent* means *equivalent up to the round-off errors*.

3.1. Preconditioned LSQR

The LSQR algorithm [13, 14] is equivalent to the conjugate gradient method applied to the normal equation, but it avoids explicit formation of the normal equation at any stage. Instead, LSQR makes use of the Golub–Kahan bidiagonalization [26]. Applied to the least squares problem

$$\min_f \| g - Af \| \tag{22}$$

with a starting vector g, the bidiagonalization procedure can be written in matrix form as

$$U_{i+1}\left(\beta_1 e_1\right) = g \quad AV_i = U_{i+1}B_i \quad A^{\mathrm{T}}U_{i+1} = V_iB_i^{\mathrm{T}} + \alpha_{i+1}v_{i+1}e_{i+1}^{\mathrm{T}},$$

where e_i denotes the ith canonical basis vector, $\alpha_i \geqslant 0$ and $\beta_i \geqslant 0$ are chosen such that $\| u_i \| = \| v_i \| = 1$ and

$$B_i = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_i \\ & & & \beta_{i+1} \end{bmatrix}, \quad U_i = [u_1, u_2, ..., u_i], \quad V_i = [v_1, v_2, ..., v_i]. \tag{23}$$

As all the quantities B_i, U_i and V_i are independent of τ, the projected least squares problem

$$\min_{y_i} \left\| \begin{bmatrix} B_i \\ \sqrt{\tau} I \end{bmatrix} y_i - \beta_1 e_1 \right\|, \tag{24}$$

is a proper generalization of $\min_{y_i} \| B_i y_i - \beta_1 e_1 \|$ for $\tau \neq 0$. The projected problem (24) is then solved using QR decomposition yielding the approximation for the solution of the original problem, $f_i = V_i y_i$.

Recall that the preconditioned LSQR solves the problem where the preconditioner is applied to the least squares problem

$$\hat{f} = \operatorname{argmin} \| g - AL^{-1}\hat{f} \|,$$

which corresponds to the split preconditioned normal equation (15) without damping. Hence, the preconditioned LSQR written out in algorithm 1 is equivalent to CG applied to the split preconditioned normal equation.

3.2. Factorization-free preconditioning

In the following, we derive an MLSQR algorithm, corresponding to the variant of MCG algorithm with the M-weighted inner products applied to the left preconditioned normal equation (20). The same algorithm (same sequence of computations) can be derived based on MCG with M^{-1}-weighted inner products applied to right preconditioned normal equation (21). We chose to present the M-weighted variant (left preconditioning) because it works with the original solution, unlike the M^{-1}-weighted one (right preconditioning) which involves a change of basis.

To this end we introduce new variables $v_i = L\tilde{v}_i$ and $w_i = L\tilde{w}_i$, cf algorithm 1. We then reformulate algorithm 1 in terms of these new variables. The only parts affected are steps 3–4 of the initialization, the bidiagonalization and the update stage. Observing that step 3 of the initialization is of the same form (with $v_0 = 0$) as step 8 of the bidiagonalization, it is sufficient to consider only the latter. As $\hat{f}_i$ is a linear combination of v_j, $j \leqslant i$, the transformed solution has to be a linear combination of $\tilde{v}_j = L^{-1}v_j$, $j \leqslant i$. Thus, the resulting algorithm will directly produce a sequence of approximate solutions $f_i = L^{-1}\hat{f}_i$ and the change of variables at step 22 of algorithm 1 cancels out.

Algorithm 1. Preconditioned LSQR, equivalent to CG applied to split preconditioned normal equation.

1: **Initialization:**
2: $\beta_1 u_1 = g$
3: $\alpha_1 v_1 = L^{-\mathrm{T}} A^{\mathrm{T}} u_1$
4: $w_1 = v_1,\ \hat{f}_0 = 0,\ \bar{\phi}_1 = \beta_1,\ \bar{\rho}_1 = \alpha_1$
5: **for** $i = 1, 2, \ldots$**do**
6: **Bidiagonalization:**
7: $\beta_{i+1} u_{i+1} = AL^{-1}v_i - \alpha_i u_i$
8: $\alpha_{i+1} v_{i+1} = L^{-\mathrm{T}} A^{\mathrm{T}} u_{i+1} - \beta_{i+1} v_i$
9: **Orthogonal transformation:**
10: $\rho_i = \left(\bar{\rho}_i^2 + \beta_{i+1}^2\right)^{1/2}$
11: $c_i = \bar{\rho}_i / \rho_i$
12: $s_i = \beta_{i+1} / \rho_i$
13: $\theta_{i+1} = s_i \alpha_{i+1}$
14: $\bar{\rho}_{i+1} = -\, c_i \alpha_{i+1}$
15: $\phi_i = c_i \bar{\phi}_i$
16: $\bar{\phi}_{i+1} = s_i \bar{\phi}_i$
17: **Update:**
18: $\hat{f}_i = \hat{f}_{i-1} + \left(\phi_i / \rho_i\right) w_i$
19: $w_{i+1} = v_{i+1} - \left(\theta_{i+1} / \rho_i\right) w_i$
20: **Break if stopping criterion satisfied**
21: **end for**
22: Transform back to original solution: $f = L^{-1}\hat{f}$

The bidiagonalization in terms of $\tilde{v}_i$, $\tilde{w}_i$ and the M-weighted inner product reads

$$\beta_{i+1}u_{i+1} = A\tilde{v}_i - \alpha_i u_i \qquad \tilde{v}_{i+1} = M^{-1}A^{\mathrm{T}}u_{i+1} - \beta_{i+1}\tilde{v}_i$$
$$\alpha_{i+1} = \left(\left(\tilde{v}_{i+1}, \tilde{v}_{i+1}\right)_M \right)^{1/2} \qquad \tilde{v}_{i+1} = \tilde{v}_{i+1}/\alpha_{i+1}.$$

Keeping just one additional vector $\tilde{p} = M\tilde{v}_{i+1}$, the factorization-free algorithm needs no multiplications by the priorconditioner M, just one solve with M per iteration. This amounts to rephrasing the bidiagonalization in the following way

$$\beta_{i+1}u_{i+1} = A\tilde{v}_i - \alpha_i u_i \qquad \tilde{p} = A^{\mathrm{T}}u_{i+1} - \beta_{i+1}\tilde{p}$$
$$\tilde{v}_{i+1} = M^{-1}\tilde{p} \qquad \alpha_{i+1} = \left(\tilde{v}_{i+1}, \tilde{p}\right)^{1/2}$$
$$\tilde{p} = \tilde{p}/\alpha_{i+1} \qquad \tilde{v}_{i+1} = \tilde{v}_{i+1}/\alpha_{i+1}.$$

To conclude, the update in terms of the new variables and the original solution reads

$$f_i = f_{i-1} + \left(\phi_i/\rho_i\right)\tilde{w}_i$$
$$\tilde{w}_{i+1} = \tilde{v}_{i+1} - \left(\theta_{i+1}/\rho_i\right)\tilde{w}_i.$$

The resulting MLSQR method is summarized in algorithm 2.

We remark, that a factorization-free preconditioned LSMR method (MLSMR) can be derived in the same way, as the implicit preconditioning only affects the bidiagonalization procedure which is the same for both methods. In [27] it is suggested that LSMR could be preferable to LSQR if early stopping is used. A comparison of the two algorithms on the priorconditioned problem i.e. comparison of MLSQR and MLSMR is of independent interest but it is out of scope of this paper. Here we focus on MLSQR because it minimizes the residual of the original least squares problem (possibly with damping) over the priorconditioned subspace $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$. In contrast MLSMR minimizes over the same subspace the residual of the split preconditioned normal equation (15), which does not immediately relate to the original problem.

Algorithm 2. MLSQR: factorization-free preconditioned LSQR.

Initialization:
$\beta_1 u_1 = g$
$\tilde{p} = A^{\mathrm{T}}u_1$
$\tilde{v}_1 = M^{-1}\tilde{p}$
$\alpha_1 = \left(\tilde{v}_1, \tilde{p}\right)^{1/2}$
$\tilde{p} = \tilde{p}/\alpha_1$, $\tilde{v}_1 = \tilde{v}_1/\alpha_1$
$\tilde{w}_1 = \tilde{v}_1$, $f_0 = 0$, $\bar{\phi}_1 = \beta_1$, $\bar{\rho}_1 = \alpha_1$
for $i = 1, 2, \ldots$ **do**
  Bidiagonalization:
  $\beta_{i+1}u_{i+1} = A\tilde{v}_i - \alpha_i u_i$
  $\tilde{p} = A^{\mathrm{T}}u_{i+1} - \beta_{i+1}\tilde{p}$
  $\tilde{v}_{i+1} = M^{-1}\tilde{p}$
  $\alpha_{i+1} = \left(\tilde{v}_{i+1}, \tilde{p}\right)^{1/2}$
  $\tilde{p} = \tilde{p}/\alpha_{i+1}$
  $\tilde{v}_{i+1} = \tilde{v}_{i+1}/\alpha_{i+1}$

(Continued).

16: **Orthogonal transformation:**
17: Steps 10–16 as in algorithm 1
18: **Update:**
19: $f_i = f_{i-1} + \left(\phi_i/\rho_i\right)\tilde{w}_i$
20: $\tilde{w}_{i+1} = \tilde{v}_{i+1} - \left(\theta_{i+1}/\rho_i\right)\tilde{w}_i$
21: **Break if stopping criterion satisfied**
22: **end for**

3.3. LSQR with regularization

Two types of regularization are relevant in our framework: Tikhonov regularization, where the parameter τ controls the amount of regularization, and early truncation of Krylov methods, in which regularization arises from the problem being projected onto a small dimensional subspace.

When priorconditioning is used, Tikhonov regularization results in a simple damped least squares problem (14). Damping for a fixed value of τ can be easily incorporated in LSQR as described in [14] at the cost of doubling the number of Givens rotations. Due to the shift invariance of Krylov spaces, different choices of the damping parameter τ result in the same Krylov subspace and Lanczos vectors V_i. If V_i are stored, the projected least squares problem (24) can be efficiently solved for multiple values of τ, which is of benefit when the value of τ is not known in advance. However, the additional storage requirements may limit the feasibility of such an approach. Some techniques for solving (24) with a variable τ are discussed in [28] using singular values and the first and last rows of the matrix of the right singular vectors of the bidiagonal matrix B_i. Those quantities can be obtained at the cost $O\left(i^2\right)$ at the ith iteration. These strategies are however only viable for a limited number of iterations i as the singular value decomposition of the bidiagonal matrix B_i can not be efficiently updated even though B_i simply expands by a row and a column in each iteration. For larger i, the algorithm described in [29] for the least squares solution of (24) at the cost of $O\left(i\right)$ for each value of τ is the preferable option.

3.4. Stopping criteria

In this paper we make use of two Krylov methods: the MLSQR method for solution of the priorconditioned problem (14) and the reference LSQR method for solution of the unpriorconditioned problem (11).

As the system in (11) is a well posed inconsistent system, we choose the criterion S2 from the original paper [13] for stopping LSQR.

S2: $\frac{\| \bar{A}^T \bar{r}_i \|}{\| \bar{A} \| \| \bar{r}_i \|} \leqslant$ ATOL (inconsistent systems),

where $\bar{r}_i := \bar{g} - \bar{A} f_i$ denotes the residual of (11) with

$$\bar{A} = \begin{bmatrix} A \\ \sqrt{\tau} L \end{bmatrix}, \qquad \bar{g} = \begin{bmatrix} g \\ 0 \end{bmatrix},$$

and ATOL is a user-defined threshold, see [13].

On the other hand it is well known that (preconditioned) LSQR and hence MLSQR applied to the least squares problem (14) monotonically decrease the residual norm $\|\bar{r}_i\|$. For $\tau = 0$, (14) is an undamped problem and hence it is ill-posed. Moreover, it holds that $\bar{r}_i = r_i$ with $r_i := g - Af_i$ being the residual of the original least squared problem. Consequently, the sequence $\|r_i\|$ is also monotonically decreasing and the Morozov discrepancy principle is a natural choice for terminating MLSQR at least when $\tau = 0$.

S4: $\| r_i \| \leqslant \eta\delta$ (ill-posed problems),

where δ is the (estimated) noise level [30], and the factor $\eta > 1$ has been included to prevent underregularization [8].

3.5. MLSQR: example 1D deconvolution problem

To demonstrate the effect of priorconditioning we revisit the linear deconvolution problem in section 2.4 with the ideal linear regularizer resulting in preconditioner $M := M_{f_{\text{true}}} = L^{\mathrm{T}}L$, and compare its solutions obtained with MLSQR applied to (14) and LSQR applied to (11). The regularization parameter value $\tau = 1$ was hand-tuned for the unpriorconditioned problem. In addition, the value $\tau = 0$ was tested. For the priorconditioned problem selecting $\tau = 0$ results in a priorconditioned solution without damping, while for the unpriorconditioned problem setting $\tau = 0$ eliminates (Tikhonov) regularization.

Figures 2(a), (b) show the solution of the priorconditioned and the unpriorconditioned problem with ($\tau = 1$) and without ($\tau = 0$) regularization. The solutions in figure 2(a) correspond to the LSQR and MLSQR algorithms terminated with criterion S4 with $\delta = 10^{-2} \| g \|$ and $\eta = 1.1$, which resulted in stopping after 9 iterations for the priorconditioned problem and after 16 iterations for the unpriorconditioned problem for both tested values of τ. Figure 2(c) shows the corresponding residual norms i.e. the discrepancy principle S4. Evaluating the solutions in figure 2(a) we conclude that S4 provides a good stopping criterion for the priorconditioned problem, but it terminates the LSQR algorithm for the regularized unpriorconditioned problem too early (considerable oscillations even for $\tau = 1$). Figure 2(b) shows the solutions obtained with the stopping criterion S2 with ATOL $= 10^{-2}$. The criterion S2 over the iterations is plotted in figure 2(d). This criterion is essentially useful before the onset of oscillations which corresponds to the last good stopping point. While, with the chosen ATOL, S2 stops the unpriorconditioned regularized problem timely (for which S2 behaves rather smoothly even beyond the value of ATOL, in particular for $\tau = 1$), it terminates the priorconditioned problem prematurely (wrong levels of the piecewise constant approximation for both values of τ). Comparing the best respective solutions for both formulations, the priorconditioned formulation after 9 iterations yields a better solution than the unpriorconditioned regularized formulation after 32 iterations.

4. Iterated MLSQR for least squares problem with nonlinear regularizer

4.1. Nonlinear solver

We recall that the lagged diffusivity fixed point iteration applied to the necessary condition (5) for the minimizer of (1) requires a solution of the linearized normal equation (8) at each iteration. Here, we propose to speed up the solution of the corresponding linear least squares

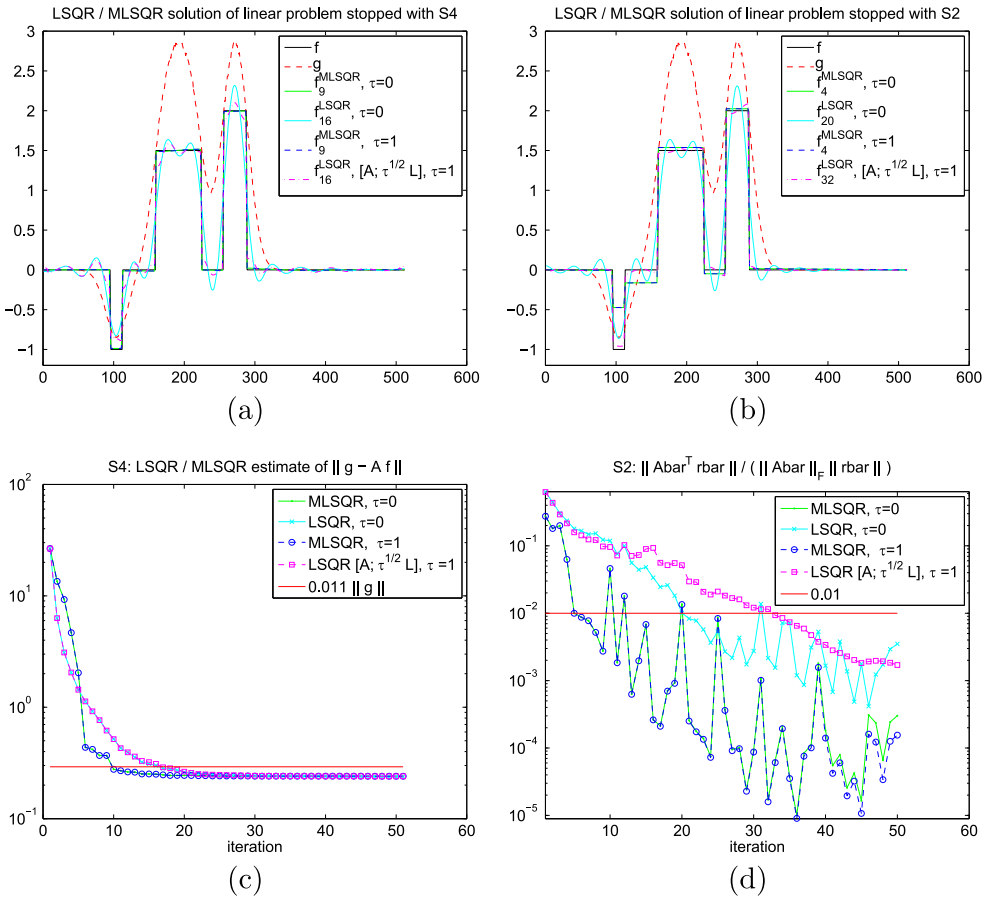

Figure 2. Upper row: solution of the linear deconvolution problem computed with LSQR (without preconditioning) and MLSQR (with preconditioning) (a) using stopping criterion S4, (b) using stopping criterion S2. Lower row: behaviour of the quantities used for stopping criteria. (c) S4: $\| g - Af \| \leqslant 1.1 \times 10^{-2}$. (d) S2: $\| \bar{A}^{\mathrm{T}} \bar{r} \| / \left(\| \bar{A} \| \| \bar{r} \| \right) \leqslant 10^{-2}$.

problem (9) using priorconditioning, i.e. solving the priorconditioned linearized problem (14) with MLSQR instead of solving the original linearized problem (11) with LSQR.

Under certain assumptions and for a parameter value τ chosen such that the Morozov discrepancy principle S4 is satisfied, the minimization problem (1) can be reformulated as the residual method [31],

$$\min_{f \in X} R(f) \text{ subject to } \| g - Af \| \leqslant \eta\delta. \tag{25}$$

Motivated by this form, we chose to monitor the value of the penalty R and stop the lagged diffusivity fixed point iteration when R does not decrease fast enough. Because each iterate f^k has approximately the same residual norm $\|r^k\| \approx \eta\delta$, the value of $\Phi(f_k)$ will in practice decrease until a feasible solution to (1) is found. The resulting iterated MLSQR method for solving the nonlinear least squares problem (1) is summarized in algorithm 3.

Algorithm 3. Iterated MLSQR method.

1: Initialize with $f^0 = 0$
2: **for** $k = 1, 2, \ldots$ **do**
3: Form the preconditioner $M_{f^{k-1}}$ using the current approximation f^{k-1}
4: Solve the linearized least squares problem (14) corresponding to the symmetrically preconditioned linearized normal equation (15), $M_{f^{k-1}} = L^{\mathrm{T}}L$

$$\left(L^{-\mathrm{T}}A^{\mathrm{T}}AL^{-1} + \tau I\right)\hat{f}^k = L^{-\mathrm{T}}A^{\mathrm{T}}g, \quad f^k = L^{-1}\hat{f}^k,$$

with MLSQR initialized with $f^k = 0$ and terminated by stopping criterion S4.
5: **Break** if $R(f^k)$ does not decrease fast enough
6: **end for**

We emphasize that it is important to stop the Krylov method in step 4 of algorithm 3 consistently: we want to avoid the situation where the trend in values of $R(f^k)$ becomes distorted by the Krylov solver (e.g. taking few Krylov iterations in one step while taking many iterations in the next will likely lead to increase in $R(f^k)$). Both discrepancy principle as well as a global upper limit on the number of the Krylov iterations provide a consistent stopping rule. In practice, it may take numerous Krylov iterations before the first few lagged diffusivity iterates reach the noise level. To circumvent this problem we combine the Morozov discrepancy principle with the global limit on the number of Krylov iterations, which does not seem to adversely affect the convergence of the iterated MLSQR algorithm, see section 5.3.

A few more comments are in order regarding algorithm 3:

- We deliberately chose to solve for f^{k+1} rather than the update $\Delta f^{k+1} := f^{k+1} - f^k$. This is because the employed regularizers encode the information on the solutions f^k and not on the updates Δf^{k+1}.
- Our simulations suggest that solutions of high quality can be obtained by setting $\tau = 0$ in (14) and regularizing solely by early stopping of the MLSQR algorithm. A clear advantage of this approach is that the data discrepancy $\|r^k\|$ is being minimized instead of $\|\bar{r}^k\|$. However, setting $\tau = 0$ invalidates the interpretation of the solution computed with algorithm 3 as the MAP estimate of the posterior distribution (10).
- If $M_{f_k}^{-1}$ is computed exactly, $\tau \neq 0$ is used and MLSQR is iterated until convergence, our scheme is equivalent to lagged diffusivity iteration. In practice, one or more of these conditions may not hold.

4.2. Inverting the preconditioner

While our preconditioner is nonstandard in the sense that it does not approximate the inverse of the forward operator it nonetheless amounts to a solution of an elliptic partial differential equation, a topic extensively handled in the literature see e.g. [32–34] for preconditioning of partial differential equations and [25, 35] of linear problems in general.

Here we use algebraic multi-grid method [32]. In particular, the use of Jacobi, SSOR or incomplete LU decomposition (ILU) smoothers results in a symmetric preconditioner if the

same number of iterations is taken in the pre- and post-smoothing stages. By using a fixed number of smoother steps and grid-to-grid operations, one obtains a fixed approximation of a matrix inverse that can be used to priorcondition Krylov methods such as MLSQR derived in section 3. The resulting preconditioner is an approximate inverse of the regularizer matrix M. Remarkably, in our experiments in section 5 even a low-cost single V-cycle multigrid approximation to M^{-1} was enough to achieve effective priorconditioning and consequently fixing the number of cycles seems unproblematic. Further analysis is necessary to establish how to construct an approximation resulting in the most effective priorconditioner and how the accuracy of such approximation affects the convergence of the nonlinear iteration.

4.3. Iterated MLSQR: example 1D deconvolution problem

We are now in a position to consider the nonlinear 1D deconvolution problem with nonlinear Perona–Malik regularization (4) (left equation). We solve the associated minimization problem (1) with algorithm 3.

In this particular example, we chose to use the undamped solution, i.e. $\tau = 0$. MLSQR was terminated with the discrepancy principle (S4) with the noise level $\delta = 10^{-2} \|g\|$ and $\eta = 1.1$. We stopped the lagged diffusivity fixed point iteration when the relative change in the functional R dropped below 15%. As the considered problem is small, we used Cholesky factorization for inverting the preconditioner.

Figure 3 shows the evolution of the first six basis vectors of $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ for $M = M_{f^k}$, $k \in \{1, 4, 7, 10\}$. For $k = 1$, the priorconditioner is discretization of a homogeneous Laplacian operator and consequently the basis vectors are smooth, poorly replicating the discontinuities of the target function. However, the priorconditioner quickly adapts to the jumps emerging in the intermediate solutions f^k as can be seen from the basis vectors for $k \geqslant 4$.

The evolution of the solution at every third lagged diffusivity step is illustrated in figure 4(a) and the solution error over outer iteration is plotted in figure 4(b). The residual norm over the inner and outer iterations is shown in figure 5(a). Notice that every lagged diffusivity iteration resets the residual norm to the value $\|g\|$. This is due to the initialization of MLSQR with $f^k = 0$ at the beginning of every outer iteration to allow the solution to develop features compatible with the new priorconditioner. The gap developing in the residual norm plot in figure 5(a) between the 5th and 6th MLSQR iterations is characteristic for the method. It means that the first five basis vectors are becoming increasingly better adapted to represent the solution, while the remaining basis vectors contribute less and less. Consequently, the updated priorconditioner allows reduction of the residual norm in fewer MLSQR iterations as shown in figure 5(c).

Figure 5(b) displays the behaviour of the functional R over the inner and outer iterations. We observe that in each lagged diffusivity iteration, while solving the linearized problem R in general increases in each MLSQR iteration until a saturation level is reached. On the other hand, the saturation level decreases as the outer iteration progresses. At a certain point, the saturation level stops decreasing (or is not decreasing fast enough), see figure 5(d), and the algorithm 3 is terminated. The plot of the error norm in figure 4(b) demonstrates that the last step taken before the stopping criterion was triggered did not perceivably improve the solution, which corroborates our choice of the stopping criterion.

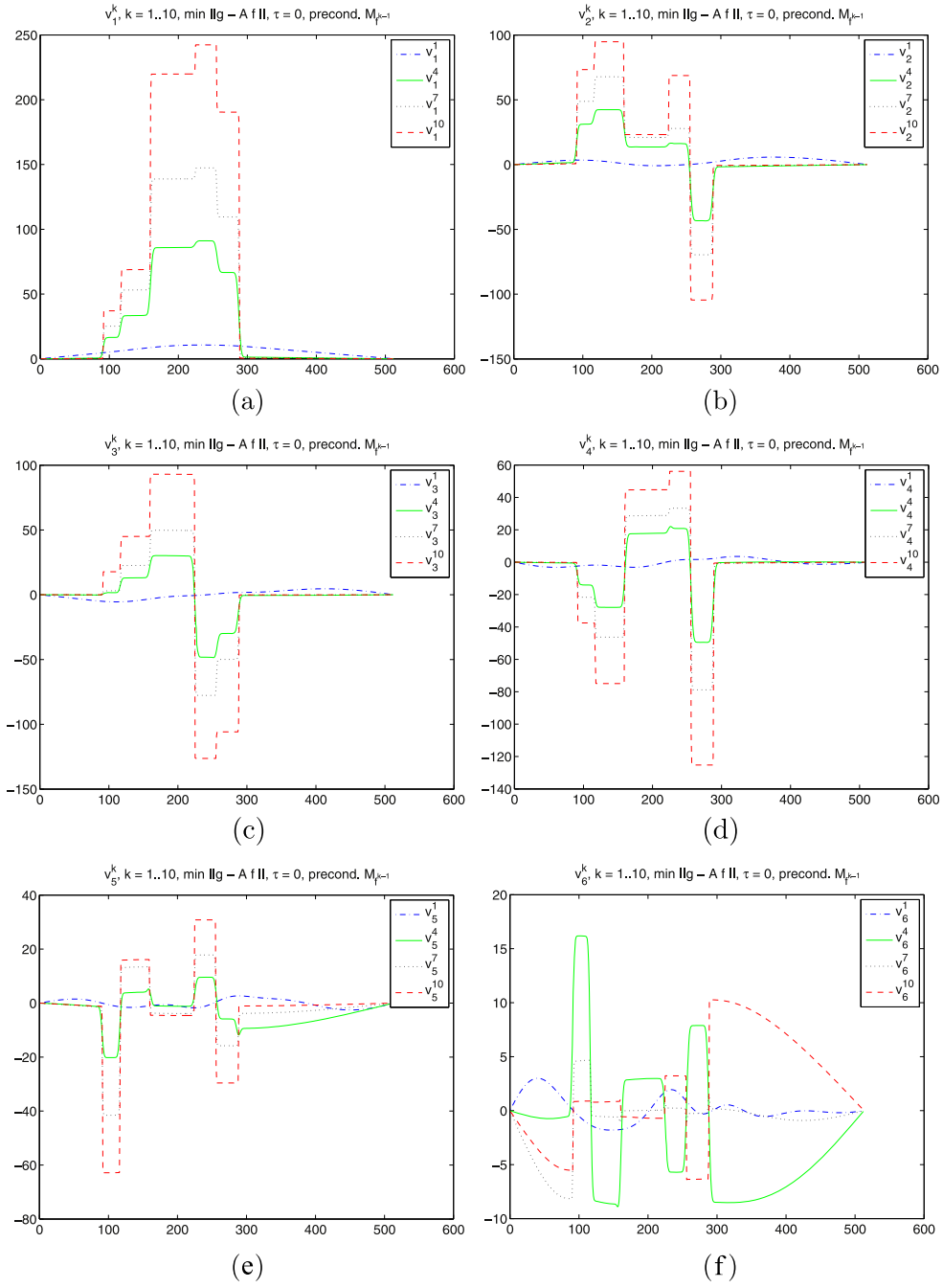

Figure 3. Evolution of the first six basis vectors of $\mathcal{K}^{M^{-1}A^{\mathrm{T}}A}$ through the outer iterations.

5. A 3D image reconstruction problem

In this section we apply our method to a large-scale 3D image reconstruction problem on an unstructured grid.

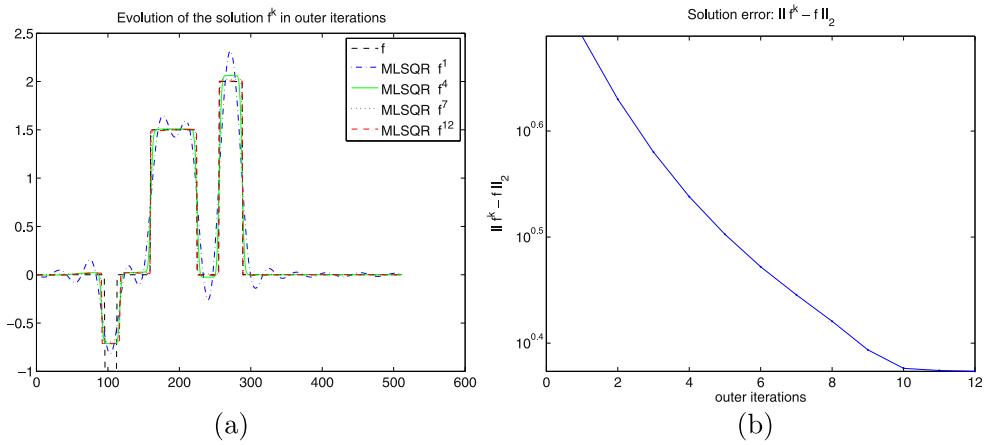

Figure 4. (a) Evolution of the solution computed with MLSQR through the outer iterations. (b) Error norm reduction over outer iterations.

5.1. fDOT

We consider fDOT, which indirectly measures the conversion of strongly scattered (i.e. diffusively propagating) light from an excitation wavelength to a (longer) emission wavelength in the presence of fluorescent markers accumulating in regions of interest. The goal is to monitor cellular and subcellular functional activity. Although fDOT is a method mostly used for small animal research [36–39], it is a promising technique with several medical applications such as detection, diagnosis and monitoring of human neoplasms, in particular breast tumours [40–42]. Due to the diffusive nature of light propagation in biological tissue, the image reconstruction is an ill-posed inverse problem [43–45].

In the linear approximation to fDOT, the forward model is described with coupled diffusion equations at the excitation and emission wavelengths, λ_e and λ_f, respectively:

$$\left[-\nabla\cdot\kappa\left(x,\lambda_e\right)\nabla+\mu_a\left(x,\lambda_e\right)\right]U\left(x,\lambda_e\right)=0,\qquad x\in\Omega, \tag{26}$$

$$\left[-\nabla\cdot\kappa\left(x,\lambda_f\right)\nabla+\mu_a\left(x,\lambda_f\right)\right]U\left(x,\lambda_f\right)=U\left(x,\lambda_e\right)h\left(x,\lambda_f\right),\qquad x\in\Omega, \tag{27}$$

with μ_a and μ_s' the absorption and the reduced scattering coefficients, respectively, $\kappa=[3(\mu_s'+\mu_a)]^{-1}$ the diffusion coefficient and h the fluorescence yield coefficient.

The diffusion equations are complemented by the respective Robin boundary conditions

$$U\left(x,\lambda_e\right)+2\zeta\kappa\left(x,\lambda_e\right)\frac{\partial U\left(x,\lambda_e\right)}{\partial\nu(x)}=\Theta_s(x)\,q,\qquad x\in\partial\Omega, \tag{28}$$

$$U\left(x,\lambda_f\right)+2\zeta\kappa\left(x,\lambda_f\right)\frac{\partial U\left(x,\lambda_f\right)}{\partial\nu(x)}=0,\qquad x\in\partial\Omega, \tag{29}$$

where ν is the outward unit normal and ζ accounts for the refractive index mismatch at the boundary. The right-hand side in (28) models the effect of the excitation light source as an inward (diffuse) photon current, a product of the source emitted photon current q and a source coupling coefficient function Θ_s (notice no source term in (26)). On the other hand, the emission photon density arises solely from the fluorescence in Ω (right-hand side of (27)), resulting in homogeneous Robin boundary condition (29).

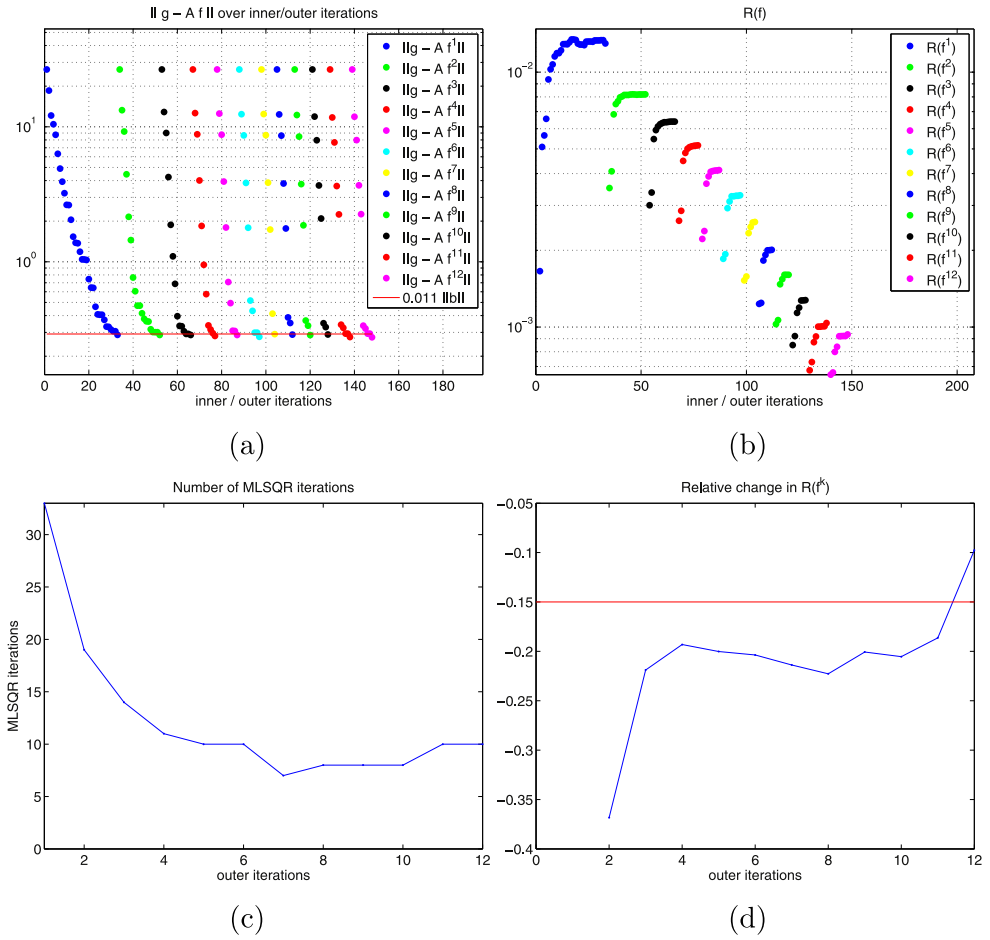

Figure 5. (a) Residual norm (S4) over lagged diffusivity/MLSQR iterations. (b) $R\left(f_i^k\right)$ over lagged diffusivity (k)/MLSQR (i) iterations. (c) Number of MLSQR iterations over lagged diffusion iterations. (d) Stopping criterion for the nonlinear iteration: $\left(R\left(f^k\right) - R\left(f^{k-1}\right)\right)/R\left(f^{k-1}\right)$ over lagged diffusivity iterations. The threshold level was chosen to be -0.15.

The measured photon density for a detector with wavelength independent detector coupling coefficient Θ_{d} is given as

$$y\left(\lambda\right) = \int_{\partial\Omega} \Theta_{\mathrm{d}}\left(x\right) \frac{U\left(x, \lambda\right)}{2\zeta} \mathrm{d}x, \quad \lambda \in \left\{\lambda_{\mathrm{e}}, \lambda_{\mathrm{f}}\right\}. \tag{30}$$

In practice several sources and detectors are deployed, resulting in a vector of measurements $y\left(\lambda\right) \in \mathbb{R}^{N_{\mathrm{s}}N_{\mathrm{d}}}$, where N_{s} and N_{d} is the number of sources and detectors, respectively. We furthermore use the following normalization, demonstrated to reduce the effects of unknown Θ_{s} and Θ_{d} [46, 47]

$$g_j = \frac{y_j\left(\lambda_{\mathrm{f}}\right)}{y_j\left(\lambda_{\mathrm{e}}\right)}. \tag{31}$$

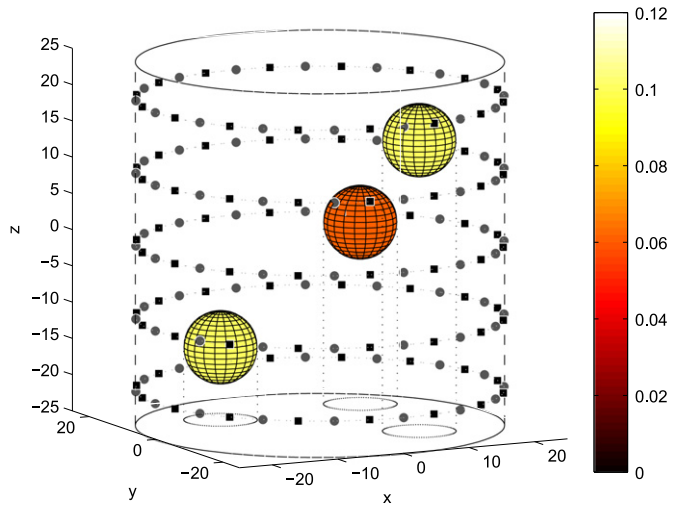

Figure 6. The cylindrical phantom with three inclusions. Source and detector locations are marked with grey points and black squares, respectively.

We now define the mapping $\mathcal{A} : h \mapsto g$ through solution of the system (26)–(31). $\mathcal{A}$ is clearly nonlinear in the parameters $\{\mu_a(x, \lambda_e), \mu_a(x, \lambda_f), \mu_s'(x, \lambda_e), \mu_s'(x, \lambda_f)\}$, but it is linear in the sought-for parameter $h(x, \lambda_f)$ (assuming that the other optical parameters are independent of h, which is reasonable for typical concentrations of fluorophores found in biomedical applications).

The adjoint mapping $\mathcal{A}^* : b \mapsto z$ for a single source-detector pair, is defined through solution of (26) with the boundary condition (28) and

$$\left[-\nabla \cdot \kappa\left(x, \lambda_f\right) \nabla + \mu_a\left(x, \lambda_f\right)\right] U^*\left(x, \lambda_f\right) = 0, \quad x \in \Omega, \tag{32}$$

with the inhomogeneous Robin boundary condition

$$U^*\left(x, \lambda_f\right) + 2\zeta\kappa\left(x, \lambda_f\right) \frac{\partial U^*\left(x, \lambda_f\right)}{\partial \nu\left(x\right)} = \frac{\Theta_d\left(x\right) b}{y\left(\lambda_e\right)}, \quad x \in \partial\Omega, \tag{33}$$

followed by the multiplication of the two solutions, yielding

$$z = U^*\left(x, \lambda_f\right) U\left(x, \lambda_e\right). \tag{34}$$

If multiple detectors are used, the right-hand side of (33) will involve summation over the detectors. Analogously, a multiple source configuration will involve N_s solves for (26), (28) and (32), (33), with (34) summing the pairwise products of U and U^* over the sources.

5.2. Simulation setup

Our test phantom is a cylinder of radius 25 mm and height 50 mm. The fluorophore distribution is represented through three spherical inclusions, one with $h = 0.06$ and the other two with $h = 0.1$, and the background value $h = 0$, see figures 6, 7(e). The remaining optical parameters are assumed homogeneous, and their values are summarized in table 1. The fluorophore was excited with each of 80 sources uniformly distributed along five rings on the boundary of the cylinder, see figure 6. The measurements were sampled by 80 detectors

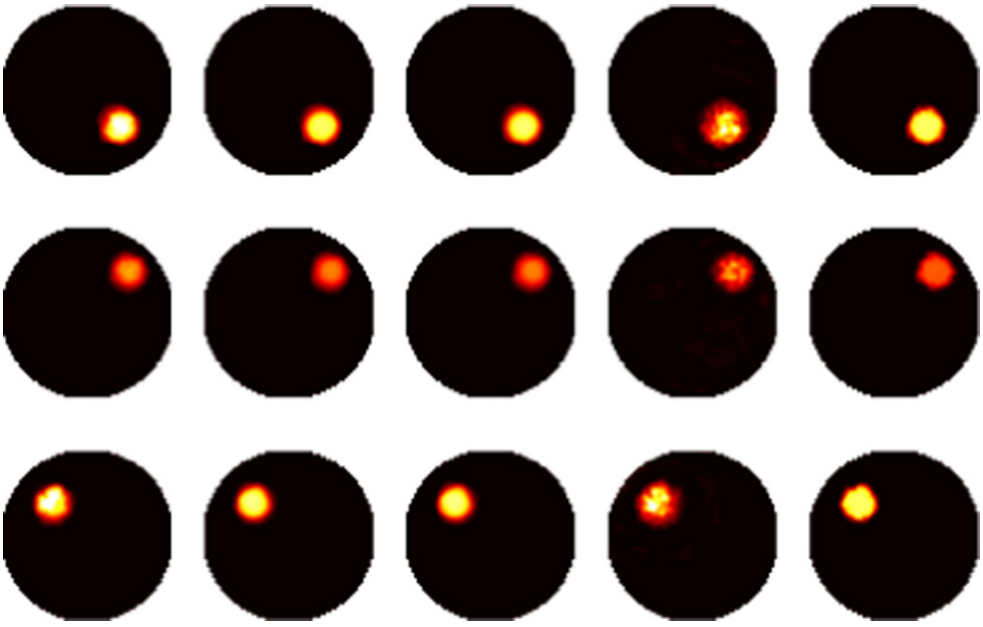

Figure 7. Transaxial planes through the centres of the spherical inclusions at $z = 15$, $z = 0$ and $z = -15$ of the solution to (a) unpriorconditioned problem, (b) priorconditioned problem with damping $\tau = 10^4$, (c) priorconditioned problem without damping $\tau = 0$, and (d) problem with no other regularization but early stopping of LSQR; (e) the phantom. (x-axis horizontal, y-axis vertical).

Table 1. Optical parameters used in the simulations.

Absorption coefficient $\mu_a(\cdot, \lambda_e) = \mu_a(\cdot, \lambda_f)$	0.05	mm^{-1}
Reduced scattering coefficient $\mu_s'(\cdot, \lambda_e) = \mu_s'(\cdot, \lambda_f)$	1	mm^{-1}
Refractive index	1.4	

placed along the same rings on the boundary, half way in between the sources, resulting in a total of 6400 measurements. All the involved coupling coefficient functions were modelled as Gaussian distributions on the boundary $\partial\Omega$. Measurement errors were simulated using additive Gaussian noise with standard deviation equal to 1% of the corresponding ideal measurement.

The fDOT problem was modelled and discretized using the finite element method (FEM) with software package TOAST [48]. The discretization with first order Lagrangian elements on a tetrahedral mesh yielded 27084 degrees of freedom which corresponds to approximately 1.5 mm resolution. The associated inverse problem was regularized with the differentiable approximation to the total variation functional in (2). The resulting matrix M_f is thus a finite element discretization of the operator $\nabla \cdot (|\nabla f|^2 + T^2)^{-1/2} \nabla$. We chose $T = 10^{-6}$ as it yields a well-conditioned matrix M for fluorescence yield coefficients with edges of height approximately equal or smaller than the expected maximum of 0.1. The preconditioner was applied

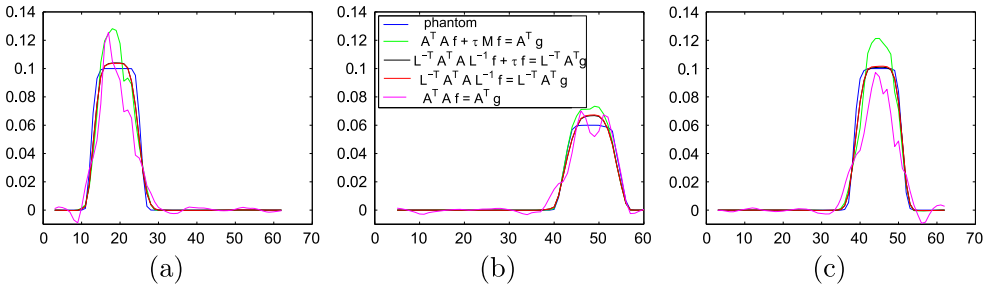

Figure 8. Cross sections along the x-axis through the centres of the inclusions of the solutions in figure 7: (a) $z = -15$, $y = 10$, (b) $z = 0$, $y = 12.5$ and (c) $z = 15$, $y = -10$.

using a computationally low-cost algebraic multigrid solver with single V-cycle and two steps of ILU(0) smoothing, implemented in IFISS [49, 50].

5.3. Solution of the fDOT problem

We solve the fDOT problem described in section 5.1 using the iterated MLSQR method in algorithm 3. The proposed method is particularly well suited for problems of this kind: fDOT is large-scale and the explicit construction of the matrix A representing the discretized forward mapping can be impractical. Instead, the action of A on a vector is obtained through solution of (26)–(31) for each source. Similarly, multiplication by A^{T} involves solving (26), (28) and (32)–(34). In our example, both of these mappings amount to 160 solves of elliptic partial differential equations. Moreover, because the FEM mesh is unstructured, the preconditioner M naturally arises in the unfactorized form.

We demonstrate the benefits of priorconditioning by comparing the performance of the iterated MLSQR method to an unpriorconditioned reference method or iterated LSQR (algorithm 3 but where in the 4th step instead the unpriorconditioned linearized problem (11) is solved with LSQR terminated with rule S2 as explained in section 3.4). The tolerance in S2 was chosen ATOL $= 10^{-3}$, and in S4, $\eta = 1.1$ and $\delta = 10^{-2} \parallel g \parallel$. The same regularization parameter value $\tau = 10^4$ was used for both methods, which was tuned to yield the best-case results for the unpriorconditioned reference method. Again, we also test $\tau = 0$ demonstrating that the proposed algorithm is very robust with respect to the choice of τ. Note that for $\tau = 0$ the reference method amounts to solution of the unregularized linear least squares problem with LSQR. As the least squares problem is ill-posed, LSQR is stopped with criterion S4.

We monitor the value of the penalty $R(f^k)$ of the intermediate solutions f^k and terminate the lagged diffusivity (outer) iteration in algorithm 3 when $R(f^k)$ ceases to decrease. For the reference iterated LSQR method, the functional $R(f^k)$ stagnates or even starts increasing after initial five outer iterations while the error norm keeps decreasing, which could be due to increasing condition number of M_{f^k}. Instead, 25 lagged diffusivity steps were taken to ensure convergence for the reference method for $\tau = 10^4$. The resulting reconstructions are depicted in figures 7 and 8. The two unpriorconditioned solutions suffer from overshooting and spurious oscillations, while their priorconditioned counterparts have the correct shape and estimate the value of the fluorescence yield coefficient h more accurately. The error norms in each lagged diffusivity iteration are plotted in figure 9(a). The iterated MLSQR attains error

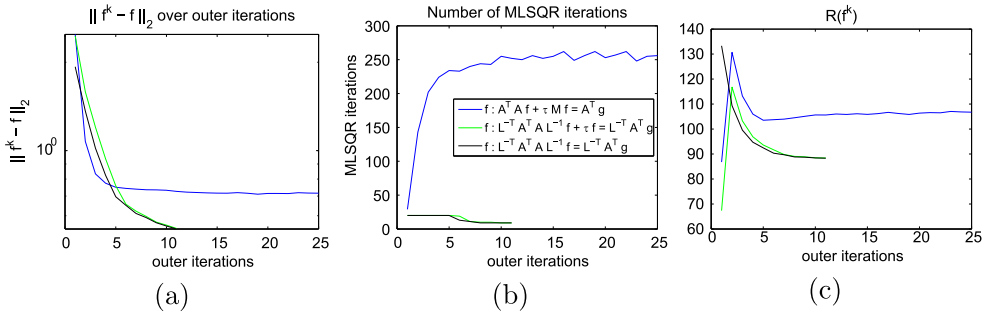

Figure 9. (a) Error norm plotted over outer iterations. (b) Number of MLSQR iterations plotted over outer iterations. (c) Penalty R plotted over outer iterations. (The unpriorconditioned method with $\tau = 0$ does not involve outer iterations. Therefore, the associated results are omitted from this figure.)

norm of 0.5428 and 0.5407 for $\tau = 10^4$ and $\tau = 0$, respectively, while the reference method 0.7158 and 1.5025.

Figure 9(b) shows the number of MLSQR iterations taken within each lagged diffusivity step. We chose to limit the number of the inner iterations for the priorconditioned algorithm to a maximum of 20 in accordance with the discussion in section 4.1. This limit was hit in the first 6 ($\tau = 10^4$) and 5 ($\tau = 0$) lagged diffusivity iterations, while MLSQR was stopped by the Morozov criterion in fewer than 20 iterations in the following lagged diffusivity steps. In fact, the motivation for limiting the maximum number of MLSQR iterations is that the first one or two lagged diffusivity steps would need an disproportionate number of MLSQR iterations to reach the residual level suggested by the Morozov discrepancy principle. The limit on Krylov iterations alleviates this issue without a noticeable effect on the final reconstruction. While after the initial phase the number of MLSQR iterations decreases to 9 for the priorconditioned algorithm, the unpriorconditioned variant requires an increasing number of Krylov steps, which we attribute to the growing condition number of the regularizer M_{f^k} over the lagged diffusivity iterations.

Figure 9(c) shows the evolution of the penalty term $R(f)$ over lagged diffusivity iterations. In the first outer iteration, the two priorconditioned solutions show a pronounced difference. The initial preconditioner is a discretization of the Laplacian and contains no information about the edges of the solution. For this reason, the damping flattens the first outer iterate particularly strongly, and consequently the associated penalty $R(f) = \int_\Omega |\nabla f(x)| \mathrm{d}x$ is of smaller magnitude than it is for the undamped variant. Notice that similar flattening takes place for the unpriorconditioned method, as well. Once the preconditioner contains some information about the edges, the behaviour of both priorconditioned cases is qualitatively the same. Apart from the first lagged diffusivity iteration, the penalty decreases over the lagged diffusivity iterations, providing further evidence to support the penalty-based stopping criterion of algorithm 3.

6. Conclusions

In this paper, we considered efficient solution of large-scale linear ill-posed inverse problems with nonlinear regularization. We devised a highly efficient matrix-free algorithm for solution of such problems combining a lagged diffusivity fixed point iteration with priorconditioning

of Krylov methods. Priorconditioning affords a way of embedding the information contained in the prior directly into the forward operator resulting in highly accelerated convergence of Krylov methods. A novel factorization-free preconditioned LSQR algorithm was presented for solving the linear priorconditioned problem which allows an implicit application of the preconditioner through efficient schemes such as multigrid. This is of particular interest for problems formulated on unstructured grids, where the preconditioner naturally occurs in an unfactorized form and the factorization is computationally infeasible. Furthermore, the relevant regularizers arise as discretizations of elliptic partial differential equations for which approaches like multigrid have been extensively studied and applied.

The presented algorithm is matrix-free, i.e. capable of solving problems where the forward mapping cannot be computed and/or stored explicitly as a matrix. In particular, in such cases the cost of the application of the forward and adjoint mappings in each step of the Krylov method overbears the additional cost of preconditioning, while the use of priorconditioning reduces the number of Krylov iterations to a fraction of those needed without priorconditioning.

We demonstrated the effectiveness of our algorithm on the 3D fDOT image reconstruction problem. For truly large-scale problems, for which the diffusion equations involved in the forward and adjoint mapping have to be solved iteratively, the cost of applying the forward and adjoint mappings is typically running hundreds of times higher than the cost of applying the preconditioner, making the reduction in the number of Krylov iterations paramount.

The proposed method opens an interesting question of using approximate inverses of regularizers as priorconditioners. Which approximations are the most effective priorconditioners and how does the accuracy affect the convergence? What is the interpretation of such priorconditioners in terms of prior distributions? Finally, the present paper deals with ill-posed problems, where the forward mapping is linear and the nonlinearity is limited to the regularization term. We intend to investigate the extension of the ideas to fully nonlinear inverse problems such as electrical impedance tomography and diffuse optical tomography.

Acknowledgments

The authors would like to thank Per Christian Hansen for stimulating discussion and the anonymous referees for their invaluable help in improving the manuscript. The authors also acknowledge the financial support of the funding bodies: MB was supported by an EPSRC Postdoctoral Fellowship (grant number EP/H02865X/1) and LH by Finnish Doctoral Programme in Computational Sciences (FICS) and Finnish Doctoral Program in Inverse Problems.

References

[1] Bertero M and Boccacci P 1998 *Introduction to Inverse Problems in Imaging* (Bristol: Institute of Physics Publishing)
[2] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* vol 375 (Berlin: Springer)
[3] Isakov V 1998 *Inverse Problems for Partial Differential Equations* vol 127 (Berlin: Springer)
[4] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica D: Nonlinear Phenom.* **60** 259–68
[5] Perona P and Malik J 1990 Scale-space and edge detection using anisotropic diffusion *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 629–39

[6] Vogel C R and Oman M E 1996 Iterative methods for total variation denoising *SIAM J. Sci. Comput.* **17** 227–38
[7] Douiri A, Schweiger M, Riley J and Arridge S 2005 An adaptive diffusion regularization method of inverse problem for diffuse optical tomography *Photon Migration and Diffuse-Light Imaging II* (Proc. SPIE vol 5859) (Munich: SPIE)
[8] Calvetti D and Somersalo E 2007 *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing* vol 2 (Berlin: Springer)
[9] Calvetti D and Somersalo E 2008 Hypermodels in the Bayesian imaging framework *Inverse Problems* **24** 034013
[10] Calvetti D 2007 Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective *J. Comput. Appl. Math.* **198** 378–95
[11] Calvetti D and Somersalo E 2005 Priorconditioners for linear systems *Inverse Problems* **21** 1397
[12] Calvetti D, McGivney D and Somersalo E 2012 Left and right preconditioning for electrical impedance tomography with structural information *Inverse Problems* **28** 055015
[13] Paige C C and Saunders M A 1982 LSQR: an algorithm for sparse linear equations and sparse least squares *ACM Trans. Math. Softw.* **8** 43–71
[14] Paige C C and Saunders M A 1982 Algorithm 583: LSQR: sparse linear equations and least squares problems *ACM Trans. Math. Softw.* **8** 195–209
[15] Eldén L 1982 A weighted pseudoinverse, generalized singular values, and constrained least squares problems *BIT Numer. Math.* **22** 487–502
[16] Hansen P C 1998 *Rank-Deficient and Discrete Ill-Posed Problems* (Philadelphia, PA: Society for Industrial and Applied Mathematics)
[17] Hilgers J W 1976 On the equivalence of regularization and certain reproducing kernel Hilbert space approaches for solving first kind problems *SIAM J. Numer. Anal.* **13** 172–84
[18] Bertaccini D and Sgallari F 2010 Updating preconditioners for nonlinear deblurring and denoising image restoration *Appl. Numer. Math.* **60** 994–1006
[19] Hanke M and Vogel C R 1999 Two-level preconditioners for regularized inverse problems I: theory *Numer. Math.* **83** 385–402
[20] Hanke M 1992 Regularization with differential operators: an iterative approach *Numer. Funct. Anal. Optim.* **13** 523–40
[21] Hansen P C and Jensen T K 2006 Smoothing-norm preconditioning for regularizing minimum-residual methods *SIAM J. Matrix Anal. Appl.* **29** 1–14
[22] Jacobsen M, Hansen P C and Saunders M A 2003 Subspace preconditioned LSQR for discrete ill-posed problems *BIT Numer. Math.* **43** 975–89
[23] Reichel L, Sgallari F and Ye Q 2012 Tikhonov regularization based on generalized Krylov subspace methods *Appl. Numer. Math.* **62** 1215–28
[24] Kilmer M, Hansen P and Español M 2007 A projection based approach to general form Tikhonov regularization *SIAM J. Sci. Comput.* **29** 315–30
[25] Saad Y 2003 *Iterative Methods for Sparse Linear Systems* 2nd edn (Philadelphia, PA: Society for Industrial and Applied Mathematics)
[26] Golub G and Kahan W 1965 Calculating the singular values and pseudo-inverse of a matrix *J. Soc. Ind. Appl. Math. B: Numer. Anal.* **2** 205–24
[27] Fong D and Saunders M 2011 LSMR: an iterative algorithm for sparse least-squares problems *SIAM J. Sci. Comput.* **33** 2950–71
[28] Björck Å 1988 A bidiagonalization algorithm for solving large and sparse ill-posed systems of linear equations *BIT Numer. Math.* **28** 659–70
[29] Eldén L 1977 Algorithms for the regularization of ill-conditioned least squares problems *BIT Numer. Math.* **17** 134–45
[30] Hanke M 1995 *Conjugate Gradient Type Methods for Ill-posed Problems* vol 327 (Harlow: Longman)
[31] Grasmair M, Scherzer O and Haltmeier M 2011 Necessary and sufficient conditions for linear convergence of l^1-regularization *Commun. Pure Appl. Math.* **64** 161–82
[32] Briggs W L, Henson V E and McCormick S F 2000 *A Multigrid Tutorial* (Philadelphia, PA: SIAM)
[33] Larsson S and Thomée V 2009 *Partial Differential Equations with Numerical Methods* vol 45 (Berlin: Springer)
[34] Thomas J W 1999 *Numerical Partial Differential Equations: Conservation Laws and Elliptic Equations* (Berlin: Springer)

[35] Van Der Vorst H A 2003 *Iterative Krylov Methods for Large Linear Systems* vol 13 (Cambridge: Cambridge University Press)
[36] Ntziachristos V 2006 Fluorescence molecular imaging *Annu. Rev. Biomed. Eng.* **8** 1–33
[37] Ntziachristos V, Bremer C and Weissleder R 2003 Fluorescence imaging with near-infrared light: new technological advances that enable in vivo molecular imaging *Eur. Radiol.* **13** 195–208
[38] Zacharakis G, Kambara H, Shih H, Ripoll J, Grimm J, Saeki Y, Weissleder R and Ntziachristos V 2005 Volumetric tomography of fluorescent proteins through small animals *in vivo PNAS* **12** 18252–7
[39] Martin A, Aguirre J, Sarasa-Renedo A, Tsoukatou D, Garofalakis A, Meyer H, Mamalaki C, Ripoll J and Planas A M 2008 Imaging changes in lymphoid organs *in vivo* after brain ischemia with three-dimensional fluorescence molecular tomography in transgenic mice expressing green fluorescent protein in T lymphocytes *Mol. Imaging* **7** 157–67
[40] Corlu A, Choe R, durduran T, Rosen M A, Schweiger M, Arridge S R, Schnall M D and Yodh A G 2007 Three-dimensional *in vivo* fluorescence diffuse optical tomography of breast cancer in humans *Opt. Express* **15** 6696–716
[41] Ntziachristos V, Yodh A G, Schnall M and Chance B 2000 Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement *PNAS* **97** 2767–72
[42] Van De Ven S *et al* 2010 A novel fluorescent imaging agent for diffuse optical tomography of the breast: first clinical experience in patients *Mol. Imaging Biol.* **12** 343–248
[43] Arridge S R and Schotland J C 2009 Optical tomography: forward and inverse problems *Inverse Probl.* **25** 123010
[44] Freiberger M, Egger H and Scharfetter H 2010 Nonlinear inversion in fluorescence optical tomography *IEEE Trans. Biomed. Eng.* **57** 2723–9
[45] Freiberger M, Clason C and Scharfetter H 2010 Total variation regularization for nonlinear fluorescence tomography with an augmented Lagrangian splitting approach *Appl. Opt.* **49** 3741–7
[46] Ntziachristos V and Weissleder R 2001 Experimental three-dimensional fluorescence reconstruction of diffuse media by use of a normalized Born approximation *Opt. Lett.* **26** 893–5
[47] Soubret A, Ripoll J and Ntziachristos V 2005 Accuracy of fluorescent tomography in the presence of heterogeneities: study of the normalized Born ratio *IEEE Trans. Med. Imaging* **24** 1377–86
[48] Schweiger M and Arridge S R 2014 The `TOAST++` software suite for forward and inverse modelling in optical tomography *J. Biomed. Opt.* **19** 040801
[49] Silvester D, Elman H and Ramagel A 2012 Incompressible flow and iterative solver software (IFISS) version 3.2 (http://manchester.ac.uk/ifiss/)
[50] Elman H C, Ramage A and Silvester D J 2012 IFISS: A computational laboratory for investigating incompressible flow problems *Technical report* MIMS Eprint 2012.81