



# Inexact Derivative-Free Optimization for Bilevel Learning

Matthias J. Ehrhardt<sup>1</sup> · Lindon Roberts<sup>2</sup>

Received: 1 July 2020 / Accepted: 8 December 2020 / Published online: 6 February 2021  
© The Author(s) 2021

## Abstract

Variational regularization techniques are dominant in the field of mathematical imaging. A drawback of these techniques is that they are dependent on a number of parameters which have to be set by the user. A by-now common strategy to resolve this issue is to learn these parameters from data. While mathematically appealing, this strategy leads to a nested optimization problem (known as bilevel optimization) which is computationally very difficult to handle. It is common when solving the upper-level problem to assume access to exact solutions of the lower-level problem, which is practically infeasible. In this work we propose to solve these problems using inexact derivative-free optimization algorithms which never require exact lower-level problem solutions, but instead assume access to approximate solutions with controllable accuracy, which is achievable in practice. We prove global convergence and a worst-case complexity bound for our approach. We test our proposed framework on ROF denoising and learning MRI sampling patterns. Dynamically adjusting the lower-level accuracy yields learned parameters with similar reconstruction quality as high-accuracy evaluations but with dramatic reductions in computational work (up to 100 times faster in some cases).

**Keywords** Derivative-free optimization · Bilevel optimization · Machine learning · Variational regularization

**Mathematics Subject Classification (2010)** 65D18 · 65K10 · 68T05 · 90C26 · 90C56

## 1 Introduction

Variational regularization techniques are dominant in the field of mathematical imaging. For example, when solving a linear inverse problem  $Ax = y$ , variational regularization can be posed as the solution to

$$\min_x \mathcal{D}(Ax, y) + \alpha \mathcal{R}(x). \quad (1)$$

Here, the data fidelity  $\mathcal{D}$  is usually chosen related to the assumed noise model of the data  $y$  and the regularizer  $\mathcal{R}$

models our a priori knowledge of the unknown solution. Many options have been proposed in the literature, see, for instance, [1,8,14,29,44] and references therein. An important parameter for any variational regularization technique is the regularization parameter  $\alpha$ . While some theoretical results and heuristic choices have been proposed in the literature, see, e.g., [8,23] and references therein or the L-curve criterion [27], the appropriate choice of the regularization parameter in a practical setting remains an open problem. Similarly, other parameters in (1) have to be chosen by the user, such as smoothing of the total variation [14], the hyperparameter for total generalized variation [9] or the sampling pattern in magnetic resonance imaging (MRI), see, e.g., [25,45,46].

Instead of using heuristics for choosing all of these parameters, here we are interested in finding these from data. A by-now common strategy to learn parameters of a variational regularization model from data is bilevel learning, see, e.g., [4,21,28,30,38,39,45] and references in [1]. Given labeled data  $(x_i, y_i)_{i=1, \dots, n}$  we find parameters  $\theta \in \Theta \subset \mathbb{R}^m$  by

---

MJE acknowledges support from the EPSRC (EP/S026045/1, EP/T026693/1), the Faraday Institution (EP/T007745/1) and the Leverhulme Trust (ECF-2019-478)..

---

✉ Matthias J. Ehrhardt  
m.ehrhardt@bath.ac.uk

✉ Lindon Roberts  
lindon.roberts@anu.edu.au

<sup>1</sup> Institute for Mathematical Innovation and Department of Mathematical Sciences, University of Bath, Bath, UK

<sup>2</sup> Mathematical Sciences Institute, Australian National University, Canberra, Australia

solving the upper-level problem

$$\min_{\theta \in \Theta} f(\theta) := \frac{1}{n} \sum_{i=1}^n \|\hat{x}_i(\theta) - x_i\|^2 + \mathcal{J}(\theta), \quad (2)$$

where  $\hat{x}_i(\theta) \in X \subset \mathbb{R}^d$  aims to recover the true data  $x_i$  by solving the lower-level problems

$$\hat{x}_i(\theta) := \arg \min_{x \in X} \Phi_{i,\theta}(x), \quad \forall i = 1, \dots, n. \quad (3)$$

The lower-level objective  $\Phi_{i,\theta}$  could be of the form  $\Phi_{i,\theta}(x) = \mathcal{D}(Ax, y_i) + \theta \mathcal{R}(x)$  as in (1), but we will not restrict ourselves to this special case. In general  $\Phi_{i,\theta}$  will depend on the data  $y_i$ .

In many situations, it is possible to acquire suitable data  $(x_i, y_i)_{i=1, \dots, n}$ . For image denoising, we may take any ground truth images  $x_i$  and add artificial noise to generate  $y_i$ . Alternatively, if we aim to learn a sampling pattern (such as for learning MRI sampling patterns, which we consider in this work), then  $x_i$  can be any fully sampled image. The same also holds for problems such as image compression, where again  $x_i$  is any ground truth image. In both these cases,  $y_i$  is subsampled information from  $x_i$  (depending on  $\theta$ ) from which the remaining information is reconstructed to get  $\hat{x}_i(\theta)$ .

While mathematically appealing, this nested optimization problem is computationally very difficult to handle since even the evaluation of the upper-level problem (2) requires the exact solution of the lower-level problems (3). This requirement is practically infeasible, and common algorithms in the literature compute the lower-level solution only to some accuracy, thereby losing any theoretical performance guarantees, see, e.g., [21,39,45]. One reason for needing exact solutions is to compute the gradient of the upper-level objective using the implicit function theorem [45], which we address by using upper-level solvers which do not require gradient computations.

In this work we propose to solve these problems using inexact derivative-free optimization (DFO) algorithms which never require exact solutions to the lower-level problem while still yielding convergence guarantees. Moreover, by dynamically adjusting the accuracy we gain a significant computational speedup compared to using a fixed accuracy for all lower-level solves. The proposed framework is tested on two problems: learning regularization parameters for ROF denoising and learning the sampling pattern in MRI.

We contrast our approach to [30], which develops a semismooth Newton method to solve the full bilevel optimality conditions. In [30] the upper- and lower-level problems are of specific structure, and exact solutions of the (possibly very large) Newton system are required. Separately, the approach in [38] replaces the lower-level problem with finitely many iterations of some algorithm and solves this

perturbed problem exactly. Our formulation is very general, and all approximations are controlled to guarantee convergence to the solution of the original variational problem.

**Aim:** Use inexact computations of  $\hat{x}_i(\theta)$  within a derivative-free upper-level solver, which makes (2) computationally tractable, while retaining convergence guarantees.

## 1.1 Derivative-Free Optimization

Derivative-free optimization methods—that is, optimization methods that do not require access to the derivatives of the objective (and/or constraints)—have grown in popularity in recent years and are particularly suited to settings where the objective is computationally expensive to evaluate and/or noisy; we refer the reader to [2,18] for background on DFO and examples of applications and to [32] for a comprehensive survey of recent work. The use of DFO for algorithm tuning has previously been considered in a general framework [3], and in the specific case of hyperparameter tuning for neural networks in [31].

Here, we are interested in the particular setting of learning for variational methods (2), which has also been considered in [39] where a new DFO algorithm based on discrete gradients has been proposed. In [39] it was assumed that the lower-level problem can be solved exactly such that the bilevel problem can be reduced to a single nonconvex optimization problem. In the present work we lift this stringent assumption.

In this paper we focus on DFO methods which are adapted to nonlinear least-squares problems as analyzed in [13,47]. These methods are called model based in that they construct a model approximating the objective at each iteration, locally minimize the model to select the next iterate, and update the model with new objective information. Our work also connects to [19], which considers model-based bilevel optimization where both the lower- and upper-level problems are solved in a derivative-free manner; particular attention is given here to reusing evaluations of the (assumed expensive) lower-level objective at nearby upper-level parameters, to make lower-level model construction simpler.

Our approach for bilevel DFO is based on dynamic-accuracy (derivative-based) trust-region methods [17, Chapter 10.6]. In these approaches, we use the measures of convergence (e.g., trust-region radius, model gradient) to determine a suitable level of accuracy with which to evaluate the objective; we start with low-accuracy requirements, and increase the required accuracy as we converge to a solution. In a DFO context, this framework is the basis of [19], and a similar approach was considered in [15] in the context of analyzing protein structures. This framework has also been recently extended in a derivative-based context to higher-order regularization methods [7,26]. We also note that there has been some work on multilevel and multi-fidelity models (in both a DFO and derivative-based context), where

an expensive objective can be approximated by surrogates which are cheaper to evaluate [11,34].

## 1.2 Contributions

There are a number of novel aspects to this work. Our use of DFO for bilevel learning means our upper-level solver genuinely expects inexact lower-level solutions. We give worst-case complexity theory for our algorithm both in terms of upper-level iterations and computational work from the lower-level problems. Our numerical results on ROF denoising and a new framework for learning MRI sampling patterns demonstrate our approach is substantially faster—up to 100 times faster—than the same DFO approach with high-accuracy lower-level solutions, while achieving the same quality solutions. More details on the different aspects of our contributions are given below.

### 1.2.1 Dynamic Accuracy DFO Algorithm for Bilevel Learning

As noted in [45], bilevel learning can require very high-accuracy solutions to the lower-level problem. We avoid this via the introduction of a dynamic accuracy model-based DFO algorithm. In this setting, the upper-level solver dynamically changes the required accuracy for lower-level problem minimizers, where less accuracy is required in earlier phases of the upper-level optimization. The proposed algorithm is similar to [19], but adapted to the nonlinear least-squares case and allowing derivative-based methods to solve the lower-level problem. Our theoretical results extend the convergence results of [19] to include derivative-based lower-level solvers and a least-squares structure, as well as adding a worst-case complexity analysis in a style similar to [13] (which is also not present in the derivative-based convergence theory in [17]). This analysis gives bounds on the number of iterations of the upper-level solver required to reach a given optimality, which we then extend to bound the total computational effort required for the lower-level problem solves. There is increasing interest, but comparatively fewer works, which explicitly bound the total computational effort of nonconvex optimization methods; see [42] for Newton-CG methods and references therein. We provide a preliminary argument that our computational effort bounds are tight with regards to the desired upper-level solution accuracy, although we delegate a complete proof to future work.

### 1.2.2 Robustness

We observe in all our results using several lower-level solvers (gradient descent and FISTA) for a variety of applications that the proposed upper-level DFO algorithm converges to similar objective values and minimizers. We also present numerical results for denoising showing that the learned parameters are

robust to initialization of the upper-level solver despite the upper-level problem being likely nonconvex. Together, these results suggest that this framework is a robust approach for bilevel learning.

### 1.2.3 Efficiency

Bilevel learning with a DFO algorithm was previously considered [39], but there a different DFO method based on discrete gradients was used, and was applied to nonsmooth problems with exact lower-level evaluations. In [39], only up to two parameters were learned, whereas here we demonstrate our approach is capable of learning many more. Our numerical results include examples with up to 64 parameters.

We demonstrate that the dynamic accuracy DFO achieves comparable or better objective values than the fixed accuracy variants and final reconstructions of comparable quality. However, our approach is able to achieve this with a dramatically reduced computational load, in some cases up to 100 times less work than the fixed accuracy variants.

### 1.2.4 New Framework for Learning MRI Sampling

We introduce a new framework to learn the sampling pattern in MRI based on bilevel learning. Our idea is inspired by the image inpainting model of [16]. Compared to other algorithms to learn the sampling pattern in MRI based on first-order methods [45], the proposed approach seems to be much more robust to initialization and choice of solver for the lower-level problem. As with the denoising examples, our dynamic accuracy DFO achieves the same upper-level objective values and final reconstructions as fixed accuracy variants but with substantial reductions in computational work.

### 1.2.5 Regularization Parameter Choice Rule with Machine Learning

Our numerical results suggest that the bilevel framework can learn regularization parameter choice rule which yields a convergent regularization method in the sense of [29,44], indicating for the first time that machine learning can be used to learn mathematically sound regularization methods.

## 1.3 Structure

In Sect. 2 we describe problems where the lower-level model (1) applies and describe how to efficiently attain a given accuracy level using standard first-order methods. Then, in Sect. 3 we introduce the dynamic accuracy DFO algorithm and present our global convergence and worst-case complexity bounds. Finally, our numerical experiments are described in Sect. 4.

### 1.4 Notation

Throughout, we let  $\| \cdot \|$  denote the Euclidean norm of a vector in  $\mathbb{R}^n$  and the operator 2-norm of a matrix in  $\mathbb{R}^{m \times n}$ . We also define the weighted (semi)norm  $\|x\|_S^2 := x^T S x$  for a symmetric and positive (semi)definite matrix  $S$ . The gradient of a scalar-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is denoted by  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , and the derivative of a vector-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is denoted by  $\partial f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ ,  $(\partial f)_{i,j} = \partial_i f_j$  where  $\partial_i f_j$  denotes the partial derivative of  $f_j$  with respect to the  $i$ th coordinate. If  $f$  is a function of two variables  $x$  and  $y$ , then  $\partial_x f$  denotes the derivative with respect to  $x$ .

### 1.5 Software

Our implementation of the DFO algorithm and all numerical testing code will be made public upon acceptance.

## 2 Lower-Level Problem

In order to have sufficient control over the accuracy of the solution to (3) we will assume that  $\Phi_{i,\theta}$  are  $L_i$ -smooth and  $\mu_i$ -strongly convex, see definitions below.

**Definition 1 (Smoothness)** A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -smooth if it is differentiable and its derivative is Lipschitz continuous with constant  $L > 0$ , i.e., for all  $x, y \in \mathbb{R}^n$  we have  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ .

**Definition 2 (Strong Convexity)** A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex for  $\mu > 0$  if  $f - \frac{\mu}{2}\| \cdot \|^2$  is convex.

Moreover, when the lower-level problem is strictly convex and smooth, with  $\Phi_i(x, \theta) := \Phi_{i,\theta}(x)$  we can equivalently describe the minimizer of  $\Phi_{i,\theta}$  by

$$\partial_x \Phi_i(\hat{x}_i(\theta), \theta) = 0. \tag{4}$$

Smoothness properties of  $\hat{x}_i$  follow from the implicit function theorem and its generalizations if  $\Phi_i$  is smooth and regular enough.

**Assumption 1** We assume that for all  $i = 1, \dots, n$  the following statements hold.

1. *Convexity:* For all  $\theta \in \Theta$  the functions  $\Phi_{i,\theta}$  are  $\mu_i$ -strongly convex.
2. *Smoothness in  $x$ :* For all  $\theta \in \Theta$  the functions  $\Phi_{i,\theta}$  are  $L_i$ -smooth.
3. *Smoothness in  $(x, \theta)$ :* The derivatives  $\partial_x \Phi_i : X \times \Theta \rightarrow X$  and  $\partial_{xx} \Phi_i : X \times \Theta \rightarrow X^2$  exist and are continuous.

**Theorem 1** Under Assumption 1 the function  $\hat{x}_i(\theta) := \arg \min_x \Phi_i(x, \theta)$  is

1. *well-defined*
2. *locally Lipschitz*
3. *continuously differentiable and*  $\partial \hat{x}_i(\theta) = -\partial_{xx} \Phi_i(\hat{x}_i(\theta), \theta)^{-1} \partial_x \partial_\theta \Phi_i(\hat{x}_i(\theta), \theta)$ .

**Proof** Ad (1) Finite and convex functions are continuous [41, Corollary 2.36]. It is easy to show that  $\mu$ -strongly convex functions are coercive. Then, the existence and uniqueness follow from classical theorems, e.g., [10, Theorem 6.31]. Ad (2) This statement follows directly from [40, Theorem 2.1]. Ad (3) This follows directly from the classical inverse function theorem, see, e.g., [22, Theorem 3.5.1].  $\square$

### 2.1 Examples

A relevant case of the model introduced above is the parameter tuning for linear inverse problems, which can be solved via the variational regularization model

$$\frac{1}{2} \|Ax - y_i\|_S^2 + \alpha \text{TV}(x), \tag{5}$$

where  $\text{TV}(x) := \sum_{j=1}^m \|\widehat{\nabla} x(j)\|$  denotes the discretized total variation, e.g.,  $\widehat{\nabla} x(j)$  is the finite forward difference discretization of the spatial gradient of  $x$  at pixel  $j$ . However, we note that (5) does not satisfy Assumption 1.

To ensure Assumption 1 holds, we instead use  $\|x\| \approx \sqrt{\|x\|^2 + v^2}$ , to approximate problem (5) by a smooth and strongly convex problem of the form

$$\hat{x}_i(\theta) := \arg \min_x \left\{ \Phi_{i,\theta}(x) = \frac{1}{2} \|A(\theta)x - y_i\|_{S(\theta)}^2 + \alpha(\theta) \text{TV}_{v(\theta)}(x) + \frac{\xi(\theta)}{2} \|x\|^2 \right\}, \tag{6}$$

with the smoothed total variation given by  $\text{TV}_{v(\theta)}(x) := \sum_{j=1}^m \sqrt{\|\widehat{\nabla} x(j)\|^2 + v(\theta)^2}$ . Here we already introduced the notation that various parts of the problem may depend on a vector of parameters  $\theta$  which usually needs to be selected manually. We will learn these parameters using the bilevel framework. For simplicity denote  $A_\theta := A(\theta)$ ,  $S_\theta := S(\theta)$ ,  $\alpha_\theta := \alpha(\theta)$ ,  $v_\theta := v(\theta)$  and  $\xi_\theta := \xi(\theta)$ . Note that  $\Phi_{i,\theta}$  in (6) is  $L_i$ -smooth and  $\mu_i$ -strongly convex with

$$L_i \leq \|A_\theta^* S_\theta A_\theta\| + \alpha_\theta \frac{\|\partial\|^2}{v_\theta} + \xi_\theta, \quad \text{and} \\ \mu_i \geq \lambda_{\min}(A_\theta^* S_\theta A_\theta) + \xi_\theta, \tag{7}$$

where  $\lambda_{\min}(A_\theta^* S_\theta A_\theta)$  denotes the smallest eigenvalue of  $A_\theta^* S_\theta A_\theta$  and  $A_\theta^*$  is the adjoint of  $A_\theta$ .

We now describe two specific problems we will use in our numerical results. They both choose a specific form for (3) which aims to find a minimizer  $\hat{x}_i(\theta)$  which (approximately)

recovers the data  $x_i$ , and so both use (2) as the upper-level problem.

### 2.1.1 Total Variation-Based Denoising

A particular problem we consider is a smoothed version of the ROF model [43], i.e.,  $A_\theta = I$ ,  $S_\theta = I$ . Then, (6) simplifies to

$$\Phi_{i,\theta}(x) = \frac{1}{2} \|x - y_i\|^2 + \alpha_\theta \text{TV}_{v_\theta}(x) + \frac{\xi_\theta}{2} \|x\|^2, \quad (8)$$

which is  $L_i$ -smooth and  $\mu_i$ -strongly convex with constants as in (7) with  $\|A_\theta^* S_\theta A_\theta\| = \|I\| = 1$  and  $\lambda_{\min}(A_\theta^* S_\theta A_\theta) = \lambda_{\min}(I) = 1$ . In our numerical examples we will consider two cases. First, we will just learn the regularization parameter  $\alpha$  given manually set  $\nu$  and  $\xi$ . Second, we will learn all three parameters  $\alpha$ ,  $\nu$  and  $\xi$ .

### 2.1.2 Undersampled MRI Reconstruction

Another problem we consider is the reconstruction from undersampled MRI data, see, e.g., [33], which can be phrased as (6) with  $A_\theta = F$  where  $F$  is the discrete Fourier transform and  $S_\theta = \text{diag}(s)$ ,  $s \in [0, 1]^d$ . Then, (6) simplifies to

$$\Phi_{i,\theta}(x) = \frac{1}{2} \|Fx - y_i\|_{S_\theta}^2 + \alpha_\theta \text{TV}_{v_\theta}(x) + \frac{\xi_\theta}{2} \|x\|^2, \quad (9)$$

which is  $L_i$ -smooth and  $\mu_i$ -strongly convex with constants as in (7) with  $\|A_\theta^* S_\theta A_\theta\| \leq 1$  and  $\lambda_{\min}(A_\theta^* S_\theta A_\theta) \geq 0$ . The sampling coefficients  $s_j$  indicate the relevance of a sampling location. The data term (9) can be rewritten as

$$\|Fx - y_i\|_{S_\theta}^2 = \sum_{s_j > 0} s_j | [Fx - y_i]_j |^2. \quad (10)$$

Most commonly the values  $s$  are binary and manually chosen. Here we aim to use bilevel learning to find a sparse  $s$  such that the images  $x_i$  can be reconstructed well from sparse samples of  $y_i$ . This approach was first proposed in [45].

## 2.2 Example Training Data

Throughout this paper, we will consider training data of artificially generated 1D images. Each ground truth image  $x_i$  is randomly generated piecewise-constant function. For a desired image size  $N$ , we select values  $C_i \in [N/4, 3N/4]$  and  $R_i \in [N/8, N/4]$  from a uniform distribution. We then define  $x_i \in \mathbb{R}^N$  by

$$[x_i]_j := \begin{cases} 1, & |j - C_i| < R_i, \\ 0, & \text{otherwise,} \end{cases} \quad \forall j = 1, \dots, N. \quad (11)$$

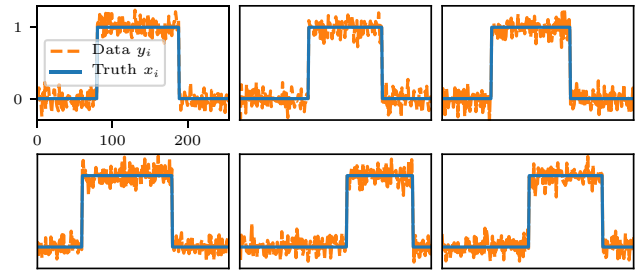


Fig. 1 Examples of training pairs  $(x_i, y_i)$  for image denoising

That is, each  $x_i$  is zero except for a single randomly generated subinterval of length  $2R_i$  centered around  $C_i$  where it takes the value 1.

We then construct our  $y_i$  by taking the signal to be reconstructed and adding Gaussian noise. Specifically, for the image denoising problem we take

$$y_i := x_i + \sigma \omega_i, \quad (12)$$

where  $\sigma > 0$  and  $\omega_i \in \mathbb{R}^N$  is randomly drawn vector of i.i.d. standard Gaussians. For the MRI sampling problem, we take

$$y_i := Fx_i + \frac{\sigma}{\sqrt{2}} \omega_i. \quad (13)$$

where  $\sigma > 0$  and  $\omega_i \in \mathbb{C}^N$  is a randomly drawn vector with real and imaginary parts both standard Gaussians.

In Fig. 1 we plot an example collection of pairs  $(x_i, y_i)$  for the image denoising problem with  $N = 256$ , and in Fig. 2 we plot the solution to (8) for the first of these  $(x_i, y_i)$  pairs for a variety of choices for the parameters  $\alpha_\theta, \epsilon_\theta, \eta_\theta$ .

## 2.3 Approximate Solutions

### 2.3.1 Gradient Descent

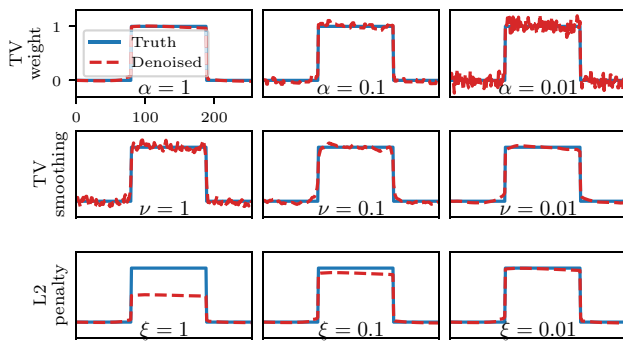
For simplicity we drop the dependence on  $i$  for the remainder of this section.

The lower-level problem (3) can be solved with gradient descent (GD) which converges linearly for  $L$ -smooth and  $\mu$ -strongly convex problems. One can show (e.g., [14]) that GD

$$x^{k+1} = x^k - \tau \nabla \Phi(x^k), \quad (14)$$

with  $\tau = 1/L$ , converges linearly to the unique solution  $x^*$  of (3). More precisely, for all  $k \in \mathbb{N}$  we have [5, Theorem 10.29]

$$\|x^k - x^*\|^2 \leq (1 - \mu/L)^k \|x^0 - x^*\|^2. \quad (15)$$



**Fig. 2** Examples of denoised data using model (8) obtained by running GD with a tolerance of  $\|x^k - x^*\| \leq 1e-6$ . The data  $(x_i, y_i)$  are the top-left image in Fig. 1. **Top:** results with  $\alpha = 1, 0.1, 0.01$  (left to right) with  $\nu = \xi = 1e-3$  throughout. **Middle:** results with  $\nu = 1, 0.1, 0.01$  (left to right) with  $\alpha = 1$  and  $\xi = 1e-3$  throughout. **Bottom:** results with  $\xi = 1, 0.1, 0.01$  (left to right) with  $\alpha = 1$  and  $\nu = 1e-3$  throughout

Moreover, if one has a good estimate of the strong convexity constant  $\mu$ , then it is better to choose  $\tau = 2/(L + \mu)$ , which gives an improved linear rate [36, Theorem 2.1.15]

$$\|x^k - x^*\|^2 \leq (1 - \mu/L)^{2k} \|x^0 - x^*\|^2. \tag{16}$$

### 2.3.2 FISTA

Similarly, we can use FISTA [6] to approximately solve the lower-level problem. FISTA applied to a smooth objective with convex constraints is a modification of [35] and can be formulated as the iteration

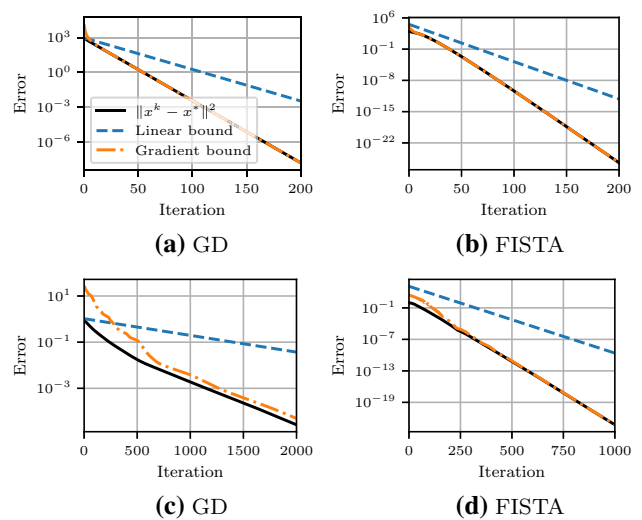
$$\begin{aligned} t_{k+1} &= \frac{1 - qt_k^2 + \sqrt{(1 - qt_k^2)^2 + 4t_k^2}}{2}, \\ \beta_{k+1} &= \frac{(t_k - 1)(1 - t_{k+1}q)}{t_{k+1}(1 - q)}, \\ z^{k+1} &= x^k + \beta_{k+1}(x^k - x^{k-1}), \\ x^{k+1} &= z^{k+1} - \tau \nabla \Phi(z^{k+1}), \end{aligned} \tag{17}$$

where  $q := \tau\mu$ , and we choose  $\tau = 1/L$  and  $t_0 = 0$  [14, Algorithm 5]. We then achieve linear convergence with [14, Theorem 4.10]

$$\Phi(x^k) - \Phi(x^*) \leq (1 - \sqrt{q})^k \left[ \frac{L}{2} (1 + \sqrt{q}) \|x^0 - x^*\|^2 \right], \tag{18}$$

and so, since  $\Phi(x^k) - \Phi(x^*) \geq (\mu/2) \|x^k - x^*\|^2$  from  $\mu$ -strong convexity, we get

$$\|x^k - x^*\|^2 \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left[ \frac{L}{\mu} \left(1 + \sqrt{\frac{\mu}{L}}\right) \|x^0 - x^*\|^2 \right]. \tag{19}$$



**Fig. 3** Comparison of a priori linear convergence bounds (16) and (19) against the a posteriori gradient bound (20). **a, b:** 200 iterations of GD and FISTA on Nesterov’s quadratic function. **c, d:** GD (2,000 iterations) and FISTA (1,000 iterations) on a 1D denoising problem

### 2.3.3 Ensuring Accuracy Requirements

We will need to be able to solve the lower-level problem to sufficient accuracy that we can guarantee  $\|x^k - x^*\|^2 \leq \epsilon$ , for a suitable accuracy  $\epsilon > 0$ . We can guarantee this accuracy by ensuring we terminate with  $k$  sufficiently large, given an estimate  $\|x^0 - x^*\|^2$ , using the a priori bounds (15) or (19). A simple alternative is to use the a posteriori bound  $\|x - x^*\| \leq \|\nabla \Phi(x)\|/\mu$  for all  $x$  (a consequence of [6, Theorem 5.24(iii)]), and terminate once

$$\|\nabla \Phi(x^k)\|^2 / \mu^2 \leq \epsilon. \tag{20}$$

To compare these two options, we consider two test problems: (i) a version of Nesterov’s quadratic [36, Section 2.1.4] in  $\mathbb{R}^{10}$ , and (ii) 1D image denoising. Nesterov’s quadratic is defined as

$$\begin{aligned} \Phi(x) &:= \frac{\tilde{\mu}(Q - 1)}{8} (x^T A x - 2x_1) + \frac{\tilde{\mu}}{2} \|x\|^2, \\ \text{where } A &:= \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & & -1 & 2 \end{bmatrix}, \end{aligned} \tag{21}$$

for  $x \in \mathbb{R}^{10}$ , with  $\tilde{\mu} = 1$  and  $Q = 100$ , which is  $\mu$ -strongly convex and  $L$ -smooth for  $\mu \approx 3$  and  $L \approx 98$ ; we apply no constraints,  $X = \mathbb{R}^{10}$ .

We also consider a 1D denoising problem as in (8) with randomly generated data  $y \in \mathbb{R}^N$  (with  $N = 100$  pixels) as per Section 2.2,  $\alpha = 0.3$ ,  $\nu = \xi = 10^{-3}$ , and  $x^*$  estimated

by running  $10^4$  iterations of FISTA. Here, the problem is  $\mu$ -strongly convex and  $L$ -smooth with  $\mu \approx 1$  and  $L \approx 1, 201$ . We estimate the true solution  $x^*$  by running FISTA for 10,000 iterations (which gives an upper bound estimate  $\|x^k - x^*\|^2 \leq 3e-26$  from (20)).

In Fig. 3, we compare the true error  $\|x^k - x^*\|^2$  against the a priori linear convergence bounds (15) or (19) with the true value of  $\|x^0 - x^*\|^2$ , and the a posteriori gradient bound (20). In both cases, the gradient-based bound (20) provides a much tighter estimate of the error, particularly for high-accuracy requirements. Thus, in our numerical results, we terminate the lower-level solver as soon (20) is achieved for our desired tolerance. The gradient-based bound has the additional advantage of not requiring an a priori estimate of  $\|x^0 - x^*\|$ . For comparison, in our results below we will also consider terminating GD/FISTA after a fixed number of iterations.

### 3 Dynamic Accuracy DFO Algorithm

#### 3.1 DFO Background

Since evaluating  $\hat{x}_i(\theta)$  in the upper-level problem (2) is only possible with some error (it is computed by running an iterative process), it is not straightforward or cheap to evaluate  $\partial \hat{x}_i(\theta)$ . Hence for solving (2) we turn to DFO techniques, and specifically consider those which exploit the nonlinear least-squares problem structure. In this section we outline a model-based DFO method for nonlinear least-squares problems [13], a trust-region method based on the classical (derivative-based) Gauss–Newton method [37, Chapter 10]. However, these approaches are based on having access to exact function evaluations, and so we augment this with a standard approach for dynamic accuracy trust-region methods [17, Chapter 10.6]; this was previously considered for general model-based DFO methods in [19].

Here, we write the upper-level problem (2) in the general form

$$\min_{\theta \in \mathbb{R}^d} f(\theta) := \frac{1}{n} \|r(\theta)\|^2 = \frac{1}{n} \sum_{i=1}^n r_i(\theta)^2, \tag{22}$$

where  $r_i(\theta) := \|\hat{x}_i(\theta) - x_i\|$  and  $r(\theta) := [r_1(\theta), \dots, r_n(\theta)]^T$ . Without loss of generality, we do not include a regularization term  $\mathcal{J}(\theta)$ ; we can incorporate this term by defining  $r_{n+1}(\theta) := \sqrt{\mathcal{J}(\theta)}$  and then taking  $r(\theta) := [r_1(\theta), \dots, r_{n+1}(\theta)]^T$ , for instance.

The upper-level objective (22) assumes access to exact evaluations of the lower-level objective  $r_i(\theta)$ , which is not achievable in practice. We therefore assume we only have access to inaccurate evaluations  $\tilde{x}_i(\theta) \approx \hat{x}_i(\theta)$ , giving

$$\tilde{r}_i(\theta) := \|\tilde{x}_i(\theta) - x_i\|, \tilde{r}(\theta) := [\tilde{r}_1(\theta), \dots, \tilde{r}_n(\theta)]^T, \text{ and } \tilde{f}(\theta) := \frac{1}{n} \|\tilde{r}(\theta)\|^2.$$

Our overall algorithmic framework is based on trust-region methods, where at each iteration  $k$  we construct a model  $m^k$  for the objective which we hope is accurate in a neighborhood of our current iterate  $\theta^k$ . Simultaneously we maintain a trust-region radius  $\Delta^k > 0$ , which tracks the size of the neighborhood of  $\theta^k$  where we expect  $m^k$  to be accurate. Our next iterate is determined by minimizing the model  $m^k$  within a ball of size  $\Delta^k$  around  $\theta^k$ .

Usually  $m^k$  is taken to be a quadratic function (e.g., a second-order Taylor series for  $f$  about  $\theta^k$ ). However, here we use the least-squares problem structure (22) and construct a linear model

$$r(\theta^k + s) \approx \tilde{r}(\theta^k + s) \approx M^k(s) := \tilde{r}(\theta^k) + J^k s, \tag{23}$$

where  $\tilde{r}(\theta^k)$  is our approximate evaluation of  $r(\theta^k)$  and  $J^k \in \mathbb{R}^{n \times d}$  is a matrix approximating  $\partial r(\theta^k)^T$ . We construct  $J^k$  by interpolation: we maintain an interpolation set  $z^0, \dots, z^d \in \mathbb{R}^d$  (where  $z^0 := \theta^k$  at each iteration  $k$ ) and choose  $J^k$  so that

$$M^k(z^t - \theta^k) = \tilde{r}(z^t), \quad \forall t = 1, \dots, d. \tag{24}$$

This condition ensures that our linear model  $M^k$  exactly interpolates  $\tilde{r}$  at our interpolation points  $z^t$  (i.e., the second approximation in (23) is exact for each  $s = z^t - \theta^k$ ). We can therefore find  $J^k$  by solving the  $d \times d$  linear system (with  $n$  right-hand sides):

$$[(z^1 - \theta^k) \dots (z^d - \theta^k)]^T g_i^k = \begin{bmatrix} \tilde{r}_i(z^1) - \tilde{r}_i(\theta^k) \\ \vdots \\ \tilde{r}_i(z^d) - \tilde{r}_i(\theta^k) \end{bmatrix}, \tag{25}$$

for all  $i = 1, \dots, n$ , where  $g_i^k \in \mathbb{R}^d$  is the  $i$ -th row of  $J^k$ . The model  $M^k$  gives a natural quadratic model for the full objective  $f$ :

$$f(\theta^k + s) \approx m^k(s) := \frac{1}{n} \|M^k(s)\|^2 = \tilde{f}(\theta^k) + (g^k)^T s + \frac{1}{2} s^T H^k s, \tag{26}$$

where  $g^k := \frac{2}{n} (J^k)^T \tilde{r}(\theta^k)$  and  $H^k := \frac{2}{n} (J^k)^T J^k$ . We compute a tentative step  $s^k$  as a(n approximate) minimizer of the trust-region subproblem

$$\min_{s \in \mathbb{R}^d} m^k(s), \quad \text{subject to } \|s\| \leq \Delta^k. \tag{27}$$

There are a variety of efficient algorithms for computing  $s^k$  [17, Chapter 7]. Finally, we evaluate  $\tilde{f}(\theta^k + s^k)$  and decide

whether to accept or reject the step (i.e., set  $\theta^{k+1} = \theta^k + s^k$  or  $\theta^{k+1} = \theta^k$ ) depending on the ratio

$$\rho^k = \frac{\text{actual reduction}}{\text{predicted reduction}} := \frac{f(\theta^k) - f(\theta^k + s^k)}{m^k(0) - m^k(s^k)}. \tag{28}$$

Although we would like to accept/reject using  $\rho^k$ , in reality we only observe the approximation

$$\tilde{\rho}^k := \frac{\tilde{f}(\theta^k) - \tilde{f}(\theta^k + s^k)}{m^k(0) - m^k(s^k)}, \tag{29}$$

and so we use this instead.

This gives us the key components of a standard trust-region algorithm. We have two extra considerations in our context: the accuracy of our derivative-free model (26) and the lack of exact evaluations of the objective.

Firstly, we require a procedure to verify if our model (26) is sufficiently accurate inside the trust-region, and if not, modify the model to ensure its accuracy. We discuss this in Section 3.2. The notion of ‘sufficiently accurate’ we use here is that  $m^k$  is as good an approximation to  $f$  as a first-order Taylor series (up to constant factors), which we call ‘fully linear.’<sup>1</sup>

**Definition 3 (Fully linear model)** The model  $m^k$  (26) is a fully linear model for  $f(\theta)$  in  $B(\theta^k, \Delta^k)$  if there exist constants  $\kappa_{\text{ef}}, \kappa_{\text{eg}} > 0$  (independent of  $\theta^k$  and  $\Delta^k$ ) such that

$$|f(\theta^k + s) - m^k(s)| \leq \kappa_{\text{ef}}(\Delta^k)^2, \tag{30}$$

$$\|\nabla f(\theta^k + s) - \nabla m^k(s)\| \leq \kappa_{\text{eg}}\Delta^k, \tag{31}$$

for all  $\|s\| \leq \Delta^k$ .

Secondly, we handle the inaccuracy in objective evaluations by ensuring  $\tilde{f}(\theta^k)$  and  $\tilde{f}(\theta^k + s^k)$  are evaluated to a sufficiently high accuracy when we compute  $\tilde{\rho}^k$  (29). Specifically, suppose we know that  $|\tilde{f}(\theta^k) - f(\theta^k)| \leq \delta^k$  and  $|\tilde{f}(\theta^k + s^k) - f(\theta^k + s^k)| \leq \delta_+^k$  for some accuracies  $\delta^k$  and  $\delta_+^k$ . Throughout, we use  $\delta^k$  and  $\delta_+^k$  to refer to the accuracies with which  $\tilde{f}(\theta^k)$  and  $\tilde{f}(\theta^k + s^k)$  have been evaluated, in the sense above. Before we compute  $\tilde{\rho}^k$ , we first ensure that

$$\max(\delta^k, \delta_+^k) \leq \eta'_1 [m^k(0) - m^k(s^k)], \tag{32}$$

where  $\eta'_1 > 0$  is an algorithm parameter. We achieve this by running the lower-level solver for a sufficiently large number of iterations.

The full upper-level algorithm is given in Algorithm 1; it is similar to the approach in [19], the DFO method [18, Algorithm 10.1]—adapted for the least-squares problem structure—and the (derivative-based) dynamic accuracy trust-region method [17, Algorithm 10.6.1].

<sup>1</sup> If  $f$  is  $L$ -smooth, then the Taylor series  $m^k(s) = f(\theta^k) + \nabla f(\theta^k)^T s$  is fully linear with  $\kappa_{\text{ef}} = L/2$  and  $\kappa_{\text{eg}} = L$  for all  $\Delta^k$ .

**Algorithm 1** Dynamic accuracy DFO algorithm for (22).

**Inputs:** Starting point  $\theta^0 \in \mathbb{R}^n$ , initial trust-region radius  $0 < \Delta^0 \leq \Delta_{\text{max}}$ .

**Parameters:** strictly positive values  $\Delta_{\text{max}}, \gamma_{\text{dec}}, \gamma_{\text{inc}}, \eta_1, \eta_2, \eta'_1, \epsilon$  satisfying  $\gamma_{\text{dec}} < 1 < \gamma_{\text{inc}}, \eta_1 \leq \eta_2 < 1$ , and  $\eta'_1 < \min(\eta_1, 1 - \eta_2)/2$ .

- 1: Select an arbitrary interpolation set and construct  $m^0$  (26).
- 2: **for**  $k = 0, 1, 2, \dots$  **do**
- 3:   **repeat**
- 4:     Evaluate  $\tilde{f}(\theta^k)$  to sufficient accuracy that (32) holds with  $\eta'_1$  (using  $s^k$  from the previous iteration of this inner repeat/until loop). Do nothing in the first iteration of this repeat/until loop.
- 5:     **if**  $\|g^k\| \leq \epsilon$  **then**
- 6:       By replacing  $\Delta^k$  with  $\gamma_{\text{dec}}^i \Delta^k$  for  $i = 0, 1, 2, \dots$ , find  $m^k$  and  $\Delta^k$  such that  $m^k$  is fully linear in  $B(\theta^k, \Delta^k)$  and  $\Delta^k \leq \|g^k\|$ . [criticality phase]
- 7:     **end if**
- 8:     Calculate  $s^k$  by (approximately) solving (27).
- 9:     **until** the accuracy in the evaluation of  $\tilde{f}(\theta^k)$  satisfies (32) with  $\eta'_1$  [accuracy phase]
- 10:    Evaluate  $\tilde{r}(\theta^k + s^k)$  so that (32) is satisfied with  $\eta'_1$  for  $\tilde{f}(\theta^k + s^k)$ , and calculate  $\tilde{\rho}^k$  (29).
- 11:    Set  $\theta^{k+1}$  and  $\Delta^{k+1}$  as:

$$\theta^{k+1} = \begin{cases} \theta^k + s^k, & \tilde{\rho}^k \geq \eta_2, \text{ or } \tilde{\rho}^k \geq \eta_1 \text{ and } m^k \\ & \text{fully linear in } B(\theta^k, \Delta^k), \\ \theta^k, & \text{otherwise,} \end{cases} \tag{33}$$

and

$$\Delta^{k+1} = \begin{cases} \min(\gamma_{\text{inc}} \Delta^k, \Delta_{\text{max}}), & \tilde{\rho}^k \geq \eta_2, \\ \Delta^k, & \tilde{\rho}^k < \eta_2 \text{ and } m^k \text{ not} \\ & \text{fully linear in } B(\theta^k, \Delta^k), \\ \gamma_{\text{dec}} \Delta^k, & \text{otherwise.} \end{cases} \tag{34}$$

- 12:    If  $\theta^{k+1} = \theta^k + s^k$ , then build  $m^{k+1}$  by adding  $\theta^{k+1}$  to the interpolation set (removing an existing point). Otherwise, set  $m^{k+1} = m^k$  if  $m^k$  is fully linear in  $B(\theta^k, \Delta^k)$ , or form  $m^{k+1}$  by making  $m^k$  fully linear in  $B(\theta^{k+1}, \Delta^{k+1})$ .
- 13: **end for**

Our main convergence result is the below.

**Theorem 2** Suppose Assumptions 2 and 3 hold. Then if

$$K > \left\lceil \left( 2 + 2 \frac{\log \gamma_{\text{inc}}}{|\log \gamma_{\text{dec}}|} \right) \frac{2(\kappa_{\text{eg}} + 1)f(\theta^0)}{(\eta_1 - 2\eta'_1)\epsilon \Delta_{\text{min}}} + 2 \frac{\log(\Delta^0/\Delta_{\text{min}})}{|\log \gamma_{\text{dec}}|} \right\rceil, \tag{35}$$

with  $\kappa_{\text{eg}}$  and  $\Delta_{\text{min}}$  given by Lemmas 1 and 8, respectively, then  $\min_{k=0, \dots, K} \|\nabla f(\theta^k)\| < \epsilon$ .

We summarize Theorem 2 as follows, noting that the iteration and evaluation counts match the standard results for model-based DFO (e.g., [13,24]).

**Corollary 1** *Suppose the assumptions of Theorem 2 hold. Then, Algorithm 1 is globally convergent, i.e.,*

$$\lim_{k \rightarrow \infty} \|\nabla f(\theta^k)\| = 0. \quad (36)$$

Also, if  $\epsilon \in (0, 1]$ , then the number of iterations before  $\|\nabla f(\theta^k)\| < \epsilon$  for the first time is at most  $\mathcal{O}(\kappa^3 \epsilon^{-2})$  and the number of evaluations of  $\tilde{r}(\theta)$  is at most  $\mathcal{O}(d\kappa^3 \epsilon^{-2})$ , where  $\kappa := \max(\kappa_{\text{ef}}, \kappa_{\text{eg}}, \kappa_H)$ .<sup>2</sup>

We note that since  $\theta^k \in \mathcal{B}$  and  $\mathcal{B}$  is bounded from Assumption 2 (and closed by continuity of  $f$ ), then by Corollary 1 and compactness there exists a subsequence of iterates  $\{\theta_{k_i}\}_{i \in \mathbb{N}}$  which converges to a stationary point of  $f$ . However, there are relatively few results which prove convergence of the full sequence of iterates for nonconvex trust-region methods (see [17, Theorem 10.13] for a restricted result in the derivative-based context).

### 3.2 Guaranteeing Model Accuracy

As described above, we need a process to ensure that  $m^k$  (26) is a fully linear model for  $f$  inside the trust region  $B(\theta^k, \Delta^k)$ . For this, we need to consider the geometry of the interpolation set.

**Definition 4** The Lagrange polynomials of the interpolation set  $\{z^0, z^1, \dots, z^d\}$  are the linear polynomials  $\ell_t$ ,  $t = 0, \dots, d$  such that  $\ell_t(z^s) = \delta_{s,t}$  for all  $s, t = 0, \dots, d$ .

The Lagrange polynomials of  $\{z^0, \dots, z^d\}$  exist and are unique whenever the matrix in (25) is invertible. The required notion of ‘good geometry’ is given by the below definition (where small  $\Lambda$  indicates better geometry).

**Definition 5 ( $\Lambda$ -poisedness)** For  $\Lambda > 0$ , the interpolation set  $\{z^0, \dots, z^d\}$  is  $\Lambda$ -poised in  $B(\theta^k, \Delta^k)$  if  $|\ell_t(\theta^k + s)| \leq \Lambda$  for all  $t = 0, \dots, d$  and all  $\|s\| \leq \Delta^k$ .

The below result confirms that, provided our interpolation set has sufficiently good geometry, and our evaluations  $\tilde{r}(\theta^k)$  and  $\tilde{r}(y^t)$  are sufficiently accurate, our interpolation models are fully linear.

**Assumption 2** *The extended level set*

$$\mathcal{B} := \{z : z \in B(\theta, \Delta_{\max}) \text{ for some } \theta \text{ with } f(\theta) \leq f(\theta^0)\}, \quad (37)$$

*is bounded, and  $r(\theta)$  is continuously differentiable and  $\partial r(\theta)$  is Lipschitz continuous with constant  $L_J$  in  $\mathcal{B}$ .*

<sup>2</sup> If we have to evaluate  $\tilde{r}(\theta)$  at different accuracy levels as part of the accuracy phase, we count this as one evaluation, since we continue solving the corresponding lower-level problem from the solution from the previous, lower accuracy evaluation.

In particular, Assumption 2 implies that  $r(\theta)$  and  $\partial r(\theta)$  are uniformly bounded in the same region—that is,  $\|r(\theta)\| \leq r_{\max}$  and  $\|\partial r(\theta)\| \leq J_{\max}$  for all  $\theta \in \mathcal{B}$ —and  $f$  (22) is  $L$ -smooth in  $\mathcal{B}$  [13, Lemma 3.2].

We note that  $\mathcal{B}$  in Assumption 2 is bounded whenever the regularizer  $\mathcal{J}$  is coercive, such as in Sect. 4.4. This may also be replaced by the weaker assumption that  $r$  and  $\partial r$  are uniformly bounded on  $\mathcal{B}$  (and  $\mathcal{B}$  need not be bounded) [13, Assumption 3.1], and there are theoretical results which give this for some inverse problems in image restoration [20]. In our numerical experiments, we enforce upper and lower bounds on  $\theta$ , which also yields the uniform boundedness of  $r$  and  $\partial r$ . Also, we note that if  $r_i(\theta) = \|\hat{x}_i(\theta) - x_i\|$  is not itself  $L$ -smooth, we can instead treat each entry of  $\hat{x}_i(\theta) - x_i$  as a separate term in (22).

**Lemma 1** *Suppose Assumption 2 holds and  $\Delta^k \leq \Delta_{\max}$ . If the interpolation set  $\{z^0 := \theta^k, z^1, \dots, z^d\}$  is  $\Lambda$ -poised in  $B(\theta^k, \Delta^k)$  and for each evaluation  $t = 0, \dots, d$  and each  $i = 1, \dots, n$  we have*

$$\|\tilde{x}_i(z^t) - \hat{x}_i(z^t)\| \leq c(\Delta^k)^2, \quad (38)$$

*for some  $c > 0$ , then the corresponding models  $M^k$  (23) and  $m^k$  (26) are fully linear models for  $r(\theta)$  and  $f(\theta)$ , respectively.*

**Proof** This is a straightforward extension of [13, Lemma 3.3], noting that

$$|\tilde{r}_i(z^t) - r_i(z^t)| \leq \|\tilde{x}_i(z^t) - \hat{x}_i(z^t)\|, \quad \forall i = 1, \dots, n, \quad (39)$$

and so (38) gives  $\|\tilde{r}(z^t) - r(z^t)\| \leq c\sqrt{n}(\Delta^k)^2$  for all  $t = 0, \dots, d$ .

We conclude by noting that for any  $\Lambda > 1$  there are algorithms available to determine if a set is  $\Lambda$ -poised, and if it is not, change some interpolation points to make it so; details may be found in [18, Chapter 6], for instance.

### 3.3 Lower-Level Objective Evaluations

We now consider the accuracy requirements that Algorithm 1 imposes on our lower-level objective evaluations. In particular, we require the ability to satisfy (32), which imposes requirements on the error in the calculated  $\tilde{f}$ , rather than the lower-level evaluations  $\tilde{r}$ . The connection between errors in  $\tilde{r}$  and  $\tilde{f}$  is given by the below result.

**Lemma 2** *Suppose we compute  $\tilde{x}_i(\theta)$  satisfying  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| \leq \delta_x$  for all  $i = 1, \dots, n$ . Then, we have*

$$|\tilde{f}(\theta) - f(\theta)| \leq 2\sqrt{\tilde{f}(\theta)} \delta_x + \delta_x^2$$

$$\text{and } |\tilde{f}(\theta) - f(\theta)| \leq 2\sqrt{f(\theta)} \delta_x + \delta_x^2. \quad (40)$$

Moreover, if  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| \leq \sqrt{\tilde{f}(\theta) + \delta_f} - \sqrt{f(\theta)}$  for  $i = 1, \dots, n$ , then  $|\tilde{f}(\theta) - f(\theta)| \leq \delta_f$ .

**Proof** Letting  $\epsilon(\theta) := \tilde{r}(\theta) - r(\theta)$ , we have

$$f(\theta) = \frac{1}{n} \|\tilde{r}(\theta) - \epsilon(\theta)\|^2, \quad (41)$$

$$= \tilde{f}(\theta) - \frac{2}{n} \epsilon(\theta)^T \tilde{r}(\theta) + \frac{1}{n} \|\epsilon(\theta)\|^2, \quad (42)$$

and hence

$$|f(\theta) - \tilde{f}(\theta)| \leq \frac{2}{n} \|\epsilon(\theta)\| \|\tilde{r}(\theta)\| + \frac{1}{n} \|\epsilon(\theta)\|^2, \quad (43)$$

$$\leq \frac{2}{n} \sqrt{n} \|\epsilon(\theta)\|_\infty \sqrt{n \tilde{f}(\theta)} + \|\epsilon(\theta)\|_\infty^2, \quad (44)$$

and the first part of (40) follows since  $\|\epsilon(\theta)\|_\infty \leq \delta_x$  from (39). The second part of (40) follows from an identical argument but writing  $\tilde{f}(\theta) = \frac{1}{n} \|r(\theta) + \epsilon(\theta)\|^2$ , and the final conclusion follows immediately from the first part of (40).  $\square$

We construct these bounds to rely mostly on  $\tilde{f}(\theta)$ , since this is the value which is observed by the algorithm (rather than the true value  $f(\theta)$ ). From the concavity of  $\sqrt{\cdot}$ , if  $\tilde{f}(\theta)$  is larger, then  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\|$  must be smaller to achieve the same  $\delta_f$ .

Lastly, we note the key reason why we require (32): it guarantees that our estimate  $\tilde{\rho}^k$  of  $\rho^k$  is not too inaccurate.

**Lemma 3** Suppose  $|\tilde{f}(\theta^k) - f(\theta^k)| \leq \delta^k$  and  $|\tilde{f}(\theta^k + s^k) - f(\theta^k + s^k)| \leq \delta_{\pm}^k$ . If (32) holds, then  $|\tilde{\rho}^k - \rho^k| \leq 2\eta'_1$ .

**Proof** Follows immediately from (29) and (28); see [17, Section 10.6.1].  $\square$

### 3.4 Convergence and Worst-Case Complexity

We now prove the global convergence of Algorithm 1 and analyze its worst-case complexity (i.e., the number of iterations required to achieve  $\|\nabla f(\theta^k)\| \leq \epsilon$  for the first time).

**Assumption 3** The computed trust-region step  $s^k$  satisfies

$$m^k(0) - m^k(s^k) \geq \frac{1}{2} \|g^k\| \min\left(\Delta^k, \frac{\|g^k\|}{\|H^k\| + 1}\right), \quad (45)$$

and there exists  $\kappa_H \geq 1$  such that  $\|H^k\| + 1 \leq \kappa_H$  for all  $k$ .

Assumption 3 is standard and the condition (45) easy to achieve in practice [17, Chapter 6.3].

Firstly, we must show that the inner loops for the criticality and accuracy phases terminate. We begin with the criticality phase and then consider the accuracy phase.

**Lemma 4** ([13], Lemma B.1) Suppose Assumption 2 holds and  $\|\nabla f(\theta^k)\| \geq \epsilon > 0$ . Then, the criticality phase terminates in finite time with

$$\min\left(\Delta_{\text{init}}^k, \frac{\gamma_{\text{dec}} \epsilon}{\kappa_{\text{eg}} + 1}\right) \leq \Delta^k \leq \Delta_{\text{init}}^k, \quad (46)$$

where  $\Delta_{\text{init}}^k$  is the value of  $\Delta^k$  before the criticality phase begins.

**Lemma 5** ([13], Lemma 3.7) Suppose Assumption 2 holds. Then, in all iterations we have  $\|g^k\| \geq \min(\epsilon, \Delta^k)$ . Also, if  $\|\nabla f(\theta^k)\| \geq \epsilon > 0$ , then  $\|g^k\| \geq \epsilon / (\kappa_{\text{eg}} + 1) > 0$ .

We note that our presentation of the criticality phase here can be made more general by allowing  $\|g^k\| \geq \epsilon_C \neq \epsilon$  as the entry test, setting  $\Delta^k$  to  $\omega^i \Delta^k$  for some  $\omega \in (0, 1)$  possibly different to  $\gamma_{\text{dec}}$ , and having an exit test  $\Delta^k \leq \mu \|g^k\|$  for some  $\mu > 0$ . All the below results hold under these assumptions, with modifications as per [13].

**Lemma 6** If Assumptions 2 and 3 hold and  $\|\nabla f(\theta^k)\| \geq \epsilon > 0$ , then the accuracy phase terminates in finite time (i.e., line 10 of Algorithm 1 is eventually called)

**Proof** From Lemma 5 we have  $\|g^k\| \geq \epsilon / (\kappa_{\text{eg}} + 1)$ , and the result then follows from [17, Lemma 10.6.1].  $\square$

We now collect some key preliminary results required to establish complexity bounds.

**Lemma 7** Suppose Assumptions 2 and 3 hold,  $m^k$  is fully linear in  $B(\theta^k, \Delta^k)$  and

$$\Delta^k \leq c_0 \|g^k\|,$$

$$\text{where } c_0 := \min\left(\frac{1 - \eta_2 - 2\eta'_1}{4\kappa_{\text{ef}}}, \frac{1}{\kappa_H}\right) > 0, \quad (47)$$

then  $\tilde{\rho}^k \geq \eta_2$ .

**Proof** We compute

$$|\rho^k - 1| = \left| \frac{(f(\theta^k) - f(\theta^k + s^k)) - (m^k(0) - m^k(s^k))}{m^k(0) - m^k(s^k)} \right|, \quad (48)$$

$$\leq \frac{|f(\theta^k + s^k) - m^k(s^k)|}{|m^k(0) - m^k(s^k)|} + \frac{|f(\theta^k) - m^k(0)|}{|m^k(0) - m^k(s^k)|}. \quad (49)$$

Since  $\Delta^k \leq \|g^k\| / \kappa_H$ , from Assumption 3 we have

$$m^k(0) - m^k(s^k) \geq \frac{1}{2} \|g^k\| \Delta^k. \quad (50)$$

From this and full linearity, we get

$$|\rho^k - 1| \leq 2 \left( \frac{2\kappa_{cf}(\Delta^k)^2}{\|g^k\|\Delta^k} \right) \leq 1 - \eta_2 - 2\eta'_1, \tag{51}$$

and so  $\rho^k \geq \eta_2 + 2\eta'_1$ ; hence,  $\tilde{\rho}^k \geq \eta_2$  from Lemma 3.  $\square$

**Lemma 8** *Suppose Assumptions 2 and 3 hold. Suppose  $\|\nabla f(\theta^k)\| \geq \epsilon$  for all  $k = 0, \dots, k_\epsilon$  and some  $\epsilon \in (0, 1)$ . Then, for all  $k \leq k_\epsilon$ ,*

$$\Delta^k \geq \Delta_{\min} := \gamma_{\text{dec}} \min \left( \Delta^0, \frac{c_0\epsilon}{\kappa_{\text{eg}} + 1}, \frac{\gamma_{\text{dec}}\epsilon}{\kappa_{\text{eg}} + 1} \right) > 0. \tag{52}$$

**Proof** As above, we let  $\Delta^k_{\text{init}}$  and  $m^k_{\text{init}}$  denote the values of  $\Delta^k$  and  $m^k$  before the criticality phase (i.e.,  $\Delta^k_{\text{init}} = \Delta^k$  and  $m^k_{\text{init}} = m^k$  if the criticality phase is not called). From Lemma 5, we know  $\|g^k\| \geq \epsilon/(\kappa_{\text{eg}} + 1)$  for all  $k \leq k_\epsilon$ . Suppose by contradiction  $k \leq k_\epsilon$  is the first iteration such that  $\Delta^k < \Delta_{\min}$ . Then from Lemma 4,

$$\frac{\gamma_{\text{dec}}\epsilon}{\kappa_{\text{eg}} + 1} \geq \Delta_{\min} > \Delta^k \geq \min \left( \Delta^k_{\text{init}}, \frac{\gamma_{\text{dec}}\epsilon}{\kappa_{\text{eg}} + 1} \right), \tag{53}$$

and so  $\Delta^k \geq \Delta^k_{\text{init}}$ ; hence,  $\Delta^k = \Delta^k_{\text{init}}$ . That is, either the criticality phase is not called, or terminates with  $i = 0$  (in this case, the model  $m^k$  is formed simply by making  $m^k_{\text{init}}$  fully linear in  $B(\theta^k, \Delta^k) = B(\theta^k, \Delta^k_{\text{init}})$ ).

If the accuracy phase loop occurs, we go back to the criticality phase, which can potentially happen multiple times. However, since the only change is that  $\tilde{r}(\theta^k)$  is evaluated to higher accuracy, incorporating this information into the model  $m^k$  can never destroy full linearity. Hence, after the accuracy phase, by the same reasoning as above, either one iteration of the criticality phase occurs (i.e.,  $m^k$  is made fully linear) or it is not called. If the accuracy phase is called multiple times and the criticality phase occurs multiple times, all times except the first have no effect (since the accuracy phase can never destroy full linearity). Thus,  $\Delta^k$  is unchanged by the accuracy phase.

Since  $\Delta_{\min} < \Delta^0_{\text{init}}$ , we have  $k \geq 1$ . As  $k$  is the first iteration such that  $\Delta^k < \Delta_{\min}$  and  $\Delta^k = \Delta^k_{\text{init}}$ , we must have  $\Delta^k_{\text{init}} = \gamma_{\text{dec}}\Delta^{k-1}$  (as this is the only other way  $\Delta^k$  can be reduced). Therefore,  $\Delta^{k-1} = \Delta^k/\gamma_{\text{dec}} < \Delta_{\min}/\gamma_{\text{dec}}$ , and so

$$\Delta^{k-1} \leq \min \left( \frac{c_0\epsilon}{\kappa_{\text{eg}} + 1}, \frac{\gamma_{\text{dec}}\epsilon}{\kappa_{\text{eg}} + 1} \right). \tag{54}$$

We then have  $\Delta^{k-1} \leq c_0\epsilon/(\kappa_{\text{eg}} + 1) \leq c_0\|g^{k-1}\|$ , and so by (7) either  $\tilde{\rho}^k \geq \eta_2$  or  $m^{k-1}$  is not fully linear. Either way, we set  $\Delta^k_{\text{init}} \geq \Delta^{k-1}$  in (34). This contradicts  $\Delta^k_{\text{init}} = \gamma_{\text{dec}}\Delta^{k-1}$  above, and we are done.  $\square$

We now bound the number of iterations of each type. Specifically, we suppose that  $k_\epsilon + 1$  is the first  $k$  such that  $\|\nabla f(\theta^k)\| \geq \epsilon$ . Then, we define the sets of iterations:

- $\mathcal{S}_\epsilon$  is the set of iterations  $k \in \{0, \dots, k_\epsilon\}$  which are ‘successful,’ i.e.,  $\tilde{\rho}^k \geq \eta_2$ , or  $\tilde{\rho}^k \geq \eta_1$  and  $m^k$  is fully linear in  $B(\theta^k, \Delta^k)$ .
- $\mathcal{M}_\epsilon$  is the set of iterations  $k \in \{0, \dots, k_\epsilon\}$  which are ‘model-improving,’ i.e.,  $\tilde{\rho}^k < \eta_2$  and  $m^k$  is not fully linear in  $B(\theta^k, \Delta^k)$ .
- $\mathcal{U}_\epsilon$  is the set of iterations  $k \in \{0, \dots, k_\epsilon\}$  which are ‘unsuccessful,’ i.e.,  $\tilde{\rho}^k < \eta_1$  and  $m^k$  is fully linear in  $B(\theta^k, \Delta^k)$ .

These three sets form a partition of  $\{0, \dots, k_\epsilon\}$ .

**Proposition 1** *Suppose Assumptions 2 and 3 hold. Then,*

$$|\mathcal{S}_\epsilon| \leq \frac{2(\kappa_{\text{eg}} + 1)f(\theta^0)}{(\eta_1 - 2\eta'_1)\epsilon\Delta_{\min}}. \tag{55}$$

**Proof** By definition of  $k_\epsilon$ ,  $\|\nabla f(\theta^k)\| \geq \epsilon$  for all  $k \leq k_\epsilon$  and so Lemma 5 and Lemma 8 give  $\|g^k\| \geq \epsilon/(\kappa_{\text{eg}} + 1)$  and  $\Delta^k \geq \Delta_{\min}$  for all  $k \leq k_\epsilon$ , respectively. For any  $k \leq k_\epsilon$  we have

$$f(\theta^k) - f(\theta^{k+1}) = \rho^k [m^k(0) - m^k(s^k)], \tag{56}$$

$$\geq \frac{1}{2}\rho^k \|g^k\| \min \left( \Delta^k, \frac{\|g^k\|}{\|H^k\| + 1} \right), \tag{57}$$

$$\geq \frac{1}{2}\rho^k \frac{\epsilon}{\kappa_{\text{eg}} + 1} \min \left( \Delta_{\min}, \frac{\epsilon}{\kappa_H(\kappa_{\text{eg}} + 1)} \right), \tag{58}$$

by definition of  $\rho^k$  and Assumption 3. If  $k \in \mathcal{S}_\epsilon$ , we know  $\tilde{\rho}^k \geq \eta_1$ , which implies  $\rho^k \geq \eta_1 - 2\eta'_1 > 0$  from Lemma 3. Therefore,

$$f(\theta^k) - f(\theta^{k+1}) \geq \frac{1}{2}(\eta_1 - 2\eta'_1) \frac{\epsilon}{\kappa_{\text{eg}} + 1} \min \left( \Delta_{\min}, \frac{\epsilon}{\kappa_H(\kappa_{\text{eg}} + 1)} \right), \tag{59}$$

$$= \frac{1}{2}(\eta_1 - 2\eta'_1) \frac{\epsilon}{\kappa_{\text{eg}} + 1} \Delta_{\min}, \tag{60}$$

for all  $k \in \mathcal{S}_\epsilon$ , where the last line follows since  $\Delta_{\min} < c_0\epsilon/(\kappa_{\text{eg}} + 1) \leq \epsilon/[\kappa_H(\kappa_{\text{eg}} + 1)]$  by definition of  $\Delta_{\min}$  (52) and  $c_0$  (47).

The iterate  $\theta^k$  is only changed on successful iterations (i.e.,  $\theta^{k+1} = \theta^k$  for all  $k \notin \mathcal{S}_\epsilon$ ). Thus, as  $f(\theta) \geq 0$  from the least-squares structure (22), we get

$$f(\theta^0) \geq f(\theta^0) - f(\theta^{k_\epsilon+1}), \tag{61}$$

$$= \sum_{k \in \mathcal{S}_\epsilon} f(\theta^k) - f(\theta^{k+1}), \tag{62}$$

$$\geq |\mathcal{S}_\epsilon| \left[ \frac{1}{2}(\eta_1 - 2\eta'_1) \frac{\epsilon}{\kappa_{eg} + 1} \Delta_{\min} \right], \tag{63}$$

and the result follows.  $\square$

We are now in a position to prove our main results.

**Proof of Theorem 2** To derive a contradiction, suppose that  $\|\nabla f(\theta^k)\| \geq \epsilon$  for all  $k \in \{0, \dots, K\}$ , and so  $\|g^k\| \geq \epsilon/(\kappa_{eg} + 1)$  and  $\Delta^k \geq \Delta_{\min}$  by Lemma 5 and Lemma 8, respectively. Since  $K \leq k_\epsilon$  by definition of  $k_\epsilon$ , we will try to construct an upper bound on  $k_\epsilon$ . We already have an upper bound on  $|\mathcal{S}_\epsilon|$  from Proposition 1.

If  $k \in \mathcal{S}_\epsilon$ , we set  $\Delta^{k+1} \leq \gamma_{\text{inc}} \Delta^k$ . Similarly, if  $k \in \mathcal{U}_\epsilon$  we set  $\Delta^{k+1} = \gamma_{\text{dec}} \Delta^k$ . Thus,

$$\Delta_{\min} \leq \Delta^{k_\epsilon} \leq \Delta^0 \gamma_{\text{inc}}^{|\mathcal{S}_\epsilon|} \gamma_{\text{dec}}^{|\mathcal{U}_\epsilon|}. \tag{64}$$

That is,  $\Delta_{\min}/\Delta^0 \leq \gamma_{\text{inc}}^{|\mathcal{S}_\epsilon|} \gamma_{\text{dec}}^{|\mathcal{U}_\epsilon|}$ , and so

$$|\mathcal{U}_\epsilon| \leq \frac{\log \gamma_{\text{inc}}}{|\log \gamma_{\text{dec}}|} |\mathcal{S}_\epsilon| + \frac{\log(\Delta^0/\Delta_{\min})}{|\log \gamma_{\text{dec}}|}, \tag{65}$$

noting we have changed  $\Delta_{\min}/\Delta^0 < 1$  to  $\Delta^0/\Delta_{\min} > 1$  and used  $\log \gamma_{\text{dec}} < 0$ , so all terms in (65) are positive. Now, the next iteration after a model-improving iteration cannot be model-improving (as the resulting model is fully linear), giving

$$|\mathcal{M}_\epsilon| \leq |\mathcal{S}_\epsilon| + |\mathcal{U}_\epsilon|. \tag{66}$$

If we combine (65) and (66) with  $k_\epsilon \leq |\mathcal{S}_\epsilon| + |\mathcal{M}_\epsilon| + |\mathcal{U}_\epsilon|$ , we get

$$k_\epsilon \leq 2(|\mathcal{S}_\epsilon| + |\mathcal{U}_\epsilon|), \tag{67}$$

$$\leq \left( 2 + 2 \frac{\log \gamma_{\text{inc}}}{|\log \gamma_{\text{dec}}|} \right) |\mathcal{S}_\epsilon| + 2 \frac{\log(\Delta^0/\Delta_{\min})}{|\log \gamma_{\text{dec}}|}, \tag{68}$$

which, given the bound on  $|\mathcal{S}_\epsilon|$  (55) means  $K \leq k_\epsilon$  is bounded above by the right-hand side of (35), a contradiction.  $\square$

**Proof of Corollary 1** The iteration bound follows directly from Theorem 2, noting that  $\Delta_{\min} = \mathcal{O}(\kappa^{-2}\epsilon)$ . This also implies that  $\liminf_{k \rightarrow \infty} \|\nabla f(\theta^k)\| = 0$  and so (36) holds from the same argument as in [18, Theorem 10.13] without modification.

For the evaluation bound, we also need to count the number of inner iterations of the criticality phase. Suppose  $\|\nabla f(\theta^k)\| < \epsilon$  for  $k = 0, \dots, k_\epsilon$ . Similar to the above, we define: (a)  $\mathcal{C}_\epsilon^U$  to be the number of criticality phase iterations corresponding to the first iteration of  $i = 0$  where  $m^k$  was

not already fully linear, in iterations  $0, \dots, k_\epsilon$  and (b)  $\mathcal{C}_\epsilon^U$  to be the number of criticality phase iterations corresponding to all other iterations  $i > 0$  (where  $\Delta^k$  is reduced and  $m^k$  is made fully linear) in iterations  $0, \dots, k_\epsilon$ .

From Lemma 8 we have  $\Delta^k \geq \Delta_{\min}$  for all  $k \leq k_\epsilon$ . We note that  $\Delta^k$  is reduced by a factor  $\gamma_{\text{dec}}$  for every iteration of the criticality phase in  $\mathcal{C}_\epsilon^U$ . Thus by a more careful reasoning as we used to reach (65), we conclude

$$\Delta_{\min} \leq \Delta^0 \gamma_{\text{inc}}^{|\mathcal{S}_\epsilon|} \gamma_{\text{dec}}^{|\mathcal{U}_\epsilon| + |\mathcal{C}_\epsilon^U|}, \tag{69}$$

$$|\mathcal{C}_\epsilon^U| \leq \frac{\log \gamma_{\text{inc}}}{|\log \gamma_{\text{dec}}|} |\mathcal{S}_\epsilon| + \frac{\log(\Delta^0/\Delta_{\min})}{|\log \gamma_{\text{dec}}|} - |\mathcal{U}_\epsilon|. \tag{70}$$

Also, after every iteration  $k$  in which the first iteration of criticality phase makes  $m^k$  fully linear, we have either a (very) successful or unsuccessful step, not a model-improving step. From the same reasoning as in Lemma 8, the accuracy phase can only cause at most one more step criticality phase in which  $m^k$  is made fully linear, regardless of how many times it is called.<sup>3</sup> Thus,

$$|\mathcal{C}_\epsilon^M| \leq 2(|\mathcal{S}_\epsilon| + |\mathcal{U}_\epsilon|). \tag{71}$$

Combining (70) and (71) with (65) and (66), we conclude that the number of times we make  $m^k$  fully linear is

$$|\mathcal{M}_\epsilon| + |\mathcal{C}_\epsilon^U| + |\mathcal{C}_\epsilon^M| \leq \left( 3 + 3 \frac{\log \gamma_{\text{inc}}}{|\log \gamma_{\text{dec}}|} \right) |\mathcal{S}_\epsilon| + 3 \frac{\log(\Delta^0/\Delta_{\min})}{|\log \gamma_{\text{dec}}|}, \tag{72}$$

$$\leq \left( 3 + 3 \frac{\log \gamma_{\text{inc}}}{|\log \gamma_{\text{dec}}|} \right) \frac{2(\kappa_{eg} + 1)f(\theta^0)}{(\eta_1 - 2\eta'_1)\epsilon \Delta_{\min}} + 3 \frac{\log(\Delta^0/\Delta_{\min})}{|\log \gamma_{\text{dec}}|}, \tag{73}$$

where the second inequality follows from Proposition 1.

If  $\epsilon < 1$ , we conclude that the number of times we make  $m^k$  fully linear before  $\|\nabla f(\theta^k)\| < \epsilon$  for the first time is the same as the number of iterations,  $\mathcal{O}(\kappa^3 \epsilon^{-2})$ . Since each iteration requires one new objective evaluation (at  $\theta^k + s^k$ ) and each time we make  $m^k$  fully linear requires at most  $\mathcal{O}(d)$  objective evaluations (corresponding to replacing the entire interpolation set), we get the stated evaluation complexity bound.  $\square$

### 3.5 Estimating the Lower-Level Work

We have from Corollary 1 that we can achieve  $\|\nabla f(\theta^k)\| < \epsilon$  in  $\mathcal{O}(\epsilon^{-2})$  evaluations of  $\tilde{r}(\theta)$ . In this section, we use the

<sup>3</sup> Of course, there may be many more initial steps of the criticality phase in which  $m^k$  is already fully linear, but no work is required in this case.

fact that evaluations of  $\tilde{r}(\theta)$  come from finitely terminating a linearly convergent procedure (i.e., strongly convex optimization) to estimate the total work required in the lower-level problem. This is particularly relevant in an imaging context, where the lower-level problem can be large scale and poorly conditioned; this can be the dominant cost of Algorithm 1.

**Proposition 2** *Suppose Assumptions 2 and 3 hold and  $\|\nabla f(\theta^k)\| \geq \epsilon$  for all  $k = 0, \dots, k_\epsilon$  and some  $\epsilon \in (0, 1]$ . Then, for every objective evaluation in iterations  $k \leq k_\epsilon$  it suffices to guarantee that  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| = \mathcal{O}(\epsilon^2)$  for all  $i = 1, \dots, n$ .*

**Proof** For all  $k \leq k_\epsilon$  we have  $\|g^k\| \geq \epsilon/(\kappa_{eg} + 1)$  and  $\Delta^k \geq \Delta_{\min}$  by Lemmas 5 and 8, respectively. There are two places where we require upper bounds on  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\|$  in our objective evaluations: ensuring  $\tilde{f}(\theta^k)$  and  $\tilde{f}(\theta^k + s^k)$  satisfy (32) and ensuring our model is fully linear using Lemma 1.

In the first case, we note that

$$m^k(0) - m^k(s^k) \geq \frac{1}{2} \frac{\epsilon}{\kappa_{eg} + 1} \min \left( \Delta_{\min}, \frac{\epsilon}{\kappa_H(\kappa_{eg} + 1)} \right), \tag{74}$$

$$= \frac{1}{2} \frac{\epsilon}{\kappa_{eg} + 1} \Delta_{\min}, \tag{75}$$

by Assumption 3 and using  $\Delta_{\min} < c_0\epsilon/(\kappa_{eg} + 1) \leq \epsilon/[\kappa_H(\kappa_{eg} + 1)]$  by definition of  $\Delta_{\min}$  (52) and  $c_0$  (47). Therefore, to ensure (32) it suffices to guarantee

$$\begin{aligned} & \max \left( |\tilde{f}(\theta^k) - f(\theta^k)|, |\tilde{f}(\theta^k + s^k) - f(\theta^k + s^k)| \right) \\ & \leq \delta_f^{\min} := \frac{1}{2} \eta'_1 \frac{\epsilon}{\kappa_{eg} + 1} \Delta_{\min}. \end{aligned} \tag{76}$$

From Lemma 2, specifically the second part of (40), this means to achieve (32) it suffices to guarantee

$$\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| \leq \sqrt{f(\theta) + \delta_f^{\min}} - \sqrt{f(\theta)}, \tag{77}$$

for all  $i = 1, \dots, n$ , where  $\theta \in \cup_{k \leq k_\epsilon} \{\theta^k, \theta^k + s^k\}$ . From Assumption 2 we have  $f(\theta) \leq f_{\max} := r_{\max}^2/n$ , and so from the fundamental theorem of calculus we have

$$\sqrt{f(\theta) + \delta_f^{\min}} - \sqrt{f(\theta)} = \int_{f(\theta)}^{f(\theta) + \delta_f^{\min}} \frac{1}{2\sqrt{t}} dt, \tag{78}$$

$$\geq \frac{\delta_f^{\min}}{2\sqrt{f(\theta) + \delta_f^{\min}}} \geq \frac{\delta_f^{\min}}{2\sqrt{f_{\max} + \delta_f^{\min}}}. \tag{79}$$

Since  $\epsilon < 1$ ,  $\delta_f^{\min}$  is bounded above by a constant and so  $\sqrt{f_{\max} + \delta_f^{\min}}$  is bounded above. Thus, (32) is achieved provided  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| = \mathcal{O}(\delta_f^{\min})$  for all  $i = 1, \dots, n$ .

For the second case (ensuring full linearity), we need to guarantee (38) holds. This is achieved provided  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| = \mathcal{O}(\Delta_{\min}^2)$  for all  $i = 1, \dots, n$ . The result then follows by noting  $\delta_f^{\min} = \mathcal{O}(\epsilon \Delta_{\min})$  and  $\Delta_{\min} = \mathcal{O}(\epsilon)$ .  $\square$

Corollary 1 and Proposition 2 say that to ensure  $\|\nabla f(\theta^k)\| < \epsilon$  for some  $k$ , we have to perform  $\mathcal{O}(d\kappa^3\epsilon^{-2})$  upper-level objective evaluations, each requiring accuracy at most  $\|\tilde{x}_i(\theta) - \hat{x}_i(\theta)\| = \mathcal{O}(\epsilon^2)$  for all  $i$ . Since our lower-level evaluations correspond to using GD/FISTA to solve a strongly convex problem, the computational cost of each upper-level evaluation is  $\mathcal{O}(n \log(\epsilon^{-2}))$  provided we have reasonable initial iterates. From this, we conclude that the total computational cost before achieving  $\|\nabla f(\theta^k)\| < \epsilon$  is at most  $\mathcal{O}(\epsilon^{-2} \log(\epsilon^{-1}))$  iterations of the lower-level algorithm. However, this is a conservative approach to estimating the cost: many of the iterations correspond to  $\|\nabla f(\theta^k)\| \gg \epsilon$ , and so the work required for these is less. This suggests the question: *can we more carefully estimate the work required at different accuracy levels to prove a lower  $\epsilon$ -dependence on the total work?* We now argue that this is not possible without further information about asymptotic convergence rates (e.g., local convergence theory). For simplicity we drop all constants and  $\mathcal{O}(\cdot)$  notation in the below.

Suppose we count the work required to achieve progressively higher accuracy levels  $1 \geq \epsilon_0 > \epsilon_1 > \dots > \epsilon_N := \epsilon$  for some desired accuracy  $\epsilon \ll 1$ . Since each  $\epsilon_i < 1$ , we assume that we require  $\epsilon_i^{-2}$  evaluations to achieve accuracy  $\epsilon_i$ , where each evaluation requires  $\log(\epsilon_i^{-1})$  computational work. We may choose  $\epsilon_0 < 1$ , since the cost to achieve accuracy  $\epsilon_0$  is fixed (i.e., independent of our desired accuracy  $\epsilon$ ), so does not affect our asymptotic bounds. Counting the total lower-level problem work—which we denote  $W(\epsilon)$ —in this way, we get

$$W(\epsilon) = W(\epsilon_0) + \sum_{i=1}^N \left( \epsilon_i^{-2} - \epsilon_{i-1}^{-2} \right) \log(\epsilon_i^{-1}). \tag{80}$$

The second term of (80) corresponds to a right Riemann sum approximating  $\int_{\epsilon_0^{-2}}^{\epsilon^{-2}} \log(\sqrt{x}) dx$ . Since  $x \rightarrow \log(\sqrt{x}) = \log(x)/2$  is strictly increasing, the right Riemann sum overestimates the integral; hence,

$$W(\epsilon) \geq W(\epsilon_0) + \frac{1}{2} \int_{\epsilon_0^{-2}}^{\epsilon^{-2}} \log(x) dx, \tag{81}$$

$$= W(\epsilon_0) + \frac{1}{2} \left[ \epsilon^{-2} (\log(\epsilon^{-2}) - 1) - \epsilon_0^{-2} (\log(\epsilon_0^{-2}) - 1) \right], \tag{82}$$

independent of our choices of  $\epsilon_1, \dots, \epsilon_{N-1}$ . That is, as  $\epsilon \rightarrow 0$ , we have  $W(\epsilon) \sim \epsilon^{-2} \log(\epsilon^{-1})$ , so our naïve estimate is tight.

We further note that this naïve bound applies more generally. Suppose the work required for a single evaluation of the lower-level objective to accuracy  $\epsilon$  is  $w(\epsilon^{-2}) \geq 0$  (e.g.,  $w(x) = \log(x)/2$  above). Assuming  $w$  is increasing (i.e., higher accuracy evaluations require more work), we get, similar to the above,

$$W(\epsilon) \geq W(\epsilon_0) + \int_{\epsilon_0^{-2}}^{\epsilon^{-2}} w(x)dx. \tag{83}$$

Since  $w$  is increasing and nonnegative, by

$$\int_{\epsilon_0^{-2}}^{\epsilon^{-2}} w(x)dx \geq \int_{(\epsilon_0^{-2} + \epsilon^{-2})/2}^{\epsilon^{-2}} w(x)dx, \tag{84}$$

$$\geq \frac{\epsilon_0^{-2} + \epsilon^{-2}}{2} w\left(\frac{\epsilon_0^{-2} + \epsilon^{-2}}{2}\right), \tag{85}$$

the naïve work bound  $W(\epsilon) \sim \epsilon^{-2}w(\epsilon^{-2})$  holds provided  $w(x) = \mathcal{O}(w(x/2))$  as  $x \rightarrow \infty$ ; that is,  $w(x)$  does not increase too quickly. This holds in a variety of cases, such as  $w(x)$  bounded, concave or polynomial (but not if  $w(x)$  grows exponentially). In particular, this holds for  $w(x) \sim \log(x)/2$  as above, and  $w(x) \sim x^{1/2}$  and  $w(x) \sim x$ , which correspond to the work required (via standard sublinear complexity bounds) if the lower-level problem is a strongly convex, convex or nonconvex optimization problem, respectively.

## 4 Numerical Results

### 4.1 Upper-Level Solver (DFO-LS)

We implement the dynamic accuracy algorithm (Algorithm 1) in DFO-LS [12], an open-source Python package which solves nonlinear least-squares problems subject to bound constraints using model-based DFO.<sup>4</sup> As described in [12], DFO-LS has a number of modifications compared to the theoretical algorithm Algorithm 1. The most notable modifications here are that DFO-LS:

- Allows for bound constraints (and internally scales variables so that the feasible region is  $[0, 1]$  for all variables);
- Does not implement a criticality phase;
- Uses a simplified model-improving step;
- Maintains two trust-region radii to avoid decreasing  $\Delta^k$  too quickly;
- Implements a ‘safety phase,’ which treats iterations with short steps  $\|s^k\| \ll \Delta^k$  similarly to unsuccessful iterations.

<sup>4</sup> Available at <https://github.com/numericalalgorithmsgroup/dfols>.

More discussion on DFO-LS can be found in [12,13].

Here, we use DFO-LS v1.1.1, modified for the dynamic accuracy framework as described above. When determining the accuracy level for a given evaluation, we require accuracy level  $\delta_x = 10(\Delta^k)^2$  for all evaluations (c.f. Lemma 1), and also (32) when checking objective decrease (29).

### 4.2 Application: 1D Image Denoising

In this section, we consider the application of DFO-LS to the problem of learning the regularization and smoothing parameters for the image denoising model (8) as described in Section 2.1.1. We use training data constructed using the method described in Sect. 2.2 with  $N = 256$  and  $\sigma = 0.1$ .

*1-parameter case* The simplest example we consider is the 1-parameter case, where we only wish to learn  $\alpha$  in (8). We fix  $\nu = \xi = 10^{-3}$  and use a training set of  $n = 10$  randomly generated images. We choose  $\alpha = 10^\theta$ , optimize over  $\theta$  within bounds  $\theta \in [-7, 7]$  with starting value  $\theta^0 = 0$ . We do not regularize this problem, i.e.,  $\mathcal{J} = 0$ .

*3-parameter case* We also consider the more complex problem of learning three parameters for the denoising problem (namely  $\alpha, \nu$  and  $\xi$ ). We choose to penalize a large condition number of the lower-level problem, thus promoting efficient solution of the lower-level problem after training. To be precise we choose

$$\mathcal{J}(\alpha, \nu, \xi) = \left(\frac{L(\alpha, \nu, \xi)}{\mu(\alpha, \nu, \xi)}\right)^2 \tag{86}$$

where  $L$  and  $\mu$  are the smoothness and strong convexity constants given in Sect. 2.1.1.

The problem is solved using the parameterization  $\alpha = 10^{\theta_1}, \nu = 10^{\theta_2}$  and  $\xi = 10^{\theta_3}$ . Here, we use a training set of  $n = 20$  randomly generated images, and optimize over  $\theta \in [-7, 7] \times [-7, 0]^2$ . Our default starting value is  $\theta^0 = (0, -1, -1)$  and our default choice of upper-level regularization parameter is  $\beta = 10^{-6}$ .

*Solver settings* We run DFO-LS with a budget of 20 and 100 evaluations of the upper-level objective  $f$  for the 1- and 3-parameter cases, respectively, and with  $\rho_{\text{end}} = 10^{-6}$  in both cases. We compare the dynamic accuracy variant of DFO-LS (given by Algorithm 1) against two variants of DFO-LS (as originally implemented in [12]):

1. Low-accuracy evaluations: each value  $\hat{x}_i$  received by DFO-LS is inaccurately estimated via a fixed number of iterations of GD/FISTA; we use 1,000 iterations of GD and 200 iterations of FISTA.
2. High-accuracy evaluations: each value  $z_i$  received by DFO-LS is estimated using 10,000 iterations of GD or 2,000 iterations of FISTA.

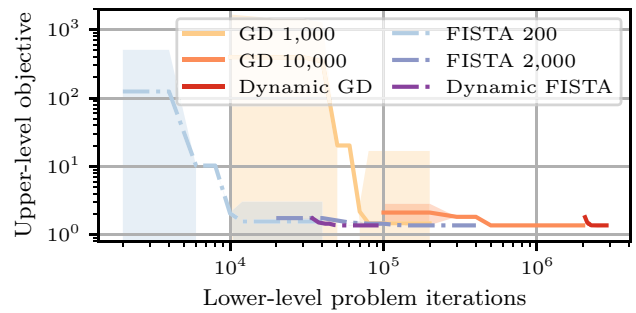
We estimate  $\delta_f$  in the plots below by taking  $\delta_r$  to be the maximum estimate of  $\|\hat{x}_i(\theta) - x_i\|$  for each  $i = 1, \dots, n$ . When running the lower-level solvers, our starting point is the final reconstruction from the previous upper-level evaluation, which we hope is a good estimate of the solution.

*1-parameter denoising results* In Fig. 4 we compare the six algorithm variants (low, high and dynamic accuracy versions of both GD and FISTA) on the 1-parameter denoising problem. Firstly in Figs. 4a, b, we show the best upper-level objective value observed against ‘computational cost,’ measured as the total GD/FISTA iterations performed (over all upper-level evaluations). For each variant, we plot the value  $\tilde{f}(\theta)$  and the uncertainty range  $\tilde{f}(\theta) \pm \delta_f$  associated with that evaluation. In Fig. 4c we show the best  $\alpha_\theta$  found against the same measure of computational cost.

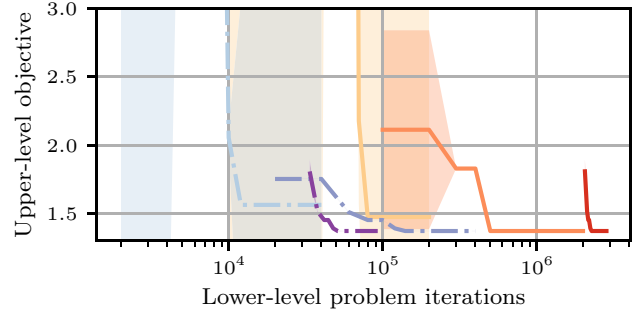
We see that both low-accuracy variants do not converge to the optimal  $\theta$ . Both high-accuracy variants converge to the same objective value and  $\theta$ , but take much more computational effort to do this. Indeed, we did not know a priori how many GD/FISTA iterations would be required to achieve convergence. By contrast, both dynamic accuracy variants find the optimal  $\theta$  without any tuning.

Moreover, dynamic accuracy FISTA converges faster than high-accuracy FISTA, but the reverse is true for GD. In Fig. 4d we show the cumulative number of GD/FISTA iterations performed after each evaluation of the upper-level objective. We see that the reason for dynamic accuracy GD converging slower than high-accuracy GD is that the initial upper-level evaluations require many GD iterations; the same behavior is seen in dynamic accuracy FISTA, but to a lesser degree. This behavior is entirely determined by our (arbitrary) choices of  $\theta^0$  and  $\Delta^0$ . We also note that the number of GD/FISTA iterations required by the dynamic accuracy variants after the initial phase is much lower than both the fixed accuracy variants. The difference between the GD and FISTA behavior in Fig. 4d is based on how the initial dynamic accuracy requirements compare to the chosen number of high-accuracy iterations (10,000 GD or 2,000 FISTA). Finally, in Fig. 5 we show the reconstructions achieved using the  $\alpha_\theta$  found by dynamic accuracy FISTA. All reconstructions are close to the ground truth, with a small loss of contrast.

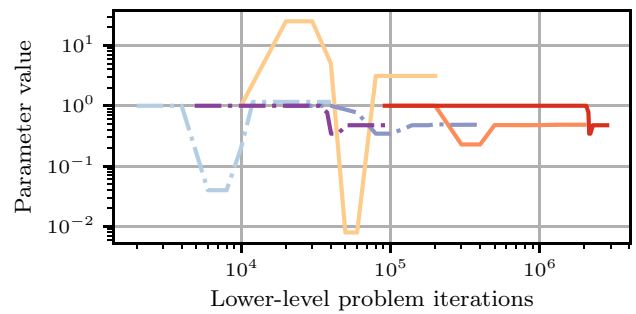
To further understand the impact of the initial evaluations and the robustness of our framework, in Fig. 4 we run the same problem with different choices  $\theta^0 \in \{-2, -1, 1\}$  (where  $\theta^0 = 0$  before). In Fig. 6 we show best  $\alpha_\theta$  found for a given computational effort for these choices. When  $\theta^0 > 0$ , the lower-level problem starts more ill-conditioned, and so the first upper-level evaluations for the dynamic accuracy variants require more GD/FISTA iterations. However, when  $\theta^0 < 0$ , we initially have a well-conditioned lower-level problem, and so the dynamic accuracy variants require many



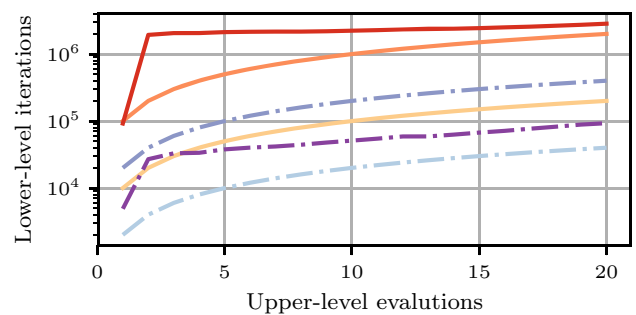
(a) Objective value  $f(\theta)$



(b) Objective value  $f(\theta)$ , zoomed in



(c) Parameter value  $\alpha_\theta$

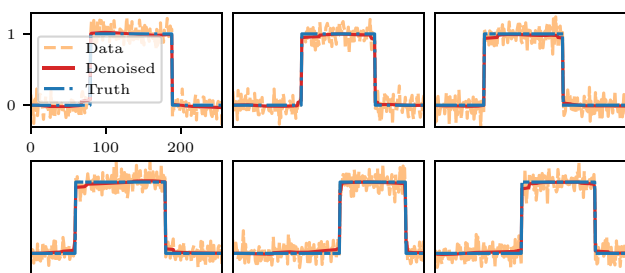


(d) Cumulative GD/FISTA iterations per upper-level evaluation

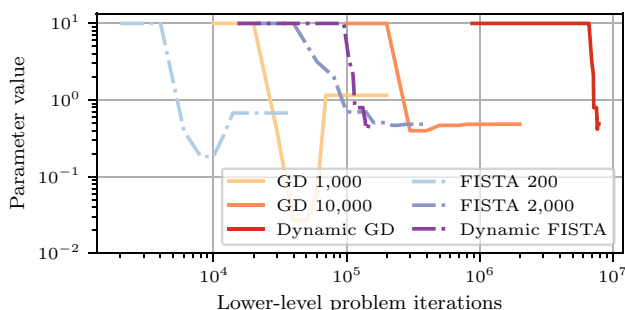
Fig. 4 Results for the 1-parameter denoising problem

fewer GD/FISTA iterations initially, and they converge at the same or a faster rate than the high-accuracy variants.

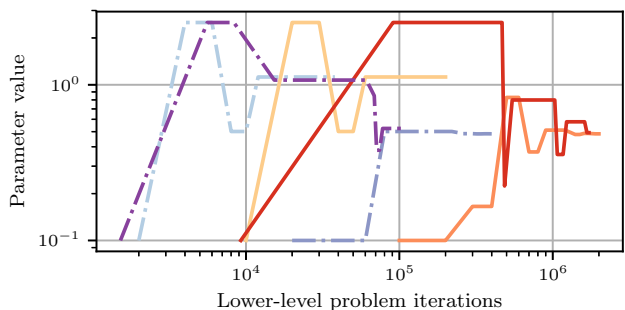
These results also demonstrate that the dynamic accuracy variants give a final regularization parameter which is robust



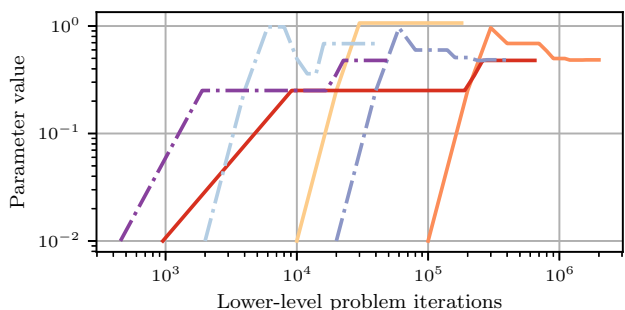
**Fig. 5** 1-parameter final reconstructions (dynamic accuracy FISTA but all except low-accuracy GD look basically the same). Reconstructions are calculated by using the final  $\theta$  returned from the given DFO-LS variant and solving (6) with 1,000 iterations of FISTA



**(a)** Start  $\theta^0 = 1$



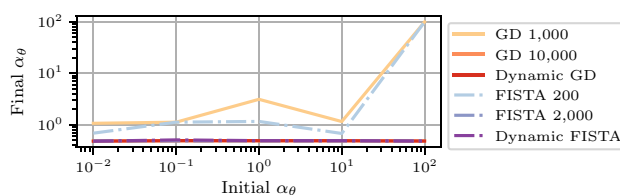
**(b)** Start  $\theta^0 = -1$



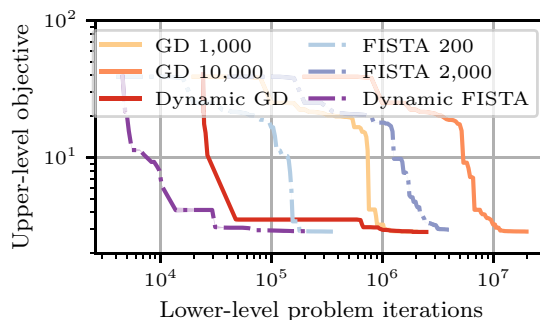
**(c)** Start  $\theta^0 = -2$

**Fig. 6** 1-parameter results: optimal  $\alpha_\theta$  found when using different initial values  $\theta^0$  (compare Fig. 4c)

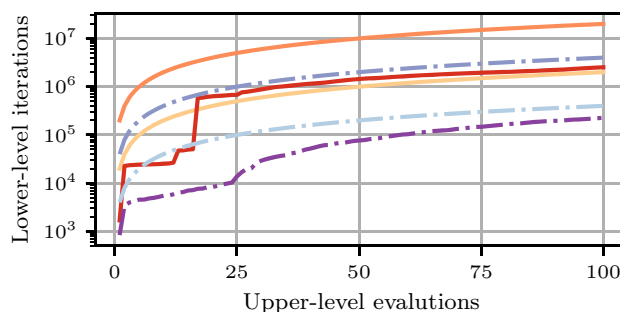
to the choice of  $\theta^0$ . In Fig. 7 we plot the final learned  $\alpha_\theta$  value compared to the initial choice of  $\alpha_\theta$  for all variants. The low-



**Fig. 7** 1-parameter results: compare optimal  $\alpha_\theta$  values found for different choices of starting points



**(a)** Objective value  $f(\theta)$



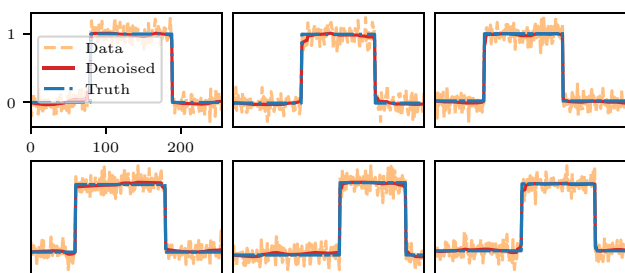
**(b)** Cumulative GD/FISTA iterations per upper-level evaluation

**Fig. 8** Results for the 3-parameter 1D denoising problem

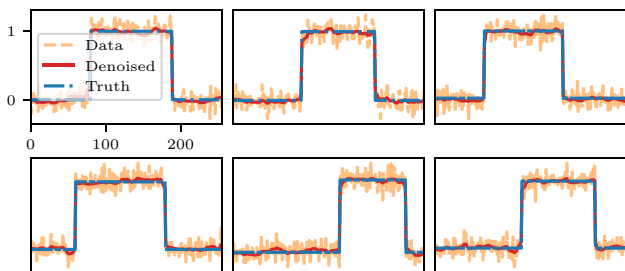
accuracy variants do not reach a consistent minimizer for different starting values, but the dynamic and high-accuracy variants both reach the same minimizer for all starting points. Thus although our upper-level problem is nonconvex, we see that our dynamic accuracy approach can produce solutions which are robust to the choice of starting point.

*3-parameter denoising results* Next, we consider the 3-parameter ( $\alpha_\theta$ ,  $\nu_\theta$  and  $\xi_\theta$ ) denoising problem.

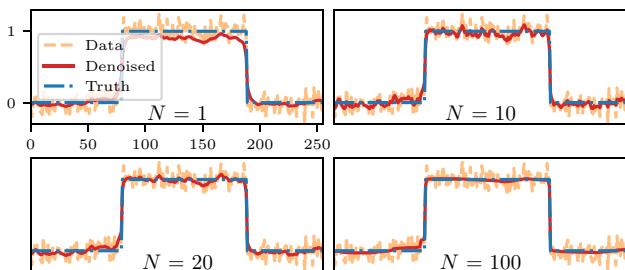
As shown in Fig. 8, both dynamic accuracy variants (GD and FISTA) achieve the best objective value at least one order of magnitude faster than the corresponding low- and high-accuracy variants. We note that (for instance) 200 FISTA iterations was insufficient to achieve convergence in the 1-parameter case, but converges here. By contrast, aside from the substantial speedup in the 3-parameter case, our approach converges in both cases without needing to select the computational effort in advance.



**Fig. 9** Example 3-parameter final reconstructions (dynamic accuracy FISTA but all other variants are similar). Reconstructions use the final  $\theta$  returned by DFO-LS and solving (6) with 1,000 FISTA iterations



**Fig. 10** Example 3-parameter final reconstructions for dynamic accuracy FISTA with  $\beta = 10^{-4}$ . Compare with reconstructions with  $\beta = 10^{-6}$  shown in Fig. 9



**Fig. 11** 3-parameter results (dynamic accuracy FISTA): reconstructions of first training image using best parameters  $\theta$  after  $N$  evaluations of upper-level objective (reconstruction based on 1,000 FISTA iterations)

The final reconstructions achieved by the optimal parameters for dynamic accuracy FISTA are shown in Fig. 9. We note that all variants produced very similar reconstructions (since they converged to similar parameter values), and that all training images are recovered with high accuracy.

Next, we consider the effect of the upper-level regularization parameter  $\beta$ . If the smaller  $\beta$  value of  $10^{-8}$  is chosen, all variants converge to slightly smaller values of  $\nu_\theta$  and  $\xi_\theta$  as the original  $\beta = 10^{-6}$ , but produce reconstructions of a similar quality. However, increasing the value of  $\beta$  yields parameters which give noticeably worse reconstructions. The reconstructions for  $\beta = 10^{-4}$  are shown in Fig. 10.

We conclude by demonstrating in Fig. 11 that, aside from reducing our upper-level objective, the parameters found by DFO-LS do in fact progressively improve the quality of the

reconstructions. The figure shows the reconstructions of one training image achieved by the best parameters found (by the dynamic accuracy FISTA variant) after a given number of upper-level objective evaluations. We see a clear improvement in the quality of the reconstruction as the upper-level optimization progresses.

### 4.3 Application: 2D Denoising

Next, we demonstrate the performance of dynamic accuracy DFO-LS on the same 3-parameter denoising problem from Sect. 4.2, but applied to 2D images. Our training data are the 25 images from the Kodak dataset.<sup>5</sup> We select the central  $256 \times 256$ -pixel region of each image, convert to monochrome and add Gaussian noise  $N(0, \sigma^2)$  with  $\sigma = 0.1$  to each pixel independently. We run DFO-LS for 200 upper-level evaluations with  $\rho_{\text{end}} = 10^{-6}$ . Unlike Sect. 4.2, we find that there is no need to regularize the upper-level problem with the condition number of the lower-level problem (i.e.,  $\mathcal{J}(\theta) = 0$  for these results).

The resulting objective decrease, final parameter values and cumulative lower-level iterations are shown in Fig. 12. All variants achieve the same (upper-level) objective value and parameter  $\alpha_\theta$ , but the dynamic accuracy variants achieve this with substantially fewer GD/FISTA iterations compared to the low- and high-accuracy variants. Interestingly, despite all variants achieving the same upper-level objective value, they do not reach a consistent choice for  $\nu_\theta$  and  $\xi_\theta$ .

In Fig. 13 we show the reconstructions achieved by the dynamic accuracy FISTA variant for three of the training images. We see high-quality reconstructions in each case, where the piecewise-constant reconstructions favored by TV regularization are evident.

Lastly, we study the impact of changing the noise level  $\sigma$  on the calibrated total variational regularization parameter  $\alpha_\theta$ . We run DFO-LS with dynamic accuracy FISTA for 200 upper-level evaluations on the same training data, but corrupted with noise level  $\sigma$  ranging from  $10^{-1}$  (as above) to  $10^{-8}$ , see Fig. 14. We see that as  $\sigma \rightarrow 0$ , so does  $\alpha_\theta$  and  $\sigma^2/\alpha_\theta$ . Note that this is a common assumption on the parameter choice rule in regularization theory to yield a *convergent* regularization method [29,44]. It is remarkable that the learned optimal parameter also has this property.

### 4.4 Application: Learning MRI Sampling Patterns

Lastly, we turn our attention to the problem of learning MRI sampling patterns. In this case, our lower-level problem is (6) with  $A(\theta) = F$ , where  $F$  is the Fourier transform, and  $S(\theta)$  is a nonnegative diagonal sampling matrix. Fol-

<sup>5</sup> Available from <http://www.cs.albany.edu/~xypan/research/snr/Kodak.html>.

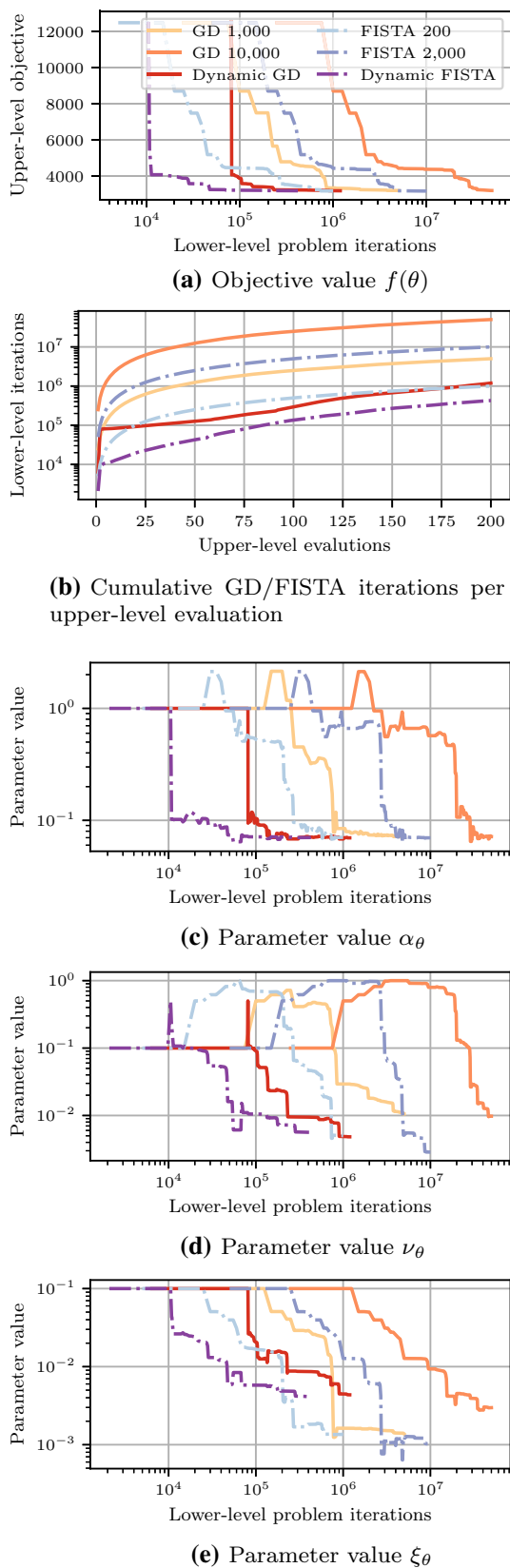


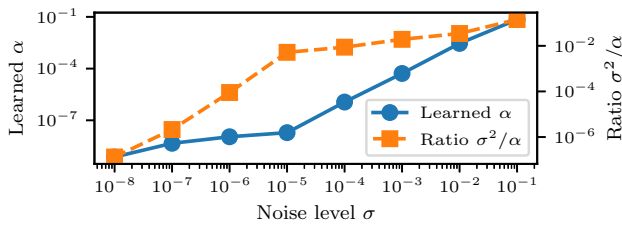
Fig. 12 Results for the 3-parameter denoising problem with 2D images



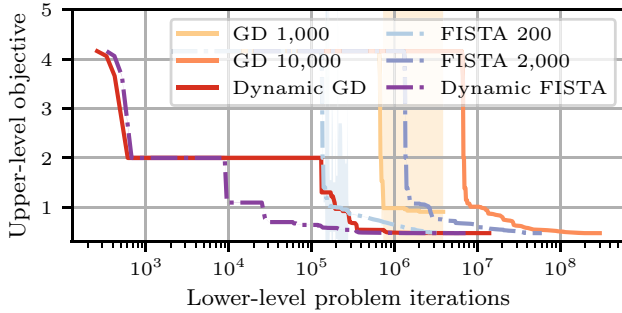
Fig. 13 Example reconstructions using denoising parameters (dynamic FISTA DFO-LS variant). Reconstructions generated with 2,000 FISTA iterations of the lower-level problem

lowing [16], we aim to find sampling parameters  $\theta \in [0, 1]^d$  corresponding to the weight associated to each Fourier mode, our sampling matrix is defined as

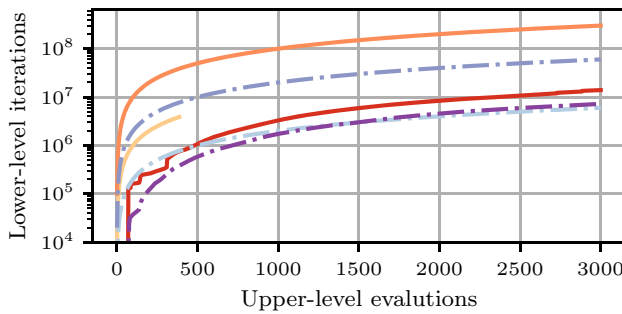
$$S(\theta) := \text{diag} \left( \frac{\theta_1}{1 - \theta_1}, \dots, \frac{\theta_d}{1 - \theta_d} \right) \in \mathbb{R}^{d \times d}. \quad (87)$$



**Fig. 14** Learned regularization parameter  $\alpha_\theta$  for 2D TV denoising with varying noise levels  $\sigma$

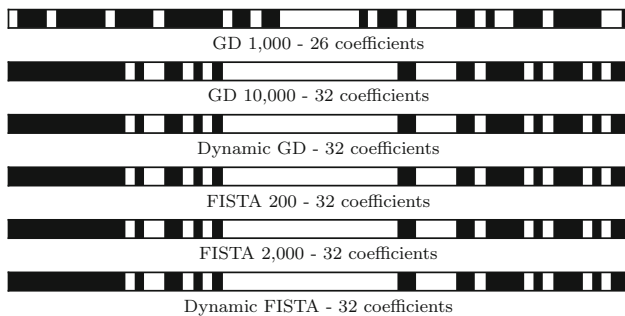


**(a)** Objective value  $f(\theta)$

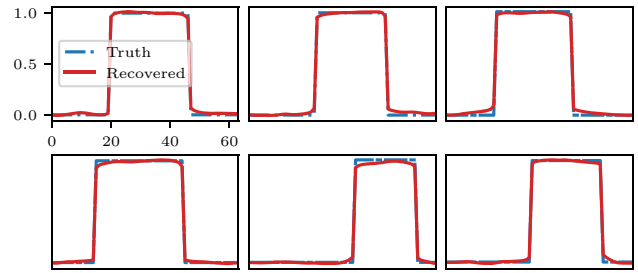


**(b)** Cumulative GD/FISTA iterations per upper-level evaluation

**Fig. 15** Results for the MRI sampling problem. Note: the low-accuracy GD variant ( $K = 1,000$ ) terminates on a small trust-region radius, as it is unable to make further progress



**Fig. 16** Final (after thresholding) MRI sampling patterns found by each DFO-LS variant. This only shows which Fourier coefficients have  $\theta_i > 0.001$ ; it does not show the relative magnitudes of each  $\theta_i$



**Fig. 17** Reconstructions using final (after thresholding) MRI sampling pattern found by the dynamic FISTA variant of DFO-LS. Results from running 2,000 FISTA iterations of the lower-level problem

The resulting lower-level problem is  $\mu$ -strongly convex and  $L$ -smooth as per (7) with  $\|A_\theta^* S_\theta A_\theta\| = \|S(\theta)\| = \max_i \theta_i / (1 - \theta_i)$  and  $\lambda_{\min}(A_\theta^* S_\theta A_\theta) = \min_i \theta_i / (1 - \theta_i)$ .

For our testing, we fix the regularization and smoothness parameters  $\alpha = 0.01$ ,  $\nu = 0.01$  and  $\xi = 10^{-4}$  in (6). We use  $n = 10$  training images constructed using the method described in Sect. 2.2 with  $N = 64$  and  $\sigma = 0.05$ . Lastly, we add a penalty to our upper-level objective to encourage sparse sampling patterns:  $\mathcal{J}(\theta) := \beta \|\theta\|_1$ , where we take  $\beta = 0.1$ . To fit the least-squares structure (22), we rewrite this term as  $\mathcal{J}(\theta) = (\sqrt{\beta} \|\theta\|_1)^2$ . To ensure that  $S(\theta)$  remains finite and  $\mathcal{J}(\theta)$  remains  $L$ -smooth, we restrict  $0.001 \leq \theta_i \leq 0.99$ .

We run DFO-LS with a budget of 3000 evaluations of the upper-level objective and  $\rho_{\text{end}} = 10^{-6}$ . As in Sect. 4.2, we compare dynamic accuracy DFO-LS against (fixed accuracy) DFO-LS with low- and high-accuracy evaluations given by a 1,000 and 10,000 iterations of GD or 200 and 1,000 iterations of FISTA.

With our  $\ell_1$  penalty on  $\theta$ , we expect DFO-LS to find a solution where many entries of  $\theta$  are at their lower bound  $\theta_i = 0.001$ . Our final sampling pattern is chosen by using the corresponding  $\theta_i$  if  $\theta_i > 0.001$ ; otherwise, we set that Fourier mode weight to zero.

In Fig. 15 we show the objective decrease achieved by each variant and the cumulative lower-level work required by each variant. All variants except low-accuracy GD achieve the best objective value with low uncertainty. However, as above, the dynamic accuracy variants achieve this value significantly earlier than the fixed accuracy variants, largely as a result of needing much fewer GD/FISTA iterations in the (lower accuracy) early upper-level evaluations. In particular dynamic accuracy GD reaches the minimum objective value about 100 times faster than high-accuracy GD. We note that FISTA with 200 iterations ends up requiring fewer lower-level iterations after a large number of upper-level evaluations, but the dynamic accuracy variant achieves its minimum objective value sooner.

We show the final pattern of sampled Fourier coefficients (after thresholding) in Fig. 16. Of the five variants which found the best objective value, all reached a similar set of

‘active’ coefficients  $\theta_i > 0.001$  with broadly similar values for  $\theta_i$  at all frequencies. For demonstration purposes we plot the reconstructions corresponding to the coefficients from the ‘dynamic FISTA’ variant in Fig. 17 (the reconstructions of the other variants were all similar). All the training images are reconstructed to high accuracy, with only a small loss of contrast near the jumps.

## 5 Conclusion

We introduce a dynamic accuracy model-based DFO algorithm for solving bilevel learning problems. This approach allows us to learn potentially large numbers of parameters, and allowing inexact upper-level objective evaluations with which we dramatically reduce the lower-level computational effort required, particularly in the early phases of the algorithm. Compared to fixed accuracy DFO methods, we often achieve better upper-level objective values and low-accuracy methods, and similar objective values as high-accuracy methods but with much less work: in some cases up to 100 times faster. These observations can be made for both lower-level solvers GD and FISTA, with different fixed accuracy requirements, for ROF denoising and learning MRI sampling patterns. Thus, the proposed approach is robust in practice, computationally efficient and backed by convergence and worst-case complexity guarantees. Although the upper-level problem is nonconvex, our numerics do not suggest that convergence to non-global minima is a point for concern here.

Future work in this area includes relaxing the smoothness and/or strong convexity assumptions on the lower-level problem (making the upper-level problem less theoretically tractable). Our theoretical analysis would benefit from a full proof that our worst-case complexity bound on the lower-level computational work is tight. Another approach for tackling bilevel learning problems would be to consider gradient-based methods which allow inexact gradient information. Lastly, bilevel learning appears to compute a regularization parameter choice strategy which yields a convergent regularization method. Further investigation is required to back these numerical results by sound mathematical theory.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your

intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Arridge, S., Maass, P., Öktem, O., Schönlieb, C.B.: Solving inverse problems using data-driven models. *Acta Numerica* **28**, 1–174 (2019)
2. Audet, C., Hare, W.: *Derivative-free and blackbox optimization*. Springer Series in Operations Research and Financial Engineering. Springer, Cham (2017)
3. Audet, C., Orban, D.: Finding optimal algorithmic parameters using derivative-free optimization. *SIAM J Optim* **17**(3), 642–664 (2006)
4. Bartels, S., Weber, N.: Parameter learning and fractional differential operators: application in image regularization and decomposition. arXiv preprint [arXiv:2001.03394](https://arxiv.org/abs/2001.03394) (2020)
5. Beck, A.: *First-order methods in optimization*, MOS-SIAM series on optimization, vol. 25. MOS/SIAM, Philadelphia (2017)
6. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm. *SIAM J Imagin Sci* **2**(1), 183–202 (2009)
7. Bellavia, S., Gurioli, G., Morini, B., Toint, P.L.: Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM J Optim* **29**(4), 2881–2915 (2020). <https://doi.org/10.1137/18M1226282>
8. Benning, M., Burger, M.: Modern regularization methods for inverse problems. *Acta Numerica* **27**, 1–111 (2018)
9. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. *SIAM J Imagin Sci* **3**(3), 492–526 (2010)
10. Bredies, K., Lorenz, D.: *Mathematical Image Processing, Applied and Numerical Harmonic Analysis*. Birkhäuser, Basel (2018). <https://doi.org/10.1007/978-3-030-01458-2>
11. Calandra, H., Gratton, S., Riccietti, E., Vasseur, X.: On High-order multilevel optimization strategies. *SIAM J Optim* **31**(1), 307–330 (2021). <https://doi.org/10.1137/19M1255355>
12. Cartis C, Fiala J, Marteau B, Roberts L (2019) Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Trans Math Softw* **45**(3): 32:1–32:41
13. Cartis, C., Roberts, L.: A derivative-free Gauss-Newton method. *Math Program Comput* **11**(4), 631–674 (2019)
14. Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. *Acta Numerica* **25**, 161–319 (2016)
15. Chen, R., Scheinberg, K., Chen, B.Y.: Aligning ligand binding cavities by optimizing superposed volume. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia, PA (2012)
16. Chen, Y., Ranftl, R., Brox, T., Pock, T.: A bi-level view of inpainting-based image compression. In: 19th Computer Vision Winter Workshop (2014)
17. Conn, A.R., Gould, N.I.M., Toint, P.L.: *Trust-region methods*, MPS-SIAM series on optimization, vol. 1. MPS/SIAM, Philadelphia (2000)
18. Conn, A.R., Scheinberg, K., Vicente, L.N.: *Introduction to derivative-free optimization*, MPS-SIAM series on optimization, vol. 8. MPS/SIAM, Philadelphia (2009)
19. Conn, A.R., Vicente, L.N.: Bilevel derivative-free optimization and its application to robust optimization. *Optim Method Softw* **27**(3), 561–577 (2012)
20. De Los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: The structure of optimal parameters for image restoration problems. *J Math Anal Appl* **434**(1), 464–500 (2016)

21. De Los Reyes, J.C., Schönlieb, C.B.: Image denoising: learning the noise model via nonsmooth PDE-constrained optimization. *Inverse Probl Imagin* **7**, 1183–1214 (2013)
22. Duistermaat, J.J., Kolk, J.A.C.: *Multidimensional real analysis I: differentiation*. Cambridge University Press, New York (2004)
23. Engl, H.W., Hanke, M., Neubauer, A.: *Regularization of inverse problems, mathematics and its applications*. Springer, NY (1996)
24. Garmanjani, R., Júdice, D., Vicente, L.N.: Trust-region methods without using derivatives: worst case complexity and the nonsmooth case. *SIAM J Optim* **26**(4), 1987–2011 (2016)
25. Gözcü, B., Mahabadi, R.K., Li, Y.H., Ilicak, E., Çukur, T., Scarlett, J., Cevher, V.: Learning-based compressive MRI. *IEEE Trans Med Imagin* **37**(6), 1394–1406 (2018)
26. Gratton, S., Simon, E., Toint, P.L.: An algorithm for the minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity. *Math Program* (2020). <https://doi.org/10.1007/s10107-020-01466-5>
27. Hansen, P.C.: Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev* **34**(4), 561–580 (1992)
28. Hintermüller, M., Papafitsoros, K., Rautenberg, C.N., Sun, H.: Dualization and Automatic Distributed Parameter Selection of Total Generalized Variation via Bilevel Optimization. *arXiv preprint arXiv:2002.05614* (2020)
29. Ito, K., Jin, B.: *Inverse problems - tikhonov theory and algorithms*. World Scientific Publishing, Singapore (2014)
30. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. *SIAM J Imagin Sci* **6**(2), 938–983 (2013)
31. Lakhmiri, D., Le Digabel, S., Tribes, C.: HyperNOMAD: Hyperparameter optimization of deep neural networks using mesh adaptive direct search. *arXiv preprint arXiv:1907.01698* (2019)
32. Larson, J.W., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. *Acta Numerica* **28**, 287–404 (2019)
33. Lustig, M., Donoho, D.L., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn Reson Med* **58**(6), 1182–1195 (2007)
34. March, A., Willcox, K.: Provably convergent multifidelity optimization algorithm not requiring high-fidelity derivatives. *AIAA J* **50**(5), 1079–1089 (2012)
35. Nesterov, Y.: A method for solving the convex programming problem with convergence rate  $o(1/k^2)$ . *Doklady Akademii Nauk SSSR* **269**(3), 543–547 (1983)
36. Nesterov, Y.: *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Dordrecht (2004)
37. Nocedal, J., Wright, S.J.: *Numerical optimization*, 2nd edn. Springer Science and Business Media, New York (2006)
38. Ochs, P., Ranftl, R., Brox, T., Pock, T.: Bilevel optimization with nonsmooth lower level problems. *SSVM* **9087**, 654–665 (2015)
39. Riis, E.S., Ehrhardt, M.J., Quispel, G.R.W., Schönlieb, C.B.: A geometric integration approach to nonsmooth, nonconvex optimization. *arXiv preprint arXiv:1807.07554* (2018)
40. Robinson, S.M.: Strongly regular generalized equations. *Math Op Res* **5**(1), 43–62 (1980)
41. Rockafellar, R.T., Wets, R.J.B.: *Variational analysis*. Springer science and business media, Berlin (2008)
42. Royer, C.W., O’Neill, M., Wright, S.J.: A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Math Program* **180**(1–2), 451–488 (2020)
43. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Phys D: Nonlinear Phenom* **60**(1), 259–268 (1992)
44. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: *Variational methods in imaging*. Springer Science and Business Media LLC, Berlin (2008)
45. Sherry, F., Benning, M., Los Reyes, J.C.D., Graves, M.J., Maierhofer, G., Williams, G., Schönlieb, C.B., Ehrhardt, M.J.: Learning the sampling pattern for MRI. *IEEE Trans Med Imagin* **39**(12), 4310–4321 (2020)
46. Usman, M., Batchelor, P.G.: *Optimized Sampling Patterns for Practical Compressed MRI*. In: *International Conference on Sampling Theory and Applications* (2009)
47. Zhang, H., Conn, A.R., Scheinberg, K.: A derivative-free algorithm for least-squares minimization. *SIAM J Optim* **20**(6), 3555–3576 (2010)

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Matthias J. Ehrhardt** received the Diploma degree (Hons.) in industrial mathematics from the University of Bremen, Germany, in 2011, and the Ph.D. degree in medical imaging from University College London, U.K., in 2015. He held a postdoctoral position with the Cambridge Image Analysis group, Department for Applied Mathematics and Theoretical Physics, University of Cambridge, U.K., from 2016 to 2018. He is currently a Prize Fellow at the Institute for Mathematical Innovation and a Leverhulme Early Career Fellow at the Department of Mathematical Sciences, both at the University of Bath, U.K. His research interests include optimization, inverse problems, computational imaging, and machine learning.



**Lindon Roberts** received a Bachelor of Computational Science (Honours) from the Australian National University in 2011 and completed his DPhil in mathematics at the University of Oxford, UK, in 2019. He is currently an MSI Fellow at the Mathematical Sciences Institute of the Australian National University. His research interests include derivative-free optimization, nonlinear optimization and machine learning.