Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Dictionary learning based regularization in quantitative MRI: A nested alternating optimization framework

Guozhi Dong[1], Michael Hintermüller [2,3], Clemens Sirotenko[2]

submitted: October 23, 2024

[1] School of Mathematics and Statistics
HNP-LAMA
Central South University
Lushan South Road 932
410083 Changsha
China
E-Mail: guozhi.dong@csu.edu.cn

[2] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
E-Mail: hint@math.hu-berlin.de

[3] Weierstrass Institute
Mohrenstrasse 39
10117 Berlin
Germany
E-Mail: michael.hintermueller@wias-berlin.de
clemens.sirotenko@wias-berlin.de

No. 3135
Berlin 2024

2020 *Mathematics Subject Classification.* 35R30, 49J52, 49J53, 92C55, 68T05.

Key words and phrases. Quantitative MRI, quantitative image reconstruction, regularization, variational methods, machine learning, non-convex and non-smooth optimization.

The work of GD is supported by the NSFC grant, No. 12471402 and the NSF grant from Hunan Province No. 2024JJ5413. This work of MH and CS is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689) and by the priority program Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization (SPP 1962).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Dictionary learning based regularization in quantitative MRI: A nested alternating optimization framework

Guozhi Dong, Michael Hintermüller, Clemens Sirotenko

ABSTRACT. In this article we propose a novel regularization method for a class of nonlinear inverse problems that is inspired by an application in quantitative magnetic resonance imaging (MRI). It is a special instance of a general dynamical image reconstruction problem with an underlying time discrete physical model. Our regularization strategy is based on dictionary learning, a method that has been proven to be effective in classical MRI. To address the resulting non-convex and non-smooth optimization problem, we alternate between updating the physical parameters of interest via a Levenberg-Marquardt approach and performing several iterations of a dictionary learning algorithm. This process falls under the category of nested alternating optimization schemes. We develop a general such algorithmic framework, integrated with the Levenberg-Marquardt method, of which the convergence theory is not directly available from the literature. Global sub-linear and local strong linear convergence in infinite dimensions under certain regularity conditions for the sub-differentials are investigated based on the Kurdyka–Łojasiewicz inequality. Eventually, numerical experiments demonstrate the practical potential and unresolved challenges of the method.

1. INTRODUCTION

Magnetic Resonance Imaging (MRI) is a well-established, non-invasive medical imaging technique. Mathematically, MRI involves solving an ill-posed inverse problem to reconstruct an image of a specific region of the body from noisy Fourier measurements. These measurements represent samples of the spatial frequency content of the tissues' magnetization. In classical MRI, the reconstructed image primarily reflects contrast information, which is heavily influenced by the scanner modality and the specific sampling pattern used. Therefore recent progress in the field concentrates on the inference of real biophysical parameters from these measurements. The methods developed to address this challenge fall within the domain of quantitative magnetic resonance imaging (qMRI) (see [24, 17, 43], and for a recent overview including data-driven reconstruction techniques, see [58]). One particularly successful approach in this field is magnetic resonance fingerprinting (MRF) [43], which serves as the starting point for this article. MRF is an advanced imaging technique that allows for the simultaneous quantification of multiple tissue properties by acquiring highly undersampled magnetic resonance data. Unlike traditional MRI methods, the method in [43] leverages a unique sequence design followed by a dictionary matching process to map the biophysical parameters of interest. However, this two-step process has two main disadvantages. First, it requires solving a highly undersampled, high-dimensional linear inverse problem. Second, it relies on the use of a pre-simulated dictionary or database, which often introduces noticeable discretization errors. Recently in [24, 57] and also [62] strategies that directly use the underlying physical model associated with the Bloch equation [12] were proposed for the reconstruction process. This leads to a nonlinear inverse problem with given data, generated according to

$$f^\delta = F(u_{true}) + \eta, \qquad u_{true} \in U_{ad}, \tag{1}$$

where the goal is to find the physical parameters u_{true} of interest in some predefined set $U_{ad} \subset X$, which is a subset of some function space X. We only have access to the available measurements f^δ, which are corrupted by noise η with an amplitude of $\delta \geq 0$. In section 2, we describe this nonlinear inverse problem and the notation in detail and derive the properties, necessary for our analysis. Since the Fourier data in MRF is usually highly undersampled, we have to apply regularization techniques in order to obtain meaningful solutions without artifacts and amplified noise. For this purpose classical variational methods that obtain an estimator by solving an often nonsmooth optimization problem such

DOI 10.20347/WIAS.PREPRINT.3135 Berlin 2024

as

$$\min_{u\in U_{ad}} \frac{1}{2}\|F(u) - f^\delta\|^2_{L^2(\Omega,\mathbb{C}^L)} + \lambda R(u) \tag{2}$$

are still of fundamental importance. Here usually $R : X \to \overline{\mathbb{R}}$ denotes a convex but often non-smooth regularization term which encodes a-priori knowledge of the solution u_{true}. The regularization parameter $\lambda > 0$ balances the effect of the regularizer R and the so called data discrepancy $\|F(u) - f^\delta\|_{L^2(\Omega,\mathbb{C}^L)}$. Classical examples include the Total Variation (TV) regularization or sparse regularization for a given basis. For TV regularization, we define $R(u) = \|Du\|_{\mathcal{M}}$, where $\|Du\|_{\mathcal{M}}$ represents the Radon norm of the distributional gradient of an image $u \in BV(\Omega)$. For more details and a rigorous mathematical description of these terms, refer to [55, 19, 8].In sparse regularization, as discussed in [44, 56], it is assumed that the true image is sparse in some basis $(\varphi_n)_{n\in\mathbb{N}} \subset H$ for a Hilbert space H. This assumption leads to the formulation of the regularizer:

$$R(u) = \sum_{n\in\mathbb{N}} |\langle u, \varphi_n\rangle_H|.$$

Methods of this kind have been intensively studied during the past decades. We refer to [56, 39, 37, 28] for classical textbooks. However, these approaches are too general, and have certain limitations to specific applications. In particular in quantitative imaging, one aims for estimating precise parameter values instead of the qualitative image contrast. Therefore one major drawback of the universal regularizers in this case is that they often introduce systematic biases into the solution, e.g. the TV regularizer may compress the function values. The second limitation is that they rely entirely on simple a-priori assumptions, such as sparsity of wavelets or Fourier coefficients. These assumptions, while useful in practice, can be too restrictive and may not fully capture the complexity and structure of the data. In many real-world scenarios, data contains hidden patterns or correlations that cannot be easily modeled by handcrafted basis functions. To address this, researchers have been increasingly interested in developing data-driven approaches, such as learning regularizers from the data itself, see [5] for an overview. While many of these approaches have practically outperformed the classical methods in (2), they often come at the price of limited interpretability and robustness. The former is often encountered in neural network based approaches, while in the latter case small changes in f^δ or hyper-parameters in the algorithm may lead to large changes in the reconstruction, cf. [30, 3]. In this work, we try to balance the interpretability and the usage of training data via dictionary learning which has been applied very successfully to imaging and in particular qualitative MRI applications in [49, 50, 52]. In subsection 2.3 we will describe dictionary learning regularization in detail. In general, the approach works on patches that are cut out of the image u and tries to solve the inverse problem simultaneously while decomposing every patch into a product of an unknown joint dictionary $D \in \mathcal{D}$ and sparse coefficients $C \in \mathcal{C}$ with some predefined sets of matrices $\mathcal{D}, \mathcal{C}$. Eventually we will end up with a non convex and non smooth optimization problem

$$\begin{aligned}\min_{u,D,C\in U_{ad}\times\mathcal{D}\times\mathcal{C}} J(u,D,C) = \frac{1}{2}\|F(u) - f^\delta\|^2_{L^2(\Omega,\mathbb{C}^L)} + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} \\ + \lambda\left(\frac{1}{2}\|Pu - DC\|^2_F + \beta\|C\|_1\right).\end{aligned} \tag{3}$$

Here, P extracts patches from the image u, vectorizes them, and assembles them into a large matrix. Additionally, $\lambda, \alpha, \beta > 0$ represent parameters (see subsection 2.3 for details on the notation).

Problems like (3) are typically solved using alternating optimization techniques, following the general pattern outlined below:

Initialize with $u_0, D_0, C_0 \in U_{ad} \times \mathcal{D} \times \mathcal{C}$.
For $k = 0, 1, \ldots$ do

1. Update physical parameter u_k by a complex update procedure.
2. Make one (computationally cheap) update step for the dictionary D_k.
3. Make one (computationally cheap) update step for the sparse coefficients C_k.
4. Go back to 1.

However, this approach might be suboptimal in some cases. To understand why, consider solving the problem

$$\min_{D\in\mathcal{D},C\in\mathcal{C}} \frac{1}{2}\|DC - Pu\|_F^2 + \beta\|C\|_1,$$

which is typically done by alternating between update steps for D and C. In this process, we effectively denoise every patch Pu. As a result, the dictionary learning process alone can be viewed as a powerful denoiser. However, when only one update step is applied to D and C, this denoising effect may not be strong enough. A natural improvement, therefore, is to perform multiple updates of D and C before updating the actual physical parameter of interest, i.e. u. This leads directly to the concept of nested optimization schemes, where a *nested* inner loop is used to solve the dictionary learning problem more thoroughly, up to near stationarity. The nested scheme can be formally stated as follows:

Initialize with $u_0, D_0, C_0 \in U_{ad} \times \mathcal{D} \times \mathcal{C}$.
For $k = 0, 1, \ldots$ do

1. Update u_k by a complex update procedure.
2. For $j = 1, 2, \ldots$ do until some stopping criterion is satisfied
   - j1. Make one (computationally cheap) update step for the dictionary D_k.
   - j2. Make one (computationally cheap) update step for the sparse coefficients C_k.
   - j3. Go back to j1.
4. Go back to 1.

In section 3 we will describe the nested optimization algorithm in detail and investigate its convergence to stationary points of the objective function introduced in subsection 2.3. We present a global convergence analysis that leads to classical sublinear convergence rates and investigate local strong convergence under the framework of the Kurdyka–Łojasiewicz (KL) inequality. For further details on the KL inequality, see also [13, 6, 7]. The one-step procedure can be examined using the convergence framework proposed in [29], which generalizes the approach introduced in [7]. Though, the nested scheme outlined above does not satisfy the necessary descent condition with respect to all variables (u, D, C). Recently, nested optimization techniques have been studied in the finite-dimensional setting [33, 34], demonstrating superior empirical performance compared to the one-step approach. However their framework does only allow for functions that are strongly convex in the nested direction. Moreover, these works assume that the KL inequality holds at every point of the objective, a condition that is challenging to verify in infinite-dimensional problems.

Contributions. Here we briefly summarize the contributions of the article.

- We present a regularization strategy based on orthogonal dictionary learning for the nonlinear inverse problem of qMRI, as introduced in section 2. This method strikes a balance between low contrast bias, data adaptivity, and interpretability. To the best of the authors knowledge, the only comparable approach in the literature is found in [40], where dictionary learning is combined with a slightly different signal model for qMRI. Their focus is on the reconstruction of the relaxivity

parameter $R_1 := 1/T_1$ and m_0 , see section 2 for the terminology. However, convergence theory is not addressed in their study.

- We address the resulting nonconvex and nonsmooth optimization problem using a nested optimization algorithm, inspired by [33, 34], but under different assumptions and with the u-variable in an infinite-dimensional space. Our problem does not fit within the framework of [34], as it lacks the partial strong convexity required there, and we do not assume that the entire objective satisfies the KL-inequality. Furthermore, while our nested update steps are more general, our conclusions are also somewhat weaker, as we do not achieve global strong convergence. It is worth noting the large body of recent work on block alternating optimization strategies, such as [7, 15, 48]. These approaches typically rely on the KL inequality in finite-dimensional settings and often require solving a global optimization problem in each update direction. Neither of these assumptions holds in our framework.

2. Dictionary learning based regularization for quantitative MRI

We will only outline very briefly the fundamental ideas of the signal generation process in MRI. For a more detailed account we refer to classical textbooks such as [17, 42, 59] and for related research papers to [23] and [24]. We will follow mostly the presentation in [24].

2.1. Brief introduction to the physics of qMRI.

In MRI, the source of contrast in an image $\Omega \subset \mathbb{R}^2$ arises from the dynamics of magnetic moments within a slice of the patient's body located at $x \in \Omega$. These magnetic moments evolve under the influence of an externally controlled magnetic field $B(t,x)$, where $t \in [0,T]$ denotes time. The evolution of the magnetic moment $m : [0,T] \times \Omega \to \mathbb{R}^3$ is commonly modeled by the Bloch equations which is given by

$$\partial_t m(x,t) = m(t,x) \times \gamma B(t,x) - \begin{pmatrix} m_1(t,x)/T_2(x) \\ m_2(t,x)/T_2(x) \\ (m_3(t,x) - m_{eq})/T_1(x) \end{pmatrix}, \quad m(0,x) = \begin{pmatrix} 0 \\ 0 \\ m_0(x) \end{pmatrix}.$$

Here, $\gamma > 0$ represents the gyromagnetic ratio, while $T_1, T_2 : \Omega \to \mathbb{R}$ denote the relaxation times, describing how quickly the three-dimensional magnetic moments relax back to the equilibrium state $m_{eq} \in \mathbb{R}^3$ after being excited by the external magnetic field $B(t,x)$. The magnetic field often consists of three parts

$$B(t,x) = B_0(x) + B_1(t,x) + (0,0,G(t)\cdot x)^T.$$

Here $B_0(x)$ denotes a very strong constant magnetic field that points into the positive z direction. $B_1(t,x)$ is only applied for a very short time and excites the magnetic moment away from the equilibrium such that a transversal component of the magnetization occurs. It is also called RF-pulse. The time between to consecutive pulses is called repetition time (TR). $G(t)$ is called gradient-field and is used for spatial frequency encoding. Under suitable assumptions one can show that the signal, that is measured at a receiver coil in a very short time after an RF-pule is applied and turned off immediately can be approximately described by

$$S(t) \approx \int_\Omega \rho(x) m_{12}(t,x) e^{-i\int_0^t G(\tau)\cdot x \,\mathrm{d}\tau} \,\mathrm{d}x,$$

where we introduced the notation $m_{12}(t,x) := m_1(t,x) + i m_2(t,x) \in \mathbb{C}$ and further physical quantity, the proton-spin density $\rho : \Omega \to \mathbb{R}$ that is commonly interpreted as the local density of (predominately) hydrogen protons or "spins" located at $x \in \Omega$. Consequently, in theory the measurement process can be modeled by a composition of the nonlinear solution operator of the Bloch-equation and the Fourier-transform

$$S(t) \approx \mathcal{F}\left(\rho(\cdot)\, m_{12}(t,\cdot)\right)\left(\int_0^t G(\tau)\,\mathrm{d}\tau\right)$$

At the echo time $t = TE$, measurements are taken, and by manipulating the gradient field $t \mapsto G(t)$, different frequencies can be sampled. In Cartesian sampling, the frequencies in two dimensions are collected along parallel lines, with one full line being sampled between two consecutive pulses; see [17] for a comprehensive overview of MRI physics and its practical applications. In classical MRI, after each pulse, the scanner must wait until the magnetization m approximately returns to the equilibrium state m_{eq} before the next RF pulse can be applied. Due to resulting time constraints and the potential for motion artifacts, only a subset of frequencies in the Fourier space (also called k-space) can be sampled. We assume that the measured signal at echo time $y_{TE} = \rho(\cdot)m_{12}(TE, \cdot)$ remains constant during the measurement process. The image y_{TE} is then reconstructed by inverting the sub sampled Fourier transform of the acquired data. Note that by sampling only a single image y_{TE}, the information about the underlying physical parameters, namely $T_1(x)$, $T_2(x)$, and $\rho(x)$, is lost, and only contrast information can be obtained. This means that while the image captures variations in signal intensity, it does not allow for direct estimation of the tissue-specific relaxation times or proton density that contribute to the signal.

In qMRI, the goal is not to recover a single image y_{TE} but to estimate the underlying spatially depended physical parameters of the Bloch equation:

$$x \to u(x) = (T_1(x), T_2(x), \rho(x))^T,$$

which we will model as an element in $L^2(\Omega, \mathbb{R}^3)$ in the forthcoming sections. Magnetic Resonance Fingerprinting (MRF) is a recent and particularly effective technique in qMRI. The idea is to collect not just one, but a sequence of highly under sampled images with a very short repetition time:

$$y_t = \rho(\cdot)m_{12}(t, \cdot) \quad t = 1, \dots, L$$

The number of collected images is denoted by $L \in \mathbb{N}$ and typically ranges between 100 and 1000 in practical applications. Here it is assumed that the data needed to form a full image is acquired at each time point. Crucially, there is no need to wait for the magnetization to return to equilibrium between pulses. In the pioneering work [43], the authors employed an Inversion Recovery Steady-State Free Precession (IR-SSFP) sequence. Using this protocol, the solution to the Bloch equations at echo times indexed by $k = 1, \dots, L$ can be approximated by a discrete dynamical system, cf. [43, 17], given by

$$m_{k+1}(\mathbb{T}(x)) = E_k(\mathbb{T}(x))R(\alpha_k)m_k(\mathbb{T}(x)) + b_k(\mathbb{T}(x)), \quad m_0(x) \in \mathbb{R}^3, \tag{4}$$

where we used the notation $\mathbb{T}(x) = (T_1(x), T_2(x))$. The matrix $R(\alpha)$ is an orthogonal rotation matrix depending on the *Flip angle* $\alpha_k \in (0, 2\pi)$ and $E_k : \mathbb{R}^2 \to \mathbb{R}^{3\times 3}, b_k : \mathbb{R} \to \mathbb{R}^3$ are given by

$$E_k(\mathbb{T}) = \begin{pmatrix} \exp(-\frac{TR_k}{T_2}) & 0 & 0 \\ 0 & \exp(-\frac{TR_k}{T_2}) & 0 \\ 0 & 0 & \exp(-\frac{TR_k}{T_1}) \end{pmatrix}, \quad b_k(\mathbb{T}) = \left[1 - \exp\left(-\frac{TR_k}{T_1}\right)\right] \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

for $\mathbb{T} = (T_1, T_2) \in \mathbb{R}^2_{>0}$ and repetition times $\mathrm{TR}_1, \dots, \mathrm{TR}_L$, which are specified by the scanning protocol.

Remark 2.1. *Note that $E_k : \mathbb{R}^2_{>0} \to \mathbb{R}^{3\times 3}$ and $b_k : \mathbb{R}^2_{>0} \to \mathbb{R}^3$ are only well-defined for elements $T_1, T_2 > 0$. However the extension to $\mathbb{R}^2$ is obvious. We define*

$$\tilde{E}_k(\mathbb{T}) = \lim_{n\to\infty} E_k(\mathbb{T}_n) \quad \tilde{b}_k(\mathbb{T}) = \lim_{n\to\infty} b_k(\mathbb{T}_n),$$

where $\mathbb{T}_n \in (0, +\infty)^2$ for every $n \in \mathbb{N}$ and $\mathbb{T}_n \to \mathrm{proj}_{(0,+\infty)^2}(\mathbb{T})$ as $n \to \infty$. One can easily check that this extension is well defined and does not depend on the choice of the sequence $(\mathbb{T}_n)_n$. Additionally, in this way we extended E_k and b_k to C^∞-functions that are defined everywhere on $\mathbb{R}^2$. We will from now on assume that E_k and b_k are defined everywhere and are smooth.

2.2. **The inverse problem of qMRI and its mathematical properties.** Using the time-discrete dynamical system in (4), we are able to write the whole data generating process, which maps the true physical parameter $u_{true} = (\rho_{true}, T_{1,true}, T_{2,true}) \in L^2(\Omega, \mathbb{R}^3)$ to the data $f^\delta \in L^2(\Omega, \mathbb{C})$, that is measured at the receiver coil, as a single equation. This equation, also known as *forward equation* or *forward model* reads

$$f^\delta = F(u_{true}) + \eta, \quad \|\eta\|_{L^2(\Omega,\mathbb{C}^L)} \leq \delta, \tag{5}$$

where $\eta \in L^2(\Omega, \mathbb{C}^L)$ is complex noise and $\delta \geq 0$ the noise level. We also made use of the *forward operator*

$$F : L^2(\Omega, \mathbb{R}^3) \to L^2(\Omega, \mathbb{C}^L), \quad u \mapsto [S_1\mathcal{F} \circ \Pi(u)_1, \ldots, S_L\mathcal{F} \circ \Pi(u)_L]. \tag{6}$$

In (6), the sequence $S_1, \ldots, S_L : L^2(\Omega, \mathbb{C}) \to L^2(\Omega, \mathbb{C})$ denote a number of sampling operators that cut out the desired frequencies that are actually sampled and the operator $\Pi : L^2(\Omega, \mathbb{R}^3) \to L^2(\Omega, \mathbb{C}^L)$ maps the physical parameter $u = (\rho, T_1, T_2) \in L^2(\Omega, \mathbb{R}^3)$ to the transversal component of the time discrete magnetization process described by the dynamical system (4), i.e.

$$\Pi(u)(x) = \rho(x)[(m_{12}(T_1(x), T_2(x))_1, \ldots, (m_{12}(T_1(x), T_2(x)))_L]^T. \tag{7}$$

The inverse problem of qMRI is then to measure noisy data $f^\delta \in L^2(\Omega, \mathbb{C}^L)$ and to find a good approximation of u_{true}. We also refer to [24] for a detailed account on this setup.

Remark 2.2 (Representation as a superposition operator)**.** *Note that the operator above* $\Pi : L^2(\Omega, \mathbb{R}^3) \to L^2(\Omega, \mathbb{C}^L)$ *can be conveniently represented as a superposition operator using the function*

$$\pi : \mathbb{R}^3 \to \mathbb{C}^L \quad \pi(u) = \rho[(m_{12}(T_1, T_2)_1, \ldots, (m_{12}(T_1, T_2))_L]^T. \tag{8}$$

In the equation above, the function π *takes as input the arguments* $u = (\rho, T_1, T_2) \in \mathbb{R}^3$*, indicating that it operates in a pixel-wise manner. For a detailed discussion on the superposition operator, we refer to* [31, 4].

Many properties of the function π will carry over to the associated superposition operator Π. Hence we start our analysis by collecting some differentiability and stability properties the mapping π in the following theorem. As usual we will make use of the notation $\mathcal{L}^k(X, Y)$ for the space of k-linear and bounded operators between two Banach spaces, X and Y. The first and second order Frechet derivatives for a function $f : X \to Y$ at $x \in X$ are denoted by $f'(x) \in \mathcal{L}(X, Y)$ and $f''(x) \in \mathcal{L}^2(X, Y)$. Moreover we make use of the norm $\|f\|_{C(X,Y)} = \sup_{x\in X} \|f(x)\|_Y$ for a continuous and bounded function $f : X \to Y$. For the sake of readability, we will again write $\mathbb{T} = (T_1, T_2) \in \mathbb{R}^2$.

Theorem 2.3 (Properties of the pointwise solution map)**.** *Let* $\pi : \mathbb{R}^3 \to \mathbb{C}^L$ *and* $m : \mathbb{R}^2 \to \mathbb{R}^{3\times L}$ *be defined as in* (8)*. Then the following statements hold*

(i) *(Differentiability) The mappings* $\pi : \mathbb{R}^3 \to \mathbb{C}^L$ *and* $m : \mathbb{R}^2 \to \mathbb{R}^{3\times L}$ *infinitely differentiable.*

(ii) *(Boundedness) There is a constant* $C = C(L, m_0) > 0$ *such that the following bounds hold true*

$$\|m\|_{C(\mathbb{R}^2,\mathbb{R}^{3\times L})} \leq C \quad \|m'\|_{C(\mathbb{R}^2,\mathcal{L}(\mathbb{R}^2,\mathbb{R}^{3\times L})} \leq C \quad \|m''\|_{C(\mathbb{R}^2,\mathcal{L}^2(\mathbb{R}^2,\mathbb{R}^{3\times L})} \leq C.$$

(iii) *(Lipschitz-properties) Denote* $u_i = (\rho_i, \mathbb{T}_i)$ *for* $i = 1, 2$*. If* $|\rho_i|) \leq b$ *for* $i = 1, 2$ *and some real number* $b > 0$*, then we find a constant* $L = L(b) > 0$ *that depends on* b*, such that*

$$\|\pi(u_1) - \pi(u_2)\|_2 \leq L_b\|u_1 - u_2\|_2$$
$$\|\pi'(u_1) - \pi'(u_2)\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{C}^L)} \leq L_b\|u_1 - u_2\|_2$$

(iv) *(Stability) There is a constant* $C_s > 0$ *such that*

$$\|\pi(u_1) - \pi(u_2)\|_2 \geq C_1\|u_1 - u_2\|_2 \tag{9}$$

Proof. Note that (i) is a direct consequence of the corresponding properties of E_k and b_k. Let us prove (ii). For this purpose we first investigate the properties of the time discrete magnetization $m : \mathbb{R}^2 \to \mathbb{R}^{3L}$. The Fréchet derivatives of m are computed iteratively using the chain rule and the sum rule. For directions $h, h_1, h_2 \in \mathbb{R}^3$, we obtain

$$m_{k+1}(\mathbb{T}) = E_k(\mathbb{T})R_k m_k(\mathbb{T}) + b_k(\mathbb{T}), \tag{10}$$

$$m'_{k+1}(\mathbb{T})[h] = E'_k(\mathbb{T})[h]R_k \cdot m_k(\mathbb{T}) + E_k(\mathbb{T})R_k \cdot m'_k(\mathbb{T})[h] + b'_k(\mathbb{T})[h], \tag{11}$$

$$\begin{aligned} m''_{k+1}(\mathbb{T})[h_1,h_2] = {} & E''_k(\mathbb{T})[h_1,h_2]R_k \cdot m_k(\mathbb{T}) + E'_k(\mathbb{T})[h_1]R_k \cdot m'_k(\mathbb{T})[h_2] \\ & + E'_k(\mathbb{T})[h_2]R_k \cdot m'_k(\mathbb{T})[h_1] + E_k(\mathbb{T})R_k \cdot m''_k(\mathbb{T})[h_1,h_2] \\ & + b''_k(\mathbb{T})[h_1,h_2]. \end{aligned} \tag{12}$$

Since $R_k : \mathbb{R}^3 \to \mathbb{R}^3$ are rotation matrices, we have $\|R_k\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^3)} = 1$. From (10) we directly see that

$$\|m_{k+1}(\mathbb{T})\|_2 \le \|E_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^{3\times 3})}\|m_k(\mathbb{T})\|_2 + \|b_k(\mathbb{T})\|_2 \le C\|m_k(\mathbb{T})\|_2 + b.$$

For constants $b, C > 0$ that are uniformly bounded in k. Here we used the boundedness of the function $t \mapsto \exp(-1/t)$ on $\mathbb{R}$. Hence it is easy to infer by iterating over all k that $\|m(\mathbb{T})\|_{\mathbb{R}^{3\times L}} \le C$ for some uniform constant $C > 0$ by induction. Similarly, using (11), we show for $\|h\|_2 = 1$ that

$$\begin{aligned} \|m'_{k+1}(\mathbb{T})[h]\|_2 &\le \|E'_k(\mathbb{T})[h]\|_2 C + \|E_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^{3\times 3})}\|m'_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^3)} + \|b'_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^3)} \\ &\le C\|h\|_2 C + \|E_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^{3\times 3})}\|m'_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^3)} + \|b'_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^3)}. \end{aligned}$$

From the definition of E_k, b_k, one easily infers that $\|E'_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^{3\times 3})} \le C$, $\|E_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^{3\times 3})} \le C$ and $\|b'_k(\mathbb{T})\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^3)} \le C$ for a generic constant $C > 0$. Thus, by induction $\|m'\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^{3\times L})} \le C$ by a possibly larger constant $C > 0$. Exactly the same argumentation applies to show that

$$\|m''(\mathbb{T})\|_{\mathcal{L}^2(\mathbb{R}^3,\mathbb{R}^{3\times L})} \le C$$

for all $\mathbb{T} \in \mathbb{R}^2$ by making again $C > 0$ larger. This proves (ii). In order to see (iii) we note that by the mean value theorem, which guarantees the existence of a constant $L_m > 0$ such that

$$\begin{aligned} \|m(\mathbb{T}_1) - m(\mathbb{T}_2)\|_2 = \| &\le L_m\|\mathbb{T}_1 - \mathbb{T}_2\|_2, \\ \|m'(\mathbb{T}_1) - m'(\mathbb{T}_2)\|_{\mathcal{L}(\mathbb{R}^3,\mathbb{R}^{3\times L})} &\le L_m\|\mathbb{T}_1 - \mathbb{T}_2\|_2. \end{aligned}$$

Hence, we have shown the Lipschitz continuity of π and π'. Consequently we obtain:

$$\begin{aligned} \|\pi(u_1) - \pi(u_2)\|_2 &= \|\rho_1 m(\mathbb{T}_1) - \rho_2 m(\mathbb{T}_2)\|_2 \\ &\le |\rho_1 - \rho_2|\|m(\mathbb{T}_1)\|_2 + |\rho_2|\|m(\mathbb{T}_1) - m(\mathbb{T}_2)\|_2 \\ &\le \|u_1 - u_2\|_2 C + bL_m\|u_1 - u_2\|_2. \end{aligned}$$

The latter estimate proves the Lipschitz continuity of π as desired. Similarly, we obtain for a direction $h = (h_\rho, h_\mathbb{T}) \in \mathbb{R}^3$ that

$$\begin{aligned} \|(\pi'(u_1) - \pi'(u_2))[h]\|_2 &\le \|h_\rho m(\mathbb{T}_1) - h_\rho m(\mathbb{T}_2)\|_2 + \|\rho_1 m'(\mathbb{T}_1)[h_\mathbb{T}] - \rho_2 m'(\mathbb{T}_2)[h_\mathbb{T}]\|_2 \\ &\le |h_\rho|\|m(\mathbb{T}_1) - m(\mathbb{T}_2)\|_2 + |\rho_1 - \rho_2|\|m'(\mathbb{T}_1)\|_2\|h_\mathbb{T}\| \\ &\quad + |\rho_2|\|(m'(\mathbb{T}_1) - m'(\mathbb{T}_2))[h_\mathbb{T}]\|_2 \\ &\le \|h\|_2 L_m\|u_1 - u_2\|_2 + \|u_1 - u_2\|_2 C\|h\|_2 \\ &\quad + bL_m\|u_1 - u_2\|_2\|h\|_2 \\ &\le (L_m + C + bL_m)\|u_1 - u_2\|_2\|h\|_2, \end{aligned}$$

which proves eventually the Lipschitz continuity in (iii) for $L_b = (L_m + C + bL_m)$. □

Note that in particular (9) guarantees that the parameters u can be in principle identified from complete measurements. The next theorem establishes properties of the corresponding superposition operator in (7) using the theory, which is developed in [31].

Theorem 2.4 (Properties of the solution operator)**.** *Consider the map $\pi : \mathbb{R}^3 \to \mathbb{C}^L$ for a bounded domain $\Omega \subset \mathbb{R}^n$ with Lipschitz boundary. Then the corresponding superposition operator, or the solution operator of the time discrete Bloch equation, introduced in (7), satisfies the following properties:*

(i) *The operator Π maps all of $L^p(\Omega, \mathbb{R}^3)$ into all of $L^q(\Omega, \mathbb{C}^L)$ for exponents $1 \leq p \leq q < \infty$.*

(ii) *(Frechet-differentiability) The operator is Frechet differentiable as a mapping between the spaces $\Pi : L^p(\Omega, \mathbb{R}^3) \to L^2(\Omega, \mathbb{C}^L)$ for $p \geq 4$. The Frechet-derivative is given by*

$$\Pi' : L^p(\Omega) \to \mathcal{L}(L^p(\Omega), L^2(\Omega)) \quad D\Pi(u)[h](x) = \pi'(u(x))h(x) \quad \text{for } p \geq 4. \tag{13}$$

(iii) *(2nd order Frechet-differentiability) The operator Π is two times Frechet differentiable as a mapping between the spaces $\Pi : L^p(\Omega, \mathbb{R}^3) \to L^2(\Omega, \mathbb{C}^L)$ for $p \geq 6$. The 2nd order Frechet-derivative is given by*

$$\Pi'' : L^p(\Omega) \to \mathcal{L}(L^p(\Omega), L^2(\Omega)) \quad \Pi''(u)[h_1, h_2](x) = \pi''(u(x))h_1(x)h_2(x) \quad \text{for } p \geq 6. \tag{14}$$

Proof. We employ the theory on abstract superposition operators developed in [31] and make use of the notation $u = (\rho, \mathbb{T}) \in \mathbb{R}^3$ as above. For (i) we have to verify the growth condition

$$\|\pi(u)\|_2 \leq c + c\|u\|_2^{\frac{p}{q}}$$

for all $u \in \mathbb{R}^3$ some $c > 0$. Since $\pi(u) = \rho \cdot m_{12}(\mathbb{T})$ we directly obtain $\|\pi(u)\|_2 = |\rho| \|m_{12}(\mathbb{T})\|_2 \leq \|u\|_2 \|m(\mathbb{T})\|_2$ and consequently the assertion (i). For (ii) we recall [31, Theorem 7]. For the continuous F-differentiability of Π we have to verify that the mapping $u \mapsto \pi'(u(\cdot))$ is continuous as a mapping from $L^p(\Omega, \mathbb{R}^3)$ to $L^r(\Omega, \mathcal{L}(\mathbb{R}^3, \mathbb{C}^L))$ for $r = pq/(p-q)$. This can be shown by verifying again the growth condition

$$\|\pi'(u)\|_{\mathcal{L}(\mathbb{R}^3, \mathbb{C}^L)} \leq c + c\|u\|_2^{\frac{p}{r}}$$

for all $u \in \mathbb{R}^3$ and some $c > 0$. Since $\|\pi'(u)[h]\|_2 = \|h_\rho m(\mathbb{T}) + \rho m'(\mathbb{T})[h_\mathbb{T}]\|_2 \leq |h_\rho| \|m(\mathbb{T})\| + |\rho| \|m'(\mathbb{T})\|_2 \|h_\mathbb{T}\|_2$ again for a direction $h = (h_\rho, h_\mathbb{T}) \in \mathbb{R}^3$. By the boundedness of $m(\mathbb{T})$ and $m'(\mathbb{T})$ we obtain $\|\pi'(u)[h]\|_2 \leq (c + c\|u\|_2)\|h\|_2$ for some $c > 0$ as desired. Since $p/r = (p-q)/q = 1$ for $p = 4$ and $q = 2$ the statement follows. In order to show (iii) we invoke theorem 9 of [31]. Completely analogous to (ii) we have to show that $u \mapsto \pi''(u(\cdot))$ is continuous as a mapping from $L^p(\Omega, \mathbb{R}^3)$ to $L^s(\Omega, \mathcal{L}^2(\mathbb{R}^3, \mathbb{C}^L))$ for $s = pq/(p-2q)$, which again can be done by verifying the growth condition

$$\|\pi''(u)\|_{\mathcal{L}^2(\mathbb{R}^3, \mathbb{C}^L)} \leq c + c\|u\|_2^{\frac{p}{s}}$$

for all $u \in \mathbb{R}^3$ and some other $c > 0$ as above. Calculating

$$\pi''(u)[h^1, h^2] = h^2_\rho m'(\mathbb{T})[h^1_\mathbb{T}] + h^1_\rho m'(\mathbb{T})[h^2_\mathbb{T}] + \rho m''(\mathbb{T})[h^1_\mathbb{T}, h^2_\mathbb{T}]$$

yields indeed the growth condition $\|\pi''(u)[h^1, h^2]\|_2 \leq (c + c\|u\|_2)\|h^1\|_2 \|h^2\|_2$ for all $h^1, h^2 \in \mathbb{R}^3$ with decomposition $h^i = (h^i_\rho, h^i_\mathbb{T})$. Hence the growth-condition is satisfied for $p/s = 1$ from which we get $1 = p/s = (p-2q)/q = (p-4)/2$ using $q = 2$. Hence for $p \geq 6$ the growth-condition is satisfied, and second order Frechet-differentiability holds. □

2.3. **Regularization by dictionary learning.** The core idea behind dictionary learning is to identify a representation system or basis directly from the data, rather than relying on predefined bases like wavelets, polynomials or Fourier basis. However, learning a basis for a whole discrete image involves a large number of degrees of freedom, making the learning process complex and prone to overfitting. To mitigate this, a common approach is to restrict the learning process to small image patches rather than the entire image. These patches are extracted from the discretized image, which helps reduce

the complexity while still capturing local image structures. Let us briefly outline how dictionary learning is used to regularize linear inverse problems before incorporating it into our nonlinear model (5). The pioneering approach in [2] used dictionary learning in a two-step procedure. In the first step, the goal is to decompose a data matrix X, containing lots of clean image patches, into a rather small dictionary $D \in \mathcal{D}$ of basis elements and corresponding sparse representations $C \in \mathcal{C}$ by solving the problem

$$\min_{D\in\mathcal{D},C\in\mathcal{C}} \frac{1}{2}\|DC - X\|_F^2 + \|C\|_0. \tag{15}$$

Here we made use of the Frobenious norm $\|\cdot\|_F$. The set $\mathcal{D}$ of admissible dictionaries is often defined to prevent multiple solutions arising from rescaling, e.g. by choosing $\mathcal{D}$ to be the set of matrices with normalized columns or the set of orthonormal matrices. On the other hand, the set $\mathcal{C}$ that defines the space for the sparse coefficients is usually the entire space. The matrix $X = [p_1, \dots, p_N]$ typically contains a large number N of vectorized image patches $p_i \in \mathbb{R}^K$, where each patch p_i represents a small region of an image. The matrix $D \in \mathcal{D} \subset \mathbb{R}^{K\times M}$ contains the dictionary elements, with $D = [\varphi_1, \dots, \varphi_M]$ consisting of columns $\varphi_i \in \mathbb{R}^K$, each of which corresponds to a learned basis vector or ätomïn the dictionary. If the number of dictionary elements M exceeds the patch dimension K (i.e., $M > K$), the dictionary is called overcomplete. The main assumption behind dictionary learning is that clean image patches can be well represented using only a few dictionary elements, meaning that

$$p_i = \sum_{j=i}^{M} \varphi_j c_{ij} \quad \text{for certain coefficients } c_{ij} \in \mathbb{R}^M \text{ with } \|c_j\|_0 \ll M, \tag{16}$$

where $\|c\|_0$ denotes the number of nonzero entries of a vector c. Writing (16) in a vector matrix notation, we obtain $X = DC$ with a sparse matrix C. Once the factorization is completed, the second step involves using these learned dictionary atoms to find an estimate of $u_{true} \in L^2(\Omega)$ from measurements $f^\delta = F(u) + \eta$ by solving a problem of the type

$$\min_{u\in L^2(\Omega),C\in\mathcal{C}} \|F(u) - f^\delta\|_{L^2(\Omega)}^2 + \frac{\lambda}{2}\left(\|P\mathbb{D}^h u - DC\|_F^2 + \beta\|C\|_0\right) \tag{17}$$

where $\mathbb{D}^h : L^2(\Omega) \to \mathbb{R}^N$ is a not further specified discretization operator. The two step approach in (17) has been successfully applied to several linear inverse problems during the past decade, cf. [5] and the references in the corresponding section. It has been recognized recently, [50, 49] that very good results can also be obtained by learning the dictionary while reconstructing the image. This approach combines the tasks of dictionary learning and image reconstruction into a single optimization framework, allowing both processes to inform and to enhance each other. This idea leads to the following optimization problem:

$$\min_{u\in L^2(\Omega),D\in\mathcal{D},C\in\mathcal{C}} \|F(u) - f^\delta\|_{L^2(\Omega)}^2 + \frac{\lambda}{2}\left(\|P\mathbb{D}^h u - DC\|_F^2 + \alpha\|C\|_0\right), \tag{18}$$

not discussing issues related to infinite dimensionality or the existence of solutions here. In [52, 51, 53] the approach in (18) has been further developed and successfully applied to the linear inverse problem of MRI, i.e. where $F = S \circ \mathcal{F}$ with an undersampling operator S and the Fourier transform $\mathcal{F}$. This resulted in a number of interpretable, robust, and data-adaptive regularization methods, see also [53] for an overview. Additionally, these methods appear to significantly reduce the contrast bias typically introduced by methods of the type (2).To address issues concerning the existence of solutions, we introduce a small strongly convex term $\|\nabla\|_{L^2(\Omega)}^2$. Additionally, to ensure the problem remains continuous, we replace the $\|\cdot\|_0$-seminorm with the convex and continuous ℓ_1-norm. This relaxation is a standard approach in the field of compressed sensing, as discussed in [**?**]. We eventually end up with

the following dictionary regularized qMRI problem, which we aim to solve in the following sections:

(P_0)

$$\min_{(u,D,C)\in U_{ad}\times\mathcal{D}\times\mathbb{R}^{K\times M}} \frac{1}{2}||F(u)-f^\delta||^2_{L^2(\Omega,\mathbb{C}^L)}+\frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)}+\frac{\lambda}{2}\left(\|P\mathbb{D}^h u-DC\|_F^2+\beta\|C\|_1\right)$$

Let us briefly fix the termininology in (P_0)

Assumption 2.5 (On the qMRI problem)**.** *We consider the following assumptions for problem* (P_0)*:*

Q1. *The space of admissible parameters is* $H_0^1(\Omega,\mathbb{R}^3)$ *where* $\Omega=[0,1]^2$ *for simplicity.*

Q2. *The set of admissible parameters is given by*

$$U_{ad}=\{u\in H_0^1(\Omega,\mathbb{R}^3): a\le u_i(x)\le b \quad \text{for a.e. } x\in\Omega\}.$$

Q3. *Let* $h=1/N$ *for some* $N\in\mathbb{N}$*. The discretization space* U^N *is given by*

$$U^N=\operatorname{span}\{N^{-2}\mathbf{1}_{\Omega_{i,j}}:\Omega\to\mathbb{R}\mid i,j=1,\dots,N\},$$

where every $\Omega_{i,j}\subset\Omega$ *is defined as* $\Omega_{i,j}=(i-1,j-1)+[0,1/N]\times[0,1/N]$*. The discretization (projection) operator* $\mathbb{D}^h: L^2(\Omega)\to U^N$*, for the meshsize* $h=1/N$ *is consequently defined separately on every patch as*

$$(\mathbb{D}^h u)|_{\Omega_{i,j}}=\frac{1}{N^2}\int_{\Omega_{ij}}u(x)\,\mathrm{d}x,$$

Q4. *The forward operator* $F: H_0^1(\Omega,\mathbb{R}^3)$ *is given by* (6)*.*

Q5. *For the space of admissible matrices we choose* $\mathcal{D}=O_k:=\{M\in\mathbb{R}^{K\times K}: M^\top M=I\}$*, which is also called* Stiefel-Manifold*, see* [1]*. The space of admissible Matrices is* $\mathcal{C}=\mathbb{R}^{K\times N^2}$*.*

Remark 2.6 (Choice of the dictionary space)**.** *We select the space of orthogonal matrices as the Ansatz space for the dictionary due to its ability to enable simple and fast update steps, while still maintaining good reconstruction quality. Various extensions can be introduced to incorporate more complex and sophisticated models, as proposed, for example, in* [53] *for linear inverse problems.*

3. A NESTED LEVENBERG-MARQUARDT TYPE OPTIMIZATION ALGORITHM

The goal in this section is to derive an optimization algorithm in order to find stationary points of the objective function (P_0) introduced in the previous section. However, in order to keep the presentation simple, we will only analyse the problem for the reconstruction of a single $u\in H_0^1(\Omega)$, instead of using the space $H_0^1(\Omega,\mathbb{R}^3)$. The convergence analysis is the same. The Hilbert space $H_0^1(\Omega)$ will be equipped with the inner product $\langle u,v\rangle_{H^1(\Omega)}=\langle u,v\rangle_{L^2(\Omega)}+\langle\nabla u,\nabla v\rangle_{L^2(\Omega)}$, which is equivalent to $\langle u,v\rangle_{H_0^1(\Omega)}=\langle\nabla u,\nabla v\rangle_{L^2(\Omega)}$ by the Poincaré inequality, cf. [8]. For regularization parameters $\alpha,\lambda>0$, we consider the following class of optimization problems on the space $X:=H_0^1(\Omega)\times Z$, where Z is a finite-dimensional abstract Hilbert space that is not further specified.

(P_1)

$$\min_{u\in H_0^1(\Omega),z\in Z} J(u,z):=\frac{1}{2}\|F(u)-f^\delta\|^2_{L^2(\Omega,\mathbb{C}^L)}+\frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)}+h(u,z)+R_1(u)+\mathcal{I}_{U_{ad}}(u)+R_2(z).$$

During the analysis we will also use the notations

$$f^\alpha(u):=f(u)+\frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)}:=\frac{1}{2}\|F(u)-f^\delta\|^2_{L^2(\Omega,\mathbb{C}^L)}+\frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)}. \tag{19}$$

Moreover, make the following assumptions for the abstract problem in (P_1).

Assumption 3.1 (Assumption on the problem class (P_1))**.** *For the problem class* (P_1)*, we make the following assumptions.*

B2.) *The function* $F : L^p(\Omega) \to L^2(\Omega, \mathbb{C}^L)$ *is two times continuously Frechet-differentiable and the Frechet-derivatives at the point* $u \in L^p(\Omega)$ *are denoted by* $D^k F(u) \in \mathcal{L}^k(L^p(\Omega), L^2(\Omega))$ *for* $k = 1, 2$. *The whole objective* $J : H_0^1(\Omega) \times Z$ *is assumed continuous on* $\mathrm{dom}(J)$.

B3.) *The set of admissible parameters* U_{ad} *has the form*

$$U_{ad} = \{u \in H_0^1(\Omega) \mid a \leq u(x) \leq b \text{ for almost every (a.e.) } x \in \Omega \},$$

for some positive real numbers $0 < a < b$.

B4.) *The space* Z *is a finite dimensional Hilbert space and the function* $h : X \to \mathbb{R}$ *satisfies:*

H1.) h *is bounded from below, proper and twice continuously differentiable on* $L^2(\Omega) \times Z$.

H2.) $u \mapsto h(u, z)$ *is convex for every element* $z \in Z$.

H3.) *The gradient* $u \mapsto \nabla_z h(u, z)$ *is Lipschitz-continuous, and there is a uniform bound* $C > 0$ *on the Lipschitz constants, i.e*

$$\|\nabla_z h(u_1, z) - \nabla_z h(u_2, z)\|_Z \leq L_{\nabla_h h(\cdot,z)} \|u_1 - u_2\|_{H_0^1(\Omega)},$$

with $L_{\nabla_h h(\cdot,z)} \leq C$ *for all* $z \in Z$.

B5.) *The function* $R_1 : H_0^1(\Omega) \to \overline{\mathbb{R}}$ *is proper, lower semi-continuous and convex. The function* $R_2 : Z \to \overline{\mathbb{R}}$ *is proper, lower semi-continuous and coercive on* Z, *i.e.* $\|z\|_Z \to \infty$ *implies* $R_2(z) \to \infty$. *Note that* R_2 *will be usually non convex.*

It is clear, that the dictionary learning regularized problem in (P_0) is a specific case of the problem class presented in (P_1) for the choice $h(u, z) = \frac{1}{2}\|P\mathbb{D}^h u - DC\|_F$ and $Z = (D, C) \in \mathbb{R}^{K\times K} \times \mathbb{R}^{K\times N}$ and $R_2(D, C) = \mathcal{I}_{O_K}(D) + \beta\|C\|_1$. Let us briefly investigate existence of solutions.

Theorem 3.2 (Existence of solutions for (P_1))**.** *Let the statements under 3.1 hold true. Then Problem* (P_1) *has a solution* $(u^*, z^*) \in U_{ad} \times Z$.

Proof. We use the direct method of calculus of variation. Let $(u_n, z_n)_n$ be an infimizing sequence, which exists since J is bounded from below. We observe, that $(\|\nabla u_n\|_{L^2})_n$ is bounded by the definition of J. Hence, by the Poincaré inequality, cf [8], we infer also the boundedness of $(\|u_n\|_{H_0^1(\Omega)})_n$. Using the coercivity of R_2 we also obtain the boundedness of $(\|z_n\|_Z)_n$. Therefore there exists a subsequence which (after relabeling) converges to some element (u^*, z^*) strongly in $L^2(\Omega) \times Z$ due to the compact embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. As the set U_{ad} is sequentially closed in the strong $L^2(\Omega)$-topology and the mapping $F : L^2(\Omega) \to L^2(\Omega)$ is continuous, we infer by classical lower semicontinuity arguments that (u^*, z^*) is a global solution of (P_1). □

3.1. **The algorithm and its global convergence analysis.** Let us now consider the optimization scheme for finding stationary points of (P_1). However, before stating the algorithm, we must clarify what we mean by stationarity.

Mathematical preliminaries. Let us briefly recall the mathematical setup needed for the analysis of non smooth and non differentiable optimization problems. For a comprehensive account, we refer to the monographs [45, 54]. Fur our purpose we will just follow the presentation in [29]. Let us consider a function $f : H \to \overline{\mathbb{R}}$, which is proper and lower semi-continuous and defined on a Hilbert space H with dual space $H^* = \mathcal{L}(H, \mathbb{R})$. We first need a suitable notion of derivative in the nons mooth setting.

Definition 3.3 (Subdifferentials)**.** *Let* $x \in \mathrm{dom}(f)$ *and* $f : H \to \overline{\mathbb{R}}$ *closed and proper.*

(i) *We call an element* $p \in H^*$ *Frechet-subgradient at* $x \in \mathrm{dom}(f)$, *if*

$$\liminf_{y\to x, y\neq x} \frac{f(y) - f(x) - \langle p, y - x\rangle_{H^*}}{\|y - x\|} \geq 0.$$

The set of Frechet-subgradients at x is called Frechet-subdifferential and is denoted by $\hat{\partial} f(x)$. If $x \notin \operatorname{dom}(f)$, we set $\hat{\partial} f(x) = \emptyset$.

(ii) *We call an element $p \in H^*$ (limiting)-subgradient at $x \in \operatorname{dom}(f)$, if there exists sequences $(x_k)_{k\in\mathbb{N}}$ and $(p_k)_{k\in\mathbb{N}}$ such that $p_k \in \hat{\partial} f(x_k)$ for every k and $x_k \to x$ and $p_k \rightharpoonup p$ as $k \to \infty$. As before we define $\partial f(x) = \emptyset$ if $x \notin \operatorname{dom}(f)$.*

We call an element $x \in H$ a stationary point of f if

$$0 \in \partial f(x).$$

For the special structure in this work, it is important to calculate subdifferentials of functions that are defined on a Cartesian product spaces and allow a composition into a differentiable and a non-smooth term.

Theorem 3.4 (Chain rule, [45, Proposition 1.107])**.** *The following two calculus rules hold true:*

(i) *Let $f(x) = g(x) + h(x)$ for $g, h : H \to \overline{\mathbb{R}}$ proper and closed. If in addition g is continuously differentiable in a neighbourhood of $x \in \operatorname{dom}(f)$, then*

$$\partial f(x) = \nabla g(x) + \partial h(x).$$

(ii) *If $f : H_1 \times H_2 \to \overline{\mathbb{R}}$ is defined on a Cartesian product of two Hilbert spaces H_1, H_2 and has the structure $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$ for $f_i : H_i \to \overline{\mathbb{R}}$ proper and closed, then*

$$\partial f(x_1, x_2) = \partial f_1(x_1) \times \partial f_2(x_2).$$

Consequently, we may compute the subdifferential of the objective $J : H_0^1(\Omega) \times Z \to \overline{\mathbb{R}}$ defined in (P_1) as:

$$\partial J(u, z) = \left\{ \nabla f^\alpha(u) + \lambda \nabla_u h(u, z) + \lambda \partial R_1(u), \lambda \nabla_z h(u, z) + \lambda \partial R_2(z) \right\}.$$

To measure the distance to stationarity, we will also need the notion of the *lazy slope*, defined as

$$\operatorname{dist}(0, \partial f(x)) := \inf_{p \in \partial f(x)} \|p\|_{H^*}.$$

While always finite, there might be situations where the infimum is not attained. The next lemma is of fundamental importance if we want to show that the limit of a sequence of almost stationary points is a stationary point. Before we state the result, let us recall, that a sequence $(x_k)_{k\in\mathbb{N}}$ converges in the f-attentive sense to $x^* \in H$ if $x_k \to x^*$ and $f(x_k) \to f(x^*)$. We write $x_k \xrightarrow{f} x^*$ in this case.

Lemma 3.5 (Stationarity of the limit)**.** *Let $x_k \in H$ for every $k \in \mathbb{N}$ and assume that the follow holds true*

$$x_k \xrightarrow{f} u^* \quad \text{and} \quad \operatorname{dist}(0, \partial f(u_k)) \to 0 \qquad \text{as } k \to \infty.$$

Then $0 \in \partial J(x^*)$, i.e. $x^* \in H$ is stationary.

Proof. For a proof we refer to [29]. □

Description of the algorithm. As already outlined in the introduction of this article, we aim at solving (P_1) by alternating between between an update step for the z-variable, denoted by *z-step*, and the u-variable, called *u-step*. We will describe both steps in detail in the following paragraphs.

Solution for the z-step. Given the iterate $u_k \in H_0^1(\Omega)$, our goal in the z-step is to solve the problem

$$(P_z^k) \qquad \min_{z\in Z} g_k(z) := h(u_k, z) + R_2(z)$$

up to limiting-stationarity. For this purpose we assume to be given an abstract algorithm which is specified by a number of transition rules $\mathcal{A}_k^n : Z \to Z$ to define the next iterate, given the current one. Formally this means $z_k^{n+1} = \mathcal{A}_k^n(z_k^n)$ starting from $z_k^0 := z_k \in Z$. This algorithm will produce a so called *descent sequence*, defined next:

Definition 3.6 (Descent sequence)**.** *A descent sequence for problem* (P_z^k) *is a sequence of points* $(z^n)_{n\in\mathbb{N}} \subset Z$ *such that the following two inequalites hold true:*

(i) *(Sufficient descent of function values) There is a* $\sigma_1 > 0$ *such that*

$$g_k(z^{n+1}) \le g_k(z^n) - \frac{\sigma_1}{2}\|z^{n+1} - z^n\|_Z^2 \quad \textit{for all } n \in \mathbb{N}. \tag{20}$$

(ii) *(Gradient inequality) There is a* $\sigma_2 > 0$*, such that*

$$\operatorname{dist}(0, \partial g_k(z_{n+1})) \le \sigma_2\|z^{n+1} - z^n\|_Z \quad \textit{for all } n \in \mathbb{N}. \tag{21}$$

Many algorithms in the literature are known to produce descent sequences, cf. [6, 7, 15, 14]. However, accelerated algorithms such as FISTA are known for their non-monotone convergence behavior and do not satisfy these properties; see [10].

Assumption 3.7 (On the update algorithm in the z-step)**.** *We assume that our algorithm for minimizing* (P_z^k)*, defined by an update mechanism* $\mathcal{A}_k^n : Z \to Z$ *produces a descent sequence for* (P_z^k)*.*

Let us briefly formalize the algorithm for the z-step.

Algorithm 1 Computation of a near stationary point for the z-step at the k-th outer loop iteration.

Get $(u_k, z_k) \in H_0^1(\Omega) \times Z$, accuracy $\eta_k > 0$.
Initialize with $z_k^0 = z_k$ and compute $z_k^1 = \mathcal{A}_k^0(z_k^0)$.
Set $n = 1$.
while $\|z_k^n - z_k^{n-1}\| \ge \eta_k$ **do**
    Compute $z_k^{n+1} = \mathcal{A}_n(z_k^n) \in Z$ such that (20) and (21) hold true for constants $\sigma_1, \sigma_2 > 0$.
    Set $n = n + 1$.
end while
Set $z_{k+1} := z_k^{n_k}$ where $n_k \in \mathbf{N}$ is the first iterate such that $\|z_k^{n_k} - z_k^{n_k-1}\| < \eta_k$.
Return $z_{k+1} = z_k^{n_k}$.

From the properties (20) and (21) we directly infer the following lemma, which describes the fundamental behaviour of every descent sequence.

Lemma 3.8. *Let* $(u_k, z_k) \in H_0^1(\Omega) \times Z$*, accuracy* $\eta_k > 0$ *be given such as in algorithm 1. For the generated sequence the following properties hold:*

(i) *The sequence of function values converges monotonically to its infimum, i.e.*

$$g_k(z_k^n) \searrow g_k^* := \inf_{n\in\mathbb{N}} g_k(z_k^n) \ge 0 \quad \textit{as } n \to \infty.$$

(ii) *The following estimate holds true for every* $N \in \mathbf{N}$

$$\sum_{n=1}^{N} \|z_k^n - z_k^{n-1}\|_Z^2 \le \frac{2(g(z_k^0) - g(z_k^N))}{\sigma_1} \le \frac{2(g(z_k^0) - g_k^*)}{\sigma_1}.$$

In particular $\|z_k^n - z_k^{n-1}\|_Z \to 0$ *and* $\operatorname{dist}(0, \partial g(z_k^n)) \to 0$ *as* $n \to \infty$ *.*

(iii) *The following estimate holds for the lazy slope and every* $N \in \mathbf{N}$

$$\min_{n=1,\dots,N} \operatorname{dist}(0, \partial g(z_k^n)) \leq \sqrt{\frac{\sigma_2^2}{2\sigma_1}\left(\frac{g(z_k^0) - g(z_k^N)}{N}\right)}.$$

Proof. Simple calculations. A proof can be for instance found in [15]. □

Remark 3.9 (Complexity of algorithm 1)**.** *Based on the above inequalities, we may also derive that*

$$\min_{n=1,\dots,N} \|z_k^n - z_k^{n-1}\|_Z^2 \leq \frac{2(g(z_k^0) - g(z_k^N))}{\sigma_1 N}.$$

Consequently, if an accuracy parameter $\eta_k > 0$ *is given, this bound implies that* $\|z_k^n - z_k^{n-1}\|_Z \leq \eta_k$ *after at most* $N(\eta_k)$ *steps, where*

$$N(\eta_k) = \frac{2(g(z_k^0) - g(z_k^{n_k}))}{\sigma_1 \eta_k^2}.$$

For example, if the accuracy parameter $\eta_k = k^{-\gamma}$ *for some* $\gamma > 0$ *is given, then the stopping index satisfies*

$$n_k \leq \frac{2(g_k(z_k) - g_k(z_{k+1}))}{\sigma_1} k^{2\gamma} \leq \frac{2(J(u_k, z_k) - J(u_{k+1}, z_{k+1}))}{\sigma_1} k^{2\gamma}. \tag{22}$$

Solution for the u-step. Let us now consider the step for u. We update the physical parameter of interest, $u \in H_0^1(\Omega)$, by performing exactly one Levenberg-Marquardt step, i.e. for a given iterate $(u_k, z_{k+1}) \in H_0^1(\Omega) \times Z$, we consider the following optimization problem:

$$u_{k+1} = \operatorname*{arg\,min}_{u \in U_{ad}} g_{\lambda_k}(u, u_k) + \frac{\alpha}{2}\|\nabla u\|_{L^2(\Omega)}^2 + h(u, z) + R_1(u). \tag{P_u}$$

Here $g_{\lambda_k}(u, v)$ is a *model function* that approximates f and is defined as follows

$$\begin{aligned} g_{\lambda_k}(u, v) &:= g(u, v) + \frac{\lambda_k}{2}\|u - v\|_{H_0^1(\Omega)}^2 \\ &:= \frac{1}{2}\|F'(u_k)[u - u_k] + F(u_k) - f^\delta\|_{L^2(\Omega)}^2 + \frac{\lambda_k}{2}\|u - v\|_{H_0^1(\Omega)}^2. \end{aligned} \tag{23}$$

We further define the *approximation error* $e_{\lambda_k} g(u, v) := g_{\lambda_k}(u, v) - f(u)$. Let us first ensure that the problem (P_u) has a solution.

Theorem 3.10 (Existence of solutions for the subproblems)**.** *Given* $z_{k+1} \in Z$, $u_k \in H_0^1(\Omega)$ *and* $\lambda_k > 0$ *the problem* (P_u) *has a unique solution.*

Proof. The proof follows similar arguments as the proof of 3.2. The uniqueness is a consequence of the λ_k-strong convexity of $g_{\lambda_k}(\cdot, u_k)$ for every $k \in \mathbb{N}$. □

Let us establish some classical estimates for the model function defined in (23).

Theorem 3.11 (Fundamental inequalities for the model-function.)**.** *Let 3.1 hold true. Then there are* $L_1, L_2 > 0$ *such that*

$$\|\nabla f(u) - \nabla f(v)\|_{H^{-1}(\Omega)} \leq L_1 \|u - v\|_{H_0^1(\Omega)}, \tag{24}$$

$$|g(u, v) - f(u)| \leq \frac{L_2}{2}\|u - v\|_{H_0^1(\Omega)}^2, \tag{25}$$

for every pair $u, v \in U_{ad}$.

Proof. We prove (24) first. For this purpose let $u, v \in U_{ad}$ and recall that

$$\nabla f(u) = F'(u)^*(F(u) - f^\delta) \in H^{-1}(\Omega).$$

By definition, we obtain for $h \in H_0^1(\Omega)$ and $u, v \in U_{ad}$

$$\begin{aligned} |\langle \nabla f(u) - \nabla f(v), h\rangle_{H^{-1}(\Omega)}| &= |\langle F(u) - f^\delta, F'(u)[h]\rangle_{L^2(\Omega)} - \langle F(v) - f^\delta, F'(v)[h]\rangle_{L^2(\Omega)}| \\ &\leq (\|F(u)\|_{L^2(\Omega)} + \|f^\delta\|_{L^2(\Omega)})\|(F'(u) - F'(v))[h]\|_{L^2(\Omega)} \\ &\quad + \|F(u) - F(v)\|_{L^2(\Omega)}\|F'(u)[h]\|_{L^2(\Omega)}. \end{aligned} \tag{26}$$

It is easy to see, using the L^∞-boundedness of U_{ad}, that $\|F(u)\|_{L^2(\Omega)} + \|f^\delta\|_{L^2(\Omega)} \leq C_1$ for some $C_1 > 0$. Furthermore we obtain, by invoking the continuity of the sampling operator, that

$$\|F(u) - F(v)\|^2_{L^2(\Omega)} \leq \int_\Omega \|\pi(u) - \pi(v)\|^2 \,\mathrm{d}x \leq C_2\|u - v\|^2_{L^2(\Omega)},$$

for some generic constant $C_2 > 0$, see also 2.3. Using a similar argumentation as in 2.4 and again the L^∞ boundedness of U_{ad}, we also find a constant $C_3 > 0$ with

$$\|F'(u)[h]\|^2_{L^2(\Omega)} \leq \int_\Omega \|\pi'(u)\|^2_{\mathcal{L}(\mathbb{R}^3,\mathbb{C}^L)}\|h\|_2^2 \,\mathrm{d}x \leq C_3\|h\|^2_{H_0^1(\Omega)}.$$

Moreover, based on the L_b-Lipschitz-continuity of π we conclude

$$\begin{aligned} \|F'(u) - F'(v)[h]\|^2_{L^2(\Omega)} &\leq \int_\Omega \|(D\pi(u) - D\pi(v))[h]\|^2 \,\mathrm{d}x \\ &\leq L_b^2 \int_\Omega \|u - v\|^2\|h\|^2 \,\mathrm{d}x \\ &\leq L_b^2\|u - v\|^2_{L^4(\Omega)}\|h\|^2_{L^4(\Omega)}. \end{aligned}$$

After involving suitable embedding constants, we obtain from (26) that (24) holds true.

Let us now prove (25). For this purpose, we let $w_1, w_2 \in L^2(\Omega)$ arbitrary and observe

$$\left|\frac{1}{2}\|w_1 - f^\delta\|^2_{L^2(\Omega)} - \frac{1}{2}\|w_2 - f^\delta\|^2_{L^2(\Omega)}\right| \leq \left|\langle w_2 - f^\delta, w_1 - w_2\rangle_{L^2(\Omega)}\right| + \frac{1}{2}\|w_1 - w_2\|^2_{L^2(\Omega)}$$

Taking $w_2 = F(v) \in L^2(\Omega)$ and $w_1 = F'(u)[v - u] + F(u) \in L^2(\Omega)$ we deduce

$$\left|\|F(u) + F'(u)[v - u] - f^\delta\|^2_{L^2(\Omega)} - \|F(v) - f^\delta\|^2_{L^2(\Omega)}\right| \leq \|F(v) - f^\delta\|_{L^2(\Omega)}B + B^2,$$

with $B := \|F(v) - F(u) - F'(u)[v - u]\|_{L^2(\Omega)}$. We estimate B from above using the inequality

$$\|F(v) - F(u) - F'(u)[v - u]\|_{L^2(\Omega)} \leq C\|v - u\|^2_{L^4(\Omega)},$$

for some $C > 0$. The latter follows from the corresponding property of π, namely

$$\|\pi(v) - \pi(u) - \pi'(u)[v - u]\|_2 \leq \int_0^1 \|\pi'(u + \tau(v - u)) - \pi'(u))\|_2\|v - u\|_2 \,\mathrm{d}\tau.$$

Since $v, u \leq b$ almost everywhere, by definition of U_{ad}, π' is Lipschitz on U_{ad}. Hence we deduce

$$B^2 \leq \int_\Omega \|\pi(v) - \pi(u) - \pi'(u)[v - u]\|_2^2 \,\mathrm{d}x \leq \frac{C_4}{2}\int_\Omega \|v - u\|_2^4 \,\mathrm{d}x = \frac{C_4}{2}\|u - v\|^4_{L^4(\Omega)},$$

for some $C_4 > 0$ that depends on b. This shows (3.1). Consequently we estimate

$$\begin{aligned} \left|\|F(u) + F'(u)[v - u] - f^\delta\|^2_{L^2(\Omega)} - \|F(v) - f^\delta\|_{L^2(\Omega)}\right| &\leq \frac{C_4^{1/2}}{2}\|F(v) - f^\delta\|_{L^2}\|v - u\|^2_{L^4(\Omega)} \\ &\quad + \frac{C_4}{2}\|v - u\|^4_{L^4(\Omega)}. \end{aligned} \tag{27}$$

Note that again $\|u-v\|_{L^4(\Omega)} \leq C_5(\|v\|_{L^\infty(\Omega)} + \|u\|_{L^\infty(\Omega)}) \leq 2C_5 b$ for some constant $C_5 > 0$ and that $\|F(v) - f^\delta\|_{L^2(\Omega)} \leq C_6$ for $v \in U_{ad}$ by the same argumentation. Combining these observations with the continuous embedding $H_0^1(\Omega) \hookrightarrow L^4(\Omega)$ in (27), we obtain a constant $C > 0$ with

$$\left| \|F(u) + F'(u)[v-u] - f^\delta\|^2_{L^2(\Omega)} - \|F(v) - f^\delta\|_{L^2(\Omega)} \right| \leq C\|v-u\|^2_{H_0^1(\Omega)}.$$

Thus the proof is finished. □

From the previous result in 3.11 we directly obtain the following majorization property:

$$f(u) \leq g(u, u_k) + \frac{L_2}{2}\|u - u_k\|^2_{H_0^1(\Omega)} \quad \text{for all } u \in U_{ad}.$$

Given a positive descent constant $C_{desc} > 0$, we may then chose the step-size $\lambda_k \geq L_2 + C_{desc}$ and deduce

$$\begin{aligned} J(u_{k+1}, z_{k+1}) &\leq J(u_k, z_{k+1}) + \frac{L_2 - \lambda_k}{2}\|u_k - u_{k+1}\|^2_{H_0^1(\Omega)} \\ &\leq J(u_k, z_{k+1}) - \frac{C_{\text{desc}}}{2}\|u_k - u_{k+1}\|^2_{H_0^1(\Omega)}, \end{aligned} \tag{28}$$

where we also used the minimization property of $u_{k+1} \in U_{ad}$.

The overall algorithm. Let us now describe the overall algorithm to find stationary points of Problem (P_1).

Algorithm 2 Computation of a stationary point of (P_1).

1: Get initial values $(u_0, z_0) \in H_0^1(\Omega) \times Z$, stopping accuracy $\varepsilon_1, \varepsilon_2 > 0$, accuracies of the nested subroutine $(\eta_k)_{k\in\mathbb{N}}$ and descent parameter $C_{\text{decs}} > 0$.
2: Set k = 0.
3: **while** no stopping criterion is satisfied, **do**
4: **z-step**: Given $(u_k, z_k) \in H_0^1(\Omega) \times Z$ and accuracy $\eta_k > 0$, compute $z_{k+1} \in Z$
5: using the abstract descent algorithm 1.
6: **u-step**: Compute a global solution $\widehat{u}(\lambda_k) \in U_{ad}$ of

$$\min_{u\in U_{ad}} g_{\lambda_k}(u, u_k) + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u, z) + R_1(u), \tag{29}$$

7: that satisfies the sufficient descent condition

$$J(\widehat{u}(\lambda_k), z_{k+1}) \leq J(u_k, z_{k+1}) - \frac{C_{\text{desc}}}{2}\|\widehat{u}(\lambda_k) - u_k\|^2_{H_0^1(\Omega)}. \tag{30}$$

8: Set $u_{k+1} = \widehat{u}(\lambda_k)$.
9: Set $k = k + 1$.
10: **if** $\|u_k - u_{k-1}\|_{H_0^1(\Omega)} < \varepsilon_1$ and $\|z_k - z_{k-1}\|_Z < \varepsilon_2$ **then**
11: stop the while loop.
12: **end if**
13: **end while**
14: Return (u_k, z_k) as reconstruction.

The algorithm above only accepts the next iterate $u(\lambda_k) \in U_{ad}$ in the u-step (29), if the descent condition (74) is satisfied. In case $L_2 > 0$ from 3.11 is known in advance, then λ_k can be chosen such that $\lambda_k \geq L_2 + C_{desc}$ to ensure that this descent property holds true. However in practice this constant is rarely known which is why we have to rely on an adaptive backtracking strategy in order to find a suitable λ_k. This strategy is described in the following algorithm.

Algorithm 3 Backtracking search for algorithm 2

1: Get $\sigma_3 \in (0,1), \tau > 1$ and $\lambda_0 > 0$
2: Set $\lambda = \lambda_0$ and solve (29) with parameter $\lambda > 0$ to obtain $u(\lambda)$.
3: **while** $J(u(\lambda), z_{k+1}) > J(u_k, z_{k+1}) - \frac{\lambda\sigma}{2}\|u(\lambda) - u_k\|^2_{H^1_0(\Omega)}$ **do**
4: Set $\lambda \leftarrow \tau\lambda$.
5: Recompute $u = u(\lambda)$ by solving (29) with the new $\lambda > 0$.
6: **end while**
7: Return $\widehat{u} = u(\lambda)$ and step size $\lambda_k := \lambda$.

Let us briefly argue that the line search strategy terminates after finitely many iterations.

Lemma 3.12. *Consider the algorithm 2 above together with the line search strategy algorithm 3 at the k-th iterate. Given u_k and the line search parameters (σ, λ_0), the iterates of the backtracking algorithm converge after at most $j_* = \log_\tau(L_2/(1-\sigma)\lambda_0)$ iterations. In particular the sequence of step sizes satisfies $C_{desc} := \sigma\lambda_0 \leq \lambda_k \leq \lambda_0\tau^{j_*}$ and is bounded in k.*

Proof. The proof is standard but we repeat it here for the convenience of the reader. Given a current step size parameter $\lambda > 0$ consider the unique solution $u(\lambda) \in H^1_0(\Omega)$ of (29). Using 3.11 we directly obtain

$$J(u(\lambda), z_{k+1}) \leq J(u_k, z_{k+1}) - \left(\frac{\lambda - L_2}{2}\right)\|u(\lambda) - u_k\|^2_{H^1_0(\Omega)}$$

with the constant $L_2 > 0$ from 3.11. Hence, whenever $\sigma\lambda \leq (\lambda - L_2)$ we also have

$$J(u(\lambda), z_{k+1}) \leq J(u_k, z_{k+1}) - \left(\frac{\sigma\lambda}{2}\right)\|u(\lambda) - u_k\|^2_{H^1_0(\Omega)},$$

which is the condition for acceptance in the backtracking algorithm above. Thus when initialized with λ_0 then the line-search stops if $\sigma\tau^j\lambda_0 \leq (\tau^j\lambda_0 - L_2)$ which happens after at most $j = \log_\tau(L_2/(1-\sigma)\lambda_0)$ steps. □

Let us now state a basic global convergence result for algorithm 2. We are now able to show the global sublinear convergence result.

Theorem 3.13 (Global convergence to stationarity)**.** *Let 3.1 and 3.7 hold true and let $(u_k, z_k)_{k\in\mathbb{N}} \subset H^1_0(\Omega) \times Z$ be a sequence that is generated by algorithm 2 with step sizes $\lambda_k \geq C_{\text{desc}} + L_2 > 0$ for $k \in \mathbb{N}$. In addition assume that the sequence of accuracy parameters are square summable, i.e. $\sum_k \eta_k^2 < +\infty$. Then the following holds*

(i) *The sequence $(u_k, z_k)_{k\in\mathbb{N}}$ is bounded in $H^1_0(\Omega) \times Z$.*
(ii) *The function values converge monotonically to its infimum and the lazy slope converges at globally sub linear rate, i.e. there is a constant $C > 0$ such that*

$$\min_{k=1,\dots,N} \operatorname{dist}(0, \partial J(u_k, z_k)) \leq C\sqrt{\frac{(J(u_0, z_0) - J(u_N, z_N)) + \sum_{k=0}^{\infty}\eta_k^2}{N}},$$

Proof. We start with (i). As in the remark after 3.11 we obtain

$$J(u_{k+1}, z_{k+1}) \leq J(u_k, z_{k+1}) - \frac{C_{desc}}{2}\|u_{k+1} - u_k\|^2_{H^1_0(\Omega)}$$

for every $k \in \mathbb{N}$. Taking into account the descent property of algorithm 1, namely

$$J(u_k, z_{k+1}) \leq J(u_k, z_k) - \frac{\sigma_1}{2}\sum_{i=0}^{(n_k-1)}\|z_k^{i+1} - z_k^i\|^2_Z$$

we directly obtain

$$J(u_{k+1}, z_{k+1}) \leq J(u_k, z_k) - \frac{C_{\text{desc}}}{2}\|u_{k+1} - u_k\|^2_{H^1_0(\Omega)}.$$

Consequently, since J is bounded from below, the sequence $(J(u_k, z_k))_k$ converges monotonically to its infimum. By the same arguments as in the existence proof 3.2, we conclude the boundedness of $(u_k, z_k)_k$ in $H^1_0(\Omega) \times Z$. In order to show (ii) we deduce from 3.11, that

$$\begin{aligned} e_{\lambda_k}(u_{k+1}, u_k) &= g_{\lambda_k}(u, u_k) - f(u) \\ &= \frac{1}{2}\|F'(u_k)[u - u_k] + F(u_k) - f^\delta\|^2_{L^2(\Omega)} + \frac{\lambda_k}{2}\|u - u_k\|^2_{H^1_0(\Omega)} - \frac{1}{2}\|F(u) - f^\delta\|^2_{L^2(\Omega)} \\ &\geq \frac{\lambda_k - L_2}{2}\|u - u_k\|^2_{H^1_0(\Omega)} \geq \frac{C_{desc}}{2}\|u - u_k\|^2_{H^1_0(\Omega)}, \end{aligned}$$

where $L_2 > 0$ is the constant from 3.11 and $C_{\text{desc}} > 0$ is the given descent constant. We now invoke the optimality system for u_{k+1} given z_{k+1}. This yields

$$0 \in \nabla_1 g(u_{k+1}, u_k) - \alpha\Delta u_{k+1} + \nabla_u h(u_{k+1}, z_{k+1}) + \lambda_k(u_{k+1} - u_k) + \partial R_1(u_{k+1}), \tag{31}$$

which follows from standard (convex) subdifferential calculus in the space $H^{-1}(\Omega)$. We set for abbreviation

$$A_k := F'(u_{k+1})^*[F(u_{k+1}) - f^\delta] - F'(u_k)^*(F'(u_k)[u_{k+1} - u_k] + F(u_k) - f^\delta) - \lambda_k(u_{k+1} - u_k) \in H^{-1}(\Omega)$$

and rewrite (31) as

$$A_k \in \nabla f(u_{k+1}) - \alpha\Delta u_{k+1} + \nabla_u h(u_{k+1}, z_{k+1}) + \partial R_1(u_{k+1}) = \partial_u J(u_{k+1}, z_{k+1}).$$

Using 3.11 and the chainrule it is not difficult to see that

$$\|A_k\|_{H^{-1}(\Omega)} \leq C_1\|u_{k+1} - u_k\|_{H^1_0(\Omega)}$$

for some $C_1 > 0$. For the z-step we obtain for arbitrary $w \in \partial R_2(z_{k+1})$

$$\begin{aligned} \|\nabla_z h(u_{k+1}, z_{k+1}) + w\|_Z &= \|\nabla_z h(u_{k+1}, z_{k+1}) - \nabla_z h(u_k, z_{k+1})\|_Z + \|\nabla_z h(u_k, z_{k+1}) + w\|_Z. \\ &\leq C_2\|u_{k+1} - u_k\|_{H^1_0(\Omega)} + \|\nabla_z h(u_k, z_{k+1}) + w\|_Z, \end{aligned}$$

where we used 3.1 H3. in the last inequality to find the constant $C_2 > 0$. We further observe by the minimization property, that

$$\begin{aligned} e_{\lambda_k}(u_{k+1}, u_k) &= g_{\lambda_k}(u_{k+1}, u_k) - f(u_{k+1}) \\ &\leq f(u_k) + \frac{\alpha}{2}\|\nabla u_k\|^2_{L^2(\Omega)} + h(u_k, z_{k+1}) + R_2(z_{k+1}) + R_1(u_k) \\ &\quad - \frac{\alpha}{2}\|\nabla u_{k+1}\|^2_{L^2(\Omega)} - h(u_{k+1}, z_{k+1}) - R_1(u_{k+1}) - R_2(z_{k+1}) - f(u_{k+1}). \end{aligned}$$

Using the fact that $h(u_k, z_{k+1}) + R_2(z_{k+1}) \leq h(u_k, z_k) + R_2(z_k)$ by the property (20), we infer

$$e_{\lambda_k}(u_{k+1}, u_k) \leq J(u_k, z_k) - J(u_{k+1}, z_{k+1}).$$

Consequently, from all previous inequalities, we deduce for every $k \in \mathbb{N}$

$$\begin{aligned} \|A_k\|^2_{H^{-1}(\Omega)} + \|\nabla_z h(u_{k+1}, z_{k+1}) + w\|^2_Z &\leq (C_1^2 + 2C_2^2)\|u_{k+1} - u_k\|^2_{H^1_0} + 2\|\nabla_z h(u_k, z_{k+1}) + w\|^2_Z \\ &\leq \frac{2(C_1^2 + 2C_2^2)}{C_{\text{desc}}} e_{\lambda_k}(u_{k+1}, u_k) + 2\|\nabla_z h(u_k, z_{k+1}) + w\|^2_Z \\ &\leq \frac{2(C_1^2 + 2C_2^2)}{C_{\text{desc}}} \left(J(u_k, z_k) - J(u_{k+1}, z_{k+1})\right) \\ &\quad + 2\|\nabla_z h(u_k, z_{k+1}) + w\|^2_Z. \end{aligned} \tag{32}$$

By the sum-rule for the convex subdifferential (cf. [9]) and 3.4 including the remark below, we conclude that

$$(A_k, \nabla_z h(u_{k+1}, z_{k+1}) + w) \in \partial J(u_{k+1}, z_{k+1})$$

and therefore, by definition of the lazy slope, also

$$\operatorname{dist}(0, \partial J(u_{k+1}, z_{k+1}))^2 \leq \frac{2(C_1^2 + 2C_2^2)}{C_{\text{desc}}} (J(u_k, z_k) - J(u_{k+1}, z_{k+1})) + 2\|\nabla_z h(u_k, z_{k+1}) + w\|_Z^2.$$

Taking the infimum over all $w \in \partial R_2(z_{k+1})$ we deduce

$$\operatorname{dist}(0, \partial J(u_{k+1}, z_{k+1}))^2 \leq \frac{2(C_1^2 + 2C_2^2)}{C_{\text{desc}}} (J(u_k, z_k) - J(u_{k+1}, z_{k+1})) + 2\eta_k.$$

Summing from 0 to $N-1$ in turn leads to

$$\sum_{k=0}^{N-1} \operatorname{dist}(0, \partial J(u_{k+1}, z_{k+1}))^2 \leq \frac{2(C_1^2 + 2C_2^2)}{C_{\text{desc}}} (J(u_0, z_0) - J(u_N, z_N)) + 2\sum_{k=0}^{N-1} \eta_k^2.$$

Consequently, with $C := \sqrt{\max(2, 2(C_1^2 + C_2^2)/C_{\text{desc}})} > 0$, we obtain

$$\min_{k=1,\dots,N} \operatorname{dist}(0, \partial J(u_k, z_k)) \leq C \sqrt{\frac{(J(u_0, z_0) - J(u_N, z_N)) + \sum_{k=0}^{N-1} \eta_k^2}{N}}, \tag{33}$$

which proves the assertion in (ii). □

3.2. **Local convergence analysis under KL-inequality.** In this section we investigate local strong convergence of the overall algorithm 2. We are particularly interested in conditions, under which we can maintain fast local convergence of the Levenberg-Marquardt algorithm. For this purpose we need slightly more restrictive assumptions, compared to the requirements in the previous subsection.

Assumption 3.14 (For the local convergence analysis)**.** *Let $(x_k)_k = (u_k, z_k)_k$ denote the sequence generated by algorithm 2. As $(x_k)_k$ is bounded, there exists a weak limit point $(u^*, z^*) \in H_0^1(\Omega) \times Z$. We assume throughout this section the following conditions:*

C1. *The operator $F : H_0^1(\Omega) \to L^2(\Omega)$ is twice continuously differentiable. This is true for the qMRI problem, for instance, if $n \leq 3$, see 2.4.*

C2. *For any weak accumulation point (u^*, z^*), there is a $\kappa > 0$ and a stationary point $u^*(z^*) \in H_0^1(\Omega)$ of $J(\cdot, z^*)$, such that*

$$\langle \nabla^2 f^\alpha(u^*(z^*))[h], h\rangle_{H^{-1}(\Omega)} \geq \frac{\kappa}{2}\|h\|_{H_0^1(\Omega)}^2 \quad \text{for every } h \in H_0^1(\Omega). \tag{34}$$

C3. *For any fixed given $u^* \in H_0^1(\Omega)$, the function $z \mapsto g(z) = h(u^*, z) + R_2(z)$ satisfies the KL-inequality with exponent $\beta = 1 - 1/q$, $q > 2$ at z^*, i.e. there are constants $\eta, C_{KL} > 0$ and a radius $\varepsilon_{KL} > 0$ such that*

$$g(z) - g(z^*) \leq C_{KL} \operatorname{dist}(0, \partial_z g(z))^q \tag{35}$$

whenever $g(z^*) < g(z) < g(z^*) + \eta$ and $z \in B_{\varepsilon_{KL}}(z^*)$. Note that here, the quantities $q, \varepsilon_{KL}, C_{KL}, \eta$ depend on u^*.

Remark 3.15 (General KL-inequality)**.** *The KL-inequaltity with a given exponent is only a special case of the general KL inequality which is investigated in* [6, 7] *and has the form*

$$\varphi'(g(z) - g(z^*)) \operatorname{dist}(0, \partial_u g(z)) \geq 0. \tag{36}$$

For a certain concave C^1-function $\varphi : (0, \eta) \to \mathbb{R}_+$, also called desingularization function. *We directly see that we obtain (35) from (36) by using $\varphi(t) = C_{KL}^{1-\beta} t^\beta$. We will not go further into details here and refer to [7] for details.*

Remark 3.16 (Discussion of 3.14.)**.** *The additional conditions, in particular C2. are necessary in order to apply the generalized implicit function theorem by Robinson, [26, 38], which is important to get a local Lipschitz continuous parameterization $z \mapsto u^*(z)$ from z to a stationary point $u^*(z)$ around $z^* \in Z$. The assumption C2. has become a standard tool in finite dimensional optimization during the recent years, cf. [6, 7, 15]. The KL-inequality can be verified for a large number of functions using the theory of real algebraic functions or more generally for functions definable in an o-minimal structure, cf. [7, 13, 47] and the references therein. Here a large KL-exponent $\beta \in (0, 1]$ (corresponding to a $q \to \infty$) is usually connected to fast local convergence behaviour as it increases the sharpness around the stationary set. Using the aforementioned toolboxes from real-algebraic geometry in finite dimensions, it is often easy to infer the existence of some KL-exponent $\beta \in (0, 1)$. However the computation of a concrete β for a certain objective is usually very difficult. We refer to [41, 61] for an overview on existing approaches. Extensions to infinite dimensions are discussed in [20, 13, 35] with applications to gradient-flow problems. In the infinite dimensional setup, the computation of KL-exponents is even more delicate, as the finite dimensional toolbox, using the theory of real algebraic geometry is to the best of the authors knowledge not available.*

Note that by the assumptions we have that $u^* \in U_{ad}$ satisfies a quadratic growth property around u^* in u-direction. However, for the subsequent part, we need uniform growth properties, for which the following lemma forms the basis.

Lemma 3.17 (Uniform Taylor bound)**.** *Let $u^* \in H_0^1(\Omega)$ such that*

$$\langle \nabla f^\alpha(u^*)[h], h \rangle_{H^{-1}(\Omega)} \geq \frac{\kappa}{2} \|h\|^2_{H_0^1(\Omega)} \quad \text{for all } h \in H_0^1(\Omega).$$

Then there is an $\varepsilon > 0$ such that

$$f^\alpha(u) \geq f^\alpha(u_0) + \langle \nabla f^\alpha(u_0), u - u_0 \rangle_{H^{-1}(\Omega)} + \frac{\kappa}{8} \|u - u_0\|^2_{H_0^1(\Omega)},$$

whenever $\|u_0 - u^*\|_{H_0^1(\Omega)} < \varepsilon$ and $\|u_0 - u\|_{H_0^1(\Omega)} < \varepsilon$.

Proof. The proof is a consequence of Taylors formula with integral remainder, i.e.

$$\begin{aligned} f^\alpha(u) = f^\alpha(u_0) + \langle \nabla f^\alpha(u_0), u - u_0 \rangle_{H^{-1}(\Omega)} + \frac{1}{2} \langle \nabla^2 f^\alpha(u_0)[u - u_0], u - u_0 \rangle_{H^{-1}(\Omega)} \\ + \int_0^1 (1 - \tau) \langle [\nabla^2 f^\alpha(u_0 + \tau(u - u_0)) - \nabla^2 f^\alpha(u_0)][u - u_0], u - u_0 \rangle_{H^{-1}(\Omega)} \, \mathrm{d}\tau. \end{aligned}$$

Adding and subtracting $\langle \nabla^2 f^\alpha(u^*)[u-u_0], u-u_0\rangle_{H^{-1}(\Omega)}$ and using the assumptions, we deduce

$$\begin{aligned} f^\alpha(u) &= f^\alpha(u_0) + \langle \nabla f^\alpha(u_0), u-u_0\rangle_{H^{-1}(\Omega)} + \frac{\kappa}{4}\|u-u_0\|^2_{H_0^1(\Omega)} \\ &\quad + \frac{1}{2}\langle [\nabla^2 f^\alpha(u_0) - \nabla^2 f^\alpha(u^*)][u-u_0], u-u_0\rangle_{H^{-1}(\Omega)} \\ &\quad + \int_0^1 (1-\tau)\langle [\nabla^2 f^\alpha(u_0+\tau(u-u_0)) - \nabla^2 f^\alpha(u_0)][u-u_0], u-u_0\rangle_{H^{-1}(\Omega)}\,\mathrm{d}\tau \\ &\geq f^\alpha(u_0) + \langle \nabla f^\alpha(u_0), u-u_0\rangle_{H^{-1}(\Omega)} + \frac{\kappa}{4}\|u-u_0\|^2_{H_0^1(\Omega)} \\ &\quad - \frac{1}{2}\|\nabla^2 f^\alpha(u_0) - \nabla^2 f^\alpha(u^*)\|_{\mathcal{L}(H_0^1(\Omega),H^{-1}(\Omega))}\|u-u_0\|^2_{H_0^1(\Omega)} \\ &\quad - \int_0^1 (1-\tau)\|\nabla^2 f^\alpha(u_0+\tau(u-u_0)) - \nabla^2 f^\alpha(u_0)\|_{\mathcal{L}(H_0^1(\Omega),H^{-1}(\Omega))}\|u-u_0\|^2_{H_0^1(\Omega)}\,\mathrm{d}\tau. \end{aligned} \tag{37}$$

As $f^\alpha : H_0^1(\Omega) \to \mathbb{R}$ is two times continuously differentiable, we may take $\varepsilon_1 > 0$, such that

$$\|\nabla^2 f^\alpha(v_1) - \nabla^2 f^\alpha(v_2)\|_{\mathcal{L}(H_0^1(\Omega),H^{-1}(\Omega))} \leq \frac{\kappa}{8} \tag{38}$$

for all $v_1, v_2 \in H_0^1(\Omega)$ such that $\|v_i - u^*\|_{H_0^1(\Omega)} < \varepsilon_1$ for $i = 1, 2$. Hence for $\varepsilon := \varepsilon_1/2$, we obtain for $\|u-u_0\|_{H_0^1(\Omega)} < \varepsilon$ and $\|u_0 - u^*\|_{H_0^1(\Omega)} < \varepsilon$ that obviously $\|u_0 - u^*\|_{H_0^1(\Omega)} < \varepsilon < \varepsilon_1$ and

$$\|u_0 + \tau(u-u_0) - u^*\|_{H_0^1(\Omega)} \leq \|u-u_0\|_{H_0^1(\Omega)} + \|u_0-u^*\|_{H_0^1(\Omega)} < \varepsilon + \varepsilon = \varepsilon_1$$

for $\tau \in (0,1)$. Hence, from (37) and(38), we obtain

$$f^\alpha(u) \geq f^\alpha(u_0) + \langle \nabla f^\alpha(u_0), u-u_0\rangle_{H^{-1}(\Omega)} + \left(\frac{\kappa}{4} - \frac{\kappa}{16} - \frac{\kappa}{16}\right)\|u-u_0\|^2_{H_0^1(\Omega)},$$

which yields the desired estimate. □

We will now show that the set-valued map, which sends z to the set of stationary points of $J(\cdot, z)$ is indeed a single valued and Lipschitz continuous map.

Lemma 3.18 (Implication Robinsons implicit function theorem)**.** *Let 3.14 hold true at a weak accumulation point $(u^*, z^*) \in H_0^1(\Omega) \times Z$ and denote again by $u^*(z^*) \in H_0^1(\Omega)$ a stationary point of $J(\cdot, z^*)$ for which (34) holds true. Then there is an $\varepsilon > 0$ such that the solution mapping*

$$\omega : Z \rightrightarrows H_0^1(\Omega) \quad \omega(z) = \partial_u J(\cdot, z)^{-1}(0)$$

admits a single-valued localization $S : Z \to H^1(\Omega)$ which satisfies $S(z^*) = u^*(z^*)$ and which is L_{z^*}-Lipschitz continuous in $B_\varepsilon(z^*)$ for some $L_{z^*} > 0$. Moreover, whenever $\|z^* - z_{k+1}\|_Z < \varepsilon$ the following inequalities hold for a constant $C_1 > 0$ independend of $k \in \mathbb{N}$:

$$\|S(z_{k+1}) - u_k\|_{H_0^1(\Omega)} \leq C_1\|u_{k+1} - u_k\|_{H_0^1(\Omega)}, \tag{39}$$

$$J(S(z_{k+1}), z_{k+1}) + \frac{\kappa}{8}\|S(z_{k+1}) - S(z^*)\|^2_{H_0^1(\Omega)} \leq J(S(z^*), z_{k+1}), \tag{40}$$

$$\|S(z_{k+1}) - S(z^*)\|_{H_0^1(\Omega)} \leq L_{z^*}\|z_{k+1} - z^*\|_Z. \tag{41}$$

Proof. For the proof we will summarize both convex terms into one $\tilde{R}_1(u) = R_1(u) + \mathcal{I}_{U_{ad}}(u)$ and use the notation R_1 for both. By assumption, $u^*(z^*) \in \omega(z^*)$ or equivalently, $u^*(z^*)$ solves the generalized equation

$$0 \in G(u, z^*) + \partial R_1(u) := \nabla f(u) - \alpha\Delta u + \nabla_u h(u, z^*) + \partial R_1(u),$$

posed in $H^{-1}(\Omega)$. We want to apply the generalized implicit function by Robinson in the version [25], Theorem 8.5, see also [26]. For this purpose consider an approximated generalized equation around $z^* \in Z$, namely

$$0 \in G_{z^*}(u) + \partial R_1(u),$$

with $G_{z^*} : H_0^1(\Omega) \to H^{-1}(\Omega)$ being defined as

$$G_{z^*}(u) := \nabla f(u^*(z^*)) + \nabla^2 f(u^*(z^*))[u - u^*(z^*)] - \alpha\Delta u + \nabla_u h(u, z^*).$$

In order to apply the generalized implicit function theorem, we need to show the strong regularity of the linearized equation. For this purpose, it is sufficient to verify that the solution map of this equation, i.e.

$$S_{z^*} : H^{-1}(\Omega) \to H_0^1(\Omega) \quad S_{z^*}(p) = (G_{z^*} + \partial R_1)^{-1}(p),$$

has a Lipschitz-continuous single valued localization around $p = 0$. This follows from the fact that $p \in G_{z^*}(u) + \partial R_1(u)$ if and only if

$$S_{z^*}(p) \in \operatorname*{arg\,min}_{u \in H_0^1(\Omega)} \langle \nabla^2 f(u^*(z^*))[u-u^*(z^*)], u-u^*(z^*)\rangle_{H^{-1}(\Omega)} + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u, z^*) - \langle p, u\rangle_{H^{-1}(\Omega)} + R_1(u).$$

As this problem is κ-strongly convex, with $\kappa > 0$ from 3.14, it is a standard result that the associated solution mapping S_{z^*} is single-valued and $1/\kappa$ -Lipschitz continuous. Moreover, we note that the approximation error

$$e(u, z) := G_{z^*}(u) - G(u, z) = \nabla f(u) - \nabla f(u^*(z^*)) - \nabla^2 f(u^*(z^*))[u - u^*(z^*)],$$

does not depend on z, which implies for the partial uniform Lipschitz-modulus $\widehat{\operatorname{Lip}}_{u^*}(G_{z^*} - G; (u^*(z^*), z^*))$ as introduced in [25, p.8], that

$$\begin{aligned}\widehat{\operatorname{Lip}}_{u^*}(G_{z^*} - G; (u^*(z^*), z^*)) &= \limsup_{\substack{u_1,u_2 \to u^*(z^*) \\ z \to z^*}} \frac{\|e(u_1, z) - e(u_2, z)\|_{H^{-1}(\Omega)}}{\|u_1 - u_2\|_{H_0^1(\Omega)}} \\ &= \limsup_{u_1,u_2 \to u^*(z^*)} \frac{\|\nabla f(u_1) - \nabla f(u_2) - \nabla^2 f(u^*(z^*))[u_1 - u_2]\|_{H^{-1}(\Omega)}}{\|u_1 - u_2\|_{H_0^1(\Omega)}}.\end{aligned}$$

Using the mean value theorem, we obtain

$$\begin{aligned}&\|\nabla f(u_1) - \nabla f(u_2) - \nabla^2 f(u^*(z^*))[u_1 - u_2]\|_{H^{-1}(\Omega)} \\ &\qquad \le \left(\int_0^1 \|\nabla^2 f(u_2 + \tau(u_1 - u_2)) - \nabla^2 f(u^*(z^*))\|_{H^{-1}(\Omega)}\, \mathrm{d}\tau\right) \|u_1 - u_2\|_{H_0^1(\Omega)},\end{aligned}$$

and henceforth

$$\begin{aligned}\widehat{\operatorname{Lip}}_{u^*}(G_{z^*} - G; (u^*(z^*), z^*)) &= \limsup_{u_1,u_2 \to u^*(z^*)} \frac{\|\nabla f(u_1) - \nabla f(u_2) - \nabla^2 f(u^*(z^*))[u_1 - u_2]\|_{H^{-1}(\Omega)}}{\|u_1 - u_2\|_{H_0^1(\Omega)}}. \\ &\le \limsup_{u_1,u_2 \to u^*(z^*)} \left(\int_0^1 \|\nabla^2 f(u_2 + \tau(u_1 - u_2)) - \nabla^2 f(u^*(z^*))\|_{H^{-1}(\Omega)}\, \mathrm{d}x\right) \to 0,\end{aligned}$$

as $u_1, u_2 \to u^*(z^*)$. Hence, by [25, Theorem 8.5], the mapping $\omega : Z \rightrightarrows H_0^1(\Omega)$ has a single valued Lipschitz continuous localization $S : Z \to H_0^1(\Omega)$ in a ball $B_\varepsilon(z^*)$ with a Lipschitz-constant $0 < L_{z^*} \lesssim 1/\kappa$. Thus, the first part of 3.18 is shown. Now, consider the first order optimality condition for deriving u_{k+1} given $z = z_{k+1}$, as in (P_u). In this way we obtain the following variational inequality:

(42) $$R_1(u) \ge -\langle \nabla f(u_{k+1}) - \alpha\Delta u_{k+1} + \nabla_u h(u_{k+1}, z_{k+1}) - q_k, u - u_{k+1}\rangle_{H^{-1}(\Omega)} + R_1(u_{k+1}),$$

for every $u \in H_0^1(\Omega)$, where we defined $q_k \in H^{-1}(\Omega)$ as

$$q_k := F'(u_{k+1})^*[F(u_{k+1}) - f^\delta] - F'(u_k)^*(F'(u_k)[u_{k+1} - u_k]) - \lambda_k(u_{k+1} - u_k)$$

viewed as an element in $H^{-1}(\Omega)$. Considering as well the first order optimality conditions for the stationary point $S(z_{k+1}) \in H_0^1(\Omega)$, we obtain again a variational inequality of the second kind,

(43)
$$R_1(u) \geq -\langle \nabla f(S(z_{k+1})) - \alpha \Delta S(z_{k+1}) + \nabla_u h(S(z_{k+1}), z_{k+1}), u - S(z_{k+1}) \rangle + R_1(S(z_{k+1})),$$

for all $u \in H_0^1(\Omega)$. Adding (42) with $u = S(z_{k+1})$ and (43) with $u = u_{k+1}$ yields

$$\begin{aligned} 0 \geq\ & \langle \nabla f(S(z_{k+1})) - \nabla f(u_k), S(z_{k+1}) - u_{k+1} \rangle_{H^{-1}(\Omega)} + \alpha \|\nabla(S(z_{k+1}) - u_{k+1})\|^2_{L^2(\Omega)} \\ & + \langle \nabla_u h(S(z_{k+1}), z_{k+1}) - \nabla_u h(u_{k+1}, z_{k+1}), S(z_{k+1}) - u_{k+1} \rangle_{H^{-1}(\Omega)} - \langle q_k, S(z_{k+1}) - u_{k+1} \rangle_{H^{-1}(\Omega)}. \end{aligned}$$

By shifting the term with q_k on the other side, we obtain the desired bound:

$$\begin{aligned} \|S(z_{k+1}) - u_{k+1}\|^2_{H_0^1(\Omega)} &\leq \|q_k\|_{H^{-1}(\Omega)} \|S(z_{k+1}) - u_{k+1}\|_{H_0^1(\Omega)} \\ &\leq \tilde{C} \|u_{k+1} - u_k\|_{H_0^1(\Omega)} \|S(z_{k+1}) - u_{k+1}\|_{H_0^1(\Omega)} \end{aligned}$$

for a constant $\tilde{C} > 0$. The estimate $\|q_k\|_{H^{-1}(\Omega)} \leq \tilde{C} \|u_{k+1} - u_k\|_{H_0^1(\Omega)}$ follows as in the proof of 3.13. If $\|u^*(z_{k+1}) - u_{k+1}\|_{H_0^1(\Omega)} = 0$, (39) is obvious. For the other case we obtain from the triangle inequality

(44)
$$\|u^*(z_{k+1}) - u_k\|_{H_0^1(\Omega)} \leq (\tilde{C} + 1) \|u_{k+1} - u_k\|_{H_0^1(\Omega)},$$

thus (39) follows with $C_1 = \tilde{C} + 1$. By possible making $\varepsilon > 0$ smaller we may deduce (40) from 3.17: By this lemma there is $\varepsilon_1 > 0$, such that

$$f^\alpha(u) \geq f^\alpha(S(z_{k+1})) + \langle \nabla f^\alpha(S(z_{k+1})), u - S(z_{k+1}) \rangle_{H^{-1}(\Omega)} + \frac{\kappa}{8} \|u - S(z_{k+1})\|^2_{H_0^1(\Omega)},$$

whenever $\|u - S(z_{k+1})\|_{H_0^1(\Omega)} < \varepsilon_1$ and $\|S(z^*) - S(z_{k+1})\|_{H_0^1(\Omega)} < \varepsilon_1$. Hence, if $\|z - z_{k+1}\|_Z < \varepsilon_1 / L_{z^*}$, we deduce

$$f^\alpha(S(z^*)) \geq f^\alpha(S(z_{k+1})) + \langle \nabla f^\alpha(S(z_{k+1})), S(z^*) - S(z_{k+1}) \rangle_{H^{-1}(\Omega)} + \frac{\kappa}{8} \|S(z^*) - S(z_{k+1})\|^2_{H_0^1(\Omega)},$$

from which (40) follows by using the convexity of $h(\cdot, z_{k+1}) + R_1$ and the first order stationarity of $S(z_{k+1})$ for $J(\cdot, z_{k+1})$. □

Remark 3.19. *In the following part of the article, we will always use the notation*

$$S : B_\varepsilon(z^*) \cap Z \to H_0^1(\Omega)$$

for the single valued localization of ω from 3.18.

Using 3.18 we show, that any weak accumulation point (u^*, z^*) of the sequence that satisfies $C2.$ of 3.14 is a stationary point of J.

Theorem 3.20 (Strong subsequential convergence to stationary points)**.** *Assume that $(x_k)_{k \in \mathbb{N}} \subset H_0^1(\Omega) \times Z$ is generated by algorithm 2 and let $x^* = (u^*, z^*) \in H_0^1(\Omega) \times Z$ be a weak accumulation point of that sequence. If $C2.$ of 3.14 holds true for a stationary point $u^*(z^*)$ of $J(\cdot, z^*)$, then $u^* = u^*(z^*)$ and (u^*, z^*) is a limiting stationary point of J.*

Proof. According to 3.18, there is a radius $\varepsilon > 0$ such that the solution map is well-defined and has a single valued localization $S : Z \to H_0^1(\Omega)$ which is L_{z^*}-Lipschitz continuous on $B_\varepsilon(z^*)$. As z^* is a weak accumulation point, there is, by finite dimensionality of Z, a subsequence $(z_{k_m})_{m \in \mathbb{N}}$ with $z_{k_m} \to z^*$ as $m \to \infty$ and with $z_{k_m} \in B_\varepsilon(z^*)$ for all $m \in \mathbb{N}$. Moreover, we obtain by (3.18)

$$\begin{aligned} \|u_{k_m} - S(z^*)\|_{H_0^1(\Omega)} &\leq \|u_{k_m} - S(z_{k_m})\|_{H_0^1(\Omega)} + \|S(z_{k_m}) - S(z^*)\|_{H_0^1(\Omega)} \\ &\leq \|u_{k_m} - u_{k_m+1}\|_{H_0^1(\Omega)} + L_{z^*} \|z_{k_m} - z^*\|_Z \to 0 \quad \text{as } m \to \infty. \end{aligned}$$

Here, we used, that $\|u_k - u_{k+1}\|_{H_0^1(\Omega)} \to 0$ as $k \to \infty$ and the choice of the subsequence. Therefore, the subsequence $(u_{k_m})_{m\in\mathbb{N}}$ also converges to $S(z^*)$ strongly and we obtain $u^* = S(z^*)$ by the uniqueness of weak limit points. We conclude

$$\begin{aligned} x_{k_m} &\to x^* \quad \text{strongly in } H_0^1(\Omega) \times Z, \\ J(u_{k_m}, z_{k_m}) &\to J(u^*, z^*) \quad \text{by continuity,} \\ \operatorname{dist}(0, \partial J(x_{k_m})) &\to 0, \quad \text{as } m \to \infty. \end{aligned}$$

By 3.5 we infer that (u^*, z^*) is a limiting stationary point of J. □

As of now, we only considered subsequential convergence of the sequence, and investigated sublinear convergence rates. However we are now interested, which conditions are necessary in order to maintain fast local convergence. However, before presenting the corresponding theorem, we need slightly more general descent inequalities. Note that 3.14 is not needed for the following lemma.

Lemma 3.21 (Fundamental inequalities)**.** *For arbitrary $u \in U_{ad}$ and step sizes $\lambda_k - L_2 \geq C_{desc}$ the following inequalities hold true for any $k \in \mathbb{N}$ and $u \in H_0^1(\Omega)$*

(45)
$$J(u, z_{k+1}) \geq J(u_{k+1}, z_{k+1}) + \left(\frac{2\lambda_k - L_2}{2}\right) \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} - \left(\frac{\lambda_k + L_2}{2}\right) \|u - u_k\|^2_{H_0^1(\Omega)},$$

(46)
$$J(u_k, z_k) \geq J(u_{k+1}, z_{k+1}) + \left(\frac{2\lambda_k - L_2}{2}\right) \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} + \sigma_1 \sum_{i=1}^{n_k - 1} \|z_k^{i+1} - z_k^i\|^2_Z.$$

Here, $L_2 > 0$ denotes again the constant in 3.11.

Proof. By the definition of J and 3.11 we obtain directly the following inequality

$$\begin{aligned} J(u, z_{k+1}) &= f(u) + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u, z_{k+1}) + R_1(u) + R_2(z_{k+1}) \\ &\geq g(u, u_k) - \frac{L_2}{2}\|u - u_k\|^2_{H_0^1(\Omega)} + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u, z_{k+1}) + R_1(u) + R_2(z_{k+1}). \end{aligned}$$

Now we invoke the λ_k-strong convexity of $g_{\lambda_k}(\cdot, u_k)$ with respect to the $H_0^1(\Omega)$-norm, i.e.

(47)
$$g_{\lambda_k}(v, u_k) \geq g_{\lambda_k}(u, u_k) + \langle \nabla_1 g_{\lambda_k}(u, u_k), v - u\rangle_{H^{-1}(\Omega)} + \frac{\lambda_k}{2}\|u - v\|^2_{H_0^1(\Omega)}$$

must hold for any $v \in H_0^1(\Omega)$. Consequently we estimate:

$$\begin{aligned}
J(u, z_{k+1}) &\geq g_{\lambda_k}(u, u_k) - \left(\frac{\lambda_k + L_2}{2}\right) \|u - u_k\|^2_{H_0^1(\Omega)} \\
&\quad + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u, z_{k+1}) + R_1(u) + R_2(z_{k+1}) \\
&\geq g_{\lambda_k}(u_{k+1}, u_k) + \frac{\lambda_k}{2}\|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} - \left(\frac{\lambda_k + L}{2}\right) \|u - u_k\|^2_{H_0^1(\Omega)} \\
&\quad + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u_{k+1}, z_{k+1}) + R_1(u) + R_2(z_{k+1}) && \text{by (47).} \\
&\geq g(u_{k+1}, u_k) + \frac{L_2}{2}\|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} + \left(\frac{2\lambda_k - L_2}{2}\right) \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} \\
&\quad - \left(\frac{\lambda_k + L_2}{2}\right) \|u - u_k\|^2_{H_0^1(\Omega)} + \frac{\alpha}{2}\|\nabla u\|^2_{L^2(\Omega)} + h(u_{k+1}, z_{k+1}) + R_1(u_{k+1}) + R_2(z_{k+1}) \\
&\geq f(u_{k+1}) + \left(\frac{2\lambda_k - L_2}{2}\right) \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} \\
&\quad - \left(\frac{\lambda_k + L_2}{2}\right) \|u - u_k\|^2_{H_0^1(\Omega)} + h(u_{k+1}, z_{k+1}) + R_1(u_{k+1}) + R_2(z_{k+1}),
\end{aligned}$$

which yields the desired estimate (45). We continue by setting $u = u_k$ in (45) and obtain

$$J(u_k, z_{k+1}) \geq J(u_{k+1}, z_{k+1}) + \left(\frac{2\lambda_k - L_2}{2}\right) \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)}.$$

Using the descent property of nested inner loop algorithm, (20) for some $\sigma_1 > 0$, we further conclude

$$J(u_k, z_{k+1}) \leq J(u_k, z_k) - \sigma_1 \sum_{i=1}^{n_k - 1} \|z_k^{i+1} - z_k^i\|^2_Z.$$

Combining the previous two inequalities yields (46). □

Let us now prove, that strong local linear convergence to a stationary point is maintained, if the function $z^* \mapsto h(u^*, z) + R(z)$ satisfies the KL-condition with exponent $q > 2$ at z^*. The proof strategy is inspired by the work [27].

Theorem 3.22 (Local convergence of (2))**.** *Let $(x_k)_k \subset H_0^1(\Omega) \times Z$ be generated by algorithm 2 such that 3.14 holds true at a weak accumulation point $x^* \in H_0^1(\Omega) \times Z$. Moreover, assume that the accuracies of the inner loop are given by $\eta_k \lesssim k^{-\gamma}$, $\gamma > 1/2$. Then, there is a radius $r > 0$ such that if $z_0 \in B_r(z^*)$, the following statements hold true:*

(i) *The function values converge linearly, i.e. there is $Q_1 \in (0, 1)$ such that*

$$J(u_{k+1}, z_{k+1}) - J(u^*, z^*) \leq Q_1 \left(J(u_k, z_k) - J(u^*, z^*)\right) \quad \text{for any } k \in \mathbb{N} \tag{48}$$

(ii) *The iterates also converge with a linear rate, i.e. there exists a $C > 0$ and a $Q_2 \in (0, 1)$ such that*

$$\sqrt{\|u_k - u^*\|^2_{H_0^1(\Omega)} + \|z_k - z^*\|^2_Z} \leq C Q_2^k \left(J(u_0, z_0) - J(u^*, z^*)\right) \quad \text{for any } k \in \mathbb{N} \tag{49}$$

Proof. As the proof is a bit technical, we divide it into parts. First note, by 3.20 we obtain that $(u^*, z^*) \in U_{ad} \times Z$ is even a strong accumulation point and also limiting-stationary for the objective J. Taking a strongly convergent subsequence and the continuity of J into account, we even infer that

$$J(x_k) \to J(x^*), \quad \text{as } k \to \infty, \tag{50}$$

which will be important in order to make the application of the desingularization $\varphi : (0, \eta) \to \mathbb{R}_+$ possible.

Step I: There is an $\varepsilon > 0$ and a $Q_1 \in (0, 1)$ independent of $k \in \mathbb{N}$ such that if $\|z_{k+1} - z^*\|_Z < \varepsilon$ implies

$$J(u_{k+1}, z_{k+1}) - J(u^*, z^*) \leq Q_1 \left(J(u_k, z_k) - J(u^*, z^*)\right). \tag{51}$$

We start with proving (51). First let us fix $\varepsilon_1 > 0$ according to 3.18 such that $z \mapsto S(z) \in H_0^1(\Omega)$ is well defined and single valued for all $z \in B_{\varepsilon_1}(z^*)$ and such that the inequalities (39), (40), (41) hold true for $\|z_{k+1} - z^*\| \leq \varepsilon_1$. An application of the general descent 3.21 with $u = S(z_{k+1})$ consequently yields

$$J(S(z_{k+1}), z_{k+1}) \geq J(u_{k+1}, z_{k+1}) + \left(\frac{2\lambda_k - L_2}{2}\right) \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} - \left(\frac{\lambda_k + L_2}{2}\right) \|S(z_{k+1}) - u_k\|^2_{H_0^1(\Omega)}$$

If $\|S(z_{k+1}) - u_k\|^2_{H_0^1(\Omega)} = 0$, nothing is to show. Otherwise we obtain from the latter inequality

$$\begin{aligned} J(u_{k+1}, z_{k+1}) - J(S(z_{k+1}), z^{k+1}) &\leq \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)} \left(\left(\frac{\lambda_k + L_2}{2}\right) C_1^2 - \left(\frac{2\lambda_k - L_2}{2}\right)\right) \\ &=: C_{L,\lambda_k} \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)}, \end{aligned} \tag{52}$$

where $C_1 > 0$ is the constant from 3.18 with $\|S(z_{k+1}) - u_k\|_{H_0^1(\Omega)} \leq C_1 \|u_k - u_{k+1}\|_{H_0^1(\Omega)}$. If $J(u^*, z_{k+1}) \leq J(u^*, z^*)$ the quadratic growth in 3.17 implies $J(u^*(z_{k+1}), z_{k+1}) \leq J(u^*, z^*)$ as well. Consequently from (52) we obtain

$$J(u_{k+1}, z_{k+1}) - J(u^*, z^*) \leq C_{L,\lambda_k} \|u_{k+1} - u_k\|^2_{H_0^1(\Omega)}. \tag{53}$$

Invoking the descent inequality (46), we find a generic constant $C_2 > 0$ such that

$$J(u_{k+1}, z_{k+1}) - J(u^*, z^*) \leq C_2 \left(J(u_k, z_k) - J(u_{k+1}, z_{k+1})\right). \tag{54}$$

If instead $J(u^*, z_{k+1}) > J(u^*, z^*)$, we deduce from the KL-inequality with exponent $\beta = 1 - 1/q \in (0, 1/2)$ and the continuity of J the existence of an $\varepsilon_{KL} > 0$ with $J(u^*, z^*) < J(u^*, z_{k+1}) < J(u^*, z^*) + \eta$ and

$$\begin{aligned} J(u^*, z_{k+1}) - J(u^*, z^*) &= h(u^*, z_{k+1}) + R_2(z_{k+1}) - h(u^*, z^*) - R_2(z^*) \\ &\leq C_{KL} \operatorname{dist}(0, \partial_z J(u^*, z_{k+1}))^q, \end{aligned}$$

whenever $\|z_{k+1} - z^*\|_Z \leq \varepsilon_{KL}$. Application of the triangle inequality yields

$$\begin{aligned} J(u^*, z_{k+1}) - J(u^*, z^*) &\leq C_{KL} \operatorname{dist}(0, \partial_z J(u^*, z_{k+1}))^q \\ &\leq C_{KL} \left(\operatorname{dist}(0, \partial_z J(u_k, z_{k+1}) + C_3 \|u^* - u_k\|_{H_0^1(\Omega)}\right)^q \\ &\leq C_{KL} (\operatorname{dist}(0, \partial_z J(u_k, z_{k+1}) + C_3 \|u^* - u^*(z_{k+1})\|_{H_0^1(\Omega)} \\ &\quad + C_3 \|u^*(z_{k+1}) - u_k\|_{H_0^1(\Omega)})^q. \end{aligned}$$

For a constant $C_3 > 0$ that bounds the Lipschitz constant of $\nabla_z h(\cdot, z_{k+1})$. From (21), we know that $\operatorname{dist}(0, \partial_z J(u_k, z_{k+1})) \leq \sigma_2 \|z_k^{n_k} - z_k^{n_k - 1}\|_Z$ and by 3.18 we have

$$\|u^*(z_{k+1}) - u_k\|_{H_0^1(\Omega)} \leq L_{z^*} \|u_{k+1} - u_k\|_{H_0^1(\Omega)}.$$

Hence, also using $\sum_{i=1}^n a_i \leq \sqrt{n}\sqrt{\sum_{i=1}^n a_i^2}$ for $a_i \geq 0$ we estimate

$$\begin{aligned} J(u^*, z_{k+1}) - J(u^*, z^*) &\leq C_{KL}\Big(\sigma_2\|z_k^{n_k} - z_k^{n_k-1}\|_Z + C_1C_3\|u_{k+1} - u_k\|_{H_0^1(\Omega)} \\ &\quad + C_3\|u^*(z_{k+1}) - u^*\|_{H_0^1(\Omega)}\Big)^q \\ &\leq C_{KL}\Big(\sigma_2^2\|z_k^{n_k} - z_k^{n_k-1}\|_Z^2 + (C_1C_3)^2\|u_{k+1} - u_k\|_{H_0^1(\Omega)}^2 \\ &\quad + C_3^2\|u^*(z_{k+1}) - u^*\|_{H_0^1(\Omega)}^2\Big)^{\frac{q}{2}} \\ &\leq C_{KL}\left(\sigma_2^2\sum_{i=0}^{n_k-1}\|z_k^{n_k} - z_k^{n_k-1}\|_Z^2 + (C_1C_3)^2\|u_{k+1} - u_k\|_{H_0^1(\Omega)}\right. \\ &\quad \left. + C_3^2\|u^*(z_{k+1}) - u^*\|_{H_0^1(\Omega)}^2\right)^{\frac{q}{2}}. \end{aligned}$$

Using the convexity of $x \mapsto x^{q/2}$ as $q > 2$ and summarizing all constants, we infer the existence of a constant $C_4 > 0$ such that

$$J(u^*, z_{k+1}) - J(u^*, z^*) \leq C_4\left(\sum_{i=0}^{n_k-1}\|z_k^{i+1} - z_k^i\|_Z^2 + \|u_{k+1} - u_k\|_{H_0^1(\Omega)}^2\right)^{\frac{q}{2}} + C_4\left(\|u^*(z_{k+1}) - u^*\|_{H_0^1(\Omega)}^2\right)^{\frac{q}{2}}.$$

Using the descent inequality (46), we deduce

(55)
$$J(u^*, z_{k+1}) - J(u^*, z^*) \leq C_4\left(J(u_k, z_k) - J(u_{k+1}, z_{k+1})\right)^{\frac{q}{2}} + C_4\left(\|u^*(z_{k+1}) - u^*\|_{H_0^1(\Omega)}^2\right)^{\frac{q}{2}}.$$

Now choose $\min(\varepsilon_{KL}, \varepsilon_1) \geq \varepsilon_2 > 0$, such that $\|z^* - z_{k+1}\|_Z < \varepsilon_2$ implies

(56)
$$\|u^*(z_{k+1}) - u^*\|_{H_0^1(\Omega)}^{q-2} \leq \frac{\kappa}{8C_4}.$$

Also taking into account (40), we obtain for $\|z_{k+1} - z^*\|_Z < \varepsilon_2$ from (55) that

(57)
$$\begin{aligned} J(u^*(z_{k+1}), z_{k+1}) - J(u^*, z^*) &\leq C_4\left(J(u_k, z_k) - J(u_{k+1}, z_{k+1})\right)^{\frac{q}{2}} \\ &\leq C_5(J(u_k, z_k) - J(u_{k+1}, z_{k+1})), \end{aligned}$$

where we introduced a constant $C_5 > 2C_4 > 0$ possibly larger then $2C_4$ (since $x^{q/2} \leq x$ only for $x \leq 1$). Now adding (52) and (57) we obtain

$$J(u_{k+1}, z_{k+1}) - J(u^*, z^*) \leq C_{L,\lambda_k}\|u_{k+1} - u_k\|_{H_0^1(\Omega)}^2 + C_5\left(J(u_k, z_k) - J(u_{k+1}, z_{k+1})\right).$$

Recall that the inequality above holds only in the case $J(u^*, z_{k+1}) > J(u^*, z^*)$. However by taking into account also the other case in (54) and again invoking the descent inequality (46), we find a generic constant $C > 0$, such that

(58)
$$J(u_{k+1}, z_{k+1}) - J(u^*, z^*) \leq C\left(J(u_k, z_k) - J(u_{k+1}, z_{k+1})\right).$$

The previous inequality (58) can be equivalently rephrased as follows, which will ensure the linear convergence of function values later on:

(59)
$$\begin{aligned} J(u_{k+1}, z_{k+1}) - J(u^*, z^*) &\leq \left(\frac{C}{1+C}\right)(J(u_k, z_k) - J(u^*, z^*)) \\ &=: Q_1\left(J(u_k, z_k) - J(u^*, z^*)\right). \end{aligned}$$

The inequalities (58) and (59) only hold under the premise, that $\|z_{k+1} - z^*\| \leq \varepsilon := \varepsilon_2$. This eventually proves (51) and therefore step I.

Step II: Show, that there is an $0 < r \leq \varepsilon$ such that $z_0 \in B_r(z^*)$ implies $z_k \in B_\varepsilon(z^*)$ for all $k \in \mathbb{N}$.

We will construct the radius $r > 0$ iteratively. We start choosing $r := \varepsilon > 0$ such that $z_0 \in B_\varepsilon(z^*)$ where $\varepsilon > 0$ is given from Step I. Then, using the stopping index $n_k \in \mathbb{N}$ in algorithm 1 we recall from 3.9 that

$$n_k \leq \frac{2k^{2\gamma}(J(x_k) - J(x_{k+1}))}{\sigma_1} \quad \text{for all } k \in \mathbb{N}. \tag{60}$$

Hence we deduce for the next iterate,

$$\begin{aligned} \|z_1 - z^*\|_Z &\leq \|z_1 - z_0\|_X + \|z_0 - z^*\|_Z \\ &\leq \sqrt{n_0 \sum_{i=0}^{n_0-1} \|z_0^{i+1} - z_0^i\|_Z^2} + \|z_0 - z^*\|_Z \\ &\leq \sqrt{2/\sigma_1}(J(x_0) - J(x_1) + \|z_0 - z^*\|_Z \\ &\leq \sqrt{2/\sigma_1}(J(x_0) - J(x^*)) + \|z_0 - z^*\|_Z, \end{aligned} \tag{61}$$

where the last inequality follows from $J(x_k) \geq J(x^*)$ for all k, by the remark at the beginning of the proof. Invoking continuity of J we may reduce $r > 0$ such that also $z_1 \in B_\varepsilon(z^*)$.
Now we assume that $z_j \in B_\varepsilon(z^*)$ for $j = 0, \ldots, k$ for some $k \in \mathbb{N}$. We take very briefly the general viewpoint of KL-functions in (36) and call the desingularization function $\varphi : (0, \eta) \to \mathbb{R}_+$ as in the convergence analysis in [7]. By eventually making $r > 0$ smaller we may assume, again by continuity of J, that $J(x_j) < J(x^*) + \eta$ for all $j \geq 0$. Hence the KL-inequality is applicable for z_j, $j = 0, \ldots, k$. We then deduce by concavity of φ that

$$\varphi(J(x_j) - J(x^*)) - \varphi(J(x_{j+1}) - J(x^*)) \geq \varphi'(J(x_j) - J(x^*))(J(x_j) - J(x_{j+1})). \tag{62}$$

Note, that for the KL-exponent $\beta = 1 - 1/q$ with $q > 2$, we have that $\varphi(t) = t^\beta$ and $\varphi'(t) \lesssim t^{-\frac{1}{q}}$. Therefore we estimate first for general $j \geq 0$

$$\begin{aligned} \varphi'(J(x_j) - J(x^*)) &\gtrsim \frac{1}{(J(x_j) - J(x^*))^{\frac{1}{q}}} \\ &\gtrsim \frac{1}{Q_1^{\frac{j}{q}}(J(x_0) - J(x^*))^{\frac{1}{q}}}, \end{aligned} \tag{63}$$

where we used $z_j \in B_\varepsilon(z^*)$ for $j = 0, \ldots, k$ and inequality (51) from Step I. From the descent inequality (46) we infer that there exists a constant $C_6 > 0$ with

$$\begin{aligned} J(x_j) - J(x_{j+1}) &\geq C_6 \|u_{j+1} - u_j\|_{H_0^1(\Omega)}^2 + C_6 \sum_{i=0}^{n_j-1} \|z_j^{i+1} - z_j^i\|_Z^2 \\ &\geq C_6 \|u_{j+1} - u_j\|_{H_0^1(\Omega)}^2 + \frac{C_6}{n_j} \|z_{j+1} - z_j\|_Z^2 \\ &\geq \frac{C_6}{n_j} \left(\|u_{j+1} - u_j\|_{H_0^1(\Omega)}^2 + \|z_{j+1} - z_j\|_Z^2 \right) = \frac{C_6}{n_j} \|x_{j+1} - x_j\|_X^2. \end{aligned} \tag{64}$$

Then it is concluded from (62), (63) and (64) that

$$\frac{C_6}{n_j} \|x_{j+1} - x_j\|_X^2 \leq Q_1^{\frac{j}{q}} (J(x_0) - J(x^*))^{\frac{1}{q}} \left(\varphi(J(x_j) - J(x^*)) - \varphi(J(x_{j+1}) - J(x^*)) \right), \tag{65}$$

and henceforth

$$\begin{aligned}\|x_{j+1}-x_j\|_X &\lesssim \sqrt{n_j}Q_1^{\frac{j}{2q}}(J(x_0)-J(x^*))^{\frac{1}{2q}}\left(\varphi(J(x_j)-J(x^*))-\varphi(J(x_{j+1})-J(x^*))\right)^{\frac{1}{2}}\\ &\leq 2n_jQ_1^{\frac{j}{q}}(J(x_0)-J(x^*))^{\frac{1}{q}}+2\left(\varphi(J(x_j)-J(x^*))-\varphi(J(x_{j+1})-J(x^*))\right)\\ &\lesssim 2j^{2\gamma}Q_1^{\frac{j}{q}}(J(x_0)-J(x^*))^{\frac{1}{q}}+2\left(\varphi(J(x_j)-J(x^*))-\varphi(J(x_{j+1})-J(x^*))\right), \qquad (66)\end{aligned}$$

where we used Young's inequality in the second step and 3.9 for the estimate of $n_j \lesssim j^{2\gamma}$ in the last. We deduce by summing from $j=0,\dots,k$ and using (66) that

$$\begin{aligned}\|z_{k+1}-x^*\|_Z &\leq \|x_{k+1}-x_0\|_X+\|z_0-z^*\|_Z\\ &\leq \sum_{j=0}^{k}\|x_{j+1}-x_j\|_X+\|z_0-z^*\|_Z\\ &\lesssim (J(x_0)-J(x^*))^{\frac{1}{q}}\sum_{j=0}^{k}j^{2\gamma}Q_1^{\frac{j}{q}}\\ &\quad+2\sum_{j=0}^{k}\left(\varphi(J(x_j)-J(x^*))-\varphi(J(x_{j+1})-J(x^*))\right)+\|z_0-z^*\|_Z\\ &\lesssim (J(x_0)-J(x^*))^{\frac{1}{q}}\sum_{j=0}^{\infty}j^{2\gamma}Q_1^{\frac{j}{q}}\\ &\quad+\varphi(J(x_0)-J(x^*))-\varphi(J(x_{k+1})-J(x^*))+\|z_0-z^*\|_Z,\end{aligned}$$

where $\sum_{j=0}^{\infty}(j)^{2\alpha}Q_1^j<+\infty$. Denote the constant hidden in the " $\lesssim$ " to be $C_7>0$ and if

$$C_7(J(x_0)-J(x^*))^{\frac{1}{q}}\sum_{j=0}^{+\infty}j^{2\gamma}Q_1^{\frac{j}{q}}+C_7\varphi(J(x_0)-J(x^*))+C_7\|z_0-z^*\|_Z<\varepsilon, \qquad (67)$$

then also $z_{k+1}\in B_\varepsilon(z^*)$. This is possible by choosing a smaller $r>0$. Eventually, by induction we have shown Step II. It is important to note, that actually no constant hidden in " $\lesssim$ " above does depend on k.

Step III: Show that the assertions of the theorem are true.

Using the results from steps I and II, the linear convergence of function values in (i) follows immediately. Now, let us prove the strong convergence of the sequence $(x_k)_k$. By Step I and Step II, we have that

$z_k \in B_\varepsilon(z^*)$ for all $k \geq 0$ and therefore for arbitrary $l \in \mathbb{N}$

$$
\begin{aligned}
\|z_{k+l} - z_k\|_Z^2 + \|u_{k+1} - u_k\|_{H_0^1(\Omega)}^2 &\leq \sum_{j=k}^{k+l} \|z_{i+1} - z_i\|_Z + \|u_{i+1} - u_i\|_{H_0^1(\Omega)}^2 \\
&\leq \sum_{j=k}^{k+l} n_j \left(\sum_{i=0}^{n_j-1} \|z_j^{i+1} - z_j^i\|_Z^2 + \|u_{j+1} - u_j\|_{H_0^1(\Omega)}^2 \right) \\
&\leq \sum_{j=k}^{k+l} n_j \left(J(x_j) - J(x^*) \right) \\
&\lesssim \left(J(x_0) - J(x^*) \right) \sum_{j=k}^{k+l} j^{2\gamma} Q_1^j \\
&= \left(J(x_0) - J(x^*) \right) Q_1^k \sum_{j=0}^{l} (j+k)^{2\gamma} Q_1^j \\
&\lesssim \left(J(x_0) - J(x^*) \right) Q_1^{\frac{k}{2}} Q_1^{\frac{k}{2}} k^{2\gamma} \sum_{j=0}^{+\infty} (j+1)^{2\gamma} Q_1^j \to 0 \quad \text{as } k \to \infty,
\end{aligned} \tag{68}
$$

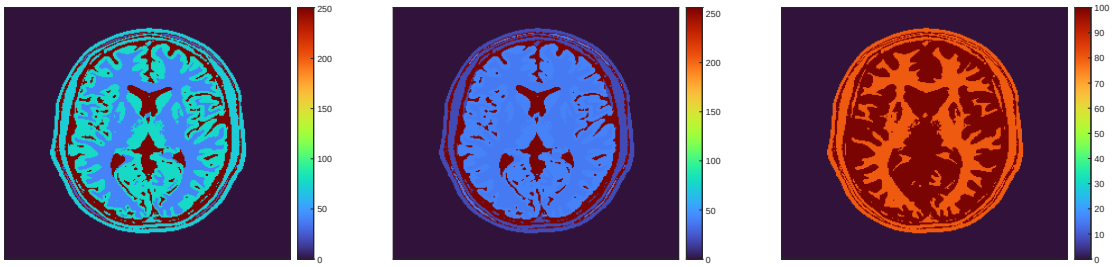
where we used again the definition of the accuracy $\eta_k \lesssim k^{-\gamma}$. Hence $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $H_0^1(\Omega) \times Z$ and converges to some element $(\tilde{u}, \tilde{z})$ strongly. As (u^*, z^*) is a weak accumulation point by assumption we infer $(\tilde{u}, \tilde{z}) = (u^*, z^*)$. The linear convergence in (iii) follows from (68), by sending $l \to \infty$ and using $C = \sup_k Q_1^{\frac{k}{2}} k^{2\gamma} \sum_{j=0}^{+\infty} (j+1)^{2\gamma} Q_1^j$ and $Q_2 = Q_1^{1/2}$ in (49). □

Remark 3.23. *Let us comment briefly on the results of 3.22. It is interesting in two different aspects and it also provides hints on possible future research directions and potential ways to improve the algorithm.*

1. *The strong monotonicity of the Hessian of f^α, as stated in 3.14, is a somewhat restrictive condition but a classical assumption in second-order and Gauss-Newton methods to achieve locally fast convergence (see the classical references [11, 46]). Without the inclusion of the learning term, i.e. if $h(u, z) + R_2(z) = 0$, the argument in [27] would guarantee locally linear convergence. However, when the learning term $h(u, z) + R_2(z)$ is included, we require an artificially high KL-exponent $q > 2$ to preserve locally fast convergence. To benefit practically from this result, it might be useful to incorporate alternative discrepancy terms $h(\cdot, \cdot)$ that exhibit sharper local curvature. This could help improve convergence behavior and enhance the overall performance of the algorithm. However this is out of the scope of the current article.*
2. *From the convergence proof above, it becomes evident that the local convergence behavior is almost entirely determined by the convergence properties of the sequence $(z_k)_k$. It is crucial to ensure that this sequence stays within a potentially very small neighborhood of the stationary point z^*. Although this is a strong assumption, and the exact radius ε in 3.22 is unknown and difficult to verify in practical applications, it highlights the importance of good initialization for the variable z. This can be achieved through appropriate pre-training. However, a detailed algorithmic and theoretical investigation of this observation lies beyond the scope of the present work.*

4. Numerical experiments

We present the numerical results of the nested alternating algorithm applied to the qMRI problem (P_0). For spatial discretization, we adopt the approach from [24], where a simple finite difference scheme is used. Further details can be found in [24]. In addition, the update steps for the variable $z = (D, C)$ must be specified. Here, we roughly follow the methods proposed in [50, 52], which apply blind dictionary learning regularization to the linear MRI problem. For the reader's convenience, the algorithm from their work, along with some of its key properties, is included in algorithm 5 in the appendix. The ground truth parameters for our numerical experiments are displayed in Figure 1. The ground truth images (physical parameters) follow a standard configuration commonly used for testing qMRI methods, and has been employed in previous studies, such as [23, 24]. To streamline the parameter search and improve the conditioning of the subproblems, we scaled both the T_1 and T_2 variables to the range $[0, 250]$ ms. Although these ranges differ from typical relaxation times, which typically lie in the ranges $T_1 \in [0, 6000]$ ms and $T_2 \in [0, 600]$ ms, this adjustment provides a simplified setting that effectively demonstrates the performance of our algorithm. while still being relevant for practical applications.



T_1-map. T_2-map ρ-map

FIGURE 1. The set of ground truth parameters T_1, T_2 (milliseconds) and proton density ρ (dimensionless).

The discrete optimization problem. For our numerical tests, we employ a uniform grid and a finite differences discretization for the variable $u = (\rho, T_1, T_2)$. This approach leads us to consider the space $U := \mathbb{R}^{n_1 \times n_2 \times 3}$, where $n_1, n_2 \in \mathbb{N}$ denote the number of pixels in each direction. In our case, we set $n_1 = n_2 = 256$. We also make use of the classical finite difference approximations for the image gradient, denoted by $\nabla^h : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2 \times 2}$, and the Laplace operator with zero boundary conditions, denoted by $\Delta^h : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$, with a mesh-size $h > 0$. If ∇^h and Δ^h are applied on elements of U, we will use the same notation and apply the operators component wise, i.e.

$$\tilde{\nabla}^h : U \to U_1 := \mathbb{R}^{3 \times n_1 \times n_2 \times 2}, \qquad \tilde{\nabla}^h u = (\nabla^h \rho, \nabla^h T_1, \nabla^h T_2),$$
$$\tilde{\Delta}^h : U \to U, \qquad \tilde{\Delta}^h u = (\Delta^h \rho, \Delta^h T_1, \Delta^h T_2),$$

for $u = (\rho, T_1, T_2) \in U$. We equip both spaces, U, U_1, with the following scaled norms, defined by:

$$\|u\|_U^2 := \frac{h^2}{M_1^2}\|\rho\|_2^2 + \frac{h^2}{M_2^2}\|T_1\|_2^2 + \frac{h^2}{M_3^2}\|T_2\|_2^2, \tag{69}$$

$$\|v\|_{U_1}^2 := \frac{h^2}{M_1^2}\|v_1\|_2^2 + \frac{h^2}{M_2^2}\|v_2\|_2^2 + \frac{h^2}{M_3^2}\|v_3\|_2^2, \tag{70}$$

where we used $M_1 = 100$ and $M_2 = M_3 = 250$ as scaling parameters. This results in the following discrete problem:

$$\min_{(u,D,C)\in X} J_d(u,D,C) := \frac{h^2}{2}\|A \circ \Pi_d(u) - f^\delta\|_2^2 + \frac{\alpha}{2}\|\nabla u\|_{U_1}^2 + \mathcal{I}_{U_{ad}}(u)$$
$$+ \sum_{i=1}^{3} \lambda^j \left(\frac{1}{2}\|P[\frac{1}{M_j}u_j] - D_j C_j\|_F^2 + \beta_j \|C_i\|_1 \right), \tag{P_2}$$

which we aim to minimize over the space $X := U \times O_K^3 \times \mathbb{R}^{M\times K\times 3}$ using the admissible set of parameters U_{ad}

$$U_{ad} = \left\{ u = (\rho, T_1, T_2) \in U \;\middle|\; \begin{array}{l} \rho_{ij} \in [0,110], (T_1)_{ij} \in [0,300], (T_2)_{ij} \in [0,300] \\ \text{for all } 1 \le i \le n_1 \text{ and } 1 \le j \le n_1 \end{array} \right\} \tag{71}$$

In (P_2), the discrete variant of the Bloch solution operator is given by

$$\Pi_d : U \to \mathbb{C}^{n_1\times n_2 \times L} \quad [\Pi_d(u)]_{ijl} = \pi(u_{ij})_l, \quad 1 \le l \le L,$$

where $\pi : \mathbb{R}^3 \to \mathbb{C}^L$ is the function, defined in (8). Moreover, the linear operator A, modelling the observation process, is defined by

$$A : \mathbb{C}^{n_1\times n_2\times L} \to \mathbb{C}^{n_1\times n_2\times L} \quad [Ay]_{ijl} = S_l \mathcal{F}[y_l] \quad 1 \le l \le L. \tag{72}$$

In the definition above, $\mathcal{F} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^{n_1\times n_2}$ denotes the normalized, discrete 2D-Fourier transform and $S_l : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^{n_1\times n_2}$ denotes a predefined sampling pattern which acts on the l-th magnetization slice as

$$S_l(y)_{i,j} = \begin{cases} y_{ij} & \text{if frequency } y_{ij} \in \mathbb{C} \text{ is sampled.} \\ 0 & \text{if } y_{ij} \text{ is not sampled.} \end{cases}$$

Regarding the sampling-pattern, we follow exactly the setting in [24]. Moreover, we make use of the linear patch extraction operator P, which cuts out small image patches and puts them into a large matrix. More precisely, we define:

$$P : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{M\times K} \quad Pu = \left[R_{11}u, R_{21}u, \dots R_{n_1 1}u, R_{12}u, R_{22}u \dots R_{n_1 2}u, \dots, R_{1n_2}u, \dots R_{n_1 n_2}u\right].$$

for $j = 1, 2, 3$. Here $R_{kl} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^M$ is an operator that extracts a patch of size $p \times p$ from an image $u \in \mathbb{R}^{n_1\times n_2}$, where the top-left corner of the patch is located at pixel (k, l). The extracted patch is then vectorized into a column vector of size $K = p^2$. We use overlapping patches, as described in [49, 52] such that exactly $M = n_1 \cdot n_2$ can be extracted. The normalization factor $1/M_j$ is introduced because, empirically, we observed improved reconstruction quality for the dictionary learning problem when the data is normalized. Additionally, this normalization significantly simplifies the process of hyperparameter tuning. The overall discrete version of algorithm 2 is eventually given by:

Algorithm 4 Computation of a stationary point of (P_2).

Get initial values $(u_0, D_0, C_0) \in U^{3N} \times O_K^3 \times \mathbb{R}^{K\times M\times 3}$, parameter $\gamma > 0$ to define the stopping accuracy for the nested subroutine, step-size parameter for the u-step $\lambda_0 > 0, \tau > 1$ and $\sigma_3 \in (0,1)$, $\varepsilon_1, \varepsilon_2 > 0$.
Set k = 0.
while no stopping criterion is satisfied **do**
  Dictionary-learning-step: Given $(u_k, D_k, C_k) \in U \times O_K^3 \times \mathbb{R}^{M\times K\times 3}$
  for $j = 1, 2, 3$ **do**:
    Use algorithm 5 with initialization $D_k^0 := (D_k)_j$, $C_k^0 := (C_k)_j$ and
    stopping accuracy $\eta_k := k^\gamma \sqrt{\|C_0\|_F^2 + \|D_0\|_F^2}$ for the problem :

$$\min_{D\in O_K, C\in\mathbb{R}^{M\times K}} \frac{1}{2}\|DC - P[\frac{1}{M_j}u_j]\|_F^2 + \frac{\beta_j}{\lambda^j}\|C\|_1,$$

    to obtain $(D_{k+1})_j \in O_K$ and $(C_{k+1})_j \in \mathbb{R}^{K\times M}$.
  end for
  u-step: Given $(u_k, D_{k+1}, C_{k+1}) \in U \times O_K^3 \times \mathbb{R}^{M\times K\times 3}$.
  for $j = 1, \ldots$ **do**:
    Take $\lambda_0 > 0$, set $\lambda_k := \lambda_0 \tau^j$ and compute a global solution $\widehat{u}(\lambda_k) \in U_{ad}$ of

$$\min_{u\in U_{ad}} g_{\lambda_k}^d(u, u_k) + \frac{\alpha}{2}\|\nabla^h u\|_{U_1}^2 + \sum_{j=1}^{3} \frac{\lambda^j}{2}\|P[\frac{1}{M_j}u] - (D_{k+1})_j (C_{k+1})_j\|_F^2, \tag{73}$$

    where $g_{\lambda_k}^d(\cdot, u_k) : U \to \mathbb{R}$ is defined as the discrete analogue of the model in (23) as

$$g_{\lambda_k}^d(u, u_k) := \frac{h^2}{2}\|A \circ \Pi_d'(u_k)[u - u_k] + F(u_k) - f^\delta\|_2^2 + \frac{\lambda_k}{2}\left(\|u - u_k\|_U^2 + \|\nabla^h(u - u_k)\|_{U_1}^2\right),$$

    until the descent condition

$$J_d(\widehat{u}(\lambda_k), D_{k+1}, C_{k+1}) \le J_d(u_k, D_{k+1}, C_{k+1}) - \frac{\sigma_3\lambda_k}{2}\left(\|\widehat{u}(\lambda_k) - u_k\|_U^2 + \|\nabla^h\widehat{u}(\lambda_k) - \nabla^h u_k\|_{U_1}^2\right), \tag{74}$$

    is satisfied.
  end for
  Set $u_{k+1} = \widehat{u}(\lambda_k)$.
  Set $k = k + 1$.
  if $\|u_k - u_{k-1}\|_U^2 + \|\nabla^h u_k - \nabla^h u_{k-1}\|_{U_1}^2 < \varepsilon_1^2$ and $\|D_k - D_{k-1}\|_F^2 + \|C_k - C_{k-1}\|_F^2 < \varepsilon_2^2$ **then**
    stop the while loop.
  end if
end while
Return $u_k \in U$ as the desired physical parameter.

Details on the implementation. The main difficulty in the implementation of algorithm 4 is the solution of the subproblem (73). We use the built-in quadratic programming solver in MATLAB, which relies on the trust-region subspace method, combining techniques from [22] and [16]. However, the subproblem is still very delicate to solve, as the Hessian involves the term:

$$(\Pi_d)'(u_k)^* A^* A (\Pi_d)'(u_k), \tag{75}$$

with A being the matrix defined in (72). Note that if $S_l = Id$ for every $1 \leq l \leq L$, i.e. if no subsampling is applied, then $A^*A = I$. If instead subsampling is used, i.e. if $S_l \neq Id$, $l = 1, \dots, L$, the matrix in (75) is generally dense and of large scale, which complicates the solution procedure of the subproblem significantly. To mitigate this issue, we follow the approximation approach, proposed in [60] and replace the term $F'(u_k)^*F'(u_k)$ by the approximation

$$(\Pi_d)'(u_k)^*A^*A(\Pi_d)'(u_k) \approx \frac{1}{r}(\Pi_d)'(u_k)^*(\Pi_d)'(u_k). \tag{76}$$

where $r \in \mathbb{N}$ is the undersampling rate. In our experiments in the next paragraph we use $r \in \{16, 32\}$. Let us comment that, while the quality of this approximation seems well documented in practical applications, a rigorous proof that quantifies potential deviations is missing.

Two experiments with different noise levels and under sampling. We test our algorithm on two artificial testcases. For both we take the groundtruth image $u_{true} \in U$ depicted in Figure 1 and simulate noisy data using the forward operator according to

$$f^\delta = A \circ \Pi_d(u_{true}) + \sigma^2 \mathcal{N}(0, I). \tag{77}$$

We utilize complex noise with a standard deviation of $\sigma^2 > 0$ and apply Cartesian subsampling patterns, following the approach in [24]. In the first experiment, we use a $16\times$ undersampling rate and set the noise standard deviation to $\sigma = 2$. In the second experiment, the undersampling rate is increased to $32\times$ and the standard deviation to $\sigma = 5$. For both experiments, we fix $L = 100$, a small yet realistic value for practical applications. Our method is compared against the BLIP reconstruction technique proposed in [23] and the Levenberg-Marquardt method from [24]. It is important to note that the Levenberg-Marquardt method guarantees convergence for zero-residual problems, where $\|F(u_{\text{true}}) - f^\delta\|_{L^2(\Omega,\mathbb{C}^L)} = 0$. For other configurations, the outcome is strongly influenced by the number of algorithm steps executed, and convergence to stationary points is not typically expected. Although numerous well-established stopping criteria exist in the inverse problems literature, we do not apply them in this study. Instead, we manually adjust the number of Levenberg-Marquardt steps for our experiments. For the linear system that must be solved at each update step of the Levenberg-Marquardt method, we employ the approximation (76). We also compare our algorithm with the dictionary learning algorithm, where, instead of using the nested update procedure, only a single update step for both D and C is performed. We refer to this algorithm as *one-step-dictionary-learning* or *dictionary learning (one step)*. The parameter settings, along with a description of the parameters for both configurations, are presented in Figure 2.

Results and observations. The results of all algorithms are visually presented in Figure 6 for the small noise case with $16\times$ undersampling and in Figure 7 for the $32\times$ undersampling with a higher noise level. The relative errors

$$\frac{\|X_{\text{reconstruction}} - X_{\text{groundtruth}}\|_2}{\|X_{\text{groundtruth}}\|_2}$$

for $X \in \{\rho, T_1, T_2\}$ are shown in Figure 3 for the low noise regime and in Figure 4 for the high noise regime. We observe that, in all experiments, the nested algorithm consistently produces the smallest relative error. While the difference compared to the one-step approach is visually almost unnoticeable, the quantitative values show improvements of up to 10 percent. Additionally, the function values are lower in comparison, as seen in Figure 5. The results for the $32\times$ undersampling are particularly promising, as the Levenberg-Marquardt method was unable to produce meaningful results in this scenario. However, one downside of the algorithm is, that the subproblems in (73) become very ill-conditioned for small mesh sizes $h > 0$ and step-size parameters λ_k. This leads to convergence issues, forcing us to set $h = 1$ and to stop the algorithm after 100 iterations in most cases for time reasons, rather than waiting for the stopping criterion in line 19 of algorithm 4 to be met. More tailored algorithms could improve performance in this regard.

List of parameters for algorithm 4		
Parameter	Description	Value
$\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{R}_{>0}$	Regularization parameters for the sparsity of C.	$0.0045 * (1,1,1)$ $0.0095 * (1,1,1)$
$\lambda = (\lambda^1, \lambda^2, \lambda^3) \in \mathbb{R}_{>0}$	Stepsize parameters in eq. (73)	$45 * (1,1,1)$ $50 * (1,1,1)$
$(\lambda_D^k, \lambda_C^k) \in \mathbb{R}_{>0}$	Step size parameter for the dictionary learning sub-problem in algorithm 5.	$(1,1)$
$\gamma > 0$	Accuracy parameter for the nested algorithm 5. See also the input of algorithm 4.	0.75
$(\lambda_0, \tau, \sigma_3) \in \mathbb{R}_{>0}$	Parameters for the backtracking search in algorithm 4.	$(1, 8, 0.5)$
$(M_1, M_2, M_3) \in \mathbb{R}_{>0}$	Scaling parameters for the norm in (70).	$(100, 260, 260)$
$p \in \mathbb{N}$	Patch size of the squared $p \times p$ patches	8
$K \in \mathbb{N}$	Size of the orthogonal dictionary $K = p^2$	64
$h \in \mathbb{R}_{>0}$	Mesh-size for the differential operators	1

FIGURE 2. The list and the description of the different parameters in algorithm 4. The upper value corresponds to the experiment with $16\times$ undersampling factor and the lower value to the experiment with $32\times$ undersampling. If only one value is provided, we used this value for both experiments.

5. CONCLUSION AND OUTLOOK

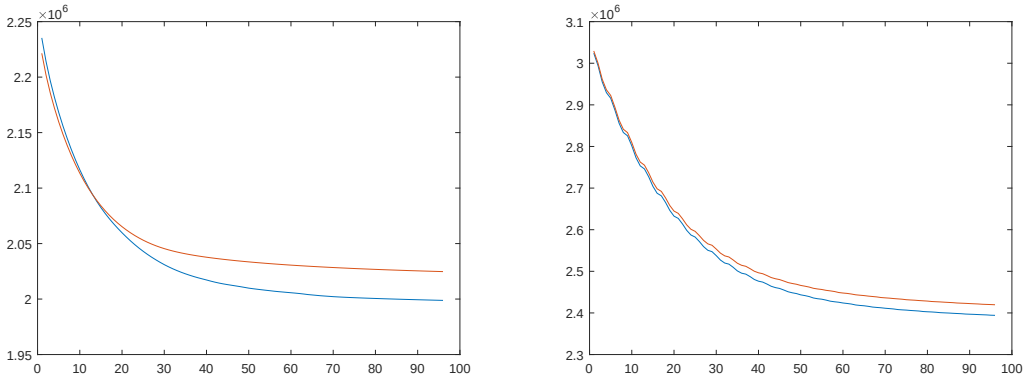
In this paper, we propose a dictionary learning-based regularization approach to solve a parameter identification problem within a general framework involving a time-discrete dynamical system as a physical prior. This is a nonlinear inverse problem, where dictionary learning is used to adapt the

Numerical results of the algorithm.				
	BLIP	Levenberg Marquardt	Dictionary learning (one step)	Dictionary learning (nested)
T_1	0.231	0.155	0.091	$\mathbf{0.086}$
T_2	0.26	0.177	0.09	$\mathbf{0.077}$
ρ	0.25	0.222	$\mathbf{0.12}$	$\mathbf{0.12}$

FIGURE 3. Results of the reconstruction algorithm for moderate $16\times$ undersampling and low noise.

Numerical results of the algorithm.				
	BLIP	Levenberg Marquardt	Dictionary learning (one step)	Dictionary learning (nested)
T_1	1.195	0.838	0.192	$\mathbf{0.184}$
T_2	0.733	0.4	0.204	$\mathbf{0.185}$
ρ	0.236	0.305	0.136	$\mathbf{0.134}$

FIGURE 4. Results of the reconstruction algorithm for moderate $32\times$ undersampling and higher noise.



(a) (b)

FIGURE 5. Comparison of function values over iteration number for the one-step approach (orange) and the nested optimization algorithm (blue). In (a), the graph corresponds to the $16\times$ undersampling regime, and in (b), to the $32\times$ undersampling. The nested method demonstrates a faster reduction in function values in both cases.

regularizer based on the given image data. Our primary motivation is to develop innovative and accurate methods for quantitative MRI, which leads to the minimization of a highly complex objective function. To address this large-scale, nonconvex, and nonsmooth optimization problem, we employed a nested alternating scheme and attempt to establish a theoretical foundation for better understanding the optimization process under regularity assumptions on the subdifferential. We demonstrate strong subsequential convergence and provide a result indicating that artificially high KL-exponents are required to maintain fast linear convergence of the algorithm. The theory does not align with classical block optimization frameworks, which are typically investigated under the assumption of a global KL inequality. In fact, block optimization schemes in the nonconvex regime are still not well understood. From an optimization standpoint, and with regard to convergence rates, most existing results rely on the use of the global KL-exponent, which is difficult to determine in many practical applications. Additionally, to the best of our knowledge, fast local convergence rates for block optimization schemes have not been well studied in the literature. For this reason, we sought in this work to examine the blocks separately and to investigate conditions on the dictionary learning problem that ensure overall fast local convergence.

Another remaining issue is that the choice of step size $\lambda_k > 0$ has not been thoroughly investigated. We anticipate that incorporating acceleration techniques, such as those developed by Nesterov, or more universal methods, could lead to faster convergence rates. Exploring these strategies could further improve the efficiency of the algorithm.

From inverse problems point of view, many questions related to regularization theory of the proposed formulation are open. For instance, the stability of the regularized solutions, and their convergence to the model solutions. This naturally raises the question of how to select the parameters λ and β in the numerical implementation, which is a critical issue but not addressed in the paper. This problem remains open even in the finite-dimensional case for linear inverse problems, and there are no available results for the infinite-dimensional, nonlinear context.

As a starting point, in the limit as $\delta \to 0$, which represents the noise-free case, in (P_0), we would formally end up with the following optimization problem

$$\min_{u,D} \|D^\top P \mathbb{D}^h u\|_{1,\gamma} \quad \text{s.t. } F(u) = f^0, \tag{78}$$

where $\|\cdot\|_{1,\gamma}$ is the Moreau-Yosida regularized ℓ_1-or Hubernorm. However, it remains unclear to what extent (P_0) is capable of identifying a good dictionary and producing a high-quality reconstruction,

even in the noise-free case. These so-called identifiability issues are challenging to analyze, even in the context of pure matrix factorization problems, as discussed in [32, 21, 36], and they appear to be unexplored in the context of nonlinear inverse problems. It is worth noting that blind-dictionary regularization shares several similarities with recently studied deep image priors, where instead of learning a dictionary, a neural network is trained jointly with the reconstruction process. Some recent recovery guarantees for this approach can be found in [18] and the references therein.

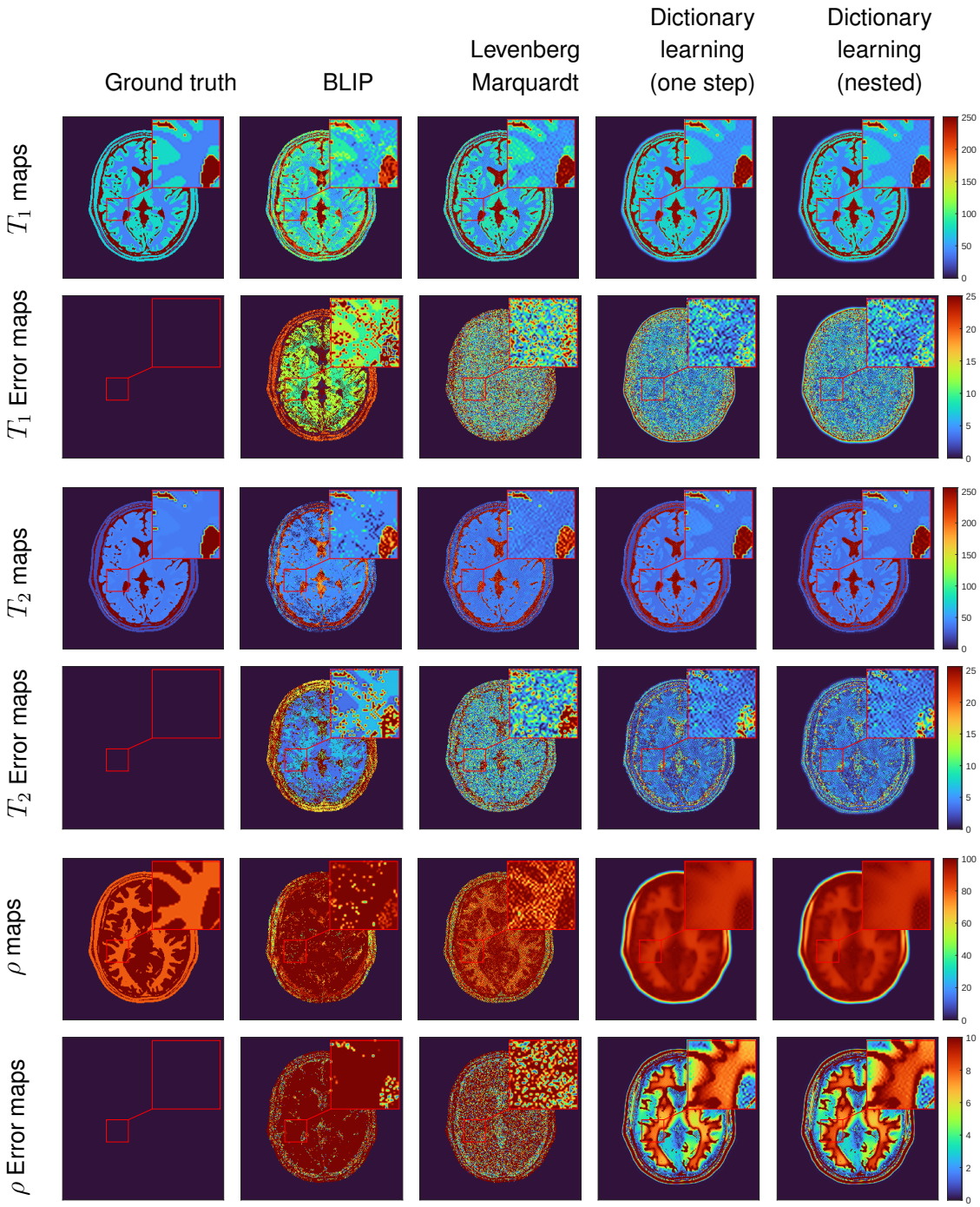


FIGURE 6. Comparison of the estimated parameters T_1 and T_2 (milliseconds) and proton density ρ (dimensionless; relative ratio). The figure compares the reconstruction quality of the BLIP method, the Levenberg-Marquardt approach from [24], and the dictionary learning approach proposed in this work for the $16\times$ undersampling case.

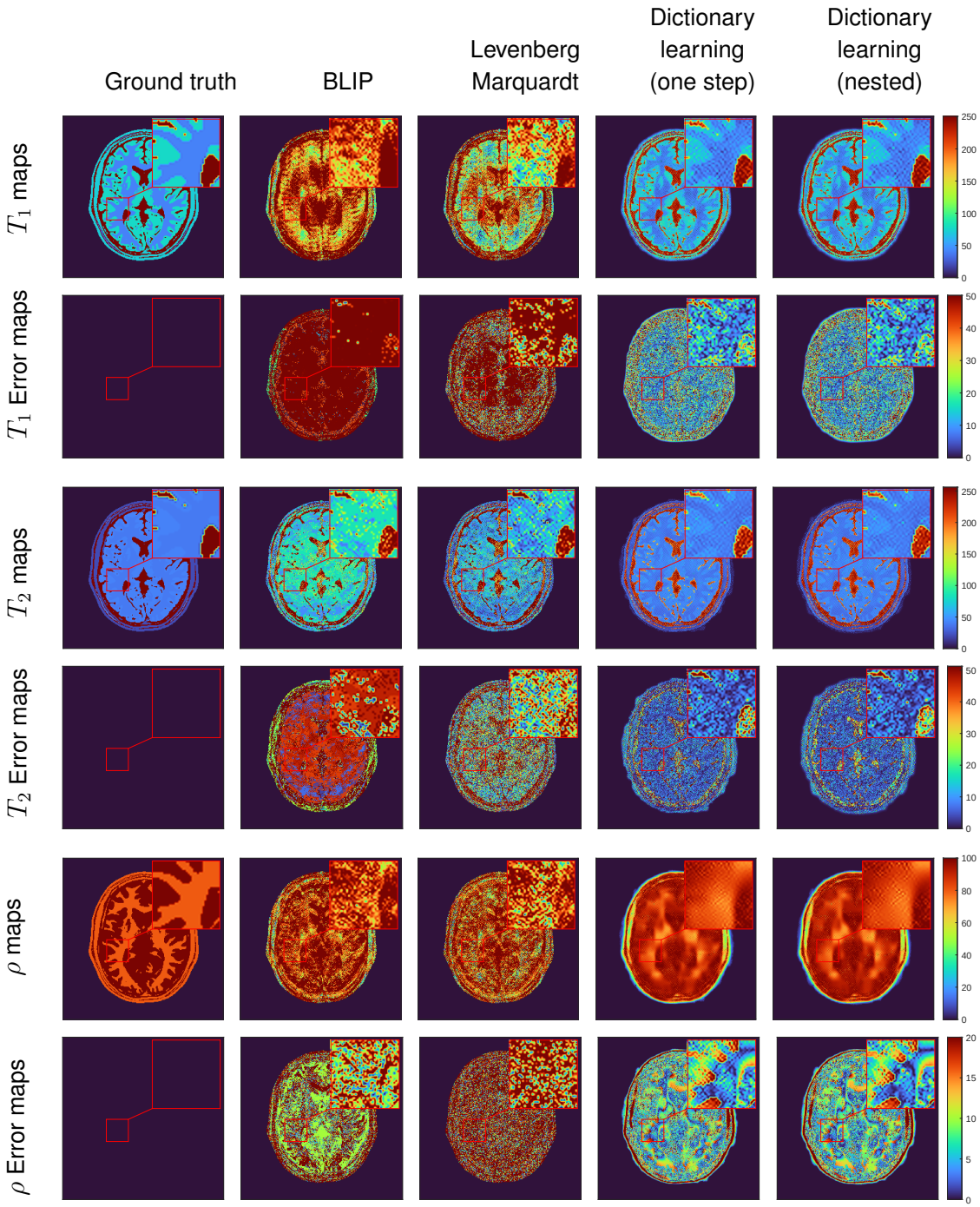


FIGURE 7. Comparison of the estimated parameters T_1 and T_2 (milliseconds) and proton density ρ (dimensionless; relative ratio). The figure compares the reconstruction quality of the BLIP method, the Levenberg-Marquardt approach from [24], and the dictionary learning approach proposed in this work for the $32\times$ undersampling case.

REFERENCES

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[2] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[3] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.
[4] Jürgen Appell and Petr P Zabrejko. Nonlinear superposition operators. *(No Title)*, 1990.
[5] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.
[6] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.
[7] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.
[8] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. SIAM, 2014.
[9] Heinz H Bauschke, Patrick L Combettes, Heinz H Bauschke, and Patrick L Combettes. *Correction to: convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.
[10] Amir Beck. *First-order methods in optimization*. SIAM, 2017.
[11] Dimitri P Bertsekas and Athena Scientific. *Nonlinear Programming*. Springer Science & Business Media, 2009.
[12] Felix Bloch. Nuclear induction. *Physical review*, 70(7-8):460, 1946.
[13] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
[14] Jérôme Bolte and Edouard Pauwels. Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41(2):442–465, 2016.
[15] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.
[16] Mary Ann Branch, Thomas F Coleman, and Yuying Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1):1–23, 1999.
[17] Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.
[18] Nathan Buskulic, Jalal Fadili, and Yvain Quéau. Convergence and recovery guarantees of unsupervised neural networks for inverse problems. *Journal of Mathematical Imaging and Vision*, pages 1–22, 2024.
[19] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76:167–188, 1997.
[20] Ralph Chill. On the łojasiewicz–simon gradient inequality. *Journal of Functional Analysis*, 201(2):572–601, 2003.
[21] Jeremy E Cohen and Nicolas Gillis. Identifiability of complete dictionary learning. *SIAM Journal on Mathematics of Data Science*, 1(3):518–536, 2019.
[22] Thomas F Coleman and Yuying Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–1058, 1996.
[23] Mike Davies, Gilles Puy, Pierre Vandergheynst, and Yves Wiaux. A compressed sensing framework for magnetic resonance fingerprinting. *Siam journal on imaging sciences*, 7(4):2623–2656, 2014.
[24] Guozhi Dong, Michael Hintermüller, and Kostas Papafitsoros. Quantitative magnetic resonance imaging: From fingerprinting to integrated physics-based models. *SIAM Journal on Imaging Sciences*, 12(2):927–971, 2019.
[25] Asen L Dontchev et al. *Lectures on variational analysis*, volume 205. Springer, 2021.
[26] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.
[27] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.
[28] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
[29] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.
[30] Martin Genzel, Jan Macdonald, and Maximilian März. Solving inverse problems with deep neural networks–robustness included? *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1119–1134, 2022.
[31] Helmuth Goldberg, Winfried Kampowsky, and Fredi Tröltzsch. On nemytskij operators in lp-spaces of abstract functions. *Mathematische Nachrichten*, 155(1):127–140, 1992.
[32] Rémi Gribonval and Karin Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

[33] Eyal Gur, Shoham Sabach, and Shimrit Shtern. Convergent nested alternating minimization algorithms for nonconvex optimization problems. *Mathematics of Operations Research*, 48(1):53–77, 2023.

[34] Eyal Gur, Shoham Sabach, and Shimrit Shtern. Nested alternating minimization with fista for non-convex and non-smooth optimization problems. *Journal of Optimization Theory and Applications*, 199(3):1130–1157, 2023.

[35] Daniel Hauer and José Mazón. Kurdyka–łojasiewicz–simon inequality for gradient flows in metric spaces. *Transactions of the American Mathematical Society*, 372(7):4917–4976, 2019.

[36] Jingzhou Hu and Kejun Huang. Global identifiability of $\backslash\ell_1$-based dictionary learning via matrix volume optimization. *Advances in Neural Information Processing Systems*, 36:36165–36186, 2023.

[37] Kazufumi Ito and Bangti Jin. *Inverse problems: Tikhonov theory and algorithms*, volume 22. World Scientific, 2014.

[38] Kazufumi Ito and Karl Kunisch. *Lagrange multiplier approach to variational problems and applications*. SIAM, 2008.

[39] Barbara Kaltenbacher, Andreas Neubauer, and Otmar Scherzer. *Iterative regularization methods for nonlinear ill-posed problems*. Walter de Gruyter, 2008.

[40] Andreas Kofler, Kirsten Miriam Kerkering, Laura Göschel, Ariane Fillmer, and Christoph Kolbitsch. Quantitative mr image reconstruction using parameter-specific dictionary learning with adaptive dictionary-size and sparsity-level choice. *IEEE Transactions on Biomedical Engineering*, 2023.

[41] Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

[42] Zhi-Pei Liang and Paul C Lauterbur. *Principles of magnetic resonance imaging*. SPIE Optical Engineering Press Bellingham, 2000.

[43] Dan Ma, Vikas Gulani, Nicole Seiberlich, Kecheng Liu, Jeffrey L Sunshine, Jeffrey L Duerk, and Mark A Griswold. Magnetic resonance fingerprinting. *Nature*, 495(7440):187–192, 2013.

[44] Stephane Mallat. *A wavelet tour of signal processing*. Academic Press, 1999.

[45] Boris S Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330. Springer Science & Business Media, 2006.

[46] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[47] Peter Ochs. *Long term motion analysis for object level grouping and nonsmooth optimization methods*. PhD thesis, PhD thesis, Albert-Ludwigs-Universität Freiburg, 2015.

[48] Duy Nhat Phan, Nicolas Gillis, et al. An inertial block majorization minimization framework for nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 24(18):1–41, 2023.

[49] Saiprasad Ravishankar and Yoram Bresler. Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE transactions on medical imaging*, 30(5):1028–1041, 2010.

[50] Saiprasad Ravishankar and Yoram Bresler. Learning sparsifying transforms. *IEEE Transactions on Signal Processing*, 61(5):1072–1086, 2012.

[51] Saiprasad Ravishankar and Yoram Bresler. Sparsifying transform learning for compressed sensing mri. In *2013 IEEE 10th international symposium on biomedical imaging*, pages 17–20. IEEE, 2013.

[52] Saiprasad Ravishankar and Yoram Bresler. Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging. *SIAM Journal on Imaging Sciences*, 8(4):2519–2557, 2015.

[53] Saiprasad Ravishankar, Jong Chul Ye, and Jeffrey A Fessler. Image reconstruction: From sparsity to data-adaptive methods and machine learning. *Proceedings of the IEEE*, 108(1):86–109, 2019.

[54] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[55] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[56] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, and Frank Lenzen. *Variational methods in imaging*, volume 167. Springer, 2009.

[57] Nick Scholand, Xiaoqing Wang, Volkert Roeloffs, Sebastian Rosenzweig, and Martin Uecker. Quantitative mri by nonlinear inversion of the bloch equations. *Magnetic Resonance in Medicine*, 90(2):520–538, 2023.

[58] Banafshe Shafieizargar, Riwaj Byanju, Jan Sijbers, Stefan Klein, Arnold J den Dekker, and Dirk HJ Poot. Systematic review of reconstruction techniques for accelerated quantitative mri. *Magnetic Resonance in Medicine*, 90(3):1172–1208, 2023.

[59] G.A. Wright. Magnetic resonance imaging. *IEEE Signal Processing Magazine*, 14(1):56–66, 1997.

[60] Gerd Wübbeler and Clemens Elster. A large-scale optimization method using a sparse approximation of the hessian for magnetic resonance fingerprinting. *SIAM Journal on Imaging Sciences*, 10(3):979–1004, 2017.

[61] Peiran Yu, Guoyin Li, and Ting Kei Pong. Kurdyka–łojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

[62] Felix F Zimmermann, Christoph Kolbitsch, Patrick Schuenke, and Andreas Kofler. Pinqi: an end-to-end physics-informed approach to learned quantitative mri reconstruction. *IEEE Transactions on Computational Imaging*, 2024.

APPENDIX: DICTIONARY LEARNING ROUTINE

Let us briefly describe the dictionary learning algorithm from [52]. Note that we need to apply this algorithm for any $u_i \in \{\rho, T_1, T_2\}$ separately. The algorithm described below aims at computing a stationary point of the objective

$$\min_{D\in O_K, C\in\mathbb{R}^{M\times K}} \frac{1}{2}\|DC - Pu\|_F^2 + \beta\|C\|_1 \tag{79}$$

for a given single image $u \in \mathbb{R}^{n_1\times n_2}$. It can be seen as a denoising algorithm for the patches collected in the matrix Pu.

Algorithm 5 Orthogonal dictionary learning by alternating optimization, [52]

1: Get $u \in \mathbb{R}^{n_1\times n_2}$, $(D_0, C_0) \in O_K \times \mathbb{R}^{M\times K}$, accuracy $\eta > 0$, sparsity regularization parameter $\beta > 0$ and step-size parameters $\lambda_D^n, \lambda_C^n > 0$, for $n \in \mathbb{N}$.
2: Set $n := 0$;
3: Initialize with $(D_n, C_n) := (D_0, C_0)$.
4: **while** no stopping criterion is satisfied **do**
5: Compute $D_{n+1} \in O_K$ by solving

$$D_{n+1} \in \operatorname*{arg\,min}_{D\in O_K} \frac{1}{2}\|DC_n - Pu\|_F^2 + \frac{\lambda_D^n}{2}\|D - D_n\|_F^2. \tag{80}$$

6: Compute $C_{n+1} \in \mathbb{R}^{M\times K}$ by solving

$$C_{n+1} \in \operatorname*{arg\,min}_{C\in\mathbb{R}^{M\times K}} \frac{1}{2}\|D_{n+1}C - Pu\|_F^2 + \frac{\lambda_C^n}{2}\|C - C_n\|_F^2 + \beta\|C\|_1. \tag{81}$$

7: Set $n = n + 1$.
8: **if** $\|D_n - D_{n-1}\|_F^2 + \|C_n - C_{n-1}\|_F^2 \le \eta_k^2$ **then**
9: stop the loop.
10: **end if**
11: **end while**
12: Return (D_n, C_n).

We collect the properties of algorithm 5 in the following lemma.

Lemma 5.1 (Properties of the dictionary learning algorithm)**.** *Consider algorithm 5. Then the following assertions are true:*

(i) *Problem* (80) *admits a closed form solution, which is given by*

$$D_{n+1} = UV^\top,$$

where $U\Sigma V^\top = (Pu)C_n^\top + \lambda_D^n D_n$ *is the singular value decomposition.*

(ii) *Problem* (81) *admits a closed form solution, which is given by*

$$C_{n+1} = \operatorname{prox}_{\beta_n\|\cdot\|_1}\left(\frac{D_{n+1}^\top Pu + \lambda_C^n C_n}{1+\lambda_C^n}\right), \quad \beta_n = \frac{\beta}{1+\lambda_C^n},$$

where the proximal operator $\operatorname{prox}_{\beta_n\|\cdot\|_1} : \mathbb{R}^{M\times K} \to \mathbb{R}^{M\times K}$ *is defined as the* soft-thresholding operator

$$[\operatorname{prox}_{\beta_n\|\cdot\|_1}(C)]_{i,j} = \begin{cases} C_{ij} - \beta_n & \text{if } C_{ij} \ge \beta \\ 0 & \text{if } -\beta_n \le C_{ij} \le \beta_n \\ C_{ij} + \beta_n & \text{if } C_{ij} \le \beta \end{cases}$$

(ii) *Let the sequences* $(\lambda_D^n, \lambda_C^n)_{n\in\mathbb{N}}$ *be bounded, i.e.* $a_D \leq \lambda_D^n \leq b_D$ *and* $a_C \leq \lambda_C^n \leq b_C$ *for* $0 < a_D, a_C, b_D, b_C$ *and all* $n \in \mathbb{N}$*. Then the sequence* $(D_n, C_n)_{n\in\mathbb{N}}$ *that is produced by algorithm 5 is a descent sequence for problem* (79) *in the sense of 3.6 with parameters* $\sigma_1 = \min(a_D, b_C)$ *and* $\sigma_2 = \max(L_C, b_C, b_D)$*, where*

$$L_C := \sup_{n\in\mathbb{N}} \|C_n\|_F < +\infty.$$

Proof. The proof of this statement is standard and we exclude it here for the sake of brevity. □