Jeffrey A. Fessler

Optimization Methods for Magnetic Resonance Image Reconstruction

Key models and optimization algorithms

©ISTOCKPHOTO.COM/GOODLIFESTUDIO

Digital Object Identifier 10.1109/MSP.2019.2943645
Date of current version: 15 January 2020

The development of compressed-sensing (CS) methods for magnetic resonance (MR) image reconstruction led to an explosion of research on models and optimization algorithms for MR imaging (MRI). Roughly 10 years after such methods first appeared in the MRI literature, the U.S. Food and Drug Administration (FDA) approved certain CS methods for commercial use, making CS a clinical success story for MRI. This article reports on several key models and optimization algorithms for MR image reconstruction. Included are both methods that the FDA has approved for clinical use and more recent methods being considered in the research community that use data-adaptive regularizers. It presents in a single survey the many algorithms devised to exploit the structure of the system model and regularizers used in MRI.

Introduction

Scope

Although the title of this article begins with "optimization methods," in practice, one first defines a model and cost function and then applies an optimization algorithm. There are several ways to partition the space of models, cost functions, and optimization methods for MRI reconstruction, such as smooth versus nonsmooth cost functions, static versus dynamic problems, and single- versus multiple-coil data. This article focuses on the static reconstruction problem because the dynamic case is rich enough to merit its own survey article. It emphasizes algorithms for multiple-coil data (parallel MRI) because modern systems all have multiple channels and advanced reconstruction methods with undersampling are most likely to be used for parallel MRI scans. The main families of parallel MRI methods include

- sensitivity encoding (SENSE) methods that model the coil sensitivities in the image domain [1]
- generalized autocalibrating partial parallel acquisition (GRAPPA) methods that model the effect of coil sensitivity in k-space [2]
- calibrationless methods that use low-rank or joint sparsity properties [3].

This article considers all three approaches, emphasizing SENSE methods for simplicity. (Figures in this article are reproduced in Jupyter notebooks with code in the open-source language Julia; these are available in the Michigan Image Reconstruction Toolbox at http://github.com/JeffFessler/MIRT.jl.)

Measurement model

The signals recorded by the sensors (receive coils) in MR scanners are linear functions of the object's transverse magnetization. That magnetization is a complicated and highly nonlinear function of the radio-frequency pulses, gradient waveforms, and tissue properties governed by the physics of the Bloch equation [4]. Quantifying tissue properties using nonlinear models is a rich topic of its own [5], but we focus here on the problem of reconstructing images of the transverse magnetization from MR measurements.

Ignoring noise, a vector $\boldsymbol{s} \in \mathbb{C}^M$ of signal samples recorded by an MR receive coil is related (typically) to a discretized version $\boldsymbol{x} \in \mathbb{C}^N$ of the transverse magnetization via a linear Fourier relationship:

$$\boldsymbol{s} = \boldsymbol{F}\boldsymbol{x}, \quad F_{ij} = \exp(-i2\pi \vec{\nu}_i \cdot \vec{\nu}_j), \quad \begin{matrix} i = 1, \dots, M \\ j = 1, \dots, N, \end{matrix} \tag{1}$$

where $\vec{\nu}_i$ denotes the k-space sample location of the ith sample (units: cycles/cm) and $\vec{\mathrm{x}}_j$ denotes the spatial coordinates of the center of the jth pixel (units: cm). In the usual case where the pixel coordinates $\{\vec{\mathrm{x}}_j\}$ and k-space sample locations $\{\vec{\nu}_i\}$ are both on appropriate Cartesian grids, matrix $\boldsymbol{F}$ is square, corresponding to the (2D or 3D) discrete Fourier transform. In this case $\boldsymbol{F}^{-1} = (1/N)\boldsymbol{F}'$, so reconstructing $\boldsymbol{x}$ from $\boldsymbol{s}$ is simply an inverse fast Fourier transform (FFT), and that approach is used in many clinical MR scans.

The reconstruction problem becomes more interesting when the k-space sample locations are on a non-Cartesian grid [6], when the scan is "accelerated" by recording $M < N$ samples, when non-Fourier effects like magnetic field inhomogeneity are considered [7] and/or when there are multiple receive coils. In parallel MRI, let $\boldsymbol{s}_l$ denote the samples recorded by the lth of L receive coils. Then one replaces the model (1) with

$$\boldsymbol{s}_l = \boldsymbol{F}\boldsymbol{C}_l\boldsymbol{x}, \tag{2}$$

where $\boldsymbol{C}_l$ is an $N \times N$ diagonal matrix containing the coil-sensitivity pattern of the l coil on its diagonal. Note that $\boldsymbol{F}$ does not depend on l; all coils see the same k-space sampling pattern. Stacking up the measurements from all coils and accounting for noise yields the following basic forward model in MRI:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_L \end{bmatrix} = \boldsymbol{y} = \underbrace{(\boldsymbol{I}_L \otimes \boldsymbol{F})\boldsymbol{C}}_{\boldsymbol{A}}\boldsymbol{x} + \boldsymbol{\varepsilon}, \quad \boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 \\ \vdots \\ \boldsymbol{C}_L \end{bmatrix}, \tag{3}$$

where $\boldsymbol{A} \in \mathbb{C}^{ML \times N}$ denotes the system matrix, $\boldsymbol{y} \in \mathbb{C}^{ML}$ denotes the measured k-space data, and $\boldsymbol{x} \in \mathbb{C}^N$ denotes the latent image. The noise in k-space is well modeled as complex white Gaussian noise. For extensions that consider other physics effects, such as relaxation and field inhomogeneity, see [8].

The goal in MR image reconstruction is to recover $\boldsymbol{x}$ from $\boldsymbol{y}$ using the model (3). All MR image reconstruction problems are underdetermined because the magnetization of the underlying object being scanned is a space-limited continuous-space function on $\mathbb{R}^3$, yet only a finite number of samples are recorded. Nevertheless, the convention in MRI is to treat the object as a finite-dimensional vector $\boldsymbol{x} \in \mathbb{C}^N$ for which the $M \geq N$ appropriate Cartesian k-space samples are considered "fully sampled" and any $M < N$ is considered "accelerated." Sampling-pattern design is a topic of ongoing interest with renewed interest in data-driven methods [9].

The matrix $\boldsymbol{F}$ in (3) is known prior to the scan because the k-space sample locations $\{\vec{\nu}_i\}$ are controlled by the pulse-sequence designer. In contrast, the coil-sensitivity maps $\{\boldsymbol{C}_l\}$ depend on the exact configuration of the receive coils for each patient. To use the model (3), one must determine the sensitivity maps from some patient-specific calibration data, e.g., by joint estimation [10], regularization [11], or subspace methods [12].

Cost functions and algorithms

Quadratic problems

When $ML \geq N$, i.e., when the total number of k-space samples acquired across all coils exceeds the number of unknown image pixel values, the linear model (3) is overdetermined. If, in addition, $\boldsymbol{A}$ is well conditioned, which depends on the sampling pattern and the coil-sensitivity maps, then it is reasonable to consider an ordinary least-squares (LS) estimator

$$\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{C}^N} \frac{1}{2} \| \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} \|_2^2 = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{y} \\ &= \left(\sum_{l=1}^{L} \boldsymbol{C}_l'\boldsymbol{F}'\boldsymbol{F}\boldsymbol{C}_l \right)^{-1} \left(\sum_{l=1}^{L} \boldsymbol{C}_l'\boldsymbol{F}'\boldsymbol{y} \right). \end{aligned} \tag{4}$$

In particular, for fully sampled Cartesian k-space data where $\boldsymbol{F}^{-1} = (1/N)\boldsymbol{F}'$, this LS solution simplifies to $\hat{\boldsymbol{x}} = (\Sigma_{l=1}^{L} \boldsymbol{C}_l'\boldsymbol{C}_l)^{-1} (\Sigma_{l=1}^{L} \boldsymbol{C}_l'\boldsymbol{F}^{-1}\boldsymbol{y})$, which is trivial to implement because each $\boldsymbol{C}_l$ is diagonal. This is known as the *optimal coil-combination approach* [13]. For regularly undersampled Cartesian data, where only every nth row of k-space is collected, the matrix $\boldsymbol{F}'\boldsymbol{F}$ has a simple block structure with $n \times n$ blocks. This structure facilitates noniterative blockwise computation known as *SENSE reconstruction* [1]. This form of LS estimation is widely used in clinical MR systems.

Regularized LS

For undersampled problems ($ML < N$) the LS solution (4) is not unique. Furthermore, even when $ML \geq N$, $\boldsymbol{A}$ is often poorly conditioned, particularly for non-Cartesian sampling. Some form of regularization is needed in such cases. Some early MRI reconstruction work used quadratically regularized cost functions leading to optimization problems of the form:

$$\hat{\boldsymbol{x}} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{C}^N} \frac{1}{2} \| \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} \|_2^2 + \beta \| \boldsymbol{T}\boldsymbol{x} \|_2^2, \tag{5}$$

where $\beta > 0$ denotes a regularization parameter and $\boldsymbol{T}$ denotes a $K \times N$ matrix transform, such as finite differences.

The conjugate gradient (CG) algorithm is well suited to such quadratic cost functions [6], [7]. The Hessian matrix $\boldsymbol{A}'\boldsymbol{A}+\beta\boldsymbol{T}'\boldsymbol{T}$ is often approximately Toeplitz [14], so a CG algorithm with circulant preconditioning is particularly effective. Although the quadratically regularized LS cost function (5) is passé in the CS era, CG is often an inner step for optimizing more complicated cost functions [15].

Edge-preserving regularization

The drawback of the quadratically regularized cost function (5) with $\boldsymbol{T}$ as finite differences is that it blurs image edges. To avoid this blur, one can replace the quadratic regularizer $\|\boldsymbol{T}\boldsymbol{x}\|_2^2$ with a nonquadratic function $\psi(\boldsymbol{T}\boldsymbol{x})$, where typically ψ is convex and smooth, such as the Huber function, a hyperbola, or the Fair potential function $\psi(z)=\delta^2(|z/\delta|-\log(1+|z/\delta|))$, among others as follows:

$$\hat{\boldsymbol{x}}=\underset{\boldsymbol{x}\in\mathbb{C}^N}{\operatorname{argmin}}\,\Psi(\boldsymbol{x})\triangleq\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2+\beta\psi(\boldsymbol{T}\boldsymbol{x}). \tag{6}$$

Such methods have their roots in Bayesian methods based on Markov random fields. The nonlinear CG algorithm is an effective optimization method for cost functions with such smooth edge-preserving regularizers. Another appropriate optimization algorithm is the optimized gradient method (OGM), a first-order method having optimal worst-case performance among all first-order algorithms for convex cost functions with Lipschitz continuous gradients [16]. The OGM has a convergence rate bound that is twice as good as that of Nesterov's fast gradient method [17]. Figure 1 compares two of these methods for the case where $\boldsymbol{T}$ represents finite differences and ψ is the Fair potential with $\delta=0.1$, which approximates total variation (TV) fairly closely while being smooth.

Sparsity models: Synthesis form

The scan time in MRI is proportional to the number of k-space samples recorded. Reducing scan time in MRI can reduce cost, improve patient comfort, and reduce motion artifacts. Reducing the number of k-space samples ML to well below N requires stronger modeling assumptions about $\boldsymbol{x}$, and sparsity models are prevalent [18], [19]. Two main categories of sparsity models are the synthesis approach and the analysis approach. In a synthesis model, one assumes $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{z}$ for some $N\times K$ matrix $\boldsymbol{B}$, where coefficient vector $\boldsymbol{z}\in\mathbb{C}^K$ should be sparse. In an analysis model, one assumes $\boldsymbol{T}\boldsymbol{x}$ is sparse, for some $K\times N$ transformation matrix $\boldsymbol{T}$.

A typical cost function for a synthesis model is

$$\hat{\boldsymbol{x}}=\boldsymbol{B}\hat{\boldsymbol{z}},\quad \hat{\boldsymbol{z}}=\underset{\boldsymbol{z}\in\mathbb{C}^K}{\operatorname{argmin}}\frac{1}{2}\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{z}-\boldsymbol{y}\|_2^2+\beta\|\boldsymbol{z}\|_1, \tag{7}$$

where the 1-norm is a convex relaxation of the ℓ_0 counting measure that encourages $\boldsymbol{z}$ to be sparse. Typically, $\boldsymbol{B}$ is a wide

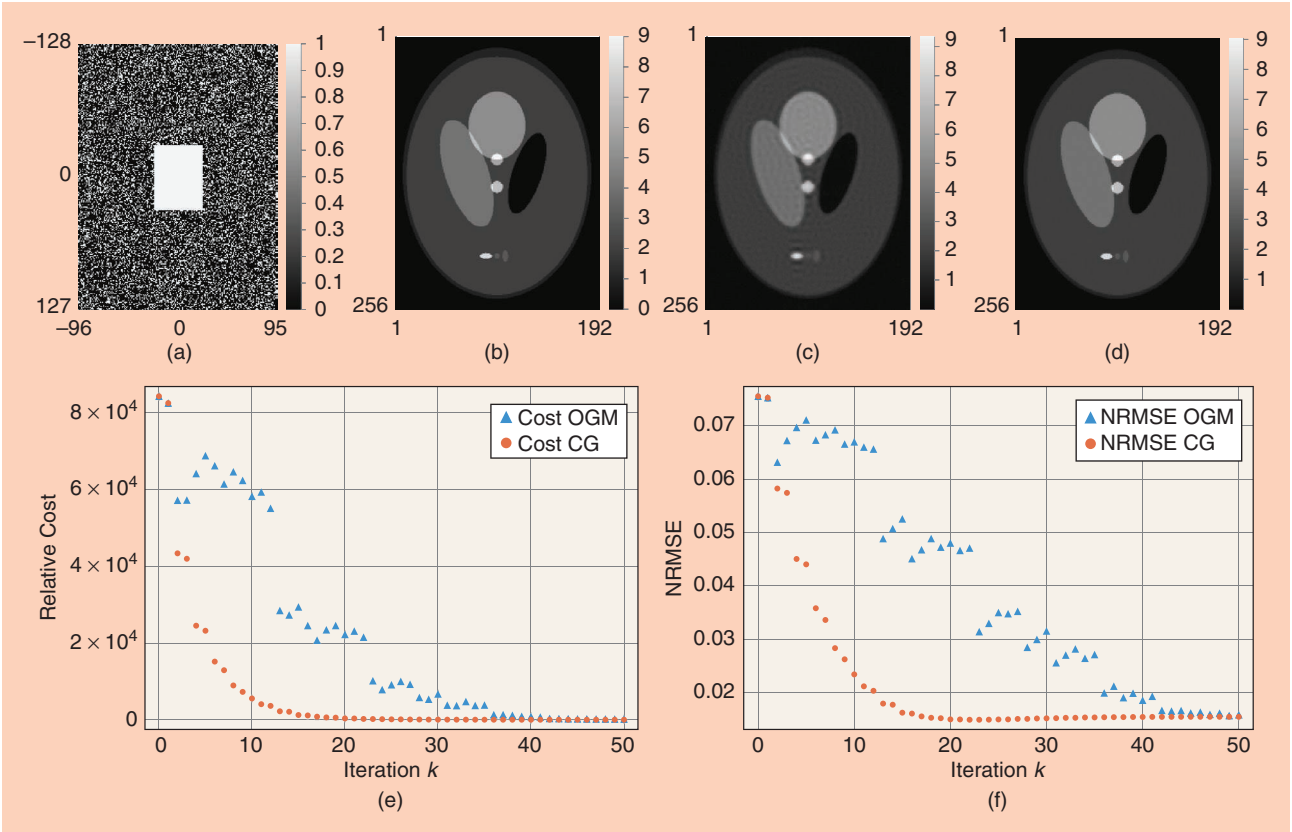

FIGURE 1. A comparison of CG and OGM convergence for single-coil MRI reconstruction with edge-preserving regularization (akin to anisotropic total variation with corner rounding). (a) A k-space sampling pattern where only 34% of the phase encodes are collected. (b) The true image. (c) The initial image from zero-filled k-space data. (d) The image with minimizer $\hat{\boldsymbol{x}}$ of (6) (both CG and OGM converge to the same limit $\hat{\boldsymbol{x}}$; CG: edge-preserving regularization). (e) A graph showing cost function $\Psi(\boldsymbol{x}_k)$ in (6). (f) A graph showing normalized root mean squared error (NRMSE) $\|\boldsymbol{x}_k-\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2$ versus iteration k.

matrix (often called an *overcomplete dictionary*) so that one can represent $\boldsymbol{x}$ well using only a fraction of the columns of $\boldsymbol{B}$. The classical approach for (7) is the iterative soft-thresholding algorithm (ISTA) [20], also known as the *proximal gradient method* (*PGM*) [21] and *proximal forward-backward splitting* [22], having the simple form

$$\boldsymbol{z}_{k+1} = \text{soft}(\boldsymbol{z}_k - \boldsymbol{D}^{-1}\boldsymbol{B}'\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{y}), \beta/\boldsymbol{d}), \tag{8}$$

where the soft-thresholding function is defined by $\text{soft}(z, c) = \text{sign}(z)\max(|z| - c, 0)$ and $\boldsymbol{D} = \text{diag}\{\boldsymbol{d}\}$ is any positive definite diagonal matrix such that $\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}$ is positive semi-definite [23].

The ISTA update (8) applies to the 1-norm in (7). If we replace that 1-norm with some other function $\psi(\boldsymbol{z})$, then one replaces (8) with the more general PGM update of the form

$$\boldsymbol{z}_{k+1} = \text{prox}_{\beta\psi}(\boldsymbol{z}_k - \boldsymbol{D}^{-1}\boldsymbol{B}'\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{y}), \beta/\boldsymbol{d}),$$

where the proximal operator of a function f is defined by

$$\text{prox}_f(\boldsymbol{v}) \triangleq \underset{\boldsymbol{x}}{\text{argmin}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + f(\boldsymbol{x}). \tag{9}$$

Traditionally $\boldsymbol{D} = \|\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\|_2\boldsymbol{I}$, but computing that spectral norm (via the power iteration) requires considerable computation for parallel MRI problems in general. However, for Cartesian sampling, $\boldsymbol{F}'\boldsymbol{F} \preceq N\boldsymbol{I}$ so it suffices to have $N\boldsymbol{B}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{B} \preceq \boldsymbol{D}$. Often the sensitivity maps are normalized such that $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}$, in which case $N\boldsymbol{B}'\boldsymbol{B} \preceq \boldsymbol{D}$ suffices. If, in addition, $\boldsymbol{B}'$ is a Parseval tight frame, then $\boldsymbol{B}'\boldsymbol{B} \preceq \boldsymbol{I}$, so using $\boldsymbol{D} = N\boldsymbol{I}$ is appropriate. For non-Cartesian sampling, nonnormalized sensitivity maps, or general choices of $\boldsymbol{B}$, finding $\boldsymbol{D}$ is more complicated [23].

Although ISTA is simple, it has an undesirably slow $O(1/k)$ convergence bound, where k denotes the number of iterations. This limitation was first overcome by the fast iterative soft-thresholding algorithm (FISTA) [24], also known as the *fast PGM* (*FPGM*), which has an $O(1/k^2)$ convergence bound. A recent extension is the proximal OGM (POGM), which has a worst-case convergence bound about twice as good as that of the FISTA/FPGM [25]. Both the FISTA and POGM are essentially as simple to implement as (8). Recent MRI studies have shown the POGM converging faster than the FISTA, as one would expect based on the convergence bounds [26], particularly when combined with adaptive restart [25]. So, the POGM (with restart) is a recommended method for optimization problems having the form (7). Figure 2 provides POGM pseudo-code for solving composite optimization problems like the MRI synthesis reconstruction model (7). Figure 3 shows that the POGM converges faster than the FISTA and ISTA for minimizing (7).

Initialize $\boldsymbol{\omega}_0 = \boldsymbol{x}_0$, $\theta_0 = 1$. Then for $k = 1$: N:

$$\theta_k = \begin{cases} \frac{1}{2}\left(1 + \sqrt{4\theta_{k-1}^2 + 1}\right), & k < N \\ \frac{1}{2}\left(1 + \sqrt{8\theta_{k-1}^2 + 1}\right), & k < N \end{cases}$$

$$\gamma_k = \frac{1}{L}\frac{2\theta_{k-1} + \theta_k - 1}{\theta_k}$$

$$\boldsymbol{\omega}_k = \boldsymbol{x}_{k-1} - \frac{1}{L}\nabla f(\boldsymbol{x}_{k-1})$$

$$\boldsymbol{z}_k = \boldsymbol{\omega}_k + \frac{\theta_{k-1} - 1}{\theta_k}(\boldsymbol{\omega}_k - \boldsymbol{\omega}_{k-1}) + \frac{\theta_{k-1}}{\theta_k}(\boldsymbol{\omega}_k - \boldsymbol{x}_{k-1}) + \frac{\theta_{k-1} - 1}{L\gamma_{k-1}\theta_k}(\boldsymbol{z}_{k-1} - \boldsymbol{x}_{k-1})$$

$$\boldsymbol{x}_k = \text{prox}_{\gamma_k g}(\boldsymbol{z}_k) = \underset{\boldsymbol{x}}{\text{argmin}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}_k\|_2^2 + \gamma_k g(\boldsymbol{x})$$

FIGURE 2. The POGM for minimizing $f(\boldsymbol{x}) + g(\boldsymbol{x})$, where f is convex with L-Lipschitz smooth gradient and g is convex. See [25] for the adaptive restart version.

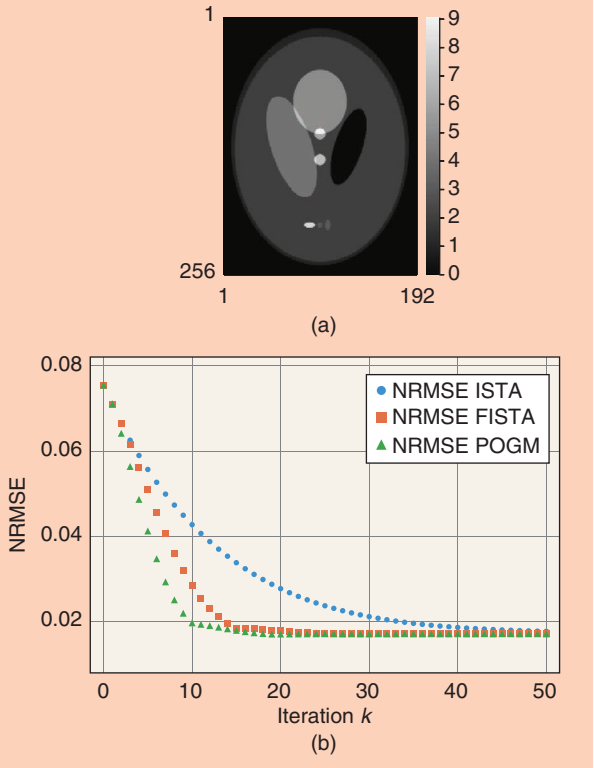

FIGURE 3. A comparison of the ISTA/PGM, FISTA/FPGM, and POGM for single-coil MRI reconstruction with orthogonal discrete wavelet transform (ODWT) sparsity regularizer using the 1-norm. (a) An image showing minimizer $\hat{\boldsymbol{x}}$ of (7) (the POGM with ODWT). (b) A graph showing NRMSE versus iteration k. The FISTA requires about 40% more iterations to converge than the POGM, which is consistent with the $2\times$ better worst-case bound of the POGM.

Sparsity models: Analysis form

A potential drawback of the synthesis formulation (7) is that $\boldsymbol{x} \approx \boldsymbol{B}\boldsymbol{z}$ may be a more realistic assumption than the strict equality $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{z}$ when $\boldsymbol{z}$ is sparse. The analysis approach avoids constraining $\hat{\boldsymbol{x}}$ to lie in any such subspace (or union of subspaces when $\boldsymbol{B}$ is wide). For an analysis-form sparsity model, a typical optimization problem involves a composite cost function consisting of the sum of a smooth term and a nonsmooth term:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\text{argmin}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \beta\|\boldsymbol{T}\boldsymbol{x}\|_1, \tag{10}$$

where $\boldsymbol{T}$ is a sparsifying operator, such as a wavelet transform, or represents finite differences, or both [18]. The expression (10) is general enough to handle combinations of multiple regularizers, such as wavelets and finite differences [19] by stacking the operators in $\boldsymbol{T}$ and possibly allowing a weight of 1-norm. When $\boldsymbol{T}$ represents finite differences, the regularizer is called *TV* [11], and combinations of TV and wavelet transforms are useful [19]. Although the details are proprietary, the FDA-approved method for CS MRI for at least one manufacturer is related to (10) [27].

When $\boldsymbol{T}$ is invertible, such as in an orthogonal wavelet transform, the optimization problem (10) can be rewritten as

$$\hat{\boldsymbol{x}} = \boldsymbol{T}^{-1}\hat{\boldsymbol{z}}, \quad \hat{\boldsymbol{z}} = \underset{\boldsymbol{z}}{\operatorname{argmin}} \frac{1}{2}\| \boldsymbol{A}\boldsymbol{T}^{-1}\boldsymbol{z} - \boldsymbol{y} \|_2^2 + \beta \| \boldsymbol{z} \|_1,$$

which is simply a special case of (7) with $\boldsymbol{B} = \boldsymbol{T}^{-1}$. Typically, $\boldsymbol{B}$ is wide, and $\boldsymbol{T}$ is tall, so this simplification is usually not possible.

In the general case (10) where $\boldsymbol{T}$ is not invertible, the optimization problem is much harder than (7) due to the nondifferentiability of the 1-norm with the matrix $\boldsymbol{T}$. The noninvertible case (with redundant Haar wavelets) is used clinically [27]. The PGM for (10) is

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{\mathcal{L}}{2}\| \boldsymbol{x} - \tilde{\boldsymbol{x}}_k \|_2^2 + \beta \| \boldsymbol{T}\boldsymbol{x} \|_1, \tag{11}$$

where $\tilde{\boldsymbol{x}}_k \triangleq \boldsymbol{x}_k - (1/\mathcal{L})\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{y})$ denotes the usual gradient update and the Lipschitz constant is $\mathcal{L} = \| \boldsymbol{A} \|_2^2$. Unfortunately, there is no simple solution for computing the proximal operator (9) in (11) in general, so inner iterative methods are required, typically involving dual formulations [24]. This challenge makes the PGM, FPGM, and POGM less attractive for (10) and has led to a vast literature on algorithms for problems like (10), with no consensus on what is best. The difficulty of (8) is the main drawback of analysis regularization, whereas a possible drawback of the synthesis regularization in (11) is that often $K \gg N$ for overcomplete $\boldsymbol{B}$.

Approximate methods

One popular "work-around" option is to "round the corner" of the 1-norm, making smooth approximations like $|z| \approx \sqrt{|z|^2 + \epsilon}$. This approximation is simply the hyperbola function, which has a long history in the edge-preserving regularization literature. All of the gradient-based algorithms mentioned for edge-preserving regularization are suitable candidates when a smooth function replaces the 1-norm. Smooth functions can shrink values toward zero, but their proximal operators never induce sparsity by setting many values exactly to zero. Whether a thresholding effect is truly essential is an open question. Hereafter we focus on methods that tackle the 1-norm directly without any such approximations.

Variable splitting methods

Variable splitting methods replace (10) with an exactly equivalent constrained minimization problem involving an auxiliary variable, such as $\boldsymbol{z} = \boldsymbol{T}\boldsymbol{x}$, e.g.,

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \min_{\boldsymbol{z}:\boldsymbol{z}=\boldsymbol{T}\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \beta\|\boldsymbol{z}\|_1. \tag{12}$$

This approach underlies the split Bregman algorithm [28], various augmented Lagrangian (AL) methods [29], and the alternating direction multiplier method (ADMM) [30]. The AL for (12) is

$$L(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \mu) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \beta\|\boldsymbol{z}\|_1 + \operatorname{real}\{\langle \boldsymbol{\gamma}, \boldsymbol{T}\boldsymbol{x} - \boldsymbol{z}\rangle\} + \frac{\mu}{2}\|\boldsymbol{T}\boldsymbol{x} - \boldsymbol{z}\|_2^2,$$

where $\gamma \in \mathbb{C}^K$ denotes the vector of Lagrange multipliers and $\mu > 0$ is an AL penalty parameter that affects the convergence rate but not the final image $\hat{\boldsymbol{x}}$. Defining the scaled dual variable $\boldsymbol{\eta} \triangleq 1/\mu\gamma$ and completing the square lead to the following scaled AL:

$$L(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\eta}, \mu) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \beta\|\boldsymbol{z}\|_1 + \frac{\mu}{2}(\|\boldsymbol{T}\boldsymbol{x} - \boldsymbol{z} + \boldsymbol{\eta}\|_2^2 - \|\boldsymbol{\eta}\|_2^2).$$

An AL approach alternates between descent updates of the primal variables $\boldsymbol{x}$, $\boldsymbol{z}$ and an ascent update of the scaled dual variable $\boldsymbol{\eta}$. The $\boldsymbol{z}$ update is simply soft thresholding:

$$\boldsymbol{z}_{k+1} = \operatorname{soft}(\boldsymbol{T}\boldsymbol{x}_k + \boldsymbol{\eta}_k, \beta/\mu).$$

The $\boldsymbol{x}$ update minimizes a quadratic function:

$$\boldsymbol{x}_{k+1} = (\boldsymbol{A}'\boldsymbol{A} + \mu\boldsymbol{T}'\boldsymbol{T})^{-1}(\boldsymbol{A}'\boldsymbol{y} + \mu\boldsymbol{T}'(\boldsymbol{z}_{k+1} + \boldsymbol{\eta}_k)).$$

A few CG iterations is a natural choice for approximating the $\boldsymbol{x}$ update. Finally, the $\boldsymbol{\eta}$ update is

$$\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k + (\boldsymbol{T}\boldsymbol{x}_{k+1} - \boldsymbol{z}_{k+1}).$$

The unit step size here ensures dual feasibility [31]. A drawback of variable splitting methods is the need to select the parameter μ. Adaptive methods have been proposed to help with this tuning [31]. One could apply the ADMM to the synthesis-regularized problem (7), though again it would require parameter tuning that is unnecessary with the POGM.

The conventional variable split in (12) ignores the specific structure of the MRI system matrix $\boldsymbol{A}$ in (3). Important properties of $\boldsymbol{A}$ include the fact that $\boldsymbol{F}'\boldsymbol{F}$ is circulant (for Cartesian sampling) or Toeplitz (for non-Cartesian sampling) and that each coil-sensitivity matrix $\boldsymbol{C}_l$ is diagonal. In contrast, the Gram matrix $\boldsymbol{A}'\boldsymbol{A}$ for parallel MRI is harder to precondition, though possible [32]. An alternative splitting that simplifies the updates is [29]:

$$\underset{\boldsymbol{x}\in\mathbb{C}^N}{\operatorname{argmin}} \min_{\boldsymbol{u}\in\mathbb{C}^{NL}, \boldsymbol{z}\in\mathbb{C}^K, \boldsymbol{v}\in\mathbb{C}^N} \frac{1}{2}\|\boldsymbol{F}_L\boldsymbol{u} - \boldsymbol{y}\|_2^2 + \beta\|\boldsymbol{z}\|_1 \quad \text{subject to } \boldsymbol{u} = \boldsymbol{C}\boldsymbol{x}, \boldsymbol{z} = \boldsymbol{T}\boldsymbol{v}, \boldsymbol{v} = \boldsymbol{x}, \tag{13}$$

where $\boldsymbol{F}_L \triangleq \boldsymbol{I}_L \otimes \boldsymbol{F}$. With this splitting, the $\boldsymbol{z}$ update again is simply soft thresholding, and the $\boldsymbol{x}$ update involves the diagonal matrix $\boldsymbol{C}'\boldsymbol{C}$, which is trivial. The $\boldsymbol{v}$ update involves the matrix $\boldsymbol{T}'\boldsymbol{T}$ that is circulant for periodic boundary conditions or is very well suited to a circulant preconditioner otherwise, using simple

FFT operations. The $\boldsymbol{u}$ update involves the matrix $\boldsymbol{F}_L'\boldsymbol{F}_L$ that is circulant or Toeplitz. This approach exploits the structure of $\boldsymbol{A}$ to simplify the updates; the primary drawback is that it requires selecting even more AL penalty parameters; condition number criteria can be helpful [29]. Another splitting with fewer auxiliary variables leads to an inner update step that requires solving denoising problems similar to (11).

Primal-dual methods

A key idea behind duality-based methods is the following:

$$\|\boldsymbol{T}\boldsymbol{x}\|_1 = \max_{\boldsymbol{z}\in\mathbb{C}^K:\|\boldsymbol{z}\|_\infty\le 1} \operatorname{real}\{\langle \boldsymbol{z}, \boldsymbol{T}\boldsymbol{x}\rangle\}.$$

Thus, the (nonsmooth) analysis-regularized problem (7) is equivalent to this constrained problem:

$$\operatorname*{argmin}_{\boldsymbol{x}} \min_{\boldsymbol{z}\in\mathcal{Z}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \beta \operatorname{real}\{\langle \boldsymbol{z}, \boldsymbol{T}\boldsymbol{x}\rangle\}, \tag{14}$$

where $\mathcal{Z} \triangleq \{\boldsymbol{z}\in C^K : \|\boldsymbol{z}\|_\infty \le 1\}$. The primal-dual methods typically alternate between updating the primal variable $\boldsymbol{x}$ and the dual variable $\boldsymbol{z}$, using more convenient alternatives to (14) that involve separate multiplication by $\boldsymbol{A}$ and by $\boldsymbol{A}'$ without requiring inner CG iterations. These methods provide convergence guarantees and acceleration techniques that lead to $O(1/k^2)$ rates [33]. A drawback of such methods is they typically require power iterations to find a Lipschitz constant, and, like AL methods, have tuning parameters that affect the practical convergence rates. Finding a simple, convergent, and tuning-free method for the analysis-regularized problem (10) remains an important open problem.

Patch-based sparsity models

Using (10) with a finite-difference regularizer is essentially equivalent to using patches of size 2×1. It is plausible that one can regularize better by considering larger patches that provide more context for distinguishing signal from noise. There are two primary modes of patch-based regularization: synthesis models and analysis methods.

A typical synthesis approach attempts to represent each patch using a sparse linear combination of atoms from some signal-patch dictionary. Let $\boldsymbol{P}_p$ denote the $d\times N$ matrix that extracts the pth of P patches (having d pixels) when multiplied by an image vector $\boldsymbol{x}$. Then the synthesis model is that $\boldsymbol{P}_p\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{z}_p$, where $\boldsymbol{D}$ is a $d\times J$ dictionary, such as the discrete cosine transform (DCT) [34], and $\boldsymbol{z}_p \in \mathbb{C}^J$ is a sparse coefficient vector for the pth patch. Under this model, a natural regularizer is

$$R(\boldsymbol{x}) = \min_{\{\boldsymbol{z}_p\}} \sum_{p=1}^{P} \frac{1}{2}\|\boldsymbol{P}_p\boldsymbol{x} - \boldsymbol{D}\boldsymbol{z}_p\|_2^2 + \alpha\|\boldsymbol{z}_p\|_1. \tag{15}$$

See [34] for an extension to the case of multiple images. The regularizer has an inner minimization over the sparse coefficients $\{\boldsymbol{z}_p\}$, so the overall problem involves both optimizing the image $\boldsymbol{x}$ and those coefficients. This structure lends itself to alternating minimization algorithms. The work in [34] used ISTA for updating $\boldsymbol{z}_p$; the results in Figure 3 suggest that the POGM may be beneficial.

A typical analysis approach for patches assumes there is a sparsifying transform $\boldsymbol{\Omega}$ such that $\boldsymbol{\Omega}\boldsymbol{P}_p\boldsymbol{x}$ tends to be sparse. For example, [35] uses a directional wavelet transform for each patch. Under this model, a natural regularizer is

$$R(\boldsymbol{x}) = \min_{\{\boldsymbol{z}_p\}} \sum_{p=1}^{P} \frac{1}{2}\|\boldsymbol{\Omega}\boldsymbol{P}_p\boldsymbol{x} - \boldsymbol{z}_p\|_2^2 + \alpha\|\boldsymbol{z}_p\|_1. \tag{16}$$

Again, a double minimization over the image $\boldsymbol{x}$ and the transform coefficients $\{\boldsymbol{z}_p\}$ is needed, so alternating minimization algorithms are natural. For alternating minimization (block coordinate descent), the update of each $\boldsymbol{z}_p$ is simply soft thresholding, and the update of $\boldsymbol{x}$ is a quadratic problem involving $\boldsymbol{A}'\boldsymbol{A} + \beta\sum_p \boldsymbol{P}_p'\boldsymbol{\Omega}'\boldsymbol{\Omega}\boldsymbol{P}_p$. When the transform $\boldsymbol{\Omega}$ is unitary and the patches are selected with periodic boundary conditions and a stride of one pixel, then this simplifies to $\boldsymbol{A}'\boldsymbol{A} + \beta\boldsymbol{I}$. A few inner iterations of the (preconditioned) CG algorithm is useful for the $\boldsymbol{x}$ update. Under these assumptions and using just a single gradient descent update for $\boldsymbol{x}$, an alternating minimization algorithm for LS with regularizer (16) simply alternates between a denoising step and a gradient step:

$$\begin{aligned}\tilde{\boldsymbol{x}}_k &= \sum_{p=1}^{P} \boldsymbol{P}_p'\boldsymbol{\Omega}' \operatorname{soft}(\boldsymbol{\Omega}\boldsymbol{P}_p\boldsymbol{x}_k, \alpha)\\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - (\boldsymbol{D} + \beta\boldsymbol{I})^{-1}(\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}) + \beta\tilde{\boldsymbol{x}}_k).\end{aligned} \tag{17}$$

For this algorithm, the cost function is monotonically nonincreasing.

Adaptive regularization

The patch dictionary $\boldsymbol{D}$ in (15) or the sparsifying transform $\boldsymbol{\Omega}$ in (16) can be chosen based on mathematical models like the DCT, or they can be learned from a population of preexisting training data and then used in (15) or (16) for subsequent patients. A third possibility is to adapt $\boldsymbol{D}$ or $\boldsymbol{\Omega}$ to each specific patient [36]. The dictionary-learning MRI approach [36] uses a nonconvex regularizer of the following form:

$$R(\boldsymbol{x}) = \min_{\boldsymbol{D}\in\mathcal{D}} \min_{\{\boldsymbol{z}_p\}} \sum_{p=1}^{P} \|\boldsymbol{P}_p\boldsymbol{x} - \boldsymbol{D}\boldsymbol{z}_p\|_2^2 + \alpha\|\boldsymbol{z}_p\|_1. \tag{18}$$

where $\mathcal{D}$ is the feasible set of dictionaries (typically constrained so that each atom has a unit norm). Now there are three sets of variables to optimize: $\boldsymbol{x}$, $\{\boldsymbol{z}_p\}$, and $\boldsymbol{D}$. This makes alternating minimization methods very suitable. The update of the image $\boldsymbol{x}$ is a quadratic optimization subproblem, the $\boldsymbol{z}_p$ update is soft thresholding, and the $\boldsymbol{D}$ update is simple when considering one atom at a time.

The transform-learning MRI approach uses a regularizer of this form:

$$R(\boldsymbol{x}) = \min_{\boldsymbol{\Omega}} \min_{\{\boldsymbol{z}_p\}} \sum_{p=1}^{P} \|\boldsymbol{\Omega}\boldsymbol{P}_p\boldsymbol{x} - \boldsymbol{z}_p\|_2^2 + \alpha\|\boldsymbol{z}_p\|_1 + \gamma r(\boldsymbol{\Omega}),$$

where $\gamma(\boldsymbol{\Omega})$ enforces or encourages properties of the sparsifying transform, such as orthogonality. Again, alternating minimization methods are well suited; the $\boldsymbol{\Omega}$ update involves (small) singular value decomposition operations.

Convolutional regularizers

An alternative to patch-based regularization is to use convolutional sparsity models [37]. A convolutional synthesis regularizer replaces (15) with

$$R(\boldsymbol{x}) = \min_{\{z_k\}} \frac{1}{2} \left\| \boldsymbol{x} - \sum_{k=1}^{K} \boldsymbol{h}_k * \boldsymbol{z}_k \right\|_2^2 + \alpha \| \boldsymbol{z}_k \|_1,$$

where $\{\boldsymbol{h}_k\}$ is a set of filters learned from training images [37] and * denotes convolution. Again, alternating minimization algorithms are a natural choice because the $\boldsymbol{x}$ update is quadratic and the $\boldsymbol{z}_k$ update is a sparse-coding problem for which proximal methods like the POGM are well suited.

A convolution analysis regularizer replaces (16) with

$$R(\boldsymbol{x}) = \min_{\{z_k\}} \sum_{k=1}^{K} \frac{1}{2} \| \boldsymbol{h}_k * \boldsymbol{x} - \boldsymbol{z}_k \|_2^2 + \alpha \| \boldsymbol{z}_k \|_1.$$

Again, alternating minimization algorithms are effective, where the $\boldsymbol{z}_k$ update is soft thresholding. One can either learn the filters $\{\boldsymbol{h}_k\}$ from good-quality (e.g., fully sampled) training data or adapt the filters for each patient by jointly optimizing $\boldsymbol{x}$, $\{\boldsymbol{h}_k\}$, and $\{\boldsymbol{z}_k\}$ using alternating minimization.

Other methods

The summation in (17) is a particular type of patch-based denoising of the current image estimate $\boldsymbol{x}_k$. There are many other denoising methods, some of which have variational formulations well suited to inverse problems, but many of which, such as nonlocal means [38] and block-matching 3D, do not [39]. One way to adapt most such denoising methods for image reconstruction is to use a plug-and-play ADMM approach [40] that replaces a denoising step like (17) with a general denoising procedure.

Non-SENSE methods

The measurement model (2) and (3) has a single latent image $\boldsymbol{x}$, viewed by each receive coil. An alternative formulation is to define a latent image for each coil $\boldsymbol{x}_l \triangleq \boldsymbol{C}_l \boldsymbol{x}$ and write the measurement model as $\boldsymbol{y}_l = \boldsymbol{F}\boldsymbol{x}_l + \boldsymbol{\varepsilon}_l$. For such formulations, the problem becomes to reconstruct the L images $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_L]$ from the measurements, while considering relationships between those images. Because multiplication by the smooth sensitivity map $\boldsymbol{C}_l$ in the image domain corresponds to convolution with a small kernel in the frequency domain, any point in k-space can be approximated by a linear combination of its neighbors in all coil data [2]. This GRAPPA modeling leads to an approximate consistency condition $\mathsf{vec}(\boldsymbol{X}) \approx \boldsymbol{G}\,\mathsf{vec}(\boldsymbol{X})$ where $\boldsymbol{G}$ is a matrix involving small k-space kernels learned from calibration data [2]. This relationship leads to spectral-aware Pareto iterative refinement optimization for supervised high-level synthesis (SPIRiT) optimization problems like the following:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X} \in \mathbb{C}^{N \times L}}{\operatorname{argmin}} \frac{1}{2} \| \boldsymbol{F}\boldsymbol{X} - \boldsymbol{Y} \|_{\text{Frob}}^2 + \beta_1 \frac{1}{2} \| (\boldsymbol{G} - \boldsymbol{I}) vec(\boldsymbol{X}) \|_2^2 + \beta_2 R(\boldsymbol{X}),$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_L] \in \mathbb{C}^{M \times L}$ and $R(\boldsymbol{X})$ is a regularizer that encourages joint sparsity because all of the images $\{\boldsymbol{x}_l\}$ have edges in the same locations. No sensitivity maps $\boldsymbol{C}$ are needed for this approach. When $\beta_2 = 0$, the problem is quadratic and CG is well suited. Otherwise, the ADMM is convenient for splitting this optimization problem into parts with easier updates. The ESPIRiT approach uses the redundancy in k-space data from multiple coils to estimate sensitivity maps from the eigenvectors of a certain block-Hankel matrix [12]; this approach helps bridge the SENSE and GRAPPA approaches while building on related signal processing tools like subspace estimation and multichannel blind deconvolution.

Conclusions

While the title of this article refers to optimization methods, it is far more important (for undersampled problems) before selecting an optimization algorithm to first select an appropriate cost function that captures useful prior information about the latent object $\boldsymbol{x}$. The literature is replete with numerous candidate models, each of which often leads to different optimization methods. Nevertheless, common ingredients arise in most formulations, such as alternating minimization (block coordinate descent) at the outer level, preconditioned CG for inner iterations related to quadratic terms, and soft thresholding or other proximal operators for nonsmooth terms that promote sparsity.

This survey has focused on 1-norm regularizers for simplicity, but (nonconvex) p "norms" with $0 \leq p < 1$ have also been investigated and appear to be beneficial, particularly for very undersampled measurements. This survey considers a single image $\boldsymbol{x}$, but many MRI scan protocols involve several images with different contrasts, and it may be useful to reconstruct them jointly, e.g., by considering common sparsity or subspace models.

There are many open problems in optimization that are relevant to MRI. The analysis form regularized problem (10) remains challenging, and further investigation of analysis versus synthesis approaches is needed. There has been considerable recent progress on finding optimal worst-case methods [16], but these optimality results are for very broad classes of cost functions, whereas the cost functions in MRI reconstruction have particular structure. Finding algorithms with optimal complexity (fastest possible convergence) for MRI-type cost functions would be valuable both for clinical practice and for facilitating research.

Finally, the current trend is to use convolutional neural network (CNN) methods to process undersampled images, to perform direct reconstruction, or to function as denoising operators. The stochastic gradient descent method currently is the universal optimization tool for training CNN models. Many "deep-learning" methods for MRI are based on network architectures that are "unrolled" versions of iterative optimization methods. like the PGM. Thus, familiarity with "classical" optimization methods for MR image reconstruction is important, even in the machine-learning era.

Acknowledgments

This work was supported in part by National Institutes of Health grants R01 EB023618, U01 EB026977, and R21 AG061839. I thank many students and postdoctoral academics for all they have taught me, Doug Noll for numerous MRI insights, and the reviewers for their detailed comments that improved the

article. For a longer version of this article with more text and references, see http://arxiv.org/abs/1903.03510.

Author

Jeffrey A. Fessler (fessler@umich.edu) received his B.S.E.E. degree from Purdue University in 1985, his M.S.E.E. degree from Stanford University in 1986, and his M.S. degree in statistics from Stanford University in 1989. He is the William L. Root Professor of Electrical Engineering and Computer Science at the University of Michigan and a professor in the Departments of Electrical Engineering and Computer Science, Radiology, and Biomedical Engineering. He became an IEEE Fellow in 2006, for contributions to the theory and practice of image reconstruction. He received the Francois Erbsmann Award for his IPMI93 presentation and the Edward Hoffman Medical Imaging Scientist Award in 2013. He has served as an associate editor of *IEEE Transactions on Medical Imaging* (*T-MI*), *IEEE Signal Processing Letters*, *IEEE Transactions on Image Processing*, and *IEEE Transactions on Computational Imaging*, and is currently serving as an associate editor of *SIAM Journal on Imaging Science*. His research interests are in statistical aspects of imaging problems, and he has supervised doctoral research in PET, SPECT, X-ray CT, MRI, and optical imaging problems.

References

[1] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Mag. Res. Med.*, vol. 42, no. 5, pp. 952–962, Nov. 1999.

[2] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Mag. Res. Med.*, vol. 47, no. 6, pp. 1202–1210, June 2002.

[3] P. J. Shin, P. E. Z. Larson, M. A. Ohliger, M. Elad, J. M. Pauly, D. B. Vigneron, and M. Lustig, "Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion," *Mag. Res. Med.*, vol. 72, no. 4, pp. 959–970, Oct. 2014.

[4] G. A. Wright, "Magnetic resonance imaging," *IEEE Sig. Proc. Mag.*, vol. 14, no. 1, pp. 56–66, Jan. 1997.

[5] D. Ma, V. Gulani, N. Seiberlich, K. Liu, J. L. Sunshine, J. L. Duerk, and M. A. Griswold, "Magnetic resonance fingerprinting," *Nature*, vol. 495, no. 7440, pp. 187–193, Mar. 2013.

[6] K. P. Pruessmann, M. Weiger, P. Boernert, and P. Boesiger, "Advances in sensitivity encoding with arbitrary k-space trajectories," *Mag. Res. Med.*, vol. 46, no. 4, pp. 638–651, Oct. 2001.

[7] B. P. Sutton, D. C. Noll, and J. A. Fessler, "Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 178–188, Feb. 2003.

[8] J. A. Fessler, "Model-based image reconstruction for MRI," *IEEE Sig. Proc. Mag.*, vol. 27, no. 4, pp. 81–89, July 2010.

[9] L. Baldassarre, Y.-H. Li, J. Scarlett, B. Gozcu, I. Bogunovic, and V. Cevher, "Learning-based compressive subsampling," *IEEE J. Sel. Top. Sig. Proc.*, vol. 10, no. 4, pp. 809–822, June 2016.

[10] L. Ying and J. Sheng, "Joint image reconstruction and sensitivity estimation in SENSE (JSENSE)," *Mag. Res. Med.*, vol. 57, no. 6, pp. 1196–1202, June 2007.

[11] K. T. Block, M. Uecker, and J. Frahm, "Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint," *Mag. Res. Med.*, vol. 57, no. 6, pp. 1086–1098, June 2007.

[12] M. Uecker, P. Lai, M. J. Murphy, P. Virtue, M. Elad, J. M. Pauly, S. S. Vasanawala, and M. Lustig, "ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA," *Mag. Res. Med.*, vol. 71, no. 3, pp. 990–1001, Mar. 2014.

[13] P. B. Roemer, W. A. Edelstein, C. E. Hayes, S. P. Souza, and O. M. Mueller, "The NMR phased array," *Mag. Res. Med.*, vol. 16, no. 2, pp. 192–225, Nov. 1990.

[14] F. Wajer and K. P. Pruessmann, "Major speedup of reconstruction for sensitivity encoding with arbitrary trajectories," in *Proc. Int. Soc. Magnetic Resonance Medicine*, 2001, p. 767. [Online]. Available: http://cds.ismrm.org/ismrm-2001/PDF3/0767.pdf

[15] J. Aelterman, H. Q. Luong, B. Goossens, A. Pizurica, and W. Philips, "Augmented Lagrangian based reconstruction of non-uniformly sub-Nyquist sampled MRI data," *Signal Process.,* vol. 91, no. 12, pp. 2731–2742, Jan. 2011.

[16] Y. Drori, "The exact information-based complexity of smooth convex minimization," *J. Complexity*, vol. 39, pp. 1–16, Apr. 2017.

[17] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O(1=k2)," *Soviet Math. Dokl*, vol. 27, no. 2, pp. 372–376, 1983. [Online]. Available: http://www.core.ucl.ac.be/~nesterov/Research/Papers/DAN83.pdf

[18] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Mag. Res. Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[19] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

[20] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[21] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bausche, R. S. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolowicz, Eds. New York: Springer, 2011, pp 185–212.

[22] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM J. Multi. Mod. Sim.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[23] M. J. Muckley, D. C. Noll, and J. A. Fessler, "Fast parallel MR image reconstruction via B1-based, adaptive restart, iterative soft thresholding algorithms (BARISTA)," *IEEE Trans. Med. Imag.*, vol. 34, no. 2, pp. 578–588, Feb. 2015.

[24] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, Nov. 2009.

[25] D. Kim and J. A. Fessler, "Adaptive restart of the optimized gradient method for convex optimization," *J. Optim. Theory Appl.*, vol. 178, no. 1, pp. 240–263, July 2018.

[26] L. El Gueddari, C. Lazarus, H. Carrie, A. Vignaud, and P. Ciuciu, "Self-calibrating nonlinear reconstruction algorithms for variable density sampling and parallel reception MRI," in *Proc. 10th Sensor Array and Multichannel Signal Processing Workshop*, 2018, pp. 415–419.

[27] J. Wetzl, C. Forman, B. J. Wintersperger, L. D'Errico, M. Schmidt, B. Mailhe, A. Maier, and A. F. Stalder, "High-resolution dynamic CE-MRA of the thorax enabled by iterative TWIST reconstruction," *Mag. Res. Med.*, vol. 77, no. 2, pp. 833–840, Feb. 2017.

[28] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[29] S. Ramani and J. A. Fessler, "Parallel MR image reconstruction using augmented Lagrangian methods," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 694–706, Mar. 2011.

[30] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programming*, vol. 55, no. 1–3, pp. 293–318, Apr. 1992.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[32] K. Koolstra, J. van Gemert, P. Boernert, A. Webb, and R. Remis, "Accelerating compressed sensing in parallel imaging reconstructions using an efficient circulant preconditioner for Cartesian trajectories," *Mag. Res. Med.*, vol. 81, no. 1, pp. 670–685, Jan. 2019.

[33] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Im. Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[34] S. Wang, S. Tan, Y. Gao, Q. Liu, L. Ying, T. Xiao, Y. Liu, X. Liu et al., "Learning joint-sparse codes for calibration-free parallel MR imaging (LINDBERG)," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 251–261, Jan. 2018.

[35] X. Qu, D. Guo, B. Ning, Y. Hou, Y. Lin, S. Cai, and Z. Chen, "Undersampled MRI reconstruction with patch-based directional wavelets," *Mag. Res. Im.*, vol. 30, no. 7, pp. 964–977, Sept. 2012.

[36] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1028–1041, May 2011.

[37] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[38] A. Buades, B. Coll, and J. Morel, "Image denoising methods. A new nonlocal principle," *SIAM Rev.*, vol. 52, no. 1, pp. 113–147, 2010.

[39] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[40] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Computat. Imaging*, vol. 3, no. 1, pp. 84–98, Mar. 2017.

SP