MRI Excitation Pulse Design and Image Reconstruction for Accelerated Neuroimaging

by

Tianrui Luo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biomedical Engineering)
in The University of Michigan
2021

Doctoral Committee:

Associate Research Scientist Jon-Fredrik Nielsen, Co-Chair
Professor Douglas C. Noll, Co-Chair
Professor Jeffrey A. Fessler
Associate Professor Nicole Seiberlich

Tianrui Luo

tianrluo@umich.edu

ORCID iD: 0000-0003-4770-530X

To my mom and dad, grandparents, and Xiaoxue

ACKNOWLEDGEMENTS

First and foremost, I want to thank my co-advisors, Dr. Jon-Fredrik Nielsen and Dr. Douglas C. Noll. I am extremely grateful for their supportiveness and patience with me. So many times I nervously showed up on our standing advising meetings with little research progress. They were never really mad but were always very understanding, helping me getting things working and moving forward. I was somewhat playful on doing research. For all but the inner volume project, the behind the scenes story was I started towards some directions, then got quickly distracted by what formed the chapters of this dissertation. Jon and Doug, with extraordinary tolerance, encouraged me on exploring and chasing my midway ideas. I must admit that with this atmosphere, I enjoyed doing research here in utmost.

I want to also thank Dr. Jeffrey A. Fessler. He was not officially my advisor, but was no less supportive, and has provided me so much advising. I learned a lot of interesting things about optimization and inverse problem from discussing with him and attending his group meeting, where I sneaked lots of food as I was both intellectually and physically hungry. He was very interested in my works, helped me making them rigorous and more impactful. Dr. Nicole Seiberlich also deserve my gratefulness for serving my dissertation committee, carefully reviewing my dissertation, and forwarding me opportunities. I also want to thank Dr. Emilie Chouzenoux for our discussion on the diagonal majorization problem of the Appendix. She has provided many feedbacks, and helped me formed a better understanding of the topic.

I had a wonderful time in the fMRI lab, and I want to thank the lovely researchers

and lab-mates here. Sydney Williams and Luis Hernandez-Garcia were the first who brought me into this lab, introduced me to MRI research. Hao Sun mentored me in so many aspects, from research all the way to investment, and has even continued guiding me years after his graduation. Jiabei Zheng, Anish Lahiri, Amos Cao, and Jonas Schollenberger, we discussed about various games, exchanged tons of memes, cheered up each other. Michelle Karker, and Melissa Haskell helped a lot on proof-reading my dissertation, fixing all sorts of issues. Shouchang Guo, Guanhua Wang, Dinank Gupta, Haowei Xiang, and Mariama Salifu helped me growing into a senior lab member; Mingjie Gao, Kathleen Ropella, Lianli Liu, Matt Muckley, Yash Shah, Gopal Nataraj, David Hong, Claire Lin, Naveen Murthy, Steven Whitaker, Cameron Blocker, and Caroline Crockett are all very delightful persons. We discussed a lot about research ideas, had lots of fun. I learned a lot about modeling, optimization, and probability from them.

Finally, I owe my deepest gratitude to my family. To my parents, and grand-parents, for so many years, I was only able to be with them for very limited days, despite how much they missed me. They cultivated my curiosities about the world, encouraged me to adventure further and further into the outside. To Xiaoxue Song, my wife, who brought me the brightest memories along these years, put me together when I was at my lowest. I would never be able to make it to this far without their unconditional love, understanding, and supports.

It has been a wonderful journey.

Tianrui Luo

Ann Arbor, Michigan

Dec. 22, 2020

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

Excitation pulse design and image reconstruction are two important topics in MR research for enabling faster imaging. On the pulse design side, selective excitations that confine signals to be within a small region-of-interest (ROI) instead of the full imaging field-of-view (FOV) can be used to reduce sampling density in the k-space, which is a direct outcome of the change in the underlying Nyquist sampling rate. On the reconstruction side, besides improving imaging algorithms' ability to restore images from less data, another objective is to reduce the reconstruction time, particularly for dynamic imaging applications.

This dissertation focuses on these two perspectives: Chapter II is devoted to the excitation pulse design. Specifically, we exploit auto-differentiation frameworks that automatically apply the chain rule on complicated computations. We derived and developed a computationally efficient Bloch-simulator and its explicit Bloch simulation Jacobian operations using such frameworks. This simulator can yield numerical derivatives with respect to pulse RF and gradient waveforms given arbitrary sub-differentiable excitation objective functions. The method does not rely on the small-tip approximation, and is accurate as long as the Bloch simulation can correctly model the spin movements due to the excitation pulses. In particular, we successfully applied this pulse design approach for jointly designing RF and gradient waveforms for 3D spatially tailored large-tip excitation objectives.

The auto-differentiable pulse design method can yield superior 3D spatially tailored excitation profiles that are useful for inner volume (IV) imaging, where one attempts to image a volumetric ROI at high spatiotemporal resolution without alias-

ing from signals outside the IV (i.e., outer volume). In Chapter III, we propose and develop a novel steady-state IV imaging strategy which suppresses aliasing by saturating the outer volume (OV) magnetizations via a 3D tailored OV excitation pulse that is followed by a signal crusher gradient. This saturation based strategy can substantially suppress the unwanted aliasing for common steady-state imaging sequences. By eliminating the outer volume signals, one can configure acquisitions for a reduced FOV to shorten the scanning time and increase spatiotemporal resolution for applications such as dynamic imaging.

In dynamic imaging (e.g., fMRI), where a time series is to be reconstructed, non-iterative reconstruction algorithms may offer savings in overall reconstruction time. Chapter IV focuses on non-iterative image reconstruction, specifically, extending the GRAPPA algorithm to general non-Cartesian acquisitions. We analyzed the formalism of conventional GRAPPA reconstruction coefficients, generalized it to non-Cartesian scenarios by using properties of the Fourier transform, and obtained an efficient non-Cartesian GRAPPA algorithm. The algorithm attains reconstruction quality that can rival classical iterative imaging methods such as conjugate gradient SENSE and SPIRiT.

In summary, this dissertation has proposed and developed multiple methods for accelerating MR imaging, from pulse design to reconstruction. While devoted to neuroimaging, the proposed methods are general and should also be useful for other applications.

CHAPTER I

Introduction

MRI is a relatively slow medical imaging modality compared to other techniques, such as CT and ultrasound. This makes the acceleration of imaging an essential topic in MRI research. Various approaches have been developed for this objective, including: selective signal excitation, where the Nyquist theorem enables an intrinsically lower sampling rate of k-space; parallel imaging, where signal redundancies from extra receive coils afford unaliased reconstructions from acquisitions that do not meet the Nyquist sampling rate; compressed sensing, where insightful signal priors, (e.g., low-rankness and sparsity), enable image retrieval from under-determined signal systems; and other approaches. In this work, we primarily develop methods and tools for selective signal excitation and parallel imaging.

1.1 Basics

This section aims to provide basic knowledge of: how magnetizations can be used to produce biomedical imaging signals; how a scanner materializes these signals with electric coils; and the coordinate frames for analyzing signals from spin magnetizations.

1.1.1 Physics and Signals

Signals in MRI are produced by electromagnetic induction: a population of hydrogen nuclear spin magnetizations precess coherently, such that, instead of cancelling each other out, the spins aggregate into a time-varying macroscopic magnetization. These spins induce signals in electric receive coils surrounding the population. Spin magnetizations are vectors in 3D space, $M = [M_{\mathsf{x}}, M_{\mathsf{y}}, M_{\mathsf{z}}]^{\mathsf{T}}$; as are 3D magnetic fields vectors $B = [B_{\mathsf{x}}, B_{\mathsf{y}}, B_{\mathsf{z}}]^{\mathsf{T}}$. To prepare such a spin population for imaging, the object of interest is first immersed into a homogeneous static magnetic field (B_0). Coherence forms as the spins within the object align with the direction of the static field direction, which is typically marked as the z-direction. These aligned spin magnetizations can be studied by exploiting their physical properties in magnetic fields and signalling.

Two key properties of spin magnetizations are Larmor precession and relaxation: M precesses about B towards the direction of $M \times B$ at frequency $\gamma\|B\|$, where γ, the gyromagnetic ratio, quantifies the linear relation between the angular frequency and the field magnitude. Meanwhile, the longitudinal part of magnetization recovers as a function of time (t), $M_{\mathsf{z}}(t) = M_0 - \exp(-t/T_1) \cdot (M_0 - M_{\mathsf{z}}(0))$; the transverse parts decay as, $[M_{\mathsf{x}}(t), M_{\mathsf{y}}(t)] = \exp(-t/T_2) \cdot [M_{\mathsf{x}}(0), M_{\mathsf{y}}(0)]$; where T_1 and T_2 are relaxation constants, and M_0 is the spin equilibrium magnitude at $t = \infty$. The Bloch equation summarizes the two properties in a differential form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} M_{\mathsf{x}} \\ M_{\mathsf{y}} \\ M_{\mathsf{z}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{T_2} & \gamma B_{\mathsf{z}} & -\gamma B_{\mathsf{y}} \\ -\gamma B_{\mathsf{z}} & -\frac{1}{T_2} & \gamma B_{\mathsf{x}} \\ \gamma B_{\mathsf{y}} & -\gamma B_{\mathsf{x}} & -\frac{1}{T_1} \end{bmatrix} \begin{bmatrix} M_{\mathsf{x}} \\ M_{\mathsf{y}} \\ M_{\mathsf{z}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{M_0}{T_1} \end{bmatrix}, \qquad [1.1]$$

or with cross product notation:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \gamma M \times B - [\frac{M_{\mathsf{x}}}{T_2}, \frac{M_{\mathsf{y}}}{T_2}, \frac{M_{\mathsf{z}} - M_0}{T_1}]^{\mathsf{T}}. \qquad [1.2]$$

When transverse magnetizations M_{x} and M_{y} are non-zero, their precession about B_0 provides the variation needed for inducing electromagnetic signals.

To yield transverse magnetizations from the spin population aligned with B_0, one introduces B_1, a second magnetic field perpendicular to the static one. The term "excitation" describes applying B_1 temporarily as a pulse. As characterized by the Bloch equation, this perpendicular B_1 will tip spins out of alignment with the z-direction. Then, signal reception starts.

1.1.2 MRI System Coils

Exploiting Faraday's law of induction, MRI systems use electrical coils to generate controllable magnetic fields. A system of superconductive coils with constant electric currents cycling parallelly are placed at the two ends of the scanner's longitudinal (z) direction, facing each other. The magnetic field generated by this coil system is relatively homogeneous in the middle, and is used as the B_0.

For imaging, to tell apart spins at different locations, we need them to act differently for signalling. MRI systems achieve this by adding to the B_0 field controllable, simple (e.g., linear) spatial inhomogeneities via gradient coils. There are generally three sets of such coils, corresponding to the x, y, and z axes, respectively.

The x coil set includes two groups of two coils facing the x-direction. The two coils within the same group are placed adjacently along direction z, but offset from the z-axis. Then, the two groups are symmetric about the z-axis in the middle of the scanner. Currents cycle oppositely in the same group of coils, and in parallel in each two symmetric coils, and their combined magnetic field generated by them varies spatially. Near the center of the four coils, the field directions are parallel to z, and the lateral variation resembles a linear gradient. Set-y is the same as set-x, except it faces the y-direction, and generates a y directional lateral gradient. For set-z, similar to B_0's setup, two z-direction facing coils are placed at the two ends

of the scanner. Their currents cycle oppositely, generating a gradient magnetic field along z direction. Currents in these gradient coils are manipulated to produce desired spatially dependent field inhomogeneities during imaging. The region for imaging is where all three gradient magnetic fields are spatially linear. The point where they each have a magnitude of 0 determines the origin of the three axes, named isocenter. These linear gradient fields' spatial slopes are typically denoted as G, a vector of three. Its inner product with an offset vector, r, relative to the isocenter yields the gradient field strength at the corresponding position.

Besides gradient coils, we need one more set of coils to apply the transverse B_1 and to excite the spins to produce signals. These are the RF coils, which are placed next to and facing the imaging region. For instance, a birdcage coil, which has superior RF homogeneity, is a single resonant system arranged as a ring surrounding the imaging region laterally. Their currents are synchronized, such that the transverse magnetic field they generate rotates at an angular frequency of $\gamma\|B_0\|$. B_1 rotating at this frequency is most effective at tipping spins. This phenomenon is more intuitive in a rotating coordinate frame.

The final components are MR signal receivers, coils that are laterally placed around the imaging region. Exploiting Faraday's law again, receive coils convert time-varying magnetic fields generated from spin magnetizations into time-varying electric currents, giving a dual role to the RF coils. On some systems, the RF coils are reused as receivers.

1.1.3 Coordinate Frames

A proper coordinate frame can significantly simplify analysis of object behaviors. For instance, our routine life on spinning earth would be too complicated if we used the solar system frame.

There are two important frames for MRI: the lab frame and the rotating frame.

The lab frame is the frame of the origin and axes described in the previous subsection 1.1.2. The rotating frame is defined relative to the lab frame: its origin and z-axes are the same as the latter, but the frame rotates about the z-axis at angular frequency $\gamma\|B_0\|$, identical to the precession frequency. Its x and y directions are oriented in cross product convention: direction of $\mathsf{x} \times \mathsf{y}$ is the same as direction z. The rotating coordinate system absorbs the intrinsic spin precession at frequency $\gamma\|B_0\|$. Consequently, B_0 is omitted in the frame, as its effect on spins is absorbed. We can thus focus on spin behaviors due to RF and gradients, etc.

In the rotating frame, relaxations are the same as in the lab frame, since both share the longitudinal (z) axis and the transverse (xy) plane. B_1 of frequency $\gamma\|B_0\|$ now points in a constant direction, causing spin magnetizations to simply rotate about it. Due to field inhomogeneities, not all hydrogen spins precess on-resonance at $\gamma\|B_0\|$ exactly. From the rotating frame perspective, these spins spontaneously precess about the z-axis in frequencies proportional to the deviation from the local field strength B_0. Thus, virtual magnetic fields of strengths equal to these differences are added to the rotating frame to explain the relative precession. Overall, by omitting B_0 and introducing virtual fields, the Bloch equation can still characterize spin behaviors in the rotating frame. Further, spin behaviors become simpler, which facilitates capturing intuition about them that are useful for controlled excitation.

1.2 Selective Excitation

Excitation is about customizing the excited spins for imaging objectives. The received signals are integrals of spin-induced signals distributed spatially across the object, just as macroscopic magnetizations are aggregated from spins. Conceptually, we divide the volume of the object to be imaged into voxels (i.e., volumetric pixels). Reconstruction, to finally obtain the object image, is to undo the integrals, to solve for the contribution of each voxel. Generally, larger volumes containing more voxels

require more signals for reconstruction. For medical imaging, the volume size is predetermined by receive coils, despite the fact that we are often interested in only part of it, e.g., a slice of a knee, a slab of a brain, etc. However, via selective excitations, to some extent, we can control which voxels are excited and contribute to induced signal. Thus, fewer acquisition samples are needed to solve for only the excited volume of interest, allowing faster imaging.

Excitation selectivity is achieved using pulses of designed RF, B_1, and gradients, G. These pulse lengths are often shorter than 10 ms, which is negligible compared to typical relaxation coefficients ($\gg 10$ ms). Ignoring relaxations, at time t, the pulse effect on spins in the rotating frame Bloch equation form is:

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = \gamma M(t) \times [B_{1\mathsf{x}}(t),\ B_{1\mathsf{y}}(t),\ G(t)^{\mathsf{T}} r + \frac{\Delta\omega(r)}{\gamma}]^{\mathsf{T}}, \qquad [1.3]$$

where $\Delta\omega(r)$ represents relative precession angular frequency due to local field inhomogeneity of B_0 imperfection at position r (off-resonance). It yields a virtual field strength $\frac{\Delta\omega(r)}{\gamma}$. Let T denote the duration of the pulse. The excitation result, $M(T)$, is the temporal integral of this differential equation, given initial state $M(0)$. Finding RF and gradient waveforms that yield satisfactory $M(T)$ for excitation objectives, e.g., selectivity, is the pulse design problem.

The pulse design problem is very challenging in general, due to the matrix differential equation. A classic simplification to the problem is the small-tip approximation: omit the off-resonance term for derivation brevity. When $M(t)$ does not deviate much from z-axis (small-tip) during excitation, M_{z} is approximately constant M_0. Then, [1.3] becomes,

$$\frac{\mathrm{d}}{\mathrm{d}t} M(t) = \gamma \begin{bmatrix} 0 & \langle G(t), r \rangle & -B_{1\mathsf{y}}(t) \\ -\langle G(t), r \rangle & 0 & B_{1\mathsf{x}}(t) \\ 0 & 0 & 0 \end{bmatrix} M(t),$$

as M_{z} is assumed to stay constant over time. This further simplifies to,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} M_{\mathsf{x}}(t) \\ M_{\mathsf{y}}(t) \end{bmatrix} = \gamma \begin{bmatrix} 0 & \left\langle G(t), r \right\rangle \\ -\left\langle G(t), r \right\rangle & 0 \end{bmatrix} \begin{bmatrix} M_{\mathsf{x}}(t) \\ M_{\mathsf{y}}(t) \end{bmatrix} + \gamma \begin{bmatrix} -B_{1\mathsf{y}}(t) \\ B_{1\mathsf{x}}(t) \end{bmatrix} M_0.$$

Let $M_{\mathsf{xy}} \coloneqq M_{\mathsf{x}} + \iota M_{\mathsf{y}}$. The simplification becomes,

$$\frac{\mathrm{d}}{\mathrm{d}t} M_{\mathsf{xy}}(t) = \iota\gamma(-\left\langle G(t), r \right\rangle M_{\mathsf{xy}} + B_1 M_0),$$

Using the integrating factor method of differential equation, this yields,

$$M_{\mathsf{xy}}(r, T) = \iota\gamma M_0 \int_{-\infty}^{T} B_1(t) e^{-\iota 2\pi \left\langle k(t), r \right\rangle} \,\mathrm{d}t, \tag{1.4}$$

where $k \coloneqq \frac{\gamma}{2\pi} \int_t^T G(\tau)\,\mathrm{d}\tau$. Conventionally, we refer to the domain hosting $k(t)$ as the excitation k-space. [1.4] establishes a Fourier transform relation between the small-tip excitation result and the pulse. Therefore, for small-tip excitation, inspecting the inverse Fourier transform of the desired selectivity profile can guide the design of pulse waveforms.

Once an excitation pulse selectively excites a region of interest, the spin population starts precessing and is ready for imaging.

1.3 Parallel Imaging

Raw MRI signals are complex-valued and modulated at the resonant spin precession frequency $\gamma\|B_0\|$. In this section, we assume demodulation has been applied. Also, to simplify signal analysis and focus on the underlying intuition, we further assume no spin relaxation during signal acquisition, and homogeneous B_0, so that spin precession differences are only caused by controlled linear gradient fields.

1.3.1 Signal Encoding and k-Space

Typical MRI signal encoding relies on the linear gradient system. Spatially linear gradient fields impose spatially linear variations of spin precession frequencies. These variations are proportional to gradient amplitudes: for two positions, r_1 and r_2, the frequency difference is $\gamma(r_1 - r_2)^\mathsf{T} G$. Hence, let $t = 0$ be the beginning of an acquisition, and τ be any time during that acquisition. The spin phase increment difference due to precession difference over time, $2\pi\gamma(r_1 - r_2)^\mathsf{T} \int_0^\tau G(t)\,\mathrm{d}t$, is proportional to the integral of gradients, and, most importantly, also spatially linear.

Consider the spin population complex image profile $p(r)$ at time $t = 0$, the phase increment at time τ becomes a spatially linear phase, $\phi_\tau(r) \coloneqq \exp(\iota 2\pi\gamma r^\mathsf{T} \int_0^\tau G(t)\,\mathrm{d}t)$, overlaid onto it. Consequently, the induction signal at $t = \tau$, expressed as an integral, is $\int p(r)\phi_\tau(r)\,\mathrm{d}r$, which is also an inner product between an image profile and a linear phase. In other words, from the Fourier transform perspective, the induction signal is the spectral value of $p(r)$ at spatial frequency, $k(\tau) \coloneqq \gamma \int_0^\tau G(t)\,\mathrm{d}t$. In MRI convention, the domain of these spectrum signals is addressed as the k-space. With knowledge of image processing, the MRI signal system can be spatially discretized and modelled as a Fourier transform:

$$\tilde{x} = Fx, \tag{1.5}$$

where, $F \in \mathbb{C}^{N_k \times N_x}$ is a Fourier transform matrix mapping images to the k-space acquired; $x \in \mathbb{C}^{N_x}$ and $\tilde{x} \in \mathbb{C}^{N_k}$ are discrete image and k-space signals, respectively. Assume a proper imaging FOV that fully contains the spin population of interest. A final image can be reconstructed from inverse Fourier transform of a Cartesian k-space dataset that satisfies the Nyquist sampling theorem.

1.3.2 Cartesian GRAPPA Parallel Imaging

Parallel MRI uses more than one receive coil to acquire data. Each of these coils is relatively more sensitive to spin inductions in a certain region. In other words, each coil sees a different coil image. Let $s_c(r)$ denote the spatial sensitivity profile of coil c. The induced signal this coil receives is thus $\int s_c(r)p(r)\phi_\tau(r)\,\mathrm{d}r$, which has the discrete space form:

$$\tilde{x}_c = F\,\mathsf{diag}(s_c)x, \quad c = 1, \dots C, \tag{1.6}$$

where $s_c \in \mathbb{C}^{N_x}$. A total number of C coils acquire C sets of $\tilde{x}_c$'s. Assume explicit or implicit knowledge of coil sensitivity profiles. These extra coils afford signal redundancies to solve [1.6] when $CN_k \geq N_x$, as N_x knowns generally suffice to solve N_x unknowns in a linear system. Parallel imaging exploits this redundancy to reduce the number of acquisitions, i.e., under-sampling, for imaging acceleration.

There are many parallel imaging techniques in MRI. For brevity, we focus on Cartesian GRAPPA, the one most relevant to this dissertation. GRAPPA does not directly reconstruct the image x from under-sampled k-space data. Instead, it attempts to estimate un-sampled k-space data by linearly combining its sampled neighboring data across coils using certain weights, w_c. In this way, Cartesian GRAPPA restores full k-space data for all coils, and obtains coil images via inverse Fourier transform. It composes final images by combining these coil images.

The way GRAPPA restores full k-spaces reveals that there are shift-invariant linear relations among k-space neighboring signals. The math behind this relies on Fourier transform properties. Here, we assume single coil "parallel" imaging for derivation brevity. This analysis extends easily to multi-coil scenarios: let Cartesian Fourier transform matrix F be unitary. Equation [1.6], by inserting $I = F^{\mathsf{H}}F$, becomes,

$$\tilde{x}_c = F\,\mathsf{diag}(s_c)F^{\mathsf{H}}Fx = \tilde{S}_c^{\mathsf{H}}\tilde{x}, \tag{1.7}$$

where $\tilde{S}_c$ is a circulant matrix whose columns are shifted duplicates of $\tilde{s}_c \coloneqq Fs_c$. In other words, a set of neighboring signals in $\tilde{x}_c$ are inner products between $\tilde{x}$ and a set of shifted copies of $\tilde{s}_c$. Let the un-sampled signal in the set correspond to shift copy $\tilde{s}_{c,0}$, and assume the set contains N_N sampled signals, corresponding to shift copies $\tilde{S}_{c,N_a} \coloneqq [\tilde{s}_{c,1}, \ldots, \tilde{s}_{c,N_a}] \in \mathbb{C}^{N_x \times N_a}$. Thus, the estimation error of the un-sampled signal is, $\tilde{x}^{\mathsf{H}}(\tilde{s}_{c,0} - \tilde{S}_{c,N_a} w_c)$. Fixing the geometry of neighboring signals, GRAPPA's shift-invariance is inherited from the shift-invariance of error vector, $\tilde{e}_c \coloneqq \tilde{s}_{c,0} - \tilde{S}_{c,N_a} w_c$. Cartesian GRAPPA is reliable for parallel imaging, as long as $\|\tilde{e}_c\|$ is negligible.

1.4 Outline and Contributions

Concluding this chapter, the rest of this dissertation is described, largely based on the introduction sections above:

Chapter II was published as a pre-print in [38]. This chapter elaborates on our novel MRI excitation pulse design approach that is based on increasingly popular auto-differentiation frameworks. Conventional pulse designs are limited, especially for multidimensional excitation, etc. In our proposed approach, we derived and developed an efficient auto-differentiable Bloch simulator. Via simulating excitation results of a pulse, it yields numerical derivatives with respect to the RF and gradient waveforms to be optimized given arbitrary pulse design loss functions. We successfully applied it on large-tip excitation designs, which are difficult to address using conventional pulse design methods.

Chapter III was based on [39, 40]. The work in this chapter proposes a novel approach to improve 3D inner volume imaging quality for steady state sequences. Existing inner volume imaging methods are impractical for steady state sequences due to unsuppressed outer volume signals. We propose to apply a recently developed method [72] on 3D spatially tailored outer volume excitation pulse design. Such pulses, combined with an immediate crusher gradient, excite and saturate the mag-

netization outside a volume-of-interest (VOI). In this way, we clear aliasing artifacts in under-sampled steady state inner volume imaging.

Chapter IV was published in [37]. In dynamic imaging, where a time-course of images are to be reconstructed, GRAPPA, a non-iterative fast reconstruction algorithm, has so far been limited to Cartesian acquisitions. We proposed a computationally efficient generalization of the conventional GRAPPA algorithm that can be applied to arbitrary non-Cartesian MRI acquisitions. Like GRAPPA, this generalization restores images non-iteratively, which can be particularly useful in saving reconstruction time for dynamic imaging. It can rival common iterative parallel imaging algorithms (e.g., cg-SENSE, SPIRiT [43, 56]) in reconstruction quality.

Chapter V summarizes the above chapters and discusses possible future extensions and directions.

Besides these MRI related works, Appendix A focuses on numerical optimization. This appendix identifies the problem of finding the optimal diagonal majorizer for the least square problem as a constrained convex semi-definite programming problem. It also derives the derivative of the diagonal majorizer, which can be useful for gradient descent algorithms for finding the optimal majorizer.

CHAPTER II

Joint Design of RF and Gradient Waveforms via Auto-differentiation for 3D Tailored Excitation in MRI

2.1 Introduction

In a magnetic resonance imaging (MRI) experiment, the dynamical system relationship between the applied radiofrequency (RF) and gradient magnetic fields, and the instantaneous spin magnetization change they induce, is concisely described by the Bloch equation[1]. While it is straightforward to calculate the magnetization pattern resulting from a given set of RF and gradient waveforms and tissue parameters, *inverting* the Bloch equation to obtain the waveforms that produce a given desired excitation pattern can be challenging.

This Bloch inversion task is conventionally called an "RF pulse design" problem, reflecting the fact that the most common way to design excitation pulses in MRI is to pre-define the gradients in some way, and then only optimize the (complex) RF waveform. Even with that simplification, the design problem remains non-linear and non-convex. Another common simplification is therefore to apply the small-tip approximation [52] that can give reasonable excitation accuracy even for flip angles

[1]This chapter was published as a pre-print in [38].

as high as 90°, at least for conventional 1D (slice-selective) excitations where the instantaneous flip angle during RF excitation remains relatively low. The small-tip approximation leads to a linear (Fourier) relationship between applied fields and the resulting magnetization pattern, and provides intuition about the excitation process by defining an "excitation k-space" trajectory and viewing RF transmission as depositing energy along that trajectory.

The more difficult problem of *jointly* optimizing both RF and gradient waveforms has been approached in various ways. Several methods are based on the small-tip approximation, and on optimizing the gradients over a restricted set of waveform shapes, such as "spoke" or "kt-point" locations in excitation k-space [9, 19, 44, 81, 82] or parameterized echo-planar or non-Cartesian trajectories [14, 26, 46, 76, 80]. A more general small-tip design approach for 3D tailored excitation was described in [72], based on a B-spline parametrization of the gradient trajectory that is not restricted to particular fixed waveform shapes. These approaches work well for small-tip excitations, but not for applications such as tailored saturation or inversion. In addition, even when the final desired flip angle is small, the *instantaneous* flip angle during RF excitation can be large and thus can violate the small-tip assumption [71]. This model mismatch can result in noticeable differences between the Bloch-simulated excitation pattern and that predicted by the small-tip model used in the design.

Another limitation of previous approaches is that the design loss functions are typically limited to certain forms such as least squares (LS) based on the complex transverse excited magnetization, although adaptations to magnitude least squares (MLS) costs have been proposed [67]. Adding hardware constraints to the design formulation adds an additional layer of complexity that is often either ignored during pulse design, or controlled indirectly via, e.g., Tikhonov regularization of the RF waveform [19].

This work[2] approaches the Bloch inversion task in a more direct and general way that is applicable to the joint design of RF and gradient waveforms for tailored multi-dimensional excitation in MRI. We temporally discretize the pulse, assuming piecewise constant gradient and RF within every time segment. Our method does not rely on the small-tip approximation, works for arbitrary sub-differentiable loss functions, and incorporates hardware constraints. Our approach contains three key elements: First, we derive analytic expressions for the Jacobian operations needed for the Bloch inversion for a unit (discrete) time step. Second, we incorporate these discrete-time Jacobian operations into an automatic differentiation framework [51], to obtain the Jacobian that relates the final magnetization pattern (at the end of the pulse) to the RF and gradient waveforms. Third, we enforce hardware limits by a change of variables that makes the optimization problem effectively unconstrained.

The paper is organized as follows. Section 2.2 gives a general form of the joint design problem, and derives the explicit Jacobians useful for accelerating the proposed auto-differentiation pulse design tools. In sections 2.3 and 2.4 we apply our pulse design tool to two large-tip excitation problems, and validate the results experimentally on a 3T MRI scanner. Sections 2.5 and 2.6 discuss and conclude this work.

2.2 Theory

2.2.1 Problem Formation

We discretize 3D space on a regular grid with a total n_M voxels ("spins"). These spins can have different parameters, e.g., T1, T2, and off-resonance. Let n_T denote the length (number of time points) of the pulse to be designed. For (single coil) joint design of blue complex RF waveform $b \in \mathbb{C}^{n_T}$ and gradient $g \in \mathbb{R}^{n_T \times 3}$ we are

[2]Open sourced, github.com/tianrluo/AutoDiffPulses

interested in tackling the following general problem:

$$\begin{aligned} \underset{g\in\mathbb{R}^{n_T\times 3},\, b\in\mathbb{C}^{n_T}}{\arg\min} \quad & \mathcal{L} \coloneqq f(M_T(g,b), M_D) + \lambda\mathcal{R} \\ \text{s. t.} \quad & \|b\|_\infty \le b_{\max} \\ & \|g\|_{\infty,\infty} \le g_{\max} \\ & \|\mathsf{D}g\|_{\infty,\infty} \le s_{\max}, \end{aligned} \tag{2.1}$$

where $\mathcal{L}$ is the loss function; $M_D \in \mathbb{R}^{n_M\times 3}$ is the target (Desired) magnetization pattern (a 3-dimensional magnetization vector at each spatial location); $M_T \in \mathbb{R}^{n_M\times 3}$ is the magnetization at the end of the pulse (time T) obtained by integrating the Bloch equation; f is the excitation error metric (e.g., a common choice is least-square error of transverse magnetization, i.e., $\|M_T[:,1{:}2] - M_D[:,1{:}2]\|_F^2$); and $\mathcal{R}$ is an optional regularizer with weight λ (a common choice is $\mathcal{R} = \|b\|_2^2$ to control peak RF amplitudes and SAR indirectly). For the constraints, we have $b_{\max}$, $g_{\max}$, and $s_{\max}$ for peak RF, gradient, and slew rate, respectively; $\mathsf{D} \in \mathbb{R}^{n_T\times n_T}$ is the temporal difference matrix divided by δ_t, i.e., $\mathsf{D}g$ takes the 1st order temporal derivative of g and yields the slew rate; and $\|\cdot\|_\infty$, and $\|\cdot\|_{\infty,\infty}$ are entry-wise norm returning the largest absolute value of the operand elements.

Problem [2.1] is challenging for two main reasons: First, the objective is non-convex with respect to its arguments, and is constrained. Second, neither $M_T(g,b)$, nor its Jacobians $\partial M_T/\partial g$ and $\partial M_T/\partial b$ that would be needed to directly minimize [2.1], have an explicit expression in g and b. To the best of our knowledge, existing methods all deal with simplifications of problem [2.1] based on, e.g., the small-tip, or spin domain models. In this work, we minimize [2.1] directly, and assume only that the temporal integration of the Bloch equation is well-approximated by a discrete-time Bloch simulator.

2.2.2 Auto-Differentiation

We propose to compute the necessary derivatives[3] using auto-differentiation [18], such that problem [2.1] can be optimized for *arbitrary* error metric f and regularization $\mathcal{R}$. Auto-differentiation tools, e.g., PyTorch [51], decouples computations into stages, and constructs the Jacobian operations at each stage. These single-stage Jacobians are eventually combined using the chain rule. For instance, with a PyTorch based Bloch simulator that computes $M_T(g, b)$, one implicitly obtains $\partial M_T/\partial g$ and $\partial M_T/\partial b$. The loss derivatives with respect to the variables we wish to optimize, i.e., $\partial \mathcal{L}/\partial g$, and $\partial \mathcal{L}/\partial b$, can then be obtained by combining these expressions with $\partial \mathcal{L}/\partial M_T$. This approach allows us to directly optimize g and b with respect to arbitrary losses.

2.2.3 Explicit Jacobian Operations

Auto-differentiation tools provide implicit Jacobian operations (also known as the default backward operations in auto-differentiation context) formed from tracking all elementary computations (e.g., addition, multiplication, etc). Such tools also allow users to substitute default Jacobian operations with their own implementations. In practice, such *explicitly* implemented Jacobian operations can be more efficient both computationally and memory-wise. Bloch simulation is typically the most computationally expensive stage in relating pulse waveforms to objective costs. Having explicit Jacobians of the Bloch simulator can therefore accelerate the computation.

To derive discrete time (δ_t) explicit Jacobians in the rotating frame, for all magnetic spins, we assume equilibrium spin magnitudes of 1, relaxation constants $e_1 \coloneqq \exp(-\delta_t/T_1)$, $e_2 \coloneqq \exp(-\delta_t/T_2)$, and gyromagnetic ratio γ. At time t, the rotating frame effective magnetic field (B-effective), $B_t \in \mathbb{R}^3$, causes the magnetic spin state $m_t \in \mathbb{R}^3$, to precess (rotate) about an axis $u_t \coloneqq B_t/\|B_t\|_2$ by angle $\phi_t \coloneqq -\gamma\delta_t\|B_t\|_2$.

[3] Or Clarke generalized subdifferentials for non-smooth objectives [8].

One iteration of discrete time Bloch simulation can be expressed as:

$$m_{t+1} = ER_t m_t + e, \tag{2.2}$$

where $E \coloneqq \mathsf{diag}([e_2, e_2, e_1])$, $e \coloneqq [0, 0, 1-e_1]^\mathsf{T}$ model the relaxations; $R_t = \cos(\phi_t)I + (1-\cos(\phi_t))u_t u_t^\mathsf{T} + \sin(\phi_t)[u_t]_\times$ models the rotation; I is the 3D identity matrix, $\mathsf{diag}([1,1,1])$; and $[u_t]_\times$ denotes the cross product matrix of u_t, i.e., $[u_t]_\times m_t = u_t \times m_t$. The rotation matrix R_t is spatially dependent, as it accounts for B-effective which incorporates applied gradients, off-resonance, etc. Relaxation terms, as they depend on the underlying tissue property, are generally also spatially dependent. To avoid notation clutter, we have not indicated those spatial dependencies in Eq. [2.2]; rather, Eq. [2.2] can be considered to hold for a single spin isochromat, with the appropriate R_t and relaxation terms for that isochromat.

One can verify the following recursive expressions for partial derivatives of the loss with respect to m_t and B_t:

$$\begin{aligned}
\tfrac{\partial \mathcal{L}}{\partial m_t} &= R_t^\mathsf{T} E \tfrac{\partial \mathcal{L}}{\partial m_{t+1}} =: h_t, \\
\tfrac{\partial \mathcal{L}}{\partial B_t} &= \gamma\delta_t/\phi_t (u_t u_t^\mathsf{T} - I)\tfrac{\partial \mathcal{L}}{\partial u_t} - \gamma\delta_t \tfrac{\partial \mathcal{L}}{\partial \phi_t} u_t, \\
\tfrac{\partial \mathcal{L}}{\partial u_t} &= \phi_t \left(\mathfrak{c}_t (m_t u_t^\mathsf{T} + m_t^\mathsf{T} u_t I) + \mathfrak{s}_t [m_t]_\times \right) E h_{t+1}, \\
\tfrac{\partial \mathcal{L}}{\partial \phi_t} &= ([u_t]_\times R_t m_t)^\mathsf{T} E h_{t+1},
\end{aligned} \tag{2.3}$$

where $\mathfrak{c}_t \coloneqq \big(1-\cos(\phi_t)\big)/\phi_t$, $\mathfrak{s}_t \coloneqq \sin(\phi_t)/\phi_t$. Given $\partial B_t/\partial g_t$ and $\partial B_t/\partial b_t$ (which are easy to compute), we obtain the necessary derivatives for the joint optimization by the chain rule:

$$\frac{\partial \mathcal{L}}{\partial g_t} = \frac{\partial \mathcal{L}}{\partial B_t} \cdot \frac{\partial B_t}{\partial g_t}, \quad \frac{\partial \mathcal{L}}{\partial b_t} = \frac{\partial \mathcal{L}}{\partial B_t} \cdot \frac{\partial B_t}{\partial b_t}.$$

Using the explicit Jacobians in [2.3] for the Bloch simulator operations halved both the computation time and memory use compared to the default implicit Jacobian

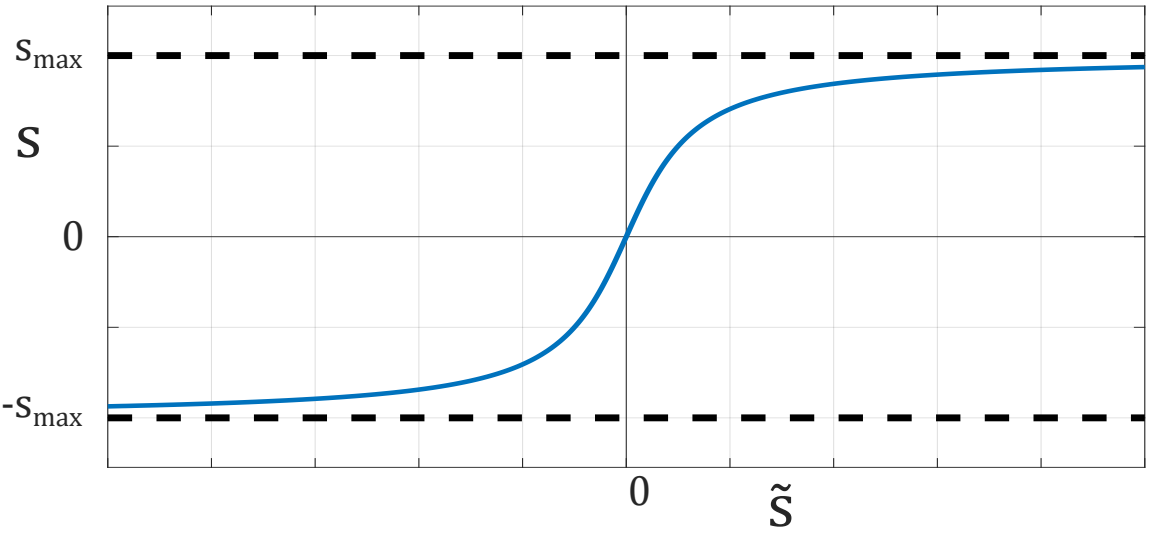

Figure 2.1: Turning constrained slew rate s into unconstrained $\tilde{s}$, by change of variable $\tan^{-1}$.

operations provided by PyTorch (v1.3).

The remaining Jacobians, such as $\partial\mathcal{L}/\partial M_T$, $\partial B_t/\partial b_t$, and $\partial B_t/\partial g_t$, typically do not involve complicated computations. Also, they can vary with different objectives, e.g., switching from LS to MLS; or with different excitation settings, e.g., uniform vs non-uniform transmit sensitivities. For program generality, we left these remaining Jacobians to be obtained implicitly by the auto-differentiation framework.

2.2.4 Constraints

Constrained optimization often requires extra effort to ensure solution feasibility, such as feasible set projection and constraint substitution with penalizations. This would involve crafting projection algorithms, and tuning penalty parameters. For problem [2.1], in the absence of convexity, we use a change of variables [4, 68] that converts the problem into an effectively unconstrained one and avoids such extra effort during optimization.

Let $s \in \mathbb{R}^{n_T \times 3}$ denote the slew rate, i.e., $s = \mathsf{D}g$. Define $\tilde{s} = \tan(\pi/2 \cdot s/s_{\max})$;

$\tilde{\rho} \coloneqq \tan(\pi/2 \cdot |b|/b_{\max})$, $\theta \coloneqq \angle b$. We automatically have $\|b\|_\infty \leq b_{\max}$ and $\|s\|_{\infty,\infty} \leq s_{\max}$ always satisfied (Fig. 2.1). Thus, we reformulate problem [2.1] as:

$$\begin{aligned} \underset{\tilde{s}\in\mathbb{R}^{n_T\times 3};\, \tilde{\rho},\, \theta\in\mathbb{R}^{n_T}}{\arg\min} \quad & \mathcal{L} \coloneqq f(M_T(g,b), M_D) + \lambda\mathcal{R} \\ \text{s. t.} \quad & \|g\|_{\infty,\infty} \leq g_{\max} \\ & \mathsf{D}g = 2s_{\max}/\pi \cdot \tan^{-1}(\tilde{s}) \\ & b = 2b_{\max}/\pi \cdot \exp(\iota\theta)\tan^{-1}(\tilde{\rho}). \end{aligned} \qquad [2.4]$$

In practice, for change of variable, $\tan^{-1}$ can be replaced with any other *strictly monotone* function, e.g., sigmoid, that maps an unconstrained domain to an interval.

Empirically, for 3D tailored pulse design, we observe that, with extended kt-points initializations [72], gradient amplitudes are well below typical max gradient constraints ($5\,\mathrm{G\,cm^{-1}}$) prior to and throughout the optimization procedure. Hence, while problem [2.4] is still constrained formally, its max gradient is practically inactive. We thus treated it as an unconstrained problem for the results shown in this paper.

2.2.5 Optimization Algorithm

We select initial waveforms g and b that satisfy the constraints. To minimize [2.4], we alternatingly update $\tilde{\rho}$, θ, and $\tilde{s}$, as shown in Algorithm 1. This alternating strategy is commonly used in existing joint design approaches [72, 80], and helps reduce the problem size for the L-BFGS algorithm used in updating the pulse. With auto-differentiation, the optimization algorithm can be formulated without reference to the specific loss function, as demonstrated with very different design problems in section 2.3. We use the L-BFGS optimizer provided by PyTorch for updating the variables within an iteration. The number of iterations may depend on pulse initializations. We empirically choose $N = 10$ for experiments in this work.

Algorithm 1 Alternating Minimization

Inputs: Variables: g, b; Number of iterations: N
Compute $\tilde{\rho}$, θ and $\tilde{s}$ from g, b
for $n = 1$ to N **do**
    Fix $\tilde{s}$; Optimize $\tilde{\rho}$, θ, using L-BFGS
    Fix $\tilde{\rho}$, θ; Optimize $\tilde{s}$, using L-BFGS
end for
Compute g and b, from $\tilde{\rho}$, θ and $\tilde{s}$
return g, b

2.3 Methods

To demonstrate the utility and generality of our approach, we designed two different kinds of 3D tailored pulses: outer-volume (OV) saturation, and inner-volume (IV) inversion.

2.3.1 3D Outer-Volume Saturation Pulse Design

Outer-volume saturation pulses can be used to limit the imaging field of view (FOV), and hence has the potential to reduce both the time needed for data acquisition as well as motion artifacts from [39, 49, 77], e.g., the chest wall or abdomen in body imaging applications. OV saturation pulses should ideally have a high flip angle in the OV region (e.g., 90 degrees), while leaving the IV unperturbed. These pulses are typically followed immediately by a gradient crusher. Since the phase of OV magnetization prior to the crusher is unimportant, we use MLS loss in design, and include a regularization term on RF power to indirectly control SAR as well as to demonstrate the generality of our approach:

$$\mathcal{L}_{90} = \| |M_T[:, 1{:}2]| - |M_D[:, 1{:}2]| \|_2^2 + \lambda \|b\|_2^2, \qquad [2.5]$$

where, $|M[:, 1{:}2]| \coloneqq \text{abs}(M[:, 1] + \iota M[:, 2])$, is a vector function computing magnitudes of spin transverse magnetizations. For the target excitation profile, we set rows in M_D to $[1, 0, 0]$ for OV spins, and $[0, 0, 1]$ for IV spins. We implemented this loss in

Figure 2.2: Schematic diagram of the imaging sequence used to characterize (validate) the 3D tailored inversion pulses.

PyTorch to obtain the Jacobian $\partial \mathcal{L}_{90}/\partial M_T$ as described in the Theory section.

In principle, small-tip based 3D tailored design approaches can also be applied to this loss by scaling the designed RF pulse to attain the desired 90° flip (although the resulting pulse may exceed peak RF limits). We therefore compare our approach with the small-tip method in [72], starting with the same initial b and g waveforms in both cases (initialized as described in 2.3.3).

2.3.2 Inner-Volume Inversion Pulse Design

Next we designed another type of excitation pulse that is difficult to design using conventional approaches: an IV *inversion* pulse. Such a pulse may be useful for, e.g., selective inversion of arterial blood for flow territory mapping in perfusion imaging. For this pulse we propose a very different excitation loss based on the *longitudinal* magnetization:

$$\mathcal{L}_{180} = \|M_T[:,3] - M_D[:,3]\|_2^2 + \lambda \|b\|_2^2. \qquad [2.6]$$

We set rows in M_D to $[0,0,1]$ for OV spins, and $[0,0,-1]$ for IV spins. We also implement this loss in PyTorch.

2.3.3 Pulse Initializations

The loss in problem [2.4] is non-convex, and the choice of initial g and b waveforms influences the final excitation result. How best to initialize these waveforms is an open

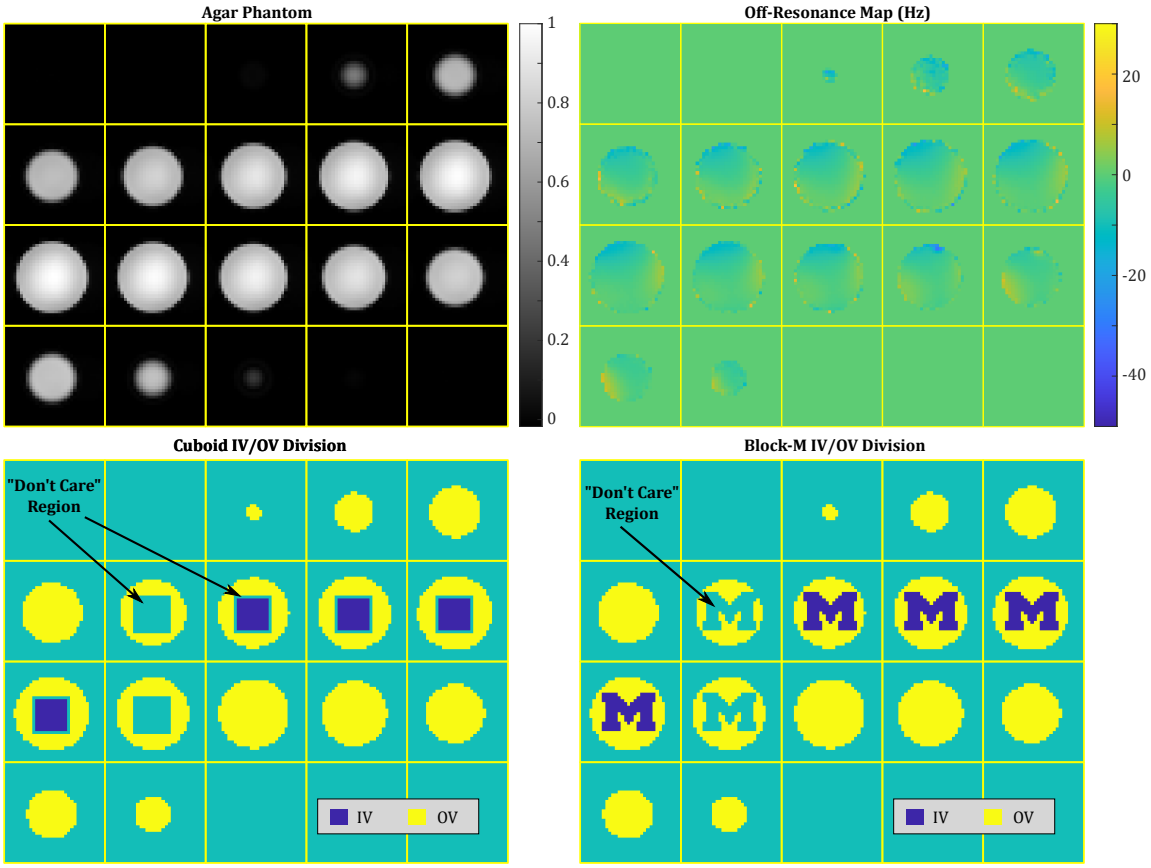

Figure 2.3: Experimental phantom, and the two target patterns (IV/OV divisions) used in our experiments. Top left: Magnitude image of the uniform Agar phantom. Top right: Observed field map, used in the pulse design to account for B0 inhomogeneity. A conservative mask that is 1-voxel-wide larger than the phantom support was used to ensure that the phantom boundary was included in the design. This expanded mask is the likely cause for the relatively large B0 values in some pixels at the edge of the mask (that are likely just outside the phantom). Bottom: Cuboid (left) and "block-M" (right) target patterns. We prescribed a "don't care" (region with arrows) at the boundary between the IV and OV regions that is excluded when calculating the design loss. For the cuboid pattern, the don't care region included the entire 3D IV/OV boundary, whereas for the block-M pattern, only the top and bottom slices (slices 6 and 11; slice numbers increase left-to-right and top-to-bottom) were included due to the low in-plane spatial resolution of the design grid.

problem. In [72], the initialization problem (in the context of small-tip 3D tailored excitation) was addressed by evaluating two popular choices for the excitation k-space trajectory, stack-of-spirals and SPINS [46], along with a novel alternative approach, "extended kt-points", that chooses gradients based on the desired (target) excitation pattern. The extended kt-points approach was shown to produce comparable or better excitation accuracy as compared with stack-of-spirals or SPINS, and was therefore chosen for the experiments in this manuscript.

Once the gradients were initialized in this way, we initialized b using the approach in [79]. These initial RF waveforms were scaled down when necessary to satisfy the $b_{\max}$ constraint.

2.3.4 B-effective Computation

The particular form of B-effective depends on the excitation objective and other application-specific components. Besides the RF and gradient waveforms, it often also contains an off-resonance map (that may vary with time) and transmit sensitivity maps. Other factors such as gradient non-linearity can also be included in B-effective. For the single transmit coil phantom studies in this work, B-effective accounts for RF, gradient, and a static off-resonance map. Specifically, at time t, let $b_t \in \mathbb{C}$ and $g_t \in \mathbb{R}^3$ denote the instantaneous RF and gradients, respectively. For position r relative to the scanner iso-center, with off-resonance $\omega(r)$, the instantaneous B-effective is:

$$B_t = [\Re(b_t), \Im(b_t), \langle g_t, r \rangle + \omega(r)/\gamma], \qquad [2.7]$$

where $\Re$ and $\Im$ extract the real and imaginary component, respectively.

Table 2.1: Pulse Duration and TR/TE

Parameters	OV90	IV180	IV180M
Pulse Duration	6.5 ms	4.8 ms	4.5 ms
TR / TE	2 s / 15 ms	3 s / minimum TE	

2.3.5 Phantom Experiments

We performed validation experiments in an Agar phantom on a GE MR750 3T scanner. Fig. 2.3 illustrates the experimental setup, including the prescribed IV and OV regions. All experiments used the same observed off-resonance map in the pulse design (Fig. 2.3). For all studies, we conducted the 3D design on a $32 \times 32 \times 20$ voxel grid of FOV $24 \times 24 \times 24\,\mathrm{cm}^3$; with RF power weighting coefficient $\lambda = 4$, and constraints: $b_{\max} = 0.25\,\mathrm{G}$, $g_{\max} = 5\,\mathrm{G\,cm^{-1}}$, $s_{\max} = 12\,\mathrm{G\,cm^{-1}\,ms^{-1}}$. We quantified excitation performance in simulations with normalized root mean squared error (NRMSE). Spins in "don't care" regions (Fig. 2.3) were excluded when calculating the NRMSE. We ran our design programs on an NVidia 2080 Ti graphics card. With the settings above, our method uses around 1.1 GB GPU RAM for all 3 designs. This includes the intrinsic GPU RAM usage of the PyTorch environment.

We performed three different experiments: 1. OV 90° excitation using the cuboid target pattern shown in Fig. 2.3 (OV90); 2. IV inversion using that same cuboid target pattern (IV180); and 3. IV inversion with a block-M target pattern (IV180M). The experiments were implemented using a vendor-agnostic platform for rapid prototyping of MR pulse sequences [33, 50]. IV dimensions were $9 \times 9 \times 4.8\,\mathrm{cm}^3$ and $9 \times 12.8 \times 4.8\,\mathrm{cm}^3$ for the cuboid and block-M target patterns, respectively. We used a single channel transmit/receive birdcage coil for all experiments and assumed uniform RF transmit sensitivity during pulse design. To mitigate Gibbs ringing artifacts, we acquired the phantom images at a matrix size of $120 \times 120 \times 48$ and then downsized in image space to match the design grid size $32 \times 32 \times 20$.

We used long TR to wait for spin full recovery from saturation and inversion. For

the OV90 experiment, as a substantial volume of the phantom is excited with large angle, we used TE=15 ms to intentionally decay signal intensity and avoid saturating amplifiers in signal receiver during acquisition. We use minimum TE for the inversion experiments.

For inversion performance validation, we use the sequence in Fig. 2.2 to obtain both phase and magnitude phantom images, with tip-down angle set to 10°. We expect a π phase difference between inverted (IV) non-inverted (OV) regions, as the excitation pulse should tip inverted and non-inverted spins in opposite directions. In addition to the IV180 and IV180M excitation pulses, we imaged the phantom using the same sequence settings (TE/TR, flip angle, matrix size) using a conventional slab-selective SLR pulse. We normalized the inversion images using this "non-inversion" image to eliminate receive coil sensitivity weighting (both magnitude and phase) in the inversion images. For completeness, the unnormalized images are shown in supplemental materials (Fig. 2.12).

2.4 Results

2.4.1 OV90

Figure 2.4 shows the OV saturation pulses obtained with the proposed method and the small-tip approach in [72], and Fig. 2.5 shows the corresponding phantom imaging results. To keep the small-tip RF pulse within peak amplitude limits, we applied VERSE [11] near the end of the pulse. Our approach required 10 min for design, longer than the Small-Tip approach (2 min). While the RF waveforms differ markedly, the (excitation) k-space trajectories are more similar, though differences are clearly observed in the 3D trajectory plot (Fig. 2.4).

We observe excellent agreement between simulated and acquired excitation patterns (Fig. 2.5). Also, the proposed method produces much lower excitation error

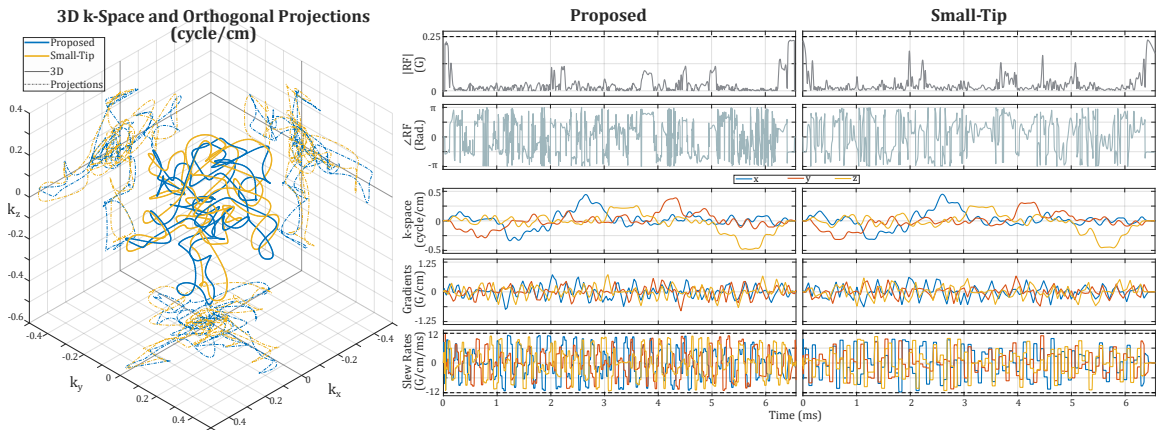

Figure 2.4: OV saturation pulses for the cuboid IV (Fig. 2.3), designed with our approach (Proposed) and the small-tip method in [72] (experiment OV90). The left panel shows the 3D k-space trajectories and their orthogonal projections: The two trajectories explore largely overlapping regions in excitation k-space. The right two panels show RF, gradient, and slew rate waveforms. Both designs satisfy the constraints, but for the small-tip design it was necessary to apply the VERSE [11] algorithm near the end of the pulse (see Discussion). Gradient peak amplitudes remain quite small ($\ll 5\,\mathrm{G\,cm^{-1}}$), whereas the gradient slew rates are frequently near their limit.

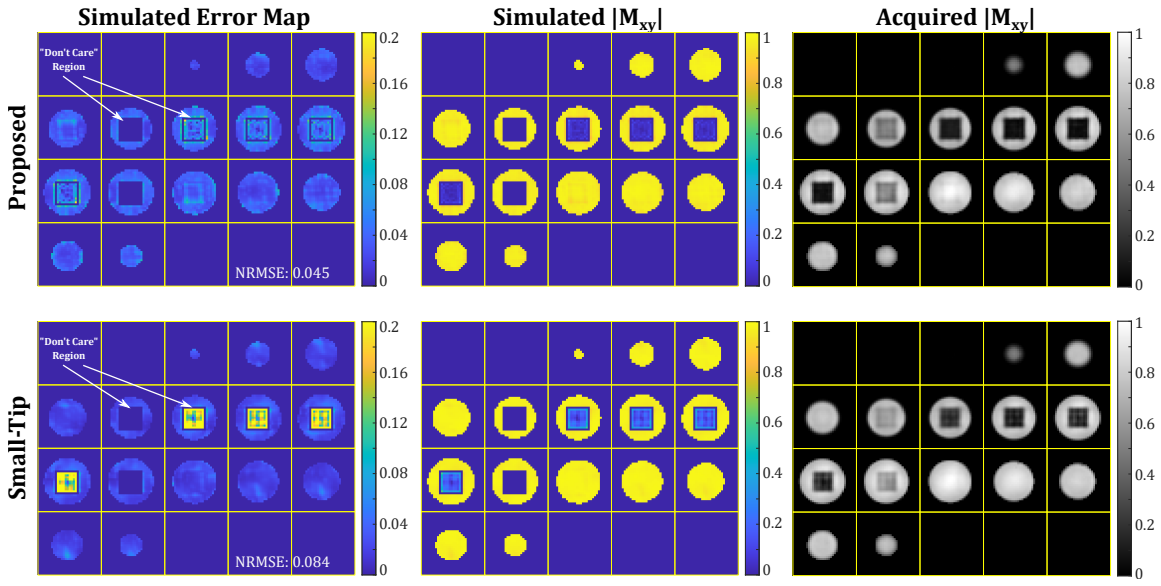

Figure 2.5: Experimental validation of the pulses shown in Fig. 2.4. The left panel shows the error map from simulation. Our approach has much smaller (-46%) NRMSE in simulation compared to Small-Tip. Acquired results (right) agree with the simulations (middle). Small-Tip approach has larger error inside the IV: This is expected, as the method produces only small-tip pulses, that we then scaled to meet the large-tip objective. The scaling increases excitation error inside IV while reducing error in the OV. Our approach directly designs large-tip pulses without this type of 'scaling' error.

than the Small-Tip design (46% lower NRMSE error overall); this is expected as the small-tip assumption is violated after scaling the RF to attain the desired 90° flip angle in the OV, which reduces the error in the OV at the expense of increased error in the IV.

2.4.2 IV180

Fig.2.6 shows the results of the cuboid IV inversion experiment. Pulse waveforms and images from simulation and phantom experiments are shown. Pulse design took 6 min. Simulations and acquired images are in excellent agreement, and indicate successful inversion within the IV with errors mainly located at the IV/OV boundary as expected. In particular, we observe dark bands along the IV/OV boundary in the magnitude image. Spins in this region are not fully inverted, resulting in low signal intensities in the magnitude image. The phase image shows an abrupt π transition at the IV/OV boundary, indicating successful IV inversion.

2.4.3 IV180M

Fig. 2.7 shows the results of the block-M IV inversion experiment. Pulse design took 6 min. Slew rates are near the limit, similar to the IV180 experiment. Simulation and acquired images again indicate successful inversion. The NRMSE is larger than in the IV180 experiment, suggesting a trade-off between geometry complexity and excitation accuracy. Excitation error is largest near the in-plane edge of the block-M, where target M_z changes sharply from 1 to -1. We again observe dark bands along the IV/OV boundary, and an abrupt phase change across that boundary, as expected.

2.5 Discussion

We have demonstrated a new approach to joint multi-dimensional excitation pulse design that directly optimizes both RF and gradient waveforms. Our approach is not

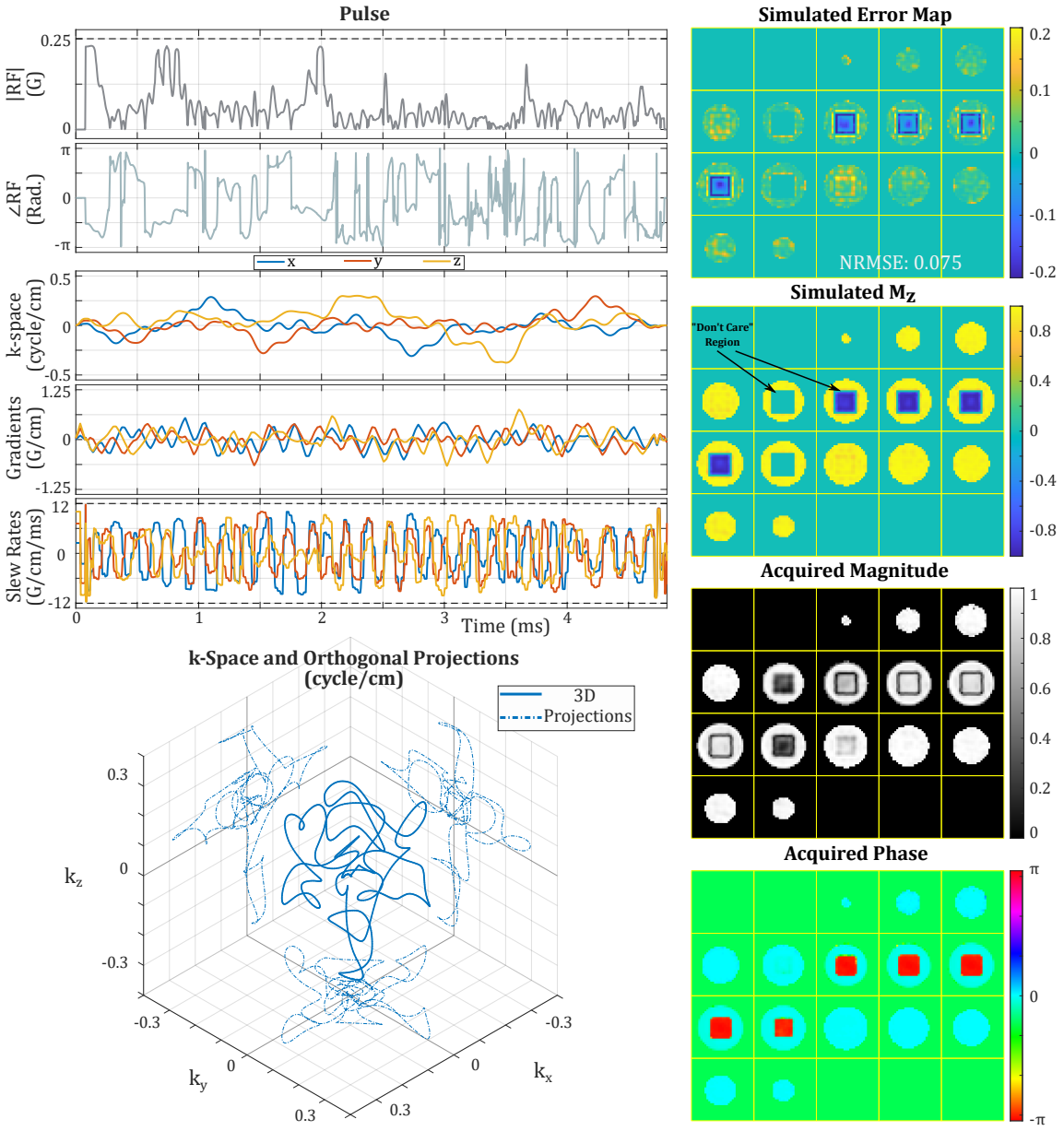

Figure 2.6: IV inversion results for the cuboid IV pattern (Experiment IV180). As desired, the 4.2 ms pulse satisfies all constraints. The gradient waveform is again far from its peak constraint of $5\,\mathrm{G\,cm^{-1}}$. Compared to the OV90 experiment, the pulse has more extreme slew rate waveforms. The acquired magnitude and phase (i.e., the "observed inversion") were obtained with the sequence in Fig. 2.2. We observe good agreement between simulated and acquired inversion patterns. The designed pulse successfully inverts the IV, as indicated by similar magnitude image intensity in the IV and OV regions (apart from transmit/receive coil shading) and a π phase shift across the IV/OV boundary. The dark bands in the acquired images at the IV/OV boundary are due to spin saturation from incomplete inversion (and overlap substantially with the prescribed "don't care" region).

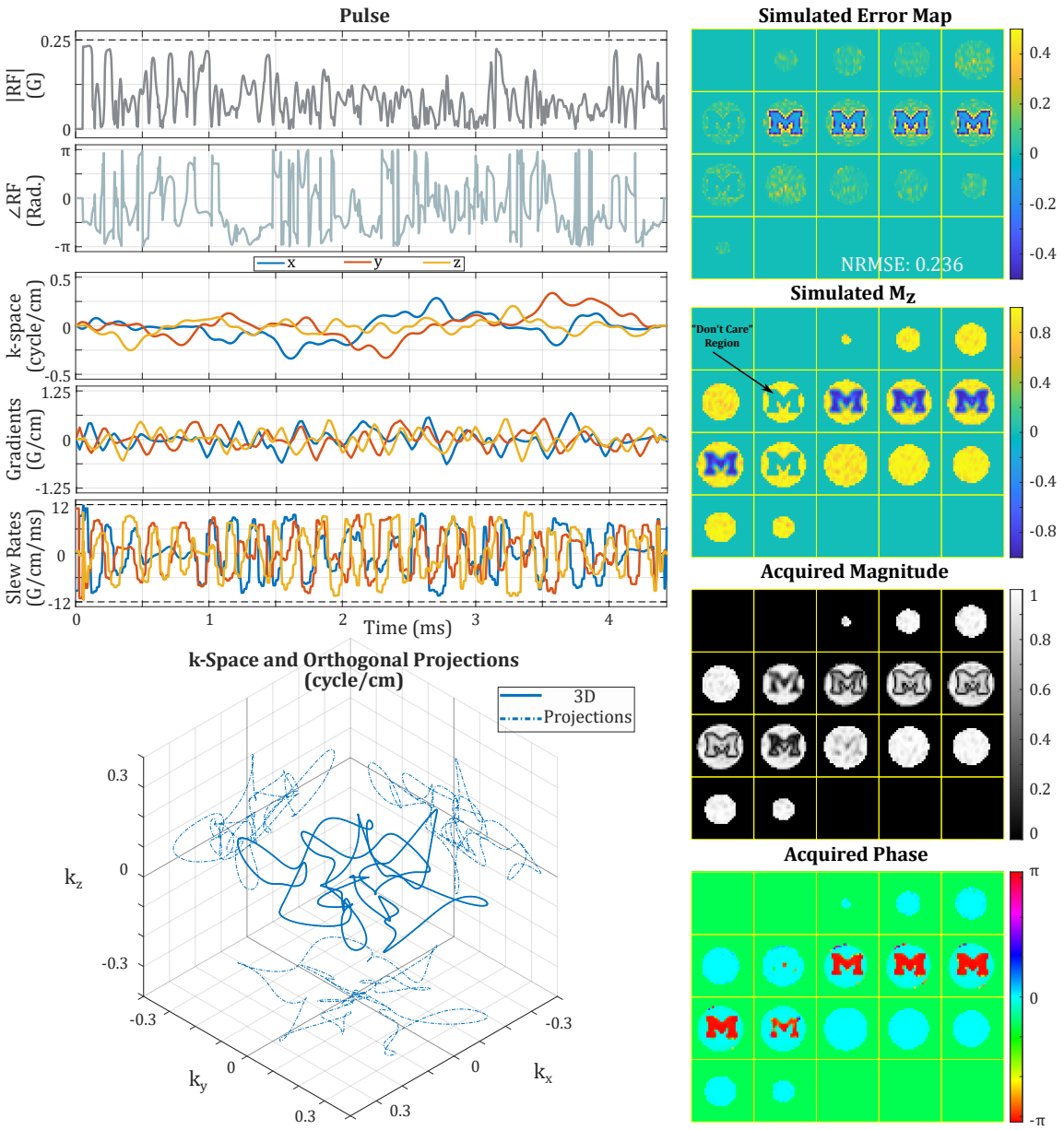

Figure 2.7: IV inversion results for the block-M target pattern (Experiment IV180M). The 4.5 ms pulse satisfies all constraints. The gradient waveforms are again well below the peak amplitude constraint of $5\,\mathrm{G\,cm^{-1}}$, and slew rates are near the constraint for significant portions of the waveform duration. As in Fig. 2.6, the pulse successfully inverts the IV. The dark bands in the acquired magnitude image at the IV/OV boundary are due to saturation effects arising from the finite resolution (excitation k-space extent) of the pulse – even though only slices 6 and 11 were included in the "don't care" region in the design due to the low in-plane spatial resolution of the design grid (see Fig. 2.3).

limited to small-tip design problems, and is compatible with quite general loss/design functions such as those that involve longitudinal and/or magnitude magnetization. We validated our approach with 3D tailored large-tip objectives. For this type of application, the "extended kt-points" small-tip initialization [72] led to excellent large-tip results.

We chose to implement our auto-differentiable Bloch simulator with B-effective as its input for its generality: one can possibly prepend to it arbitrary functions that compute B-effective from various parameters, such as multi-coil parallel transmit (pTx) sensitivities, spin movements, and even non-linear response of gradient amplifiers , gradient delays, etc. This choice that favors generality may require more memory than software designs that take RF and gradients as inputs directly, and may require more expensive hardware with adequate memory for high-dimensional design problems. In particular, an interface that uses RF, gradient and spin location inputs requires a memory size proportional to $(N_T + N_T \times 3 + N_M \times 3)$, whereas our interface requires memory proportional to $(N_M \times 3 \times N_T + N_M \times 3)$. Our implementation can find use in different scenarios for proof-of-principle designs that one could then follow by customized simulators that meet specific computational requirements.

For the Bloch simulator, one may alternatively consider using the hard pulse approximation, which splits the instantaneous rotation matrix R_t into two rotations: RF rotation, and transversal rotation due to the applied gradients and off-resonance. The hard pulse approximation is the basis for the Shinnar-Le Roux (SLR) pulse design algorithm, and is crucial for the development of that algorithm. In our case, however, such splitting actually increases the number of elementary computations: when multiplying a vector, RF and transversal rotations require 9 and 4 multiplications, respectively, while direct multiplication by R_t requires only 9. We therefore believe that the hard pulse approximation does not confer any particular advantages on our approach.

Apart from the explicit Jacobians introduced here, additional steps may be taken to reduce computation time. Computation time is primarily determined by pulse length, and not on the grid size (number of voxels) since computations are done voxel-wise and can be easily parallelized to within GPU RAM limits. Apart from increasing the simulation expense, longer pulses may also slow down Algorithm 1, since we used L-BFGS for updating RF and gradients. In the future, to shorten the optimization time for online pulse design tasks, it may be helpful to use coarser δ_t in the Bloch simulation (here we used 4 µs to match our scanner's hardware dwell time), or parameterize the gradient waveforms to reduce the optimization problem size (e.g., using B-splines as in [72]).

For the experiments presented, we used voxel resolution $7.5 \times 7.5 \times 12\,\mathrm{mm}^3$ and grid size $32 \times 32 \times 20$ for the pulse design. For more complex target excitation patterns and/or a larger FOV (e.g., as in the ISMRM parallel transmit pulse design challenge [20]), it may be desirable to increase the spatial resolution (maximum extent in excitation k-space) and/or grid size for finer excitation accuracy control. For instance, with a larger grid size, we would have space for in-plane "don't care" region for the IV180M experiment, which may help reduce excitation error. A larger grid size will increase the memory usage in simulation, for which the use of multiple graphics cards may be needed to parallelize simulations across voxels.

In the OV90 experiment (Figs. 2.4–2.5), we were able to apply VERSE [11, 27] to the pulse designed with the small-tip approach [72] to avoid violating the RF amplitude limit (after scaling to attain 90° flip angle in the OV). However, this adjustment was possible only because the 6.5 ms pulse happened to exceed peak RF only near the end of the pulse, allowing us to apply the VERSE strategy in a relatively straightforward way. In the more general case, where peak RF is exceeded during the middle of the pulse, it is more difficult to apply the VERSE technique to 3D RF pulses such as those designed here. We found empirically that in the OV90 experiment, shorter

pulses designed with the small-tip approach tended to exceed peak RF during one or more intermediate intervals (after scaling), and that we were therefore unable to carry out an effective experimental evaluation for the purposes of the comparison presented here (Figs. 2.4–2.5). The proposed approach avoids this difficulty because peak RF is constrained as described in 2.2.4; our approach was in fact able to design a shorter (4 ms) OV saturation pulse with the same excitation error as in Fig. 2.5 (not shown).

Pulse design problems are in general non-convex in terms of b and g. Due to a lack of theoretical tools for non-convex problem convergence analysis, it is unclear how to best design an optimization algorithm for such problems *a priori*. In Algorithm 1, instead of simultaneously updating both RF and gradient waveforms, we chose to update them alternatingly as often done in existing small-tip joint designs [5, 44, 72, 80]. In supplemental Figs. S6-S7, we compared the alternating scheme with the simultaneous scheme, and found empirically that the alternating scheme optimizes faster than the simultaneous scheme for the specific problem settings we have in this work. Unfortunately, the non-convexity prevents us from fully comprehending this behavior, and we make no claims that the alternating approach used here is optimal over the many possible alternatives. Iteration stopping criteria for updating b and g are also commonly chosen *ad hoc*. Besides limiting the maximum number of iterations as we have done in Algorithm 1, another option can be setting a threshold to assert large-enough loss decreases and/or updates of the variables at each iteration. However, due to the $\tan^{-1}$ change of variables, a minuscule update of b and s near their limits will be mapped to a vast difference in the optimization variables $\tilde{\rho}$ and $\tilde{s}$, respectively. As an alternative to the updates of variables, one can threshold the norms of variable derivatives, or the change of b and s as the iteration stopping criteria.

Our approach may remind readers of optimal control (OC) based pulse design methods [10, 21]. Comparing the OC formulation with our approach, we can make

the following observations (ignoring the penalization and relaxation terms): Eq. [2.2] is the forward propagation of spin states in OC (state equation); the first identity in Eq. [2.3] is the backward propagation of OC Lagrange multiplier (costate equation); the second identity in Eq. [2.3] is the derivative for iteratively optimizing B-effective as the control. Being one step in the computation of excitation losses, our auto-differentiable Bloch simulator enables reusing the forward and backward iterations regardless of the actual design loss function chosen, and propagating the derivatives to the actual controls, i.e., the RF and gradient waveforms. In future works, it may be beneficial to employ tools from control researches. For instance, optimization algorithms from OC may accelerate or replace algorithm 1.

Like many other pulse design works [5, 44, 72, 80], our approach assumes that the off-resonance map is known. This assumption may be violated in, e.g., the presence of motion or in mixed fat/water voxels. Addressing such model mismatch/uncertainties requires improving design robustness in these perspectives. Noting the relation of our tool to the control framework, we anticipate incorporation of robust control methods can strengthen our design against off-resonance map estimation mismatches.

A major advantage of our approach is that it enables designs involving arbitrary loss functions, enabling novel design formulations that have so far not been tractable. For example, we demonstrated in [2.6] a loss involving only longitudinal magnetization. Other possibilities may include the addition of constraints or regularization terms involving specific absorption rate (SAR) or peripheral nerve stimulation (PNS). Another important feature is that the method back-propagates derivatives throughout the Bloch simulator, which may facilitate development of neural network based pulse design approaches.

A limitation of our method is that it only works for fixed pulse length, as determined by the initial waveforms. As shown in the pulse plots in Figs. 2.4, 2.6, and 2.7, there are temporal intervals where neither the RF, gradients, nor slew rates are

hitting their constraints. This may suggest that the pulses can be shortened without sacrificing excitation accuracy. Pulse shortening can be formed as a minimum-time pulse design problem [10] in the OC context. Noting the relation of our method to the OC approach, for future works, we expect employing existing OC tools to be helpful in overcoming the fixed length limitation.

2.6 Conclusion

In this work, we have proposed a novel approach based on auto-differentiation tools for the joint design of RF and gradient waveforms, and validated it with multi-dimensional spatially tailored excitation tasks in MRI. Using short ($<$5 ms) excitation pulses and single (body) coil RF transmission, we demonstrated experimentally that even a fairly complex 3D spatial pattern (block-M) can be selectively inverted. Our method is not limited to specific design objectives. To reduce computation time and memory requirements, we derived explicit Jacobians for the Bloch simulator, as the simulation steps are typically the most computationally demanding. We used a change of variables to enforce hardware limits, enabling use of simpler unconstrained optimization. We anticipate that the proposed method will be useful for a broad range of excitation pulse design problems in MRI.

2.7 Supporting Information

2.7.1 Additional simulation results

For completeness, we present here both the initial pulses designed as described in III-C in the main text, and the corresponding optimized pulses obtained with the proposed approach. In the case of OV90, we also include the small-tip pulse. In each case, we show the simulated excitation pattern for each pulse.

In addition to IV90, IV180, IV180M, we include here a cuboid IV inversion pulse

based on the B0 field map acquired in the brain of a healthy volunteer. This is done to demonstrate the feasibility of designing an IV inversion pulse with the proposed approach using a more realistic B0 map than that shown in Fig. 3. In addition, for that simulation experiment we compare the optimized pulse with a pulse obtained by only optimizing the RF waveform, i.e., keeping the gradient waveforms fixed at their initial shapes. This is done to assess the relative importance of also optimizing the gradient waveforms.

The key takeaways from these figures are: (1) The excitation patterns produced by the initial pulses are substantially inferior to the optimized patterns. (2) The initial and optimized excitation k-space trajectories tend to be similar, suggesting that a local minimum is obtained. (3) The initial and optimized RF waveforms (amplitude and phase), on the other hand, differ markedly from each other, suggesting relatively weak dependence on initial RF waveform. This may be due to the fact that each RF sample can be optimized independently of the other samples, unlike gradient waveforms that are subject to slew rate constraints. (4) Despite the similarity between the initial and optimized gradient waveforms, the optimized waveforms produce a more accurate excitation than the pulse obtained by optimizing only the RF waveform (Fig. 2.11).

2.7.1.1 OV90, IV180, IV180M, and Brain Simulation

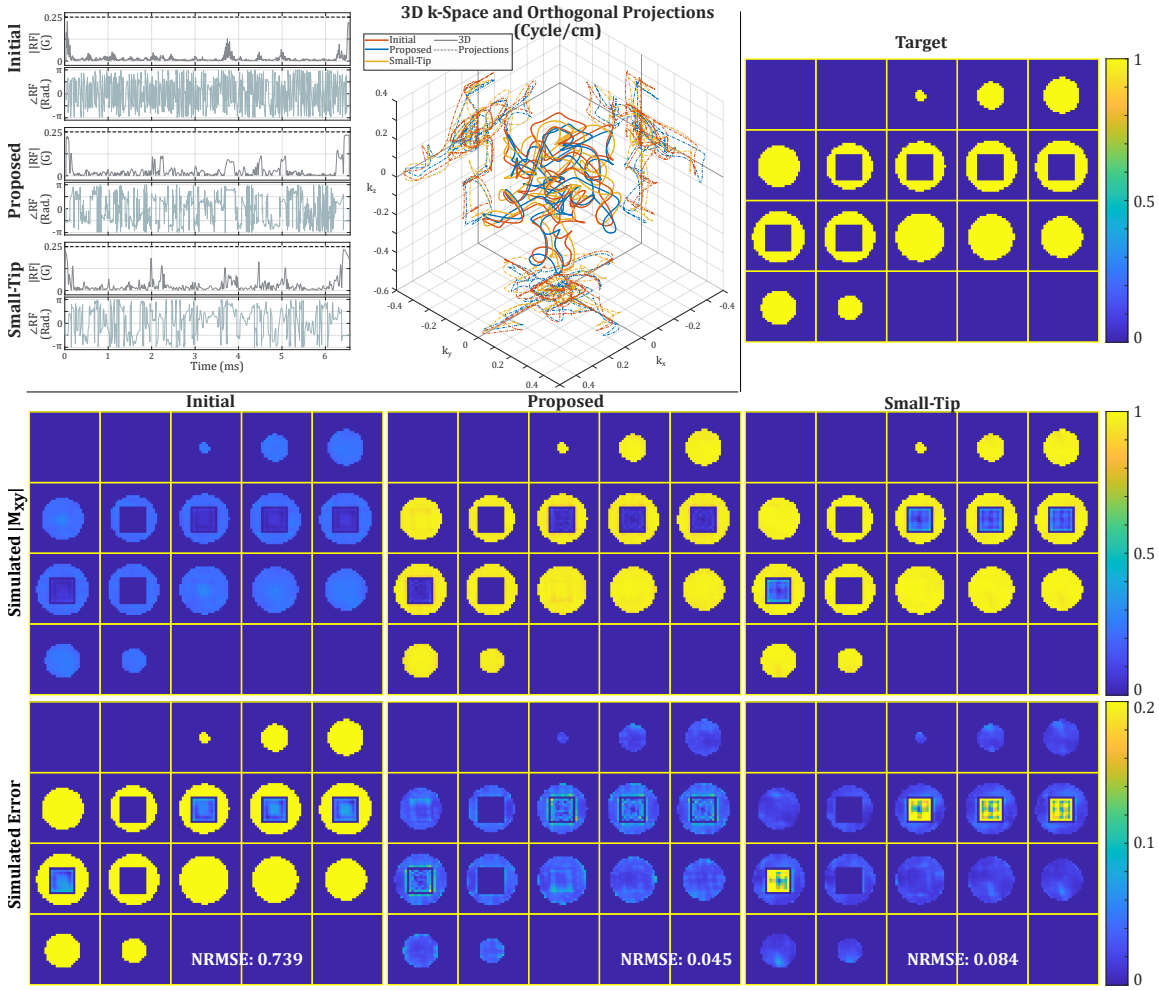

Figure 2.8: OV saturation pulses (and results) for the cuboid IV, designed with our approach (Proposed) and Sun's small-tip method (experiment OV90 in the main manuscript). On the 3D k-space plot, the three trajectories explore largely overlapping regions in excitation k-space. Our approach has much smaller (-46%) NRMSE in simulation compared to Small-Tip. Small-Tip approach has larger error inside the IV: This is expected, as the method produces only small-tip pulses, that we then scaled to meet the large-tip objective. The scaling increases excitation error inside IV while reducing error in the OV. Our approach directly designs large-tip pulses without this type of 'scaling' error.

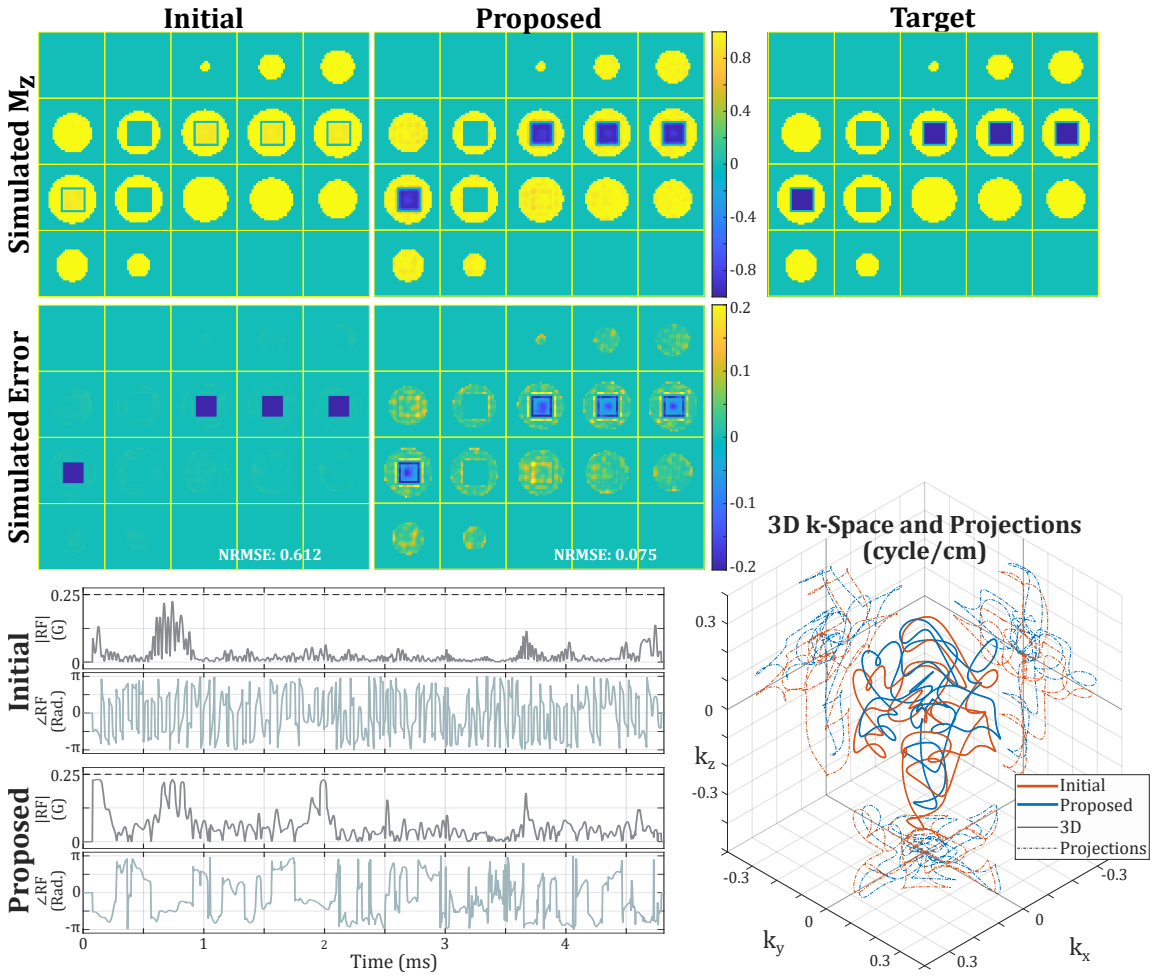

Figure 2.9: Cuboid IV inversion (experiment IV180 in the main manuscript). The lower left panel compares the RF waveforms of the initial and optimized (Proposed) pulses. The lower right panel compares the k-space trajectories of the two pulses.

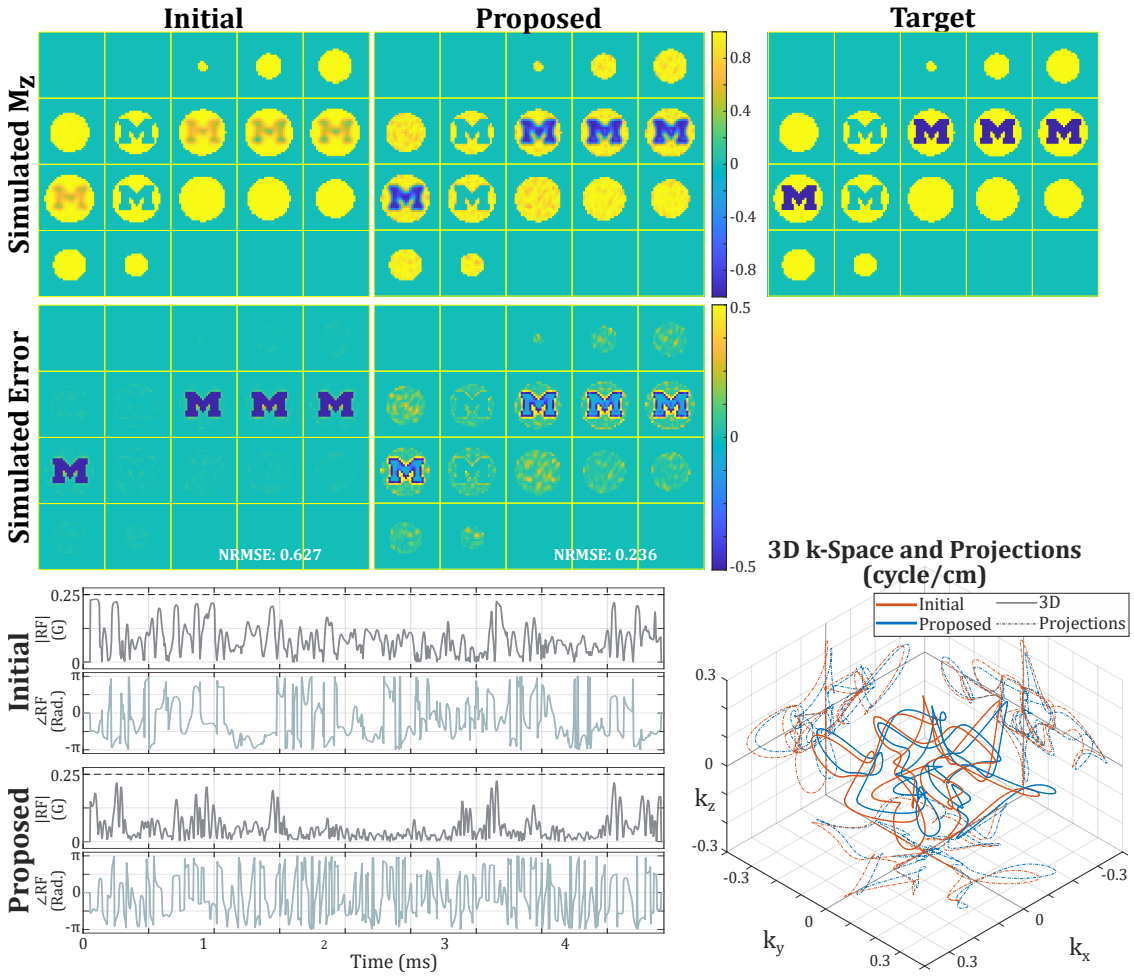

Figure 2.10: Block-M IV inversion pulse (experiment IV180M in the main manuscript). The lower left panel compares the RF waveforms of the initial and optimized (Proposed) pulses. The lower right panel compares the k-space trajectories of the two pulses.

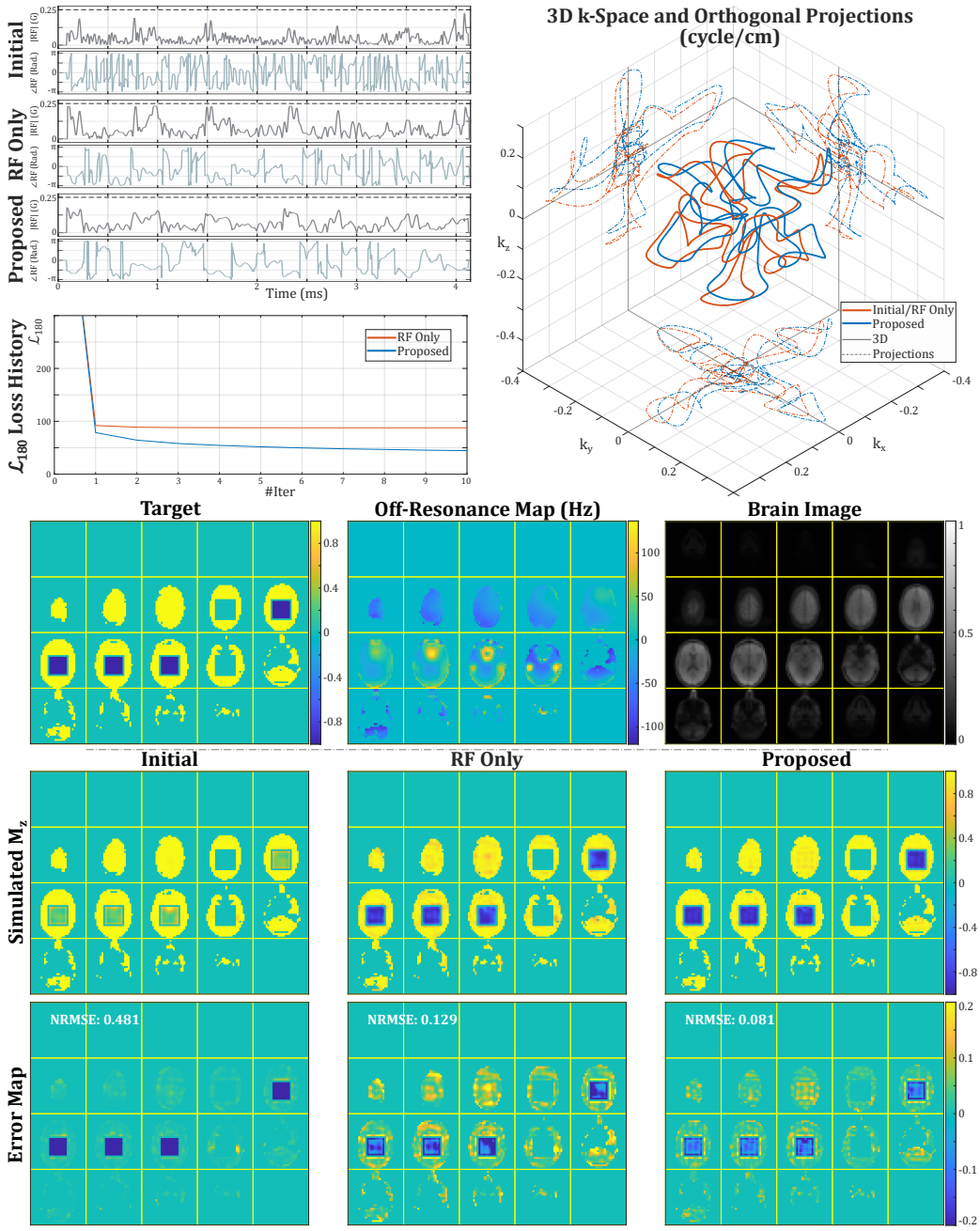

Figure 2.11: Cuboid IV inversion based on a B0 (off-resonance) map obtained in a volunteer. Three pulses are compared: (1) Initial, (2) 'RF Only', obtained by keeping the gradients fixed at their initial shapes and optimizing only the RF waveform with the proposed auto-differentiation approach, and (3) the proposed jointly optimized pulse. While our optimized k-space trajectory is similar to the initial k-space trajectory, the jointly optimized pulse (Proposed) attains an excitation accuracy (NRMSE: 8.1%) that is 37% better than the 'RF Only' pulse (NRMSE: 12.9%). This improvement is also reflected in the convergence ($\mathcal{L}_{180}$ loss history) plot.

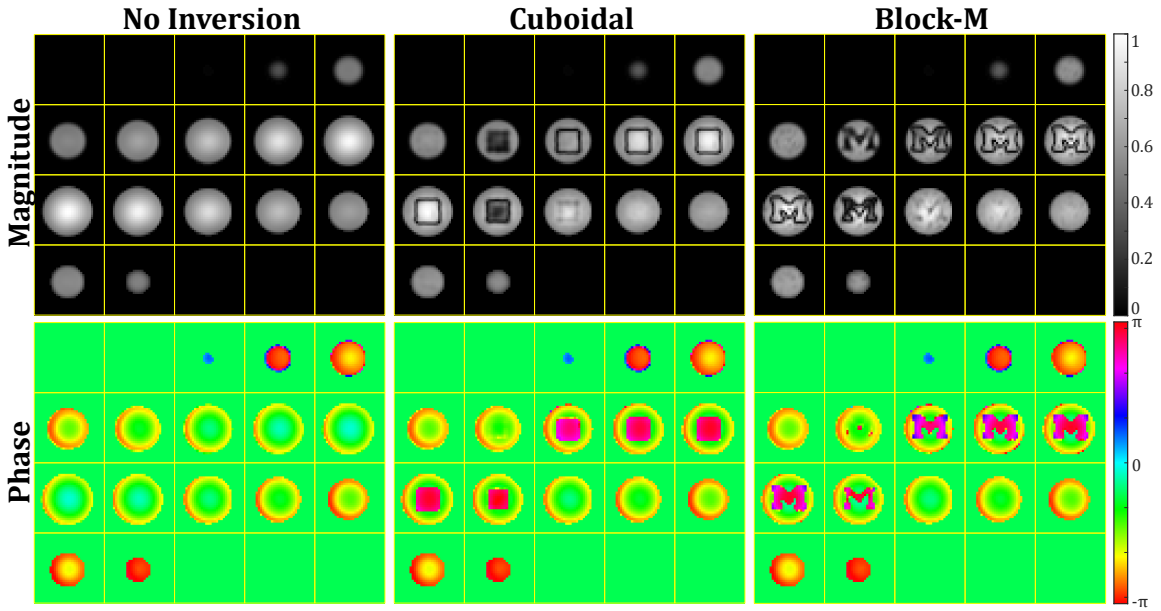

Figure 2.12: Images from the IV180 and IV180M experiments. "Cuboidal" and "Block-M" are the raw images of the IV180 and IV180M experiments, respectively. "No Inversion" was acquired using the same sequence (TR/TE, flip-angle, readout trajectory), except the inversion pulse in the sequence has RF amplitude set to 0. The 3 sets of magnitude and phase images shown here share the same image intensity and phase variations due to receiver coil sensitivity.

2.7.2 Unnormalized Inversion Images

In Fig. 2.12, we show the unnormalized images for the 3D spatially tailored inversion experiments (IV180, IV180M). In the main text, we normalized the "Cuboidal" and "Block-M" images by element-wise division by the "No Inversion" image, which eliminates the image intensity and phase variations due to receiver coil sensitivity.

2.7.3 Alternating vs Simultaneous Minimization

Here we compare the alternating optimization used in the main text with a simultaneous update scheme that optimizes b (RF waveform) and g (the three gradient waveforms) together at each iteration rather than fixing one and optimizing the other. We observe empirically that for the $\mathcal{L}_{90}$ and $\mathcal{L}_{180}$ losses defined in the main text, with

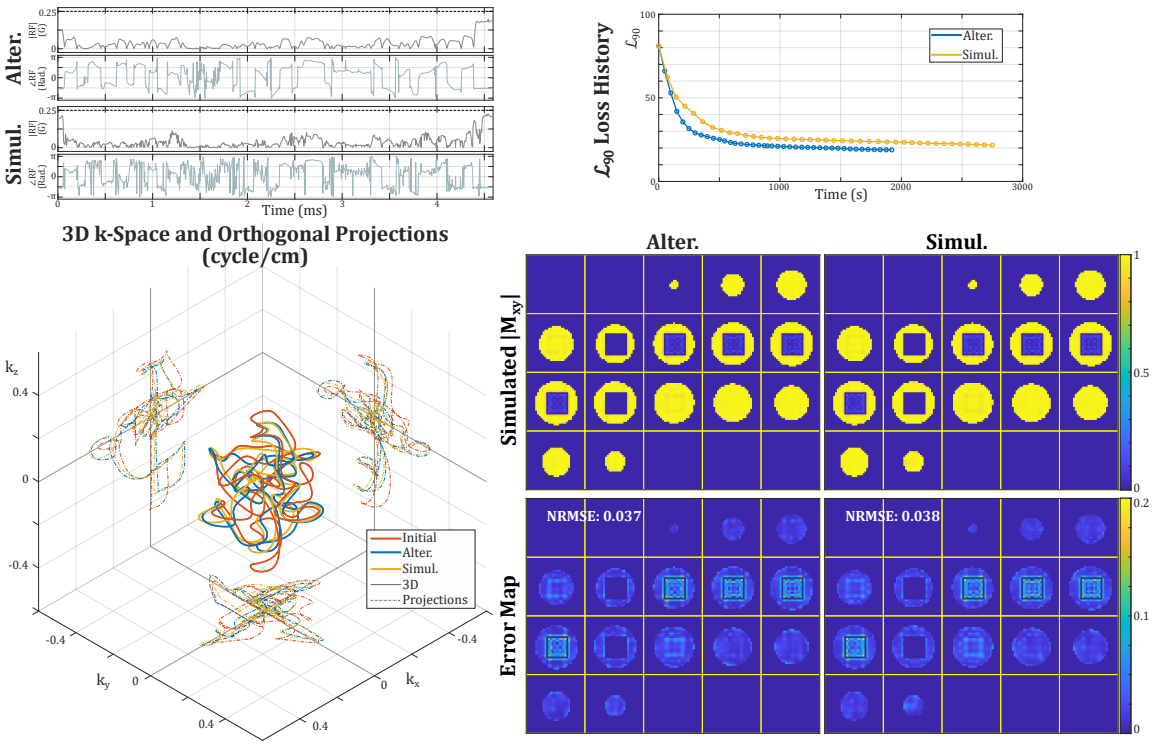

Figure 2.13: Comparison of alternating (Alter.) and Simultaneous (Simul.) minimization. The two approaches find similar but different local minima according to the RF and k-space plots (Left Panel). On the Top Right, the loss for the first 40 iterations is plotted. The computation time for each iteration is slightly longer for the simultaneous L-BFGS updates. The simultaneous updating scheme converges slightly slower, while the eventual excitation performance of the two schemes is comparable (Lower Right).

extended kt-points initializations, the alternating update decreases the design losses faster than the simultaneous update. However, the two objectives are both non-convex in terms of b and g, which makes this behavior difficult to analyze. We therefore cannot claim that this alternating scheme will outperform simultaneous updates in the general case (i.e., for all other design problems).

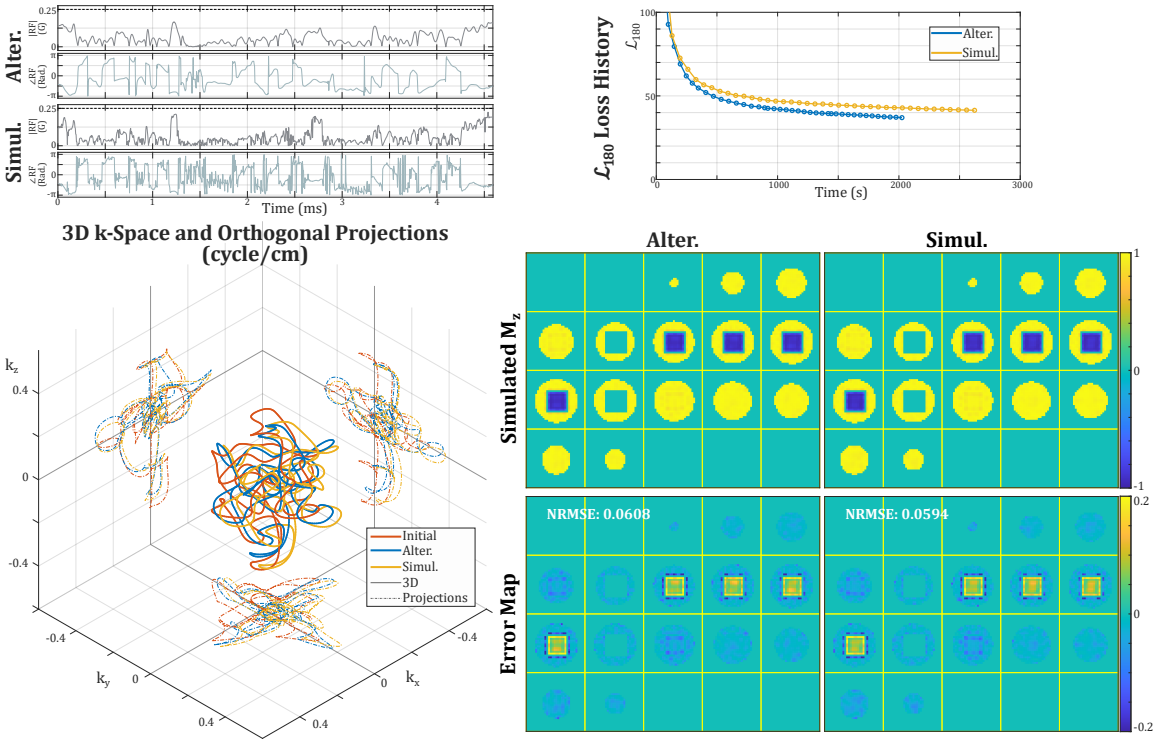

Figure 2.14: Same comparison as in Fig. 2.13, for the IV180 design. The simultaneous approach attains a slightly better inversion, but ends with a higher loss value, in which the RF power penalization is included.

2.7.4 Impact of a small Gradient Delay

On modern MRI scanners, the physically realized gradient waveforms are typically slightly misaligned in time relative to the RF waveform, even after the vendor's built-in gradient delay correction is applied. This delay is on the order of the gradient sampling (dwell, or raster) time, which on our scanner is 4 µs. To assess robustness against such delays, we simulated the excitation produced by the OV90 and IV180 pulses for delays of 4 µs and -4 µs. As shown in Fig. 2.15, such delays can degrade the performance of our designed pulses by 2-3 %.

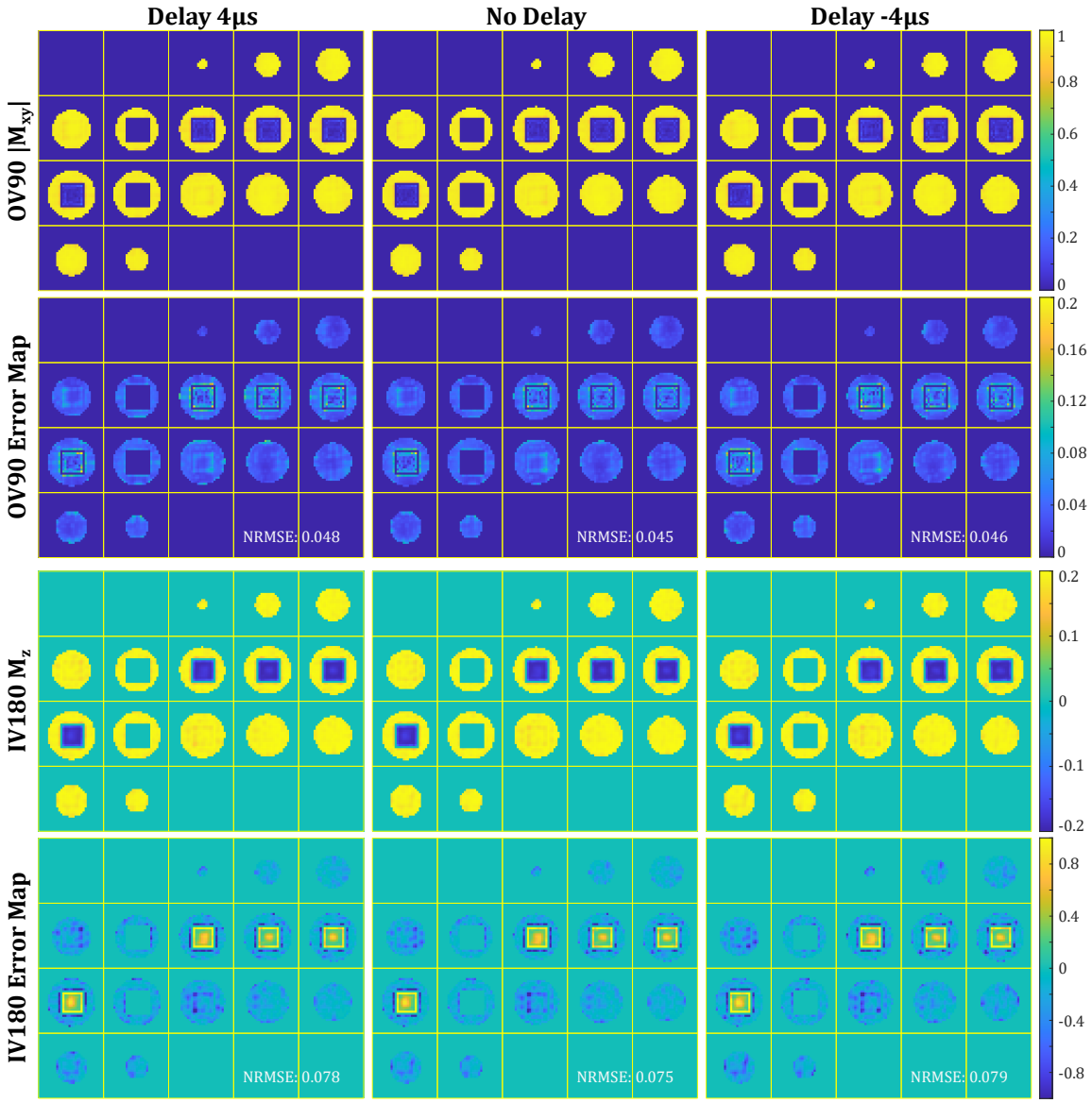

Figure 2.15: Simulated excitation performances of our designed OV90 and IV180 pulses under different delays of the applied gradient fields.

CHAPTER III

Outer Volume Saturation Facilitated Inner Volume Imaging with Application to Functional MRI

3.1 Introduction

3D volumetric steady-state imaging is an alternative to simultaneous multi-slice (SMS) [66] imaging that is free from slice profile artifacts, and is less vulnerable to spin-history effects due to, e.g., blood inflow [16]. In dynamic imaging, such as functional MRI (fMRI), these advantages can yield improved SNR [70], which may be exploited to increase spatiotemporal resolution in such applications. Yet, this SNR benefit requires multi-shot scanning to acquire sufficient 3D k-space signals for reconstruction, which aggravates imaging susceptibility to between-shot physiological signal fluctuations.

A major area of interest to mitigate such physiological noise artifacts is to enable under-sampling, i.e., reducing the number of shots needed for acquisition and reconstruction. Common approaches for this objective are based on parallel imaging [23, 57], and compressed sensing [41]. These two techniques are from the reconstruction perspective. They exploit signal redundancies that were afforded by extra receiving coils and from insightful image priors (e.g., transform sparsity), respectively, such that a small subset of k-space signals are sufficient to solve the imaging

inverse problem. The two techniques can also be combined and/or incorporated with other image priors (e.g., low-rankness [25]) using model-based image reconstruction to attain aggressive under-sampling.

Inner volume (IV) imaging is another approach to reduce the necessary shot numbers in volumetric imaging acquisition. It is from the excitation perspective. For IV imaging, the FOV is divided by the boundary of region of interest (ROI). The region inside the ROI is specified as the IV, and the outside is the OV. Similar to the way SMS restricts each free induction decay (FID) to be from a set of separated slices, such that 2D k-space readouts can fulfill acquisitions, IV imaging attempts to confine the excitation to a restricted ROI, e.g., visual cortex, via 3D spatially tailored excitations. Such spatial confinement can relax the Nyquist sampling distance criteria in the k-space, requiring fewer shots in the acquisition than needed for imaging the whole field-of-view (FOV). This imaging strategy is particularly useful for applications that have an ROI which is much smaller than the FOV, for instance, sensory system functional MRI.

Despite the sampling benefit, for confined ROI applications like sensory system fMRI, IV imaging is still not widely used for accelerating volumetric acquisitions, largely due to its implementation difficulties. Classical 3D IV structural imaging methods are based on combinations of 1D selective adiabatic refocusing pulses. This approach has an extensively long spin preparation time, and is large-tip based. Its requirement for more than one RF pulse to implement makes for a complicated steady state, and is unsuitable for steady-state dynamic imaging. Lately, multi-dimensional small-tip excitation pulse design has been increasingly studied. These works rely on an approximate Fourier relation between the excitation profile and the pulse that was established by the small-tip approximation. In particular, Sun et al. proposed a 3D tailored RF pulse design algorithm that jointly designs excitation RF and gradient waveforms that satisfy hardware constraints [72].

Sun's approach can obtain a high peak flip-angle ratio between IV and outer volume (OV) using conventional single transmit coil MR systems for single-shot 3D IV excitation [73]. However, despite the high ratio, residual excitation remains in the OV in small flip angle, and is problematic. Due to the sub-linear dependence of the signal amplitudes on flip-angles in common steady-state imaging sequences, small flips in the OV produce "amplified" signal relative to the IV signal level. These stubborn OV signals undermine the excitation confinement, limiting the shot number reduction for accelerating acquisition.

In this work, to improve 3D IV imaging signal profile for common steady-state sequences, such as balanced steady-state free precession (bSSFP), spoiled gradient echo (SPGR), and small-tip fast recovery (STFR), we propose an approach that works by prepending the 3D tailored IV excitation with a complementary tailored OV saturation. We show that this sequence modification can suppress the OV signal significantly without disturbing the IV signal noticeably. We demonstrate this IV imaging strategy with a visual cortex IV fMRI study.

3.2 Methods

Fig. 3.1 illustrates how we conduct steady-state OV magnetization saturation. It shows the saturation strategy for three different steady state sequences: bSSFP, SPGR, and STFR. Common across all three sequences, each TR starts with an IV excitation pulse that generates an FID in the ROI. At the end of TR, a 3D tailored OV pulse is played immediately followed by a spoiler gradient. bSSFP and STFR add an extra negative gradient spoiler balancer and tip-up pulse, respectively. The negative gradient spoiler balances the overall gradient moment each TR for obtaining bSSFP signal contrast. At steady-state, OV magnetizations are thus saturated right before entering the formation of FID. In this way, while the IV pulse still has residual excitation outside the IV, the OV spins are already largely saturated by the tailored

OV pulse and the spoiler gradient, contributing little signal to the FID.

We design both the IV excitation pulse and the OV saturation pulse using the methods proposed in [72]. The method starts with a pulse initialization procedure named extended kt-points [72]: a set of phase-encoding (PE) point locations are greedily collected from the k-space given a desired excitation profile. The visiting order of these locations is then determined using the traveling salesman algorithm. The method connects the PE locations using the algorithm described in [42] and yields a smooth initial k-space trajectory. The initial RF is then computed for this trajectory by solving a small-tip RF design problem [79]. This initial RF and k-space trajectory are then optimized alternatingly, i.e., fixing the RF while optimizing the k-space trajectory and vice versa, to minimize ℓ_2 excitation errors to the target excitation profiles.

3.2.1 Experiments

To demonstrate the IV imaging performance of our proposed strategy, we conduct two sets of in vivo experiments on healthy volunteers on a 3T GE MR750 scanner with a single channel transmit coil and 8-channel receiving head coil. In the first set, we use Cartesian readout and compare the steady-state signal profile of IV imaging with and without our proposed OV saturation pulse. We set acquisition parameters as: TR = 25.3 ms; TE = 1.5 ms; IV flip angle 14°; OV flip angle 60°; image matrix size $120 \times 120 \times 40$; FOV = $24 \times 24 \times 20\,\text{cm}^3$. We use the same division of IV and OV across all three sequences to compare the performances across sequence types.

Fig. 3.2 illustrates the target (design) 3D excitation, showing the IV, OV, transition, and background regions, respectively, of this experiment. The target pattern of the tailored OV pulse is essentially the complement of that of the IV pulse: the IV excitation pulse aims to only excite IV spins, and the OV pulse aims to only excite OV spins. A transition region between the OV and IV region is prescribed as a "don't

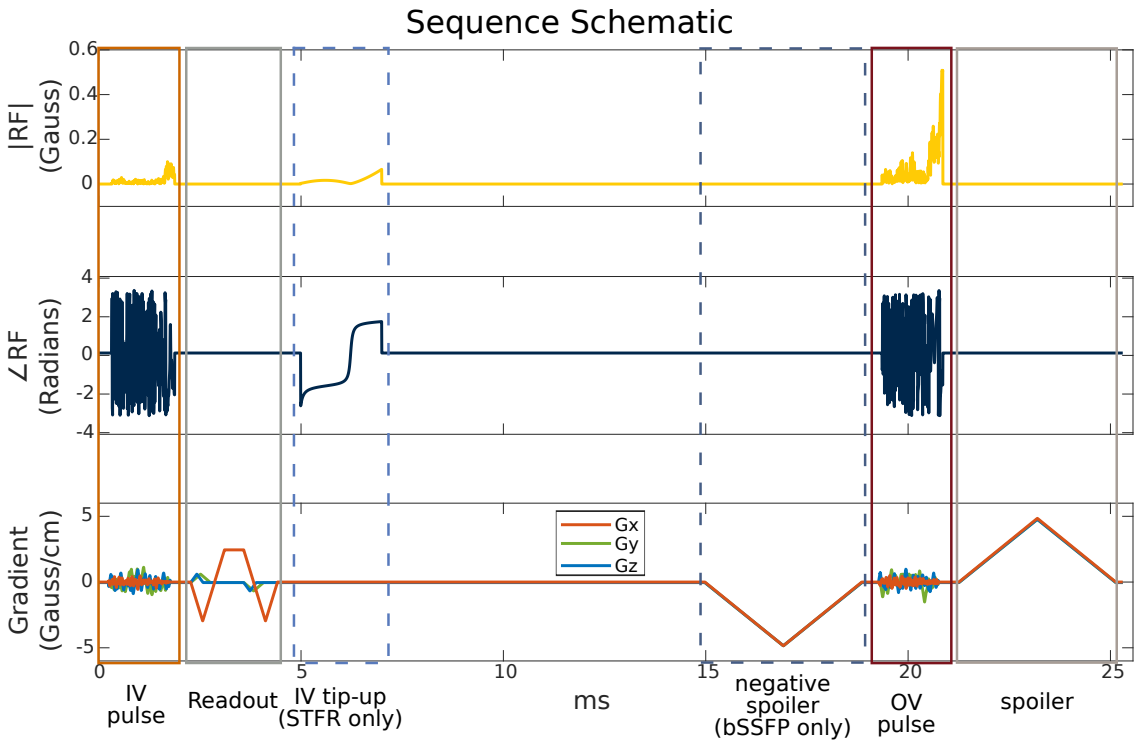

Figure 3.1: Sequence diagram of bSSFP, SPGR, and STFR. The TR and TE are set the same for all three sequences to assess their performances in the OV saturated IV imaging. 3D tailored IV and OV pulses are placed at the beginning and the end of each TR, respectively. So for steady-state imaging, the OV saturation pulse is placed right before the IV excitation pulse, annihilating OV spin magnetizations. For bSSFP and STFR, an extra sequence segment of negative spoiler gradient balancer and tip-up pulse is marked with dashed rectangular, respectively. Before the OV saturation pulse is an interval of free time that extends the overall TR to a typical duration in fMRI imaging, so the non-functional signal profile obtained can help assess the OV saturation strategy performance in steady-state dynamic imaging.

care" region, i.e., it is excluded from the cost function in the pulse design procedure.

The second experiment used a stack-of-spirals-in readout to acquire bSSFP fMRI time series with checkerboard flashing visual stimulus. Due to the banding phenomenon, the bSSFP blood oxygen level dependent (BOLD) fMRI signal level varies with the off-resonance level in voxels [48, 60]. Specifically, the blood oxygenation level affects the off-resonance frequencies in voxels. For a voxel near the bSSFP transition band, a slight change of its off-resonance frequency will cause a large difference in its steady-state spin magnetization magnitude. This change is then reflected on its image signal level, and can amplify the BOLD activity to be imaged. Transition bands are typically very narrow with a width of a few voxels. The BOLD activities we obtain in this experiment are thus a mixtures of pass band and transition band signal variation level.

We conduct three acquisitions for this experiment: 1) high-resolution full brain structural image; 2) high-resolution visual cortex IV brain structural image; 3) low-resolution visual cortex fMRI images. We set acquisition parameters as: TR = 38 ms; TE = 21 ms; IV flip angle 36°; OV flip angle 70°; image matrix sizes $240 \times 240 \times 240$ (high-resolution) and $48 \times 48 \times 20$ (low-resolution); FOV = $24 \times 24 \times 24\,\mathrm{cm}^3$.

The two structural scans are fully-sampled, with a readout composed of 240 platters, each comprised of 12 constant density interleaves. The fMRI scan readout has 20 platters, each comprised of 3 variable density interleaves. It is prospectively under-sampled with factor R = 3 in a rotating manner (rotated stack-of-spirals), with 1 out of 3 interleaves acquired in each platter. The sampling pattern does not change across temporal frames.

For the fMRI scanning, we presented 20 s on-and-off flashing full-field visual stimuli for 5 min. We used the non-Cartesian GRAPPA algorithm from Chapter IV [37] for under-sampled reconstruction before computing fMRI activation correlation maps.

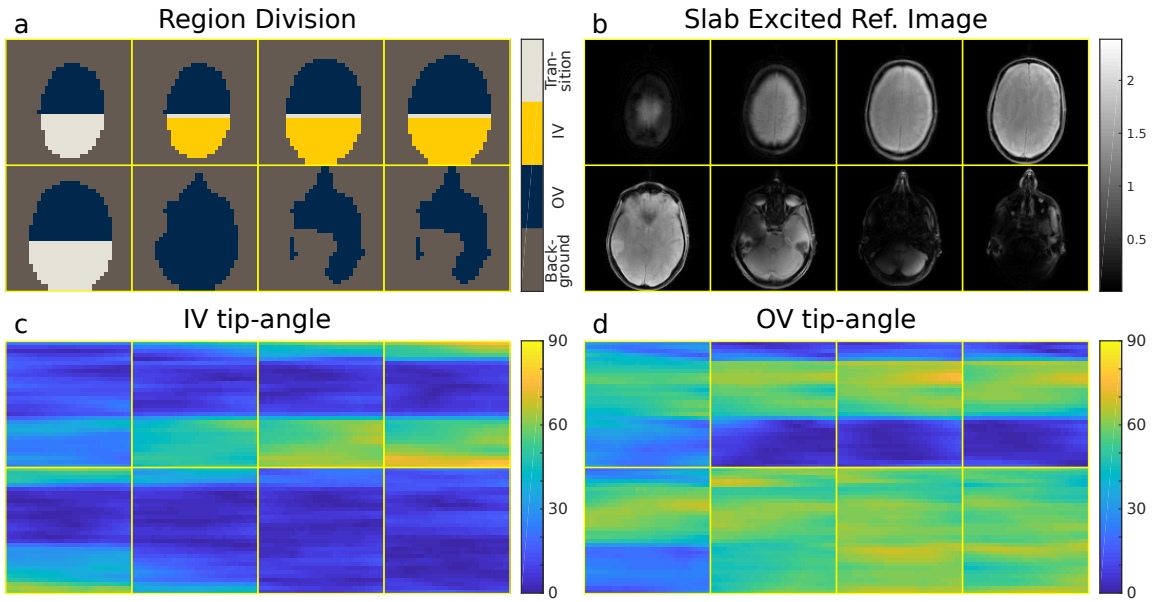

Figure 3.2: Region division and the simulated flip angles for IV and OV respectively. (a) The whole FOV is divided into 4 regions. The Background region is excluded in the pulse design. The Transition region is weighted as 0 in this experiment. IV and OV are the target regions for the 3D IV and OV pulses, respectively. (b) Reference image acquired with slab excitation. (c,d) Bloch-simulated one-shot IV and OV excitations, respectively.

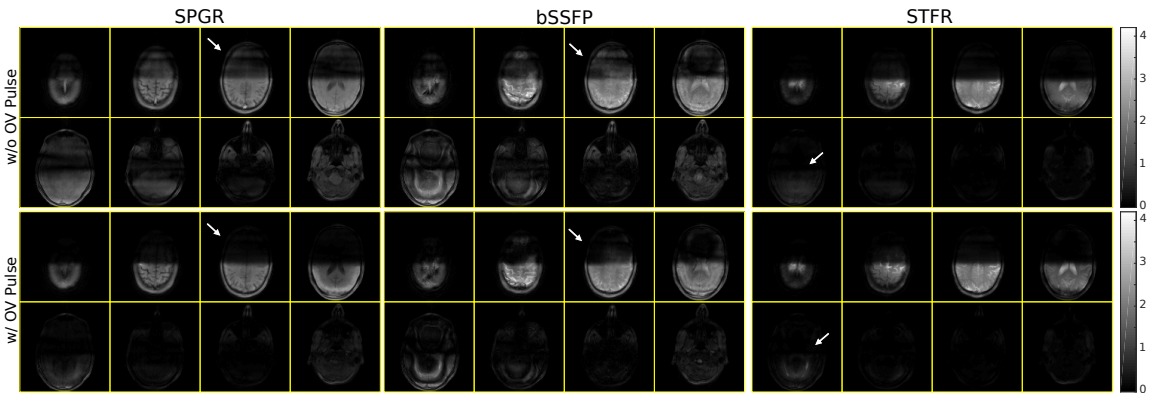

Figure 3.3: Comparison of 3D SPGR/bSSFP/STFR IV imaging both with and without the proposed 3D tailored OV suppression pulse. The suppression pulse removes the majority of the OV signal for SPGR and bSSFP. For STFR, we observe much lower OV signal than in SPGR and bSSFP, both before and after inserting the proposed OV suppression pulse. This is expected, as the tip-up pulse in STFR effectively acts as an OV signal suppression pulse.

3.3 Results

Fig. 3.3 shows the imaging results of the three sequences obtained with and without the tailored OV suppression RF pulse. We quantify the inner volume image quality with the steady-state average signal intensity ratio of IV over OV. The ratios with and without the tailored OV saturation are: SGPR, 15.17/5.38; bSSFP, 13.54/6.52; STFR, 20.31/15.69. We observe good OV suppression with the tailored OV pulse, though some signal in inferior slices remains. For the STFR result, we observe good OV suppression even without the proposed OV suppression pulse, due to intrinsic OV suppression from the tip-up pulse used in STFR.

Fig. 3.4 shows images obtained with whole-brain excitation and IV excitation, respectively, and demonstrates the spatial selectivity of our sequence. These slices are regularly spaced. With OV saturation, the excited region is mainly confined to the visual cortex, as intended.

Fig 3.5 shows the bSSFP functional activation results. Time-course correlations

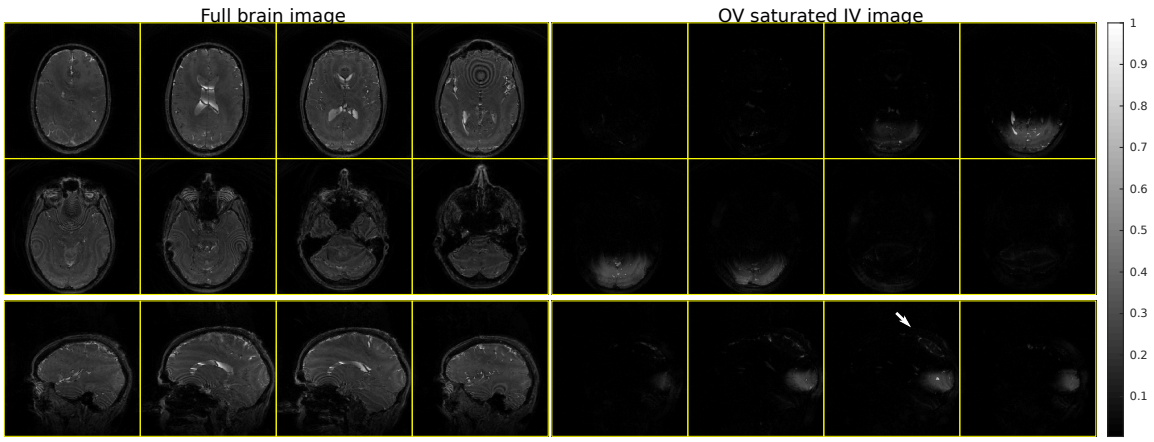

Figure 3.4: Structural IV bSSFP image, 1mm isotropic resolution. There are unsuppressed OV signals (arrow) near the top of the brain (in the sagittal images); we expect that a larger OV flip-angle can mitigate this problem.

with the stimulus are thresholded at 0.57. We observe high correlations within the visual cortex as expected. The averaged time-course plot is computed using the voxels with correlations above a threshold of 0.57. We observe $\sim$6 % signal changes.

3.4 Discussion and Conclusion

We have proposed an approach for OV signal suppression in 3D IV imaging based on saturating the OV steady-state signal using a 3D tailored RF pulse combined with spoiler/echo-shifting gradient(s). We demonstrated the efficacy of this approach for SPGR, bSSFP, and STFR sequences.

We also demonstrated that it is possible to perform IV bSSFP fMRI with good OV suppression, using tailored IV excitation and OV saturation pulses of approximately 4 ms duration each, which is short enough for use with the fMRI sequence used here. Including a pair of balanced gradient spoilers ensures that the sequence is balanced for IV spins, and unbalanced for OV spins.

For future work, it is of interest to quantifically comparing fMRI activation acquired with and without IV imaging, using metrics such as, ROC plots [69], etc.

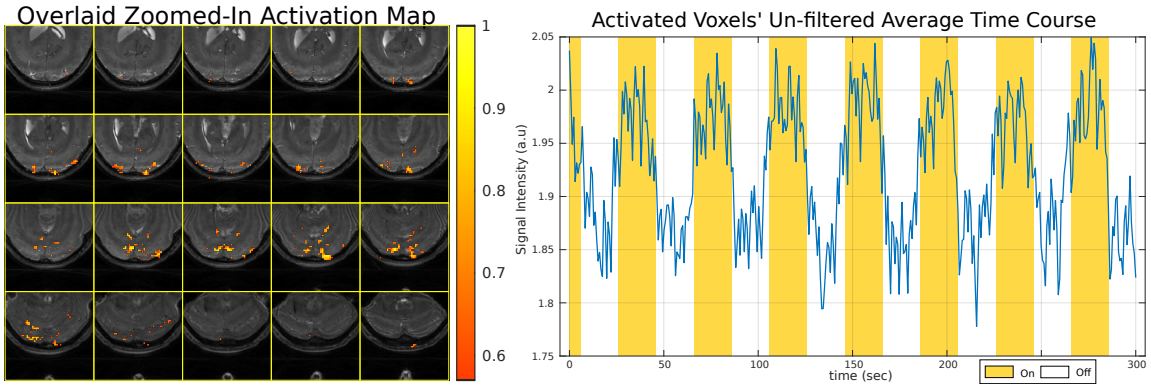

Figure 3.5: (Left) Temporally-filtered absolute functional activity map overlaid onto the zoomed-in high-resolution structural image. We primarily observe activations in the visual cortex as expected. (Right) Time course averaged across voxels with functional activation (correlation) above 0.57. Yellow region indicates visual stimuli are on. We observe $\sim$6 % signal changes between stimulus on and off.

CHAPTER IV

A GRAPPA Algorithm for Arbitrary 2D/3D Non-Cartesian Sampling Trajectories with Rapid Calibration

4.1 Introduction

Multi-shot 3D volumetric imaging is a potential alternative to single-shot simultaneous multislice (SMS) [66] imaging in, e.g., fMRI [70][1]. Compared to 2D (slice-selective) acquisitions, 3D volumetric imaging is free from slice profile artifacts, has reduced spin-history effects (due to, e.g., in-flow or motion) [16, 61], and can provide improved image SNR [35, 70]. However, a drawback of multi-shot acquisitions is increased susceptibility to physiological signal fluctuations between shots. To mitigate physiological noise artifacts, it is beneficial to accelerate the acquisition using parallel imaging (PI) such that the total acquisition time per volume is reduced (to, e.g., one-half the heartbeat interval [53, 54]). In addition, non-Cartesian readout trajectories are desirable due to their high sampling efficiency.

Iterative parallel imaging (PI) methods such as cg-SENSE [56, 57] and SPIRiT [43] are often chosen due to their ability to handle arbitrary readout trajectories. However, in applications such as fMRI where a time-series of several hundred images

[1]This chapter was published in [37].

must be reconstructed, the total reconstruction time using iterative methods can become prohibitive (since each image is reconstructed independently). To speed up reconstruction, several groups have adopted dedicated computational hardware, e.g., general purpose GPUs [6], but working with such hardware usually requires specialized programming expertise making such methods less widely transferable. Therefore, there is currently a need for an easy-to-implement and robust reconstruction technique for arbitrary non-Cartesian trajectories that scales well with the total number of image frames (i.e., short overall reconstruction time for a time-series) even on common hardware (e.g., CPUs).

The foremost non-iterative alternative to the above-mentioned ways is GRAPPA [23, 55, 74], which (like SPIRiT) does not require explicit knowledge of receive coil sensitivity maps. Unfortunately, using GRAPPA with non-Cartesian trajectories has so far been somewhat awkward. Incomplete remedies have been explored for certain kinds of readouts. For example, [22, 28] splits stack-of-stars or stack-of-spirals readouts into segments, and then performs conventional Cartesian GRAPPA within each segment. However, this procrustean approach cannot be used with arbitrary 3D non-Cartesian trajectories, and requires a large amount of ACS data for kernel estimation. In addition, a trade-off must be made between trajectory segmentation and reconstruction quality. Alternatively, one can grid a Cartesian dataset from a non-Cartesian one and apply ordinary GRAPPA thereafter (based on GROG [63, 64]). However, reliable and accurate gridding may be problematic for under-sampled trajectories due to its 1D interpolation nature that limits its effectiveness for interpolation beyond 1 k-space sample distance. Another approach that has been used in some applications is through-time GRAPPA [65], which alternates sampling patterns along time and extends the reconstruction along the temporal dimension. However, its calibration may become ill-conditioned when dynamic imaging contrast is predominantly stationary, e.g., as in fMRI. Finally, PARS [78] and kSPA [34] are non-iterative k-space

reconstruction methods that are closely related to GRAPPA, however they both rely on explicit knowledge of receive coil sensitivity maps. PARS needs to calibrate a large number of kernels which can be time-consuming in practice, while kSPA requires extra tuning of the order of its polynomial approximations. Neither of the approaches has become widely adopted.

We propose a conceptually simple method for generalizing GRAPPA to arbitrary 3D non-Cartesian PI acquisitions, and provide an efficient algorithm for its calibration [36]. For each unsampled k-space location (with a distinct local sampling constellation), our method assigns a unique GRAPPA kernel, whose calibration is efficiently implemented by utilizing the phase-shift property of the FT and the NUFFT [17, 75]. Like Cartesian GRAPPA, our method does not require explicit coil sensitivity information, and reconstruction per image volume (once weights have been calibrated) is rapid. Apart from choice of GRAPPA kernel size and Tikhonov regularization coefficient (which is also typically used in conventional GRAPPA), our proposed method is fully automatic and does not require manual parameter selection based on, e.g., segments or other trajectory-dependent aspects. Thus, our method (once coded) requires minimal user expertise and should be broadly applicable to arbitrary non-Cartesian PI applications. We demonstrate our method in 3D rotated stack-of-stars [84] and rotated stack-of-spirals [15] structural (T1-weighted) imaging, and 3D rotated stack-of-spirals fMRI, using the same implementation parameters for all three cases.

4.2 Methods

4.2.1 Background: General Principle of GRAPPA

It is helpful to consider two aspects of the GRAPPA kernel: the local sampling *constellation*, and the associated *weights*. The constellation captures the relative po-

sitions between the unsampled ("center") location that we wish to reconstruct and its sampled local neighbors, and the weights are the coefficients for later reconstruction. Depending on the sampling trajectory, there are usually many different constellations in one dataset, and thus multiple sets of weights will be required for the reconstruction.

GRAPPA works by first identifying all distinct constellations. Then, for each constellation, to solve for its weights, GRAPPA collects all combinations of data from the ACS region whose relative positions match that of the constellation. Figure 4.1 (b) shows a simple two-neighbor (upper-left and lower-right) illustrative example for reconstructing an unsampled (center) location. For each combination, the neighbors across N_c coils that surround its center form one row of matrix $\tilde{A} = [\tilde{A}_1, \ldots, \tilde{A}_{N_c}]$, where the tilde indicates that these signals reside in k-space. The submatrix $\tilde{A}_c \in \mathbb{C}^{N_k \times N_n}$ collects the neighbors from the c^{th} coil. Here N_k is the number of combinations (within the ACS region) one can collect that match the constellation being calibrated, and N_n is the number of neighbors inside the constellation. In addition, $\tilde{b}_c$ denotes the vector of center values from the c^{th} coil. The weights are obtained by solving the least square (LS) problem

$$\arg\min_{w_c} \|\tilde{A}w_c - \tilde{b}_c\|_2^2 + \lambda\|w_c\|_2^2 \Rightarrow w_c^\star = (\tilde{A}^H\tilde{A} + \lambda I)^{-1}\tilde{A}^H\tilde{b}_c, \qquad [4.1]$$

where $w_c \in \mathbb{C}^{N_c \cdot N_n}$ denotes the vector of weights for reconstructing an unsampled value for the c^{th} coil, and λ is the Tikhonov regularization coefficient.

When forming $\tilde{A}$ and $\tilde{b}_c$ in this way, the columns of $\tilde{A}$ contain the (vectorized) signals from partially overlapping rectangles, e.g., as shown in Fig. 4.1 (b). In the limit of a very large ACS region, these rectangles resemble shifted replicas of each other, apart from differences at the boundary.

4.2.2 Proposed Non-Cartesian GRAPPA

The procedure described above is well suited for Cartesian acquisitions, where all locations, whether sampled or not, are on a Cartesian grid. Hence, for any peripheral constellation, one can always find matching combinations within the ACS region, and arrange them into the LS form Eq. [4.1]. For non-Cartesian sampling, given a fixed size, constellations can be efficiently identified using kd-tree [3]: Specifically, this algorithm forms a tree-structure given all k-space points of interest; It returns all neighboring points within a given distance (i.e., prescribed GRAPPA kernel-size) when queried with a certain (unsampled) location. Following constellation identification, the calibration can no longer be directly conducted in the conventional way. However, while there may be no direct match from the "on-grid" (Cartesian) ACS region for an off-grid non-Cartesian GRAPPA constellation, we propose to obtain such matches through the phase-shift property of the DFT. We first inverse FT the on-grid ACS data, and then modulate the resulting image by a linear phase corresponding to the wanted off-grid shift. This phase-modulated image is then Fourier transformed back to k-space. The whole procedure is equivalent to a periodic sinc interpolation. A simple example of our approach is illustrated in Fig. 4.1 (c), for a typical constellation β from a spiral readout, where two off-grid (black) neighbors are assumed sampled and selected to reconstruct the unsampled center. The key idea is to "synthesize" the corresponding non-Cartesian (off-grid) combinations from an on-grid ACS dataset. These interpolated ACS data are again arranged as in Eq. [4.1], yielding the associated GRAPPA weights.

Following the calibration phase, we use the GRAPPA weights (one set for each unique constellation) to reconstruct unsampled k-space coil data and restore a full non-Cartesian k-space. We then reconstruct coil images using the NUFFT adjoint method with Voronoi density compensation [17, 58]. Final images are obtained using sum-of-squares [23] or linear [59] coil image combinations.

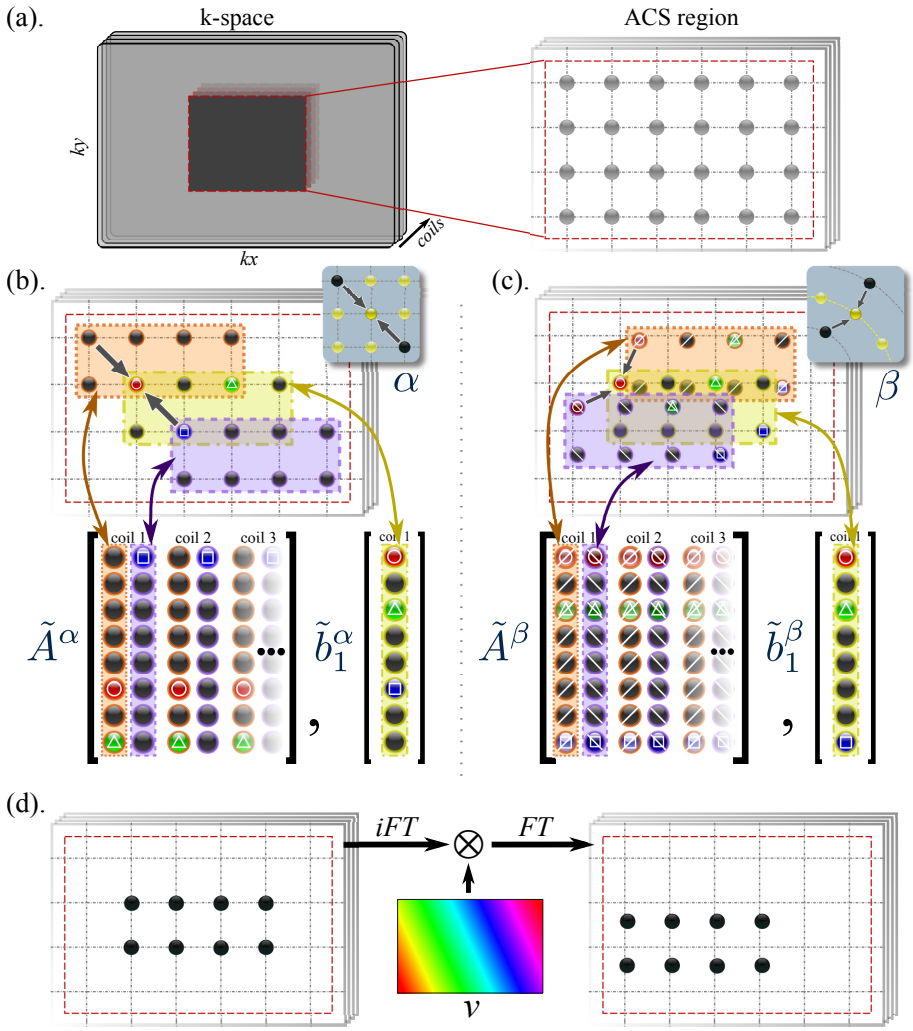

Figure 4.1: Comparison of (b) conventional Cartesian GRAPPA, and (c,d) the proposed non-Cartesian GRAPPA calibration. In both methods, the ACS region consists of densely (Nyquist) sampled (or gridded) Cartesian data, illustrated in (a). (b) Conventional (Cartesian) GRAPPA reconstruction using the constellation "α", that is composed of two sampled neighbors and an unsampled center. Several "combinations" are identified from the ACS region that match the desired constellation (α); in (b), there are eight such combinations, whose center points form the yellow rectangle in (b). Each combination gives rise to one row in $\tilde{A}$, and one element in $\tilde{b}_c$ (cf. Eq. [4.1]). As a result, the orange and purple rectangles form columns of $\tilde{A}$ (after vectorizing), and $\tilde{b}_c$ is formed from the yellow rectangle. (c,d) Proposed non-Cartesian GRAPPA reconstruction. For a non-Cartesian (local) constellation "β", we synthesize off-grid ACS data by using the phase-shift property of the Fourier transform (d). The GRAPPA weights for "β" are then obtained as in conventional GRAPPA (cf. Eq. [4.1]).

4.2.3 Efficient Implementation of the Proposed Method

Our non-Cartesian approach must calibrate a different set of GRAPPA weights for every distinct constellation. In the most general case, the number of distinct constellations equals the number of unsampled points. The workload for many applications may be less, since commonly used trajectories such as rotated stack-of-stars, stack-of-spirals, and stack-of-cones [24], often possess certain k-space regularities (e.g., regular sampling along k_z) that reduce the number of distinct constellations that must be calibrated. Nevertheless, depending on acquisition parameters such as image matrix size and under-sampling factor, calibration can still be computationally demanding or even impractical. To address this, we propose the following approximate algorithm for efficient implementation of our method. Specifically, we accelerate calibration by avoiding any explicit interpolation to ACS datasets.

In the formation of $\tilde{A}$ and $\tilde{b}_c$ in Eq. [4.1], the signals at the boundary of the ACS region are normally excluded, as forming matching combinations for them would require signals residing outside the region. In our fast algorithm, boundaries are included, by adopting circulant boundary conditions. In other words, we allow the ACS region to wrap around as needed to complete a given constellation. The underlying assumption of this approach is that when the ACS region is sufficiently large, the impact of these wrap-around combinations is marginal. With the circulant boundary assumption, the rectangles described above are now fully overlapping in a circularly-shifted manner, and the columns formed from them, that assemble the matrix $\tilde{A}_c$ and the center vector $\tilde{b}_c$, are now circularly shifted replicates. This allows efficient calibration as follows.

Figure 4.2 illustrates our algorithm for efficient weight calibration. We denote unitary $F \in \mathbb{C}^{N_k \times N_k}$ as the (2D or 3D) DFT of the ACS region size. The solution to

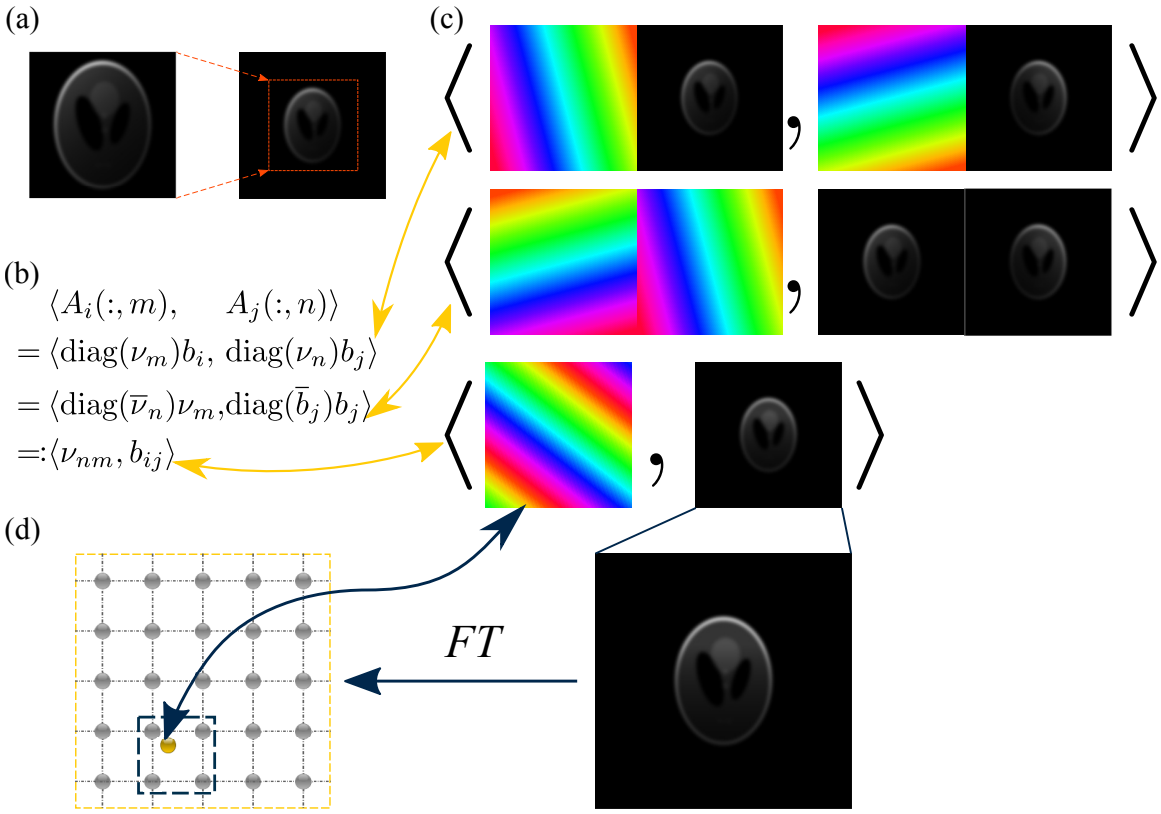

Figure 4.2: Illustration of our method for fast computation of $A^H A$ (Eq. [4.2]). (a) A (low-resolution) coil image is reconstructed from the ACS data, and this image is then zero-padded to increase the spatial frequency (Fourier) sampling density. (b,c) The $(m,n)^{\text{th}}$ element of $A_i^H A_j$ is a pixel-wise multiplication of coil images b_i and b_j, each modulated by a linear phase ν_m and ν_n, respectively. The result is a single "coil-product" image b_{ij} modulated by the combined phase image v_{nm}. (d) The frequency component corresponding to the combined linear phase is efficiently calculated by interpolating the Fourier transform of the zero-padded coil-product images, in NUFFT-like fashion.

(1) can now be written as:

$$\begin{aligned} w_c^\star &= (\tilde{A}^H\tilde{A} + \lambda I)^{-1}\tilde{A}^H\tilde{b}_c = (A^H F^H F A + \lambda I)^{-1} A^H F^H F b_c \\ &= (A^H A + \lambda I)^{-1} A^H b_c, \end{aligned} \quad [4.2]$$

where $A = [A_1, \ldots, A_{N_c}]$ and b_c are in the image domain (tilde symbol removed). In particular, b_c is a low-resolution coil image obtained by inverse FT of all of the ACS data. Then, we represent $A^H A$ and $A^H b_c$ with block matrices:

$$A^H A = \begin{pmatrix} A_1^H A_1 & \cdots & A_1^H A_{N_c} \\ \vdots & \ddots & \vdots \\ A_{N_c}^H A_1 & \cdots & A_{N_c}^H A_{N_c} \end{pmatrix}, \quad A^H b_c = \begin{pmatrix} A_1^H b_c \\ \vdots \\ A_{N_c}^H b_c \end{pmatrix}. \quad [4.3]$$

Due to the circulant attribute, the columns in A_c are essentially b_c modulated by different linear phases, i.e., $\mathsf{diag}(\nu) b_c$, where $\nu \in \mathbb{C}^{N_k}$ is a phase vector. Analytically, these phase vectors are formed following the Fourier relation. That is, for a k-space neighbor at position p relative to its constellation center as origin, the qth element in its corresponding v, which modulates image domain location q, is $[v]_q = \exp(\iota 2\pi p^T q)$. Let $m, n = 1, \ldots, N_n$. In MATLAB notation:

$$\begin{aligned} A_i(:,m)^H A_j(:,n) &= b_i^H\, \mathsf{diag}(\nu_m)^H\, \mathsf{diag}(\nu_n) b_j = (\nu_m^H\, \mathsf{diag}(\nu_n))(\mathsf{diag}(b_i)^H b_j) \\ A_i(:,m)^H b_j &= (\nu_m^H)(\mathsf{diag}(b_j)^H b_j), \end{aligned} \quad [4.4]$$

where $i, j = 1, \ldots, N_c$ are coil indices. In other words. the elements of $A_i^H A_j$ and $A_i^H b_j$ are inner products between various linear phases and fixed (low-resolution) "coil product" images, i.e., these elements are (spatial) frequency components of $\mathsf{diag}(b_i)^H b_j$. Analytically, that is, if for any image domain location q, the element, $[v_m]_q$, can be expressed as $\exp(\iota 2\pi p^T q)$ with the same p, then $A_i(:,m)^H b_j = [F\, \mathsf{diag}(b_j)^H b_j]_p$, the

frequency p component. An important property of this formalism is that b_c, and hence all coil product images, are shared across all constellations; therefore we pre-compute and cache the FT of $\mathsf{diag}(b_i)^H b_j$. For non-Cartesian constellation calibration involving off-grid frequency components of $\mathsf{diag}(b_i)^H b_j$, we prepare and cache the dense frequency spectrum of $\mathsf{diag}(b_i)^H b_j$ by zero-padding and Fourier transforming (which only needs to be done once), and then linearly interpolate to the desired off-grid frequency (as commonly done in NUFFT [17]). This algorithm reduces the complexity of computing $A^H A$ and $A^H b_c$ from $\Theta(N_k N_n^2 N_c^2 + N_k N_n N_c)$ to $\Theta(N_n^2 N_c^2 + N_n N_c)$, where N_k can reach several thousand (e.g., for a 3D ACS region of size $20 \times 20 \times 20$).

We implemented the proposed calibration in C as it requires efficient indexing for NUFFT interpolation. Compared to the MATLAB inner-product approach (Eq. [4.4]), our NUFFT interpolation based approach can speed up the calibration about 20-fold (results not shown). Once the calibration is done, the non-iterative reconstruction stage for each image frame requires only a few seconds, as opposed to minutes using iterative methods.

4.2.4 Algorithm and Implementation Details

The proposed non-Cartesian GRAPPA reconstruction procedure comprises the following list:

Calibration:

1. Prepare a Cartesian ACS dataset: This can be from either direct Cartesian acquisition at the center of the k-space, or from gridding a Nyquist sampled non-Cartesian k-space center. An ACS dataset of size $20 \times 20 \times 20 \times N_c$ has been found sufficiently large in this work. For consistent Tikhonov regularization behavior, normalize the ACS dataset by the mean square root energy across coils.

2. Identify all the constellations for calibration: From the complete readout trajectories, build a kd-tree including all sampled k-space locations. For each un-sampled k-space point, query the kd-tree with the prescribed distance (i.e., GRAPPA kernel size). This will identify all the neighbors of each center, and constellations can then be formed with the relative positions between neighbors and centers. For the un-sampled points located near the center of (the highly oversampled) k-space, the signals from close neighbors are nearly linearly dependent, and can cause ill-conditioning in calibration when Tikhonov coefficient is not chosen carefully. It is hence helpful to sift the crowded neighbors, i.e., by grouping the neighbors by Nyquist Δk rounded relative positions, and arbitrarily picking one from each group to keep as a neighbor. By sifting the neighbors such that the relative distances of the remaining neighbors become sufficiently large, we avoid any linear-dependence issue in calibration. The unique constellations identified in this step are to be calibrated. The number of neighboring points in each constellations varies from a few points (e.g., 3), when it is in the peripheral k-space, to $5 \times 5 \times 5 = 125$ capped by sifting with kernel size of 5, when it is in the center k-space.

3. Maintain a table of un-sampled points and the constellations they correspond to. With the recorded relative positions, GRAPPA coefficients are calibrated by the algorithm described in 4.2.3. With a normalized ACS, Tikhonov coefficient $\lambda = 5 \times 10^{-7}$ is used in all experiments in the following sections, which was found to produce consistent reconstruction quality.

Reconstruction:

1. Index the un-sampled points from the under-sampled dataset by the table maintained in the calibration step. For each un-sampled point, extract its sampled neighbors in the constellation it corresponds to, and reconstruct the missing

signal using the calibrated coefficients.

2. Use NUFFT adjoint method to resolve the reconstructed coil images, and combine them using either sum-of-squares or linear combinations to yield the final images.

Matlab and C code are opensourced on GitHub[2].

4.2.5 Experiments

To evaluate our proposed algorithm from Section 4.2.3 in terms of reconstruction time and image quality, we first acquired 3D fully-sampled stack-of-stars and stack-of-spirals spoiled GRE datasets in healthy volunteers on a GE 3T scanner with an 8-channel receive-only head coil. Both the fully-sampled stack-of-stars and stack-of-spirals readouts contained 20 k_z platters. For the stack-of-stars dataset, each platter contained 315 spokes. Its image matrix size and FOV were $200 \times 200 \times 20$ and $24 \times 24 \times 10\,\mathrm{cm}^3$, respectively. For the stack-of-spirals dataset, each platter contained 12 interleaves. The image matrix size and FOV were $240 \times 240 \times 20$ and $24 \times 24 \times 10\,\mathrm{cm}^3$, respectively. 8 out of 20 slices are aliased in z direction, due to the excitation profile being slightly larger than $\mathrm{FOV_z}$. Other acquisition parameters were: TR=15 ms, minimum TE, and flip-angle 8°.

We retrospectively under-sampled these two non-Cartesian acquisitions. Spokes and interleaves in-plane were regularly skipped to simulate different acceleration factors. For 3D reconstruction experiments, in the through-plane direction, we rotated the under-sampling pattern to improve reconstruction quality [15, 84]. This rotation produces a through-plane under-sampling factor that is separately the same as the in-plane and the overall under-sampling factor. While this rotation may complicate existing non-iterative methods (e.g., [15]), it does not impact our proposed non-Cartesian method (from the algorithmic viewpoint; it does of course impact the

[2]Code available at https://github.com/tianrluo/NonCrtGRAPPA

constellations that will be identified for a given kernel size). For 2D reconstruction experiments, prior to the in-plane regular under-sampling, an inverse Fourier transform in k_z is applied to convert the 3D k-space into 2D.

To compare with iterative methods (cg-SENSE), and evaluate the feasibility of reducing total reconstruction time for a non-Cartesian time-series acquisition, we acquired a prospectively 3D under-sampled (acceleration factor R=3) multi-interleaf rotated stack-of-spiral-in fMRI dataset of the motor cortex, with finger tapping stimulus. The readout has 20 k_z platters, each containing 9 interleaves (under-sampled to 3 interleaves). Other sequence parameters are: TR=36.2 ms, TE=26.3 ms, flip-angle 8°, image matrix size $88 \times 88 \times 20$ with 131 temporal frames, FOV $22 \times 22 \times 6\,\mathrm{cm}^3$. In this case, the excitation bandwidth was set to match $\mathrm{FOV_z}$. In [74], it was suggested that GRAPPA kernels can be pre-calibrated using a separate dataset with possibly different contrasts. Accordingly, in the subject undergoing fMRI we also acquired a fully-sampled stack-of-spiral-out dataset with the same FOV as the fMRI acquisition but with a different contrast: TR=30 ms, minimum TE, flip-angle 8°, image matrix size $220 \times 220 \times 20$. Again, the experiments were conducted on a GE 3T scanner using an 8-channel receive-only head coil.

In all experiments, our proposed non-Cartesian GRAPPA only used the k-space center data for autocalibration. With these data, we compute low-resolution coil-product images of size $20 \times 20 \times n_z$, (n_z slices); and zero-pad them to $128 \times 128 \times n_z$ ($\sim$5x larger than ACS size, rounded up to power of 2) to attain the dense frequency spectrum needed for our efficient algorithm. A smaller padding size suffers from reduced accuracy, and a larger padding size trades off efficiency for only marginal improvement in quality. The reconstructed coil images of our method are linearly combined into final images. The sensitivity maps used in coil-image linear combination, cg-SENSE, and g-factor simulations are estimated using the methods de-

veloped in [1]. cg-SENSE reconstruction was implemented using MIRT[3], and was preconditioned with density compensation function [58] for acceleration. A quadratic roughness penalization was used to avoid overfitting the readouts [56]. SPIRiT reconstruction was done using its reference implementation[4]. We used the g-factor to measure noise amplification. Since a direct g-factor calculation for non-Cartesian imaging is intractable, we pseudo-replicated noisy reconstructions [59] in simulation (1000 times) using coil noise covariance information measured *in vivo*.

4.3 Results

Figures 4.3 and 4.4 compare our non-Cartesian GRAPPA with cg-SENSE on the retrospectively under-sampled 3D rotated stack-of-stars and rotated stack-of-spiral spoiled GRE datasets, respectively. For each comparison, one slice that is rich in structural detail is displayed (out of total 20 slices). The GRAPPA kernel size is 5×5×5 in the units of the Nyquist sampling distance. Figures 4.5 and 4.6 compare our non-Cartesian GRAPPA with SPIRiT on the retrospectively under-sampled 2D star and spiral spoiled GRE datasets, respectively. The kernel size used in both non-Cartesian GRAPPA and SPIRiT is 7x7. As shown in both sets of comparisons, our non-Cartesian GRAPPA can rival cg-SENSE and SPIRiT in reconstruction quality in terms of error map and g-factor map. There is no modification to our algorithm or implementation across the comparisons, demonstrating the generality of our method.

From these two comparisons, besides noise, we observe edge-like (high-frequency) reconstruction errors with our method, which is also the case for the cg-SENSE and SPIRiT reconstructed images. This error behavior is expected, since for high-frequency (peripheral) k-space regions the non-Cartesian readout generally becomes sparser, such that fewer (sampled) neighbors are available for reconstruction. In addi-

[3]Michigan Image Reconstruction Toolbox, http://web.eecs.umich.edu/~fessler/code
[4]SPIRiT, http://people.eecs.berkeley.edu/~mlustig/Software.html

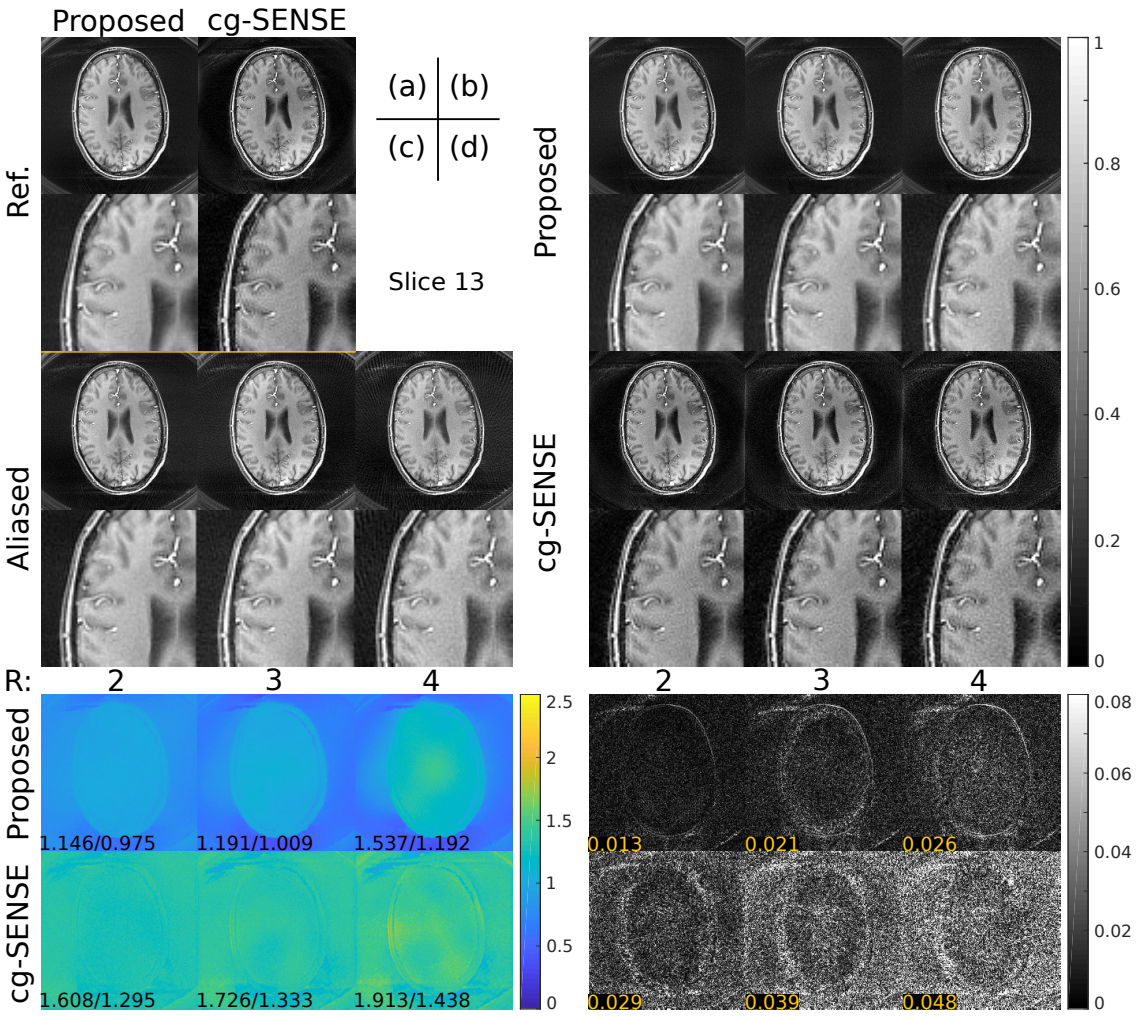

Figure 4.3: 3D reconstruction comparison between the proposed non-Cartesian GRAPPA and cg-SENSE, from retrospectively under-sampled rotated stack-of-stars dataset. One out of a total twenty slices are shown. Coil images reconstructed with the proposed method are linearly combined into final images. (a) (fully-sampled) reference images and under-sampled aliased images (acceleration factor R=2,3,4) with zoom-in details. The reference for the proposed method is reconstructed with NUFFT adjoint [17, 58]. (b) reconstructed under-sampled images of our proposed method and cg-SENSE. (c) g-factor maps of the two methods. The reported numbers are "max/mean" g-factors within the whole 3D head. (d) error map of the two methods. The reported numbers are "root mean squared error (RMSE)" within the whole 3D head. Our non-Cartesian GRAPPA algorithm rivals cg-SENSE in reconstruction quality. As expected, reconstruction error (gray-scale windowed 12.5x) is largest near object edges, since there are fewer samples available for reconstruction at higher spatial frequencies. The g-factor maps for the two methods are within a similar range, as observed in previous work [59].

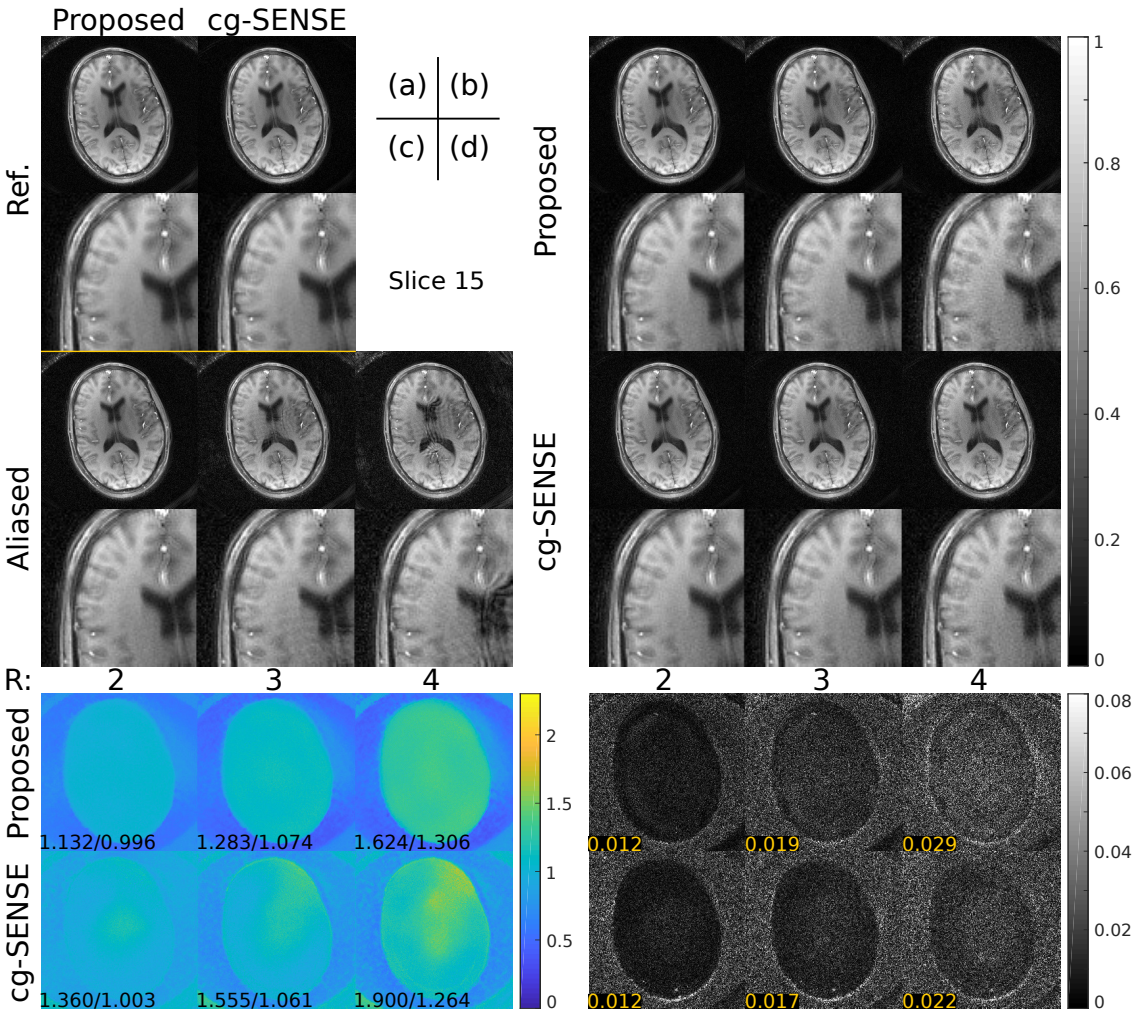

Figure 4.4: A similar comparison as in Fig. 4.3, but for a retrospectively undersampled rotated stack-of-spirals dataset. One slice out of twenty is shown. Without modifying our algorithm and implementation from Fig. 3, our non-Cartesian GRAPPA algorithm again achieves comparable reconstruction quality to cg-SENSE (error images' gray-scale windowed 12.5x).

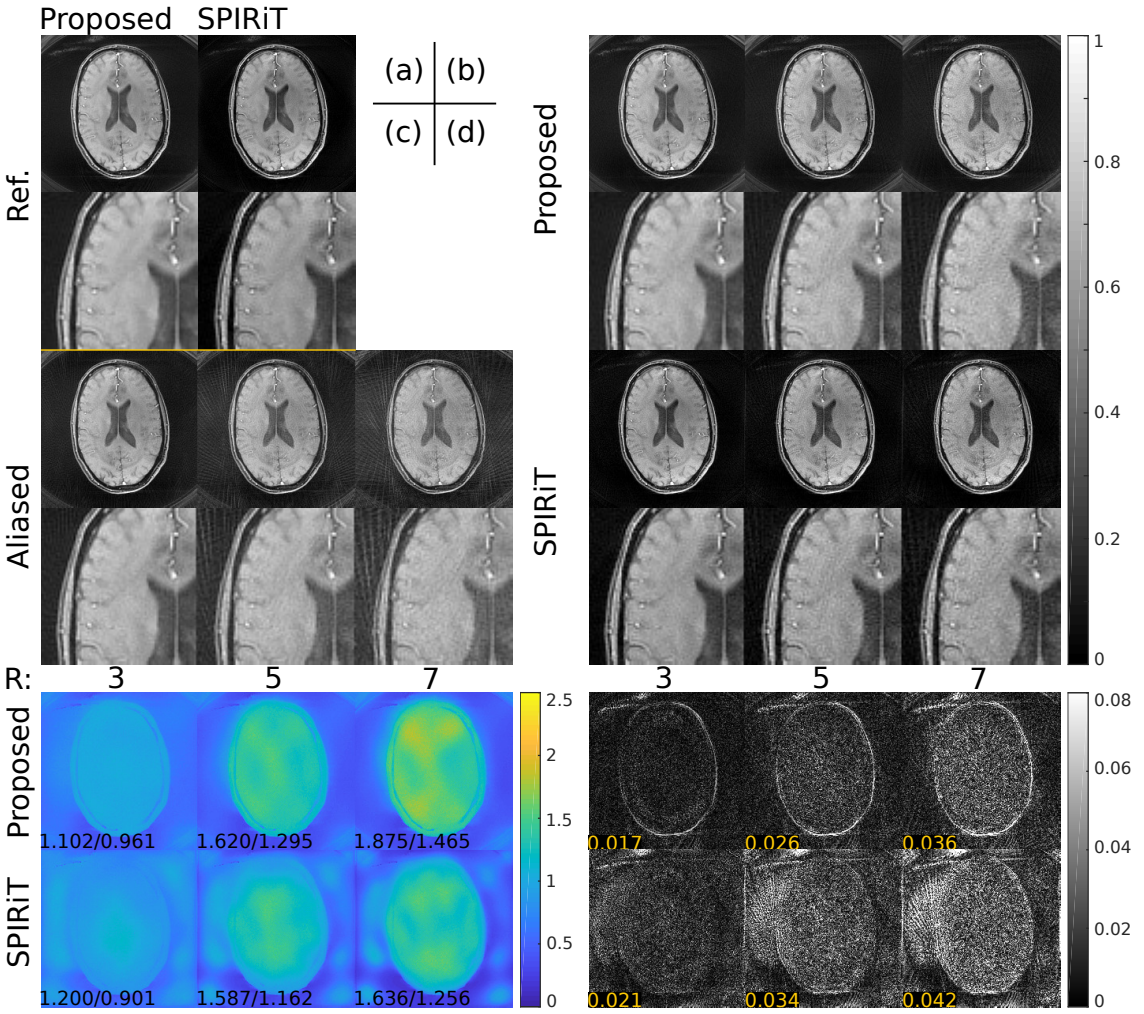

Figure 4.5: A similar comparison as in Fig. 4.3, but in 2D, between the proposed non-Cartesian GRAPPA and SPIRiT, with large retrospective undersampling factors (R=3,5,7). Without modifying our algorithm and implementation, our method achivews comparable reconstruction quality to SPIRiT. The "max/mean" g-factors, and RMSE are computed within the 2D head region.

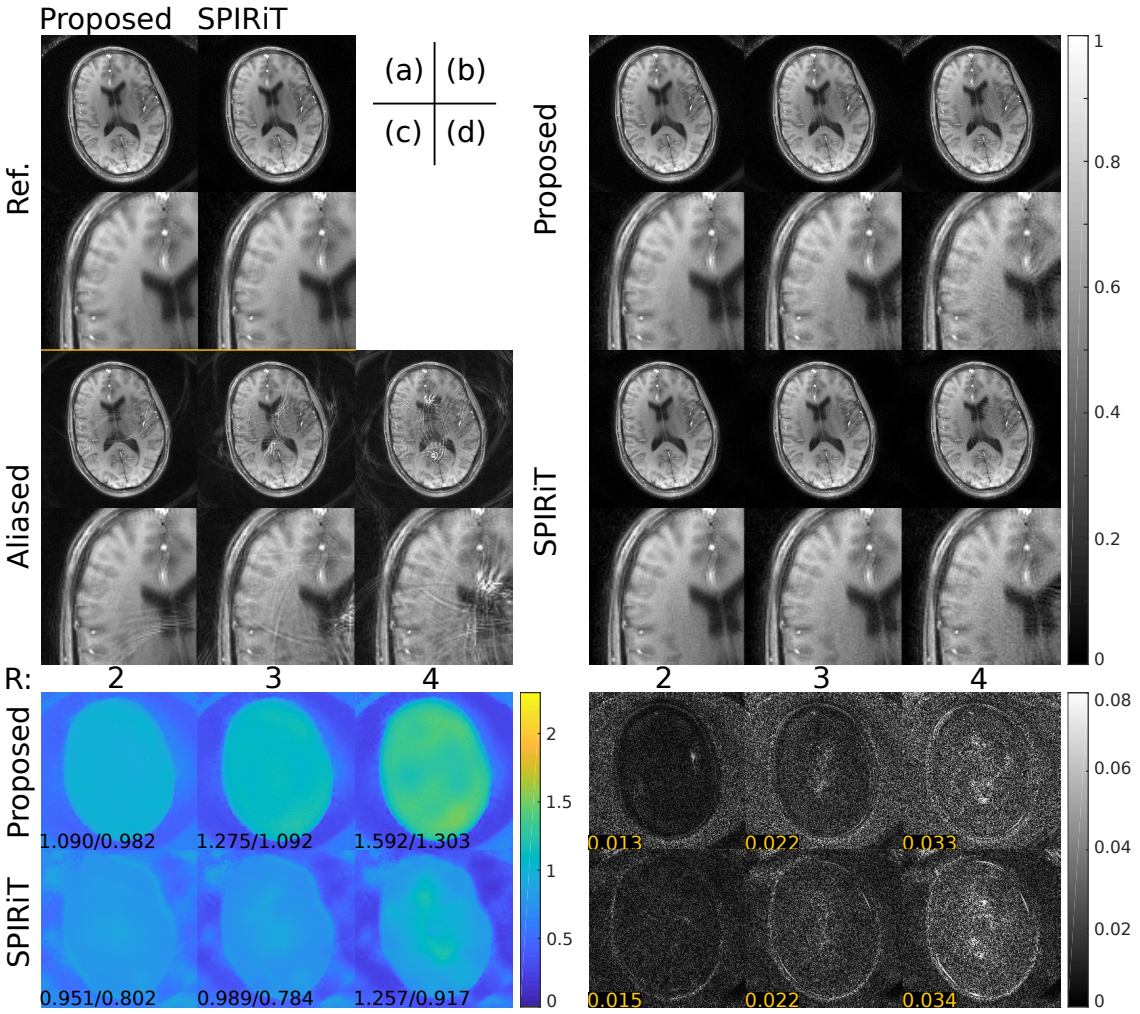

Figure 4.6: A similar comparison as in Fig. 4.5, but for retrospectively under-sampled spiral dataset, with retrospective under-sampling acceleration factor (R=2,3,4). Our proposed method still attains comparable quantities in terms of "max/mean" g-factor, and RMSE within 2D head image support. The reconstruction errors still contains high-frequency components as expected.

tion, Supporting Fig. S1 shows that in our non-Cartesian GRAPPA calibration, the influence of adopting a circulant boundary condition on reconstructed image quality is small.

In the pseudo-replica simulated g-factor maps, we observe that there are regions in the non-Cartesian GRAPPA case that fall below 1, as observed by others in Cartesian imaging [59]. We do notice that certain regions for the cg-SENSE case also fall below one. While Pruessmann *et al.* [57] have rigorously proven the g-factor to always be greater than 1 for unfolding (non-iterative) SENSE, the derivation does not apply to non-Cartesian iterative SENSE (e.g., cg-SENSE). Supporting Fig. S2 addresses this "g-factor smaller than one" behavior of cg-SENSE by illustrating a simulated 2D SENSE example where the g-factor falls below 1 analytically, obtained by calculating noise covariance directly. Basically, when under-sampling improves sampling density uniformity (e.g., over-sampled center k-space from fully-sampled trajectory), it is possible for g-factor to fall below 1.

Figure 4.7 compares the functional response obtained from the prospectively under-sampled image time-series reconstructed using our proposed algorithm, against that obtained from cg-SENSE. Three slices out of 20 are displayed. For cg-SENSE, different amounts of roughness penalization are evaluated. In this finger tapping experiment, as expected, we observe activations in the motor cortex, and, for cg-SENSE, the activation maps become increasingly blurred as roughness penalization increases. While the ground truth activation map is unknown, in terms of Dice similarity coefficient (dsc) (computed from all active voxels), our result most resembles the outcome of cg-SENSE at quadratic roughness penalty level $\lambda = 5$. However, our method is considerably faster: for this dataset it required 1 min for calibration and 2 min for time-series reconstruction, significantly shorter than the total cg-SENSE reconstruction time (15 min) (timed using MATLAB R2016a on a RHEL7.4 server with 2 Intel E5-2630V4 processors. The NUFFT operations of cg-SENSE are also

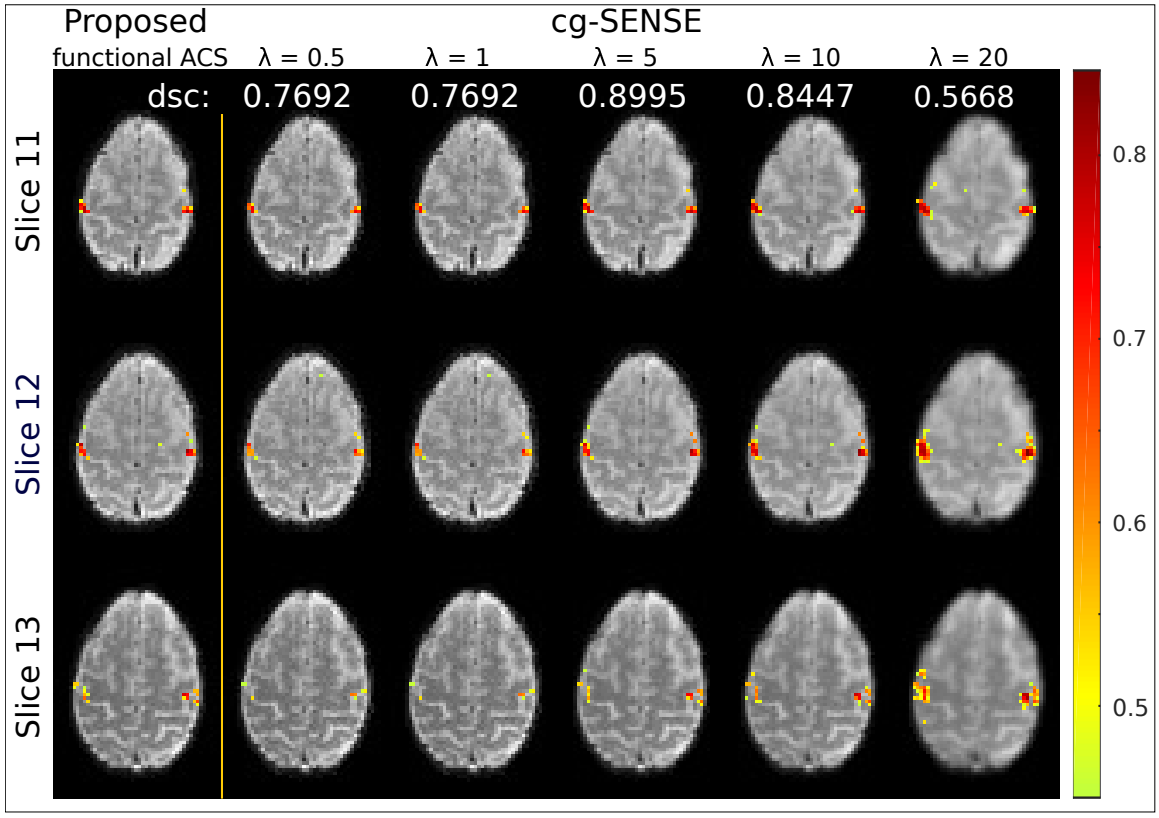

Figure 4.7: Functional imaging results from a *prospectively* under-sampled (R=3) rotated stack-of-spirals 3D acquisition, (dsc: Dice coefficient). The volunteer performed a block (stimulus on/off) finger-tapping task, that is known to reliably activate motor cortex. Shown are activation maps in three consecutive slices (out of 20) covering the active region of motor cortex. Images from left to right are reconstructed with: Proposed non-Cartesian GRAPPA with kernels calibrated using the functional imaging ACS dataset (a fully-sampled k-space center was obtained by combining the first 3 under-sampled frames of the fMRI time series); cg-SENSE with different levels of l2-roughness regularization parameter λ. While the ground truth activation map is unknown, in terms of Dice coefficient, our result most resembles the outcome of cg-SENSE at quadratic roughness penalty level $\lambda = 5$; and the activity maps from these two sets matches well with our expectation (motor cortex), while other activity maps are either incomplete or blurred. Reconstruction times were 1 min for calibration plus 2 min for reconstruction for the proposed method, and 15 min for cg-SENSE.

implemented in C subroutines.). Supporting Fig. S3 presents the activation map reconstructed using GRAPPA kernels calibrated with the separate distinct-contrast higher resolution dataset. As expected from [74], we obtain an essentially identical reconstruction compared to the reconstruction using the original GRAPPA kernels (obtained by combining three subsequent fMRI data frames to form a full-sampled k-space), with a dsc equal to 1. The image from the separate dataset is presented in Supporting Fig. S3.

4.4 Discussion

Our proposed non-Cartesian GRAPPA reconstruction is general and applicable to arbitrary readout trajectories, without modifying or adapting the code. Here we demonstrated its use with rotated stack-of-stars and rotated stack-of-spirals, using 3D kernels. As stated above, a convenient feature of these trajectories in the context of the proposed method is that the under-sampling pattern is regular along k_z, which reduces the number of different (unique) constellations. However, in the extreme case, the number of different kernels can equal the number of unsampled k-space points, which can require large computer memory. For instance, in 3D high-resolution imaging with 32 receive coils, there are about 400,000 3D kernels to be calibrated, each of which could contain about 50 neighbors on average, resulting in approximately 76 GB memory/storage demand for the coefficients (for single float precision). Coil compression [7, 83] can reduce this demand. Another mitigating approach may be to "interleave" calibration with reconstruction of each distinct constellation in a time-series; in other words, one can calibrate and cache the GRAPPA weights for a single constellation (or group of constellations) at a time, reconstruct all the data corresponding to this constellation throughout the whole time-series, then discard these weights and move on to the next constellation.

We have shown that using circulant boundary conditions (concerning the ACS

region) permits fast GRAPPA weight calibration, without degrading image quality. From our retrospectively under-sampled reconstruction experiments (in Supporting Information), we observe that for certain calibration set-ups (e.g., ACS region size, GRAPPA kernel size, and Tikhonov regularizer coefficient), allowing the ACS to wrap around can sometimes even improve reconstruction quality slightly (in terms of image error and g-factor). This behavior can be dependent on specific subjects and sampling trajectories. However, we certainly do not expect this behavior to be a general features of our method; whether the image error is slightly improved or worsened may depend on various factors such as the sampling trajectory used, or the subject or receive coil configuration. Overall, the circulant boundary condition appears to have a marginal influence on reconstructed image quality.

We demonstrated our method by reconstructing retrospectively unsampled k-space locations. While these locations are a natural choice, our method is not limited to reconstruct only the "unsampled" locations; for example, one possibility would be to reconstruct points on a Cartesian grid. The optimal choice of k-space locations to reconstruct with respect to, e.g., image quality and overall computation time, is an open problem.

k-Space reconstruction methods, such as GRAPPA and SPIRiT, are known to be robust against aliasing when FOV is limited (i.e., smaller than the object being imaged). Our proposed non-Cartesian GRAPPA is exactly equivalent to ordinary GRAPPA when the sampling is Cartesian; and hence would be expected to also be robust to limited FOV. In fMRI, limited-FOV is usually not an issue, as brains are typically small enough to be fully-covered in scanning. However, this issue could be significant in cardiac and other body imaging, where the region of interest is not that isolated from other parts of body. We will study the performance of our method in cardiac imaging in the future.

4.5 Conclusion

This paper introduced a flexible and rapid non-iterative reconstruction method that is a true non-Cartesian generalization of the canonical GRAPPA method. The method works with arbitrary sampling trajectories, and may be particularly beneficial in applications such as dynamic imaging (e.g., fMRI) where a large number of images must be reconstructed.

4.6 Supporting Information

4.6.1 Calibration Boundary Condition: Circulant vs Trimmed

Fig. 4.8 examines the influence of adopting circulant ACS boundary conditions on 2D reconstruction quality. In general, with proper Tikhonov regularization, the circulant boundary assumption produces similar reconstruction error as conventional ("trimmed") boundaries. The g-factor behavior is more complicated: depending on the choice of regularization parameter λ, circulant boundaries can do either slightly better or slightly worse than trimmed boundaries. In this comparison, reconstruction error with $\lambda = 5 \times 10^{-7}$ is low and difference is small between circulant and trimmed boundary conditions. This λ represents a compromise between g-factor and image/k-space error. We used this value throughout our experiments in the main text.

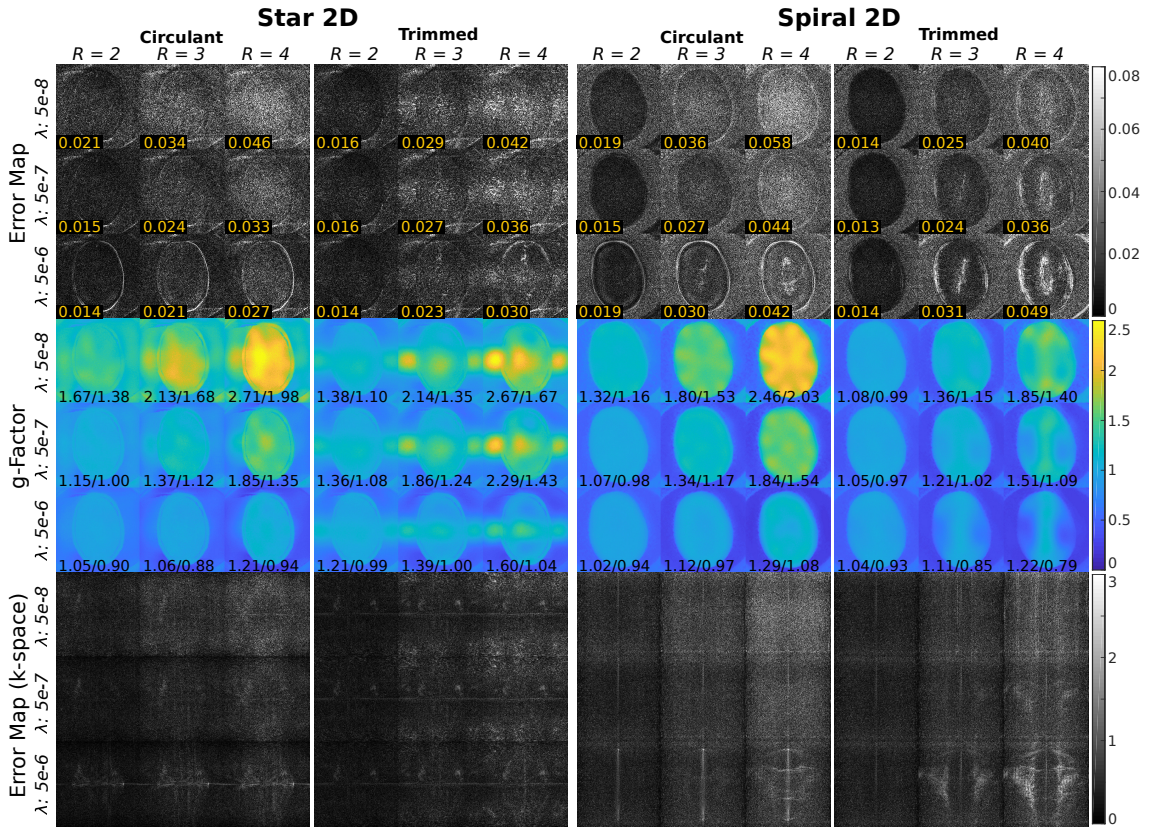

Figure 4.8: Reconstruction quality comparison of our non-Cartesian GRAPPA method using different ACS boundary conditions (circulant and trimmed), for 2D star (left panels) and spiral (right panels) datasets. The top row plots the absolute error maps and the digits are their averages within the object support. The center row plots the g-factor maps and the digits are their max/average g-factors within the support. The bottom row plots the absolute error viewed from k-space. We observe that circulant boundaries can produce similar reconstruction error as trimmed boundaries. Moreover, for certain Tikhonov regularization setups, circulant boundary outperforms trimmed boundaries.

4.6.2 g-Factor for SENSE with Over-Sampled Center of k-Space

Oversampling, which occurs often near the k-space center for non-Cartesian acquisitions, is a key reason why the g-factor can fall below 1 for SENSE. To see this, consider a simple 2-pixel example with two noise-uncorrelated coil sensitivities of [1, 0.5] and [0.3, 1], respectively. For the "fully sampled" acquisition, we sample three times at k=0, and once at k=1; for the "under-sampled" acquisition, we sample once

at both k=0 and k=1. The acceleration factor, R, in this example is thus 2. Following the definition of g-factor, a simple calculation would yield $g = \sqrt{3/4} < 1$.

For a more realistic example, in Fig. 4.9 we simulate a sampling pattern where the phase-encoding direction is over-sampled by a factor of 4.1 near the k-space center. The sensitivity maps are estimated from an *in vivo* data set; a low-resolution (64×64) matrix size is chosen to enable direct calculation. In this sampling pattern, "full"-sampling consists of both blue and red locations, while under-sampling only contains the blue locations. Again, we observe that g falls below 1 in some parts of the analytically calculated g-factor map. In summary, we conclude that oversampling, which is common for general non-Cartesian sampling, can cause the g-factor to fall below 1.

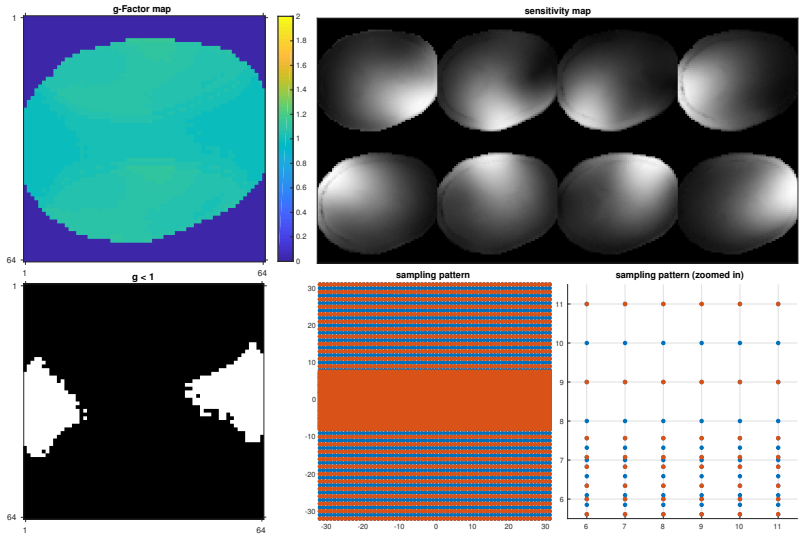

Figure 4.9: An illustration of a reconstruction with g-factor smaller than one (in some regions of the image), using realistic (*in vivo*) sensitivity maps. Here, "full" sampling consists of both blue and red locations, while "under-sampling" only contains the blue locations. The central k-space region is oversampled, as is typically the case in non-Cartesian acquisitions. The center k-space oversampling ratio 4.1 in this example produces off-grid sampling. In this example, the g-factor is just below 1.0 near the right and left parts of the image (white regions in the binary black/white image on the lower left).

4.6.3 GRAPPA calibration using ACS with different contrast

Talagala *et al.* observed that GRAPPA kernels calibrated with one acquisition can faithfully reconstruct other datasets of possibly different contrasts and resolution configurations (Ref. [13] of the maintext). This protocol can be useful for dynamic imaging, e.g., fMRI, where a structural dataset is commonly acquired alongside a number of functional activity datasets. To evaluate the use of an ACS dataset with contrast different from the to-be-reconstructed (undersampled) images, we tested reconstructing a finger-tapping dataset with kernels calibrated using a separate structural dataset. The fMRI results are shown in Fig. 7 of the main text. Here, we again show the results from Fig 7, but in addition we also show fMRI activation maps obtained with the proposed method based on ACS data from the structural dataset. The acquisition parameters are in the **Experiments** section of the main text.

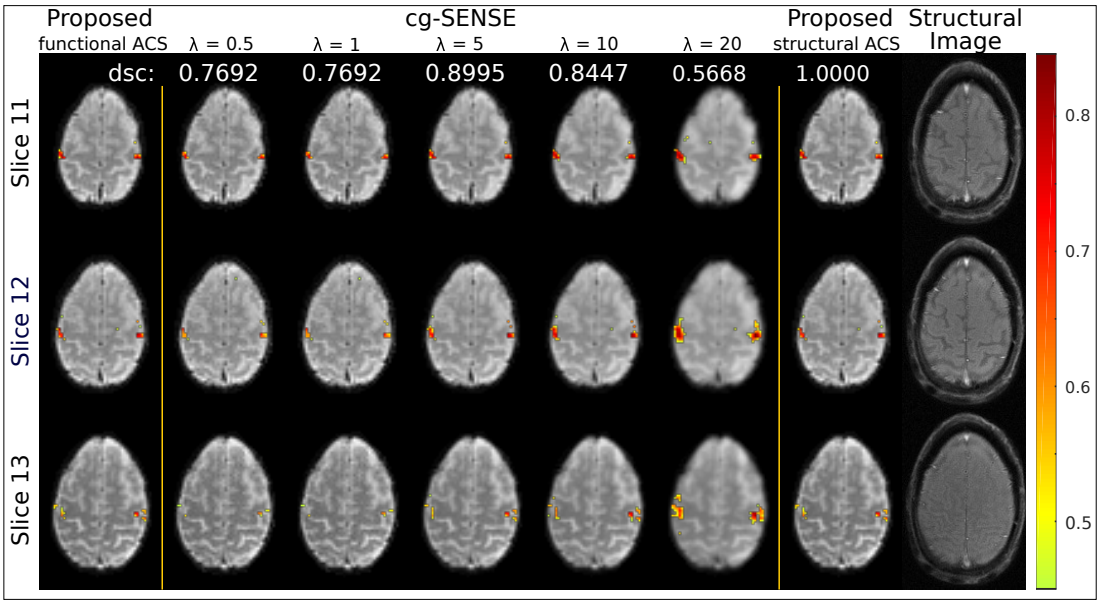

Figure 4.10: Reconstruction quality demonstration of GRAPPA kernel calibrated with ACS of a different contrast. Dice coefficients (dsc) are labeled for convenience of assessment. This figure is the same as Fig. 7, except with two extra columns (from left two right): Proposed non-Cartesian GRAPPA with kernels calibrated using the structural imaging ACS dataset; The high-resolution *structural* image of the same subject, acquired along with the fMRI scanning.

CHAPTER V

Future Works

Joint design of excitation RF and gradient waveforms remains a challenging problem. We can now obtain numerical derivatives for waveform optimization purposes via the proposed auto-differentiation method, without compromising accuracy by using simplified excitation modeling, such as the small-tip approximation. Nonetheless, due to the intrinsic non-convexity of pulse design problems in general, the pulse initialization step remains difficult and crucial, as it determines the eventual performance of the optimized pulses. We have shown that the extended kt-points initialization [72] can be a good choice for 3D spatially-tailored excitation objectives. For future works, it is of interest to validate different initializations for various excitation objectives, such as pre-phased simultaneous multi-slice excitations that is of interest for signal recovery in BOLD fMRI.

Another interesting direction from the pulse design perspective is to explore the design of error and penalty functions for various excitation objectives, and to incorporate scanning hardware modeling into the auto-differentiable Bloch simulation. We have demonstrated using longitudinal M_z error in the design loss, which is an intuitive choice for inversion pulse design. Future studies may quantify SAR and PNS to be included as penalties in the pulse design objective function to directly address such concerns for in-vivo MRI scanning. Moreover, to more accurately model excitation

procedures, we may incorporate parameters such as estimated eddy currents, gradient delays and non-linearity, etc, into the differentiable Bloch simulation.

Finally, as deep learning is gaining increasing popularity, we may explore the possibility of pulse design using deep neural networks. Specifically, we can build a neural network that takes as input a desired excitation profile and outputs a pulse. Then we Bloch simulate and quantify the excitation performance of the pulse, and obtain derivatives with respect to the pulse, which will be backpropagated to the pulse design neural network to train it.

Future directions for outer volume (OV) saturated inner volume (IV) imaging can be broad, e.g., in body imaging, where reducing artifacts from abdominal or chest motion can speed up acquisition. Specifically, experiments in Chapter II have demonstrated large-tip OV excitation pulse performances. The OV magnetization saturation strategy illustrated in Chapter III can fit in common steady-state imaging sequences. In principle, we can combine OV saturation with simultaneous multi-slice (SMS) imaging. This combination will likely have drawbacks such as slice profile artifacts or in-flow effects, but may enable very highly accelerated dynamic imaging. This may be of interest for high spatiotemporal resolution fMRI studies, helping to investigate the temporal activity order across brain regions.

For image reconstruction algorithm studies, where deep neural network based approaches are gaining increased attention, a direct extension of the non-Cartesian GRAPPA method proposed in Chapter IV may be challenging. Practical future directions for this method may lie in finding proper dynamic imaging applications.

APPENDIX

APPENDIX A

Optimal Diagonal Majorizer

Majorized minimization is an optimization method that is often used when the objective function is complicated, but can be majorized by easy-to-minimize surrogate functions. In other words, it systematically poses a series of sub-problems of minimizing simple functions (majorizers) which are everywhere no smaller than the objective function, such that, the sequence of solutions to these sub-problems converge to a minimizer of the original objective. Diagonal majorizers are separable quadratic functions, e.g., $\|D^{\frac{1}{2}}x\|_2^2$, where D is a diagonal matrix. They are of particular interest for, e.g., ℓ_1 constrained or regularized least square problems that arise in compressed sensing. The corresponding sub-problems possess closed form solutions.

The choice of a diagonal majorizer affects the convergence rate, i.e., the number of sub-problems to be solved to get close to a minimizer of the original objective. In this work, we show that the design of an optimal diagonal majorizer for least square terms, which maximizes the convergence rate, can be formed as a convex optimization problem. We establish the equivalence of this problem to other (unconstrained) problems, study optimality conditions of the problem, and provide the (sub)derivatives for minimizing it. When problem size is small, it can be solved using `cvx`[1].

[1] A software for disciplined convex programming: http://cvxr.com

A.0 Notations and Basics

This section defines notations and gives basic background knowledge for this chapter. We denote $\mathbb{S}^n$ as the space of Hermitian matrices of size $n \times n$.

Theorem A.0.1. *The non-zero eigenvalues are invariant under cyclic order permutations of a series of matrix products.*

Proof. Without loss of generality, consider two properly sized matrices, A, B, and their products AB, BA. Study the ith non-zero eigen-pair of AB, $\lambda_i x_i = ABx_i$. Left multiply both side by B yields, $\lambda_i Bx_i = BABx_i$. Let $\hat{x}_i \coloneqq Bx_i$, we have $\lambda_i \hat{x}_i = BA\hat{x}_i$, i.e., λ_i and $\hat{x}_i$ compose the ith non-zero eigen-pair of BA. □

Lemma A.0.2. *The eigenvalues of the product of two positive semidefinite matrices are all non-negative.*

Proof. Without loss of generality, consider two arbitrary positive semidefinite matrices, H, and D, Because H is Hermitian, it has a decomposition, $H = AA^{\mathsf{H}}$. Hence, the lemma holds true from theorem A.0.1, as, $\lambda_i(DH) = \lambda_i(A^{\mathsf{H}}DA) \in \mathbb{R}$. □

Theorem A.0.3 (Theorem 8.7 of Magunus's Matrix Differential Calculus [45])**.** *Let $X_0 \in \mathbb{S}^n$ have a simple eigen-pair $\lambda_0 u_0 = X_0 u_0$, where $\|u_0\|_2 = 1$, (i.e., λ_0 has a geometric multiplicity of 1). A real-valued function λ and a vector function u are defined for all X in some neighbourhood $N(X_0) \subset \mathbb{R}^{n\times n}$ of X_0, such that, $\lambda(X_0) = \lambda_0$, $u(X_0) = u_0$, and, $\lambda(X)u(X) = Xu(X)$, for $X \in N(X_0)$. Moreover, the differential at X_0 with respect to the elements of the matrix is $\mathrm{d}\lambda = u_0^{\mathsf{T}}(\mathrm{d}X)u_0$.*

Proof. $\lambda(X)$ and $u(X)$'s existence on $N(X_0)$ is a result of the implicit function theorem. For the differential, $Xu = \lambda u \implies (\mathrm{d}X)u + X\,\mathrm{d}u = \mathrm{d}\lambda u + \lambda\,\mathrm{d}u$. Left multiply both side by u^{T}, and evaluate at X_0: $u_0^{\mathsf{T}}(\mathrm{d}X)u_0 + u_0^{\mathsf{T}}X_0\,\mathrm{d}u = u_0^{\mathsf{T}}u_0\,\mathrm{d}\lambda + \lambda u_0^{\mathsf{T}}\,\mathrm{d}u \implies \mathrm{d}\lambda = u_0^{\mathsf{T}}(\mathrm{d}X)u_0$. □

Lemma A.0.4. *Let K be a positive integer, and $j \leq K$. The i-th eigenvalue of properly sized matrices product, $\lambda_i(\prod_{k=1}^{K} X_k)$, is 1-degree positive homogeneous in X_k. The derivative $\nabla_{X_j}\lambda_i(\prod_{k=1}^{K} X_k)$ is 0-degree positive homogeneous in X_j.*

Proof. Without loss of generality, $\lambda_i((\alpha A)B) = \alpha\lambda_i(AB)$, where $\alpha \in \mathbb{R}$; A, B are two properly sized matrices. $\nabla_{X_j}\lambda_i(\alpha\prod_{k=1}^{K} X_k) = \nabla_{X_j}\lambda_i(\prod_{k=1}^{K} X_k)$. □

Lemma A.0.5. *Consider a linear matrix function $\mathcal{H}(d) \coloneqq \sum_j M_j d_j$, where $M_i \in \mathbb{S}^n$, $d \in \mathbb{R}^n$. Let $\mathcal{H}(d)$ attain a simple ith largest eigen-pair at d_0, $\mathcal{H}(d_0)u_i = \lambda_i u_i$, (i.e., λ_i is the ith largest eigenvalue). Applying the previous Theorem A.0.3, the derivative of λ_i with respect to each element d_j is, $\frac{\mathsf{d}}{\mathsf{d}d_j}\lambda_i = u_i^{\mathsf{T}} M_j u_i$. The gradient with respect to d is assembled as, $[\nabla_d \lambda_i \mathcal{H}(d)]_j = u_i^{\mathsf{T}} M_j u_i$*

The above statements can be extended to complex domain in the usual way, i.e., replacing transpose with Hermitian transpose. The gradient extends to Clarke generalized gradient for non-simple eigen-pair. For non-simple largest, smallest eigenvalue, which is convex, concave of d, respectively, the Clarke generalized gradient specialize to sub-, sup- gradient, respectively.

A.1 Motivation

Consider the squared ℓ_2 cost, which appears as a term in many optimization problems (assuming full column-rank for tall system matrix A):

$$\arg\min_x f(x) = \frac{1}{2}\|Ax - y\|_2^2 \Longrightarrow x^\star \coloneqq (A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}}y\,. \tag{A.0}$$

In practice, the $(A^{\mathsf{H}}A)^{-1}$ part can be computationally intractable, and gradient based iterative algorithms with, $\lim_{k\to\infty} x^k \to x^\star$, are commonly used. Consider a gradient step from x^k of size τ_k,

$$x^{k+1} \leftarrow x^k - \tau_k \nabla f_x(x^k)\,. \tag{A.1}$$

It can be viewed as a majorized minimization step:

$$x^{k+1} = \arg\min_x g_k(x) \coloneqq f(x^k) + \left\langle x - x^k, \nabla f_x(x^k) \right\rangle + \underbrace{\frac{1}{2\tau_k}\left\| x - x^k \right\|_2^2}_{\phi_k(x)}, \qquad \text{[A.2]}$$

where "a typical condition ensuring convergence of x^k to a minimizer $x^\star$ of $f(x)$ is to require that $\tau_k \in \left(0, 1/\left\|A^{\mathsf{H}}A\right\|_2\right]$" [2].

This condition of τ_k makes $g_k(x)$ a majorizer of $f(x)$, i.e., $g_k(x) \geq f(x), \forall x$, with equality holds at x^k. In this perspective, one can design the 2nd order term, $\phi_k(x)$, to accelerate the convergence. For instance, let $\phi_k(x) = \frac{1}{2}\left\|x - x^k\right\|_M^2$, where $M \succeq H \coloneqq A^{\mathsf{H}}A \succeq 0$, [A.1] becomes, (by completing the square):

$$x^{k+1} \leftarrow x^k - M^{-1}\nabla f_x(x^k)\,. \qquad \text{[A.3]}$$

In fact, the $\phi_k(x)$ in [A.2] can be viewed as $\frac{1}{2}\left\|x - x^k\right\|_{\frac{1}{\tau_k}\cdot I}^2$, where $\frac{1}{\tau_k} \geq \lambda_{\max}(H)$, ($\lambda_{\max}(H)$ is the largest eigenvalue of H and the Lipschitz constant of f).

The ideal ϕ_k for the fastest convergence would be H, which gives the "exact" majorization, and a Newton step from x^k to x^{k+1}. For the ℓ_2 cost in [A.0], this will solve the problem exactly. However, the inversion, $H^{-1} = (A^{\mathsf{H}}A)^{-1}$, can cost $\mathcal{O}(n^3)$ for $A \in \mathbb{C}^{m\times n}$, and is unaffordable for large scale problems.

A practical choice of M is diagonal matrice (denoted as D hereafter), as its inversion costs $\mathcal{O}(n)$ in both time and storage complexities. In addition, it is separable in x, and updates can be parallelized. This is also known as the separable quadratic surrogates (SQS). Another favorable choice can be KDK^{H}, where efficient ways exist to compute Kx, e.g., FFT, FWT [47]. This note focus on designing D.

A.2 Spectral Norm Minimization

A.2.1 v-form, u-form, r-form, the derivatives

In [A.0], the gradient at x^k is: (without loss of generality, assume $y \equiv Ax^\star \in \mathcal{R}(A)$)

$$\nabla_x f(x^k) = A^{\mathsf{H}}(Ax^k - b) = H(x^k - x^\star).$$

Substitute $r^k \coloneqq x^k - x^\star$ and $\nabla_x f(x^k)$ into [A.3] (with $D \equiv D_d \coloneqq \mathsf{diag}(d), d \in \mathbb{R}^n_{++}$):

$$\begin{aligned} & x^{k+1} - x^\star = (x^k - x^\star) - D_d^{-1}\nabla f_x(x^k) \\ \iff & x^{k+1} - x^\star = (I - D_d^{-1}H)(x^k - x^\star) && \text{[A.4]} \\ \implies & r^{k+1} \qquad\quad = (I - D_d^{-1}H)r^k. && \text{[A.5]} \end{aligned}$$

Let $v^k \coloneqq D_d^{1/2} r^k, u^k \coloneqq H^{1/2} r^k$, we have the following transformation of variables:

$$D_d^{-1/2} v^k = r^k = H^{-1/2} u^k\,, \tag{A.6}$$

and the updates can be expressed, in two ways, as:

$$v^{k+1} = (I - \underbrace{D_d^{-1/2} H D_d^{-1/2}}_{H_{v(d)}})v^k, \qquad u^{k+1} = (I - \underbrace{H^{1/2} D_d^{-1} H^{1/2}}_{H_{u(d)}})u^k. \tag{A.7}$$

[A.5] and [A.7] are power iterations, whose convergence rates are determined by their spectral radii[2], $\rho(I - H_{v(d)})$ and $\rho(I - H_{u(d)})$, respectively. Smaller radius yields faster convergence. Exploiting, $D_d \succeq H$, which implies $I \succeq H_{v(d)}, H_{u(d)} \succ 0$, the spectral radii are:

$$\begin{aligned} \rho(I - H_{v(d)}) &= 1 - \lambda_{\min}(H_{v(d)}) \\ &= 1 - \frac{1}{\lambda_{\max}(H_{v(d)}^{-1})}, \end{aligned} \qquad \begin{aligned} \rho(I - H_{u(d)}) &= 1 - \lambda_{\min}(H_{u(d)}) \\ &= 1 - \frac{1}{\lambda_{\max}(H_{u(d)}^{-1})}. \end{aligned}$$

[2]The largest singular value, which can also be denoted as Schatten ∞-norm, $\|\cdot\|_\infty$.

Hence, for accelerating the convergence with a diagonal majorizer, we can form the following two equivalent convex minimization problems:

$$\begin{aligned} \arg\min_{d} \quad & \lambda_{\max}(H_{v(d)}^{-1}) \\ \text{s.t.} \quad & \lambda_{\min}(H_{v(d)}^{-1}) \geq 1, \end{aligned} \qquad \begin{aligned} \arg\min_{d} \quad & \lambda_{\max}(H_{u(d)}^{-1}) \\ \text{s.t.} \quad & \lambda_{\min}(H_{u(d)}^{-1}) \geq 1, \end{aligned} \qquad \text{[P0]}$$

where both $\lambda_{\min}(H_{v(d)}^{-1}) \geq 1$ and $\lambda_{\min}(H_{u(d)}^{-1}) \geq 1$ imply that $D_d \succeq H$. We denote $d^\star$ as a minimizer to problem [P0].

We call the two objectives in [P0] as v-form and u-form, respectively. Besides sharing eigenvalues, the eigenvectors of $H_{u(d)}^{-1}$ and $H_{v(d)}^{-1}$ are mutually transformable, as shown by the following lemma:

Lemma A.2.1. *Let $v_{[l,d]}, u_{[l,d]} \in \mathbb{C}^n$ be the lth eigenvector of $H_{v(d)}^{-1}, H_{u(d)}^{-1}$ respectively. The following identity holds, up to sign ambiguity:*

$$\sqrt{\lambda_{[l,d]}} D_d^{-1/2} v_{[l,d]} = r_{[l,d]} \coloneqq H^{-1/2} u_{[l,d]},$$

where $\lambda_{[l,d]}$ is the lth eigenvalue of both $H_{v(d)}^{-1}$ and $H_{u(d)}^{-1}$.

Define the element-wise magnitude square of a vector, $|\cdot|^2 : \mathbb{C}^n \mapsto \mathbb{R}^n$, as $|x|^2 = \mathsf{diag}(\mathsf{conj}(x))x$, we obtain the derivatives of eigenvalues with respect to d in the following Lemma.

Lemma A.2.2. $\nabla_d \lambda_{\max}(H_{u(d)}^{-1}) = \left|r_{[1,d]}\right|^2$, $\nabla_d \lambda_{\min}(H_{u(d)}^{-1}) = \left|r_{[n,d]}\right|^2$. *The derivatives are positive homogeneous in d of degree* 0.

Proof. These are directly from A.0.5, by letting, $M_i = (H^{-1/2}[i,:])^{\mathsf{H}}(H^{-1/2}[i,:])$, where $(H^{-1/2}[i,:])$ is the i-th row of $H^{-1/2}$. The 0-degree positive homogeneity is a direct result of Lemma A.0.4. □

A.2.2 The KKT Condition of [P0]

Denote convex set $\mathcal{D}_H \coloneqq \{d | D_d \succeq H\} = \left\{ d \middle| \lambda_{\min}(H_{u(d)}^{-1}) \geq 1 \right\}$. We give the KKT condition of the constraint convex problem [P0] in its u-form. The Lagrangian function of [P0] is:

$$L(d, \mu) = \lambda_{\max}(H_{u(d)}^{-1}) + \mu(1 - \lambda_{\min}(H_{u(d)}^{-1})), \quad \mu \geq 0. \tag{A.8}$$

Lemma A.2.3. *The KKT conditions of the convex problem* [P0] *are:*

$$\begin{aligned}
&D_{d^\star} \succeq H, \quad \mu^\star \geq 0, && \text{(primal, dual feasibility)},\\
&\mu^\star(1 - \lambda_{\min}(H_{u(d^\star)}^{-1})) = 0, && \text{(complementary slackness)},\\
&0 \in \partial_d \, \lambda_{\max}(H_{u(d)}^{-1}) + \partial_d(-\mu^\star \, \lambda_{\min}(H_{u(d)}^{-1})), && \text{(stationarity)}.
\end{aligned}$$

The primal, dual feasibility can be assumed; The complementary slackness is from observing that, for any $d \in \mathcal{D}_H$, $d/\, \lambda_{\min}(H_{u(d)}^{-1})$ is not suboptimal to d, which suggests that $\lambda_{\min}(H_{u(d^\star)}^{-1}) = 1$; ∂_d is the notation of subgradient set. The stationarity means there exist two subgradients respectively from sets, $\partial_d \, \lambda_{\max}(H_{u(d)}^{-1})$ and $\partial_d(-\mu^\star \, \lambda_{\min}(H_{u(d)}^{-1}))$, that add up to 0. When assuming simple geometric multiplicities of both $\lambda_{\max}$ and $\lambda_{\min}$, with Lemma A.2.2, the stationarity becomes,

$$\left| r_{[1,d^\star]} \right|^2 - \mu^\star \left| r_{[n,d^\star]} \right|^2 = 0. \tag{A.9}$$

Beside the KKT conditions , we also have the following geometrical Corollary.

Corollary A.2.3.1. *For any d such that, $\lambda_{\min}(H_{u(d)}^{-1}) = 1$, $\left| r_{[n,d]} \right|^2$ defines a supporting hyperplane at d for $\mathcal{D}_H$, and so is $\left| r_{[1,d^\star]} \right|^2$, at a minimizer $d^\star$.*

Proof. $\lambda_{\min}(H_{u(d)}^{-1}) = 1$ implies the following equality:

$$\lambda_{\min}(H_{u(d)}^{-1}) - 1 = 0 \iff v_{[n,d]}^{\mathsf{H}} \left(H_{v(d)}^{-1} - I \right) v_{[n,d]} = 0 \iff r_{[n,d]}^{\mathsf{H}} \left(D_d - H \right) r_{[n,d]} = 0;$$

and, further, the following hyperplane definition inequality:

$$\forall e \in \mathcal{D}_H, \qquad \left\langle e - d, \left| r_{[n,d]} \right|^2 \right\rangle = r_{[n,d]}^{\mathsf{H}} \left(D_e - H - (D_d - H) \right) r_{[n,d]}$$
$$= r_{[n,d]}^{\mathsf{H}} \left(D_e - H \right) r_{[n,d]} \geq 0.$$

For a minimizer $d^\star$, fixing $u_{[1,d^\star]}$, we have,

$$\forall d \in \mathcal{D}_H, \qquad \left\langle d - d^\star, \left| r_{[1,d^\star]} \right|^2 \right\rangle = u_{[1,d^\star]}^{\mathsf{H}} H^{-1/2} (D_d - D_{d^\star}) H^{-1/2} u_{[1,d^\star]}$$
$$= u_{[1,d^\star]}^{\mathsf{H}} H_{u(d)}^{-1} u_{[1,d^\star]} - u_{[1,d^\star]}^{\mathsf{H}} H_{u(d^\star)}^{-1} u_{[1,d^\star]} \geq 0,$$

which is also a hyperplane definition inequality. □

A.3 Equivalent Problems

We are now ready to install some equivalent problems of [P0].

A.3.1 Condition Number Minimization

Lemma A.2.3 has shown that $\lambda_{\min}(H_{u(d)}^{-1}) = 1$ is necessary for d to be optimal. Setting it as a constraint on d leads to a set of problems equivalent to P0:

$$\operatorname*{arg\,min}_{d:\lambda_{\min}(H_{v(d)}^{-1})=1} \lambda_{\max}(H_{v(d)}^{-1}), \qquad \operatorname*{arg\,min}_{d:\lambda_{\min}(H_{u(d)}^{-1})=1} \lambda_{\max}(H_{u(d)}^{-1}).$$

$\lambda_{\min}(H_{u(d)}^{-1}) = 1$ implies that, $\lambda_{\max}(H_{u(d)}^{-1}) = \mathsf{cond}(H_{u(d)}^{-1})$, and $\mathsf{cond}(H_{u(d)}^{-1})$ is positive homogeneous in d of degree 0, we can reform the problem as:

$$\operatorname*{arg\,min}_{d\in\mathbb{R}^n_{++},\, \|d\|_1=1} \mathsf{cond}(H_{v(d)}^{-1}), \qquad \operatorname*{arg\,min}_{d\in\mathbb{R}^n_{++},\, \|d\|_1=1} \mathsf{cond}(H_{u(d)}^{-1}). \qquad \text{[P1]}$$

The feasible set of P1 is friendly for projected gradient descent with line-search: the constraints $d \in \mathbb{R}^n_{++}$ and $\|d\|_1 = 1$ combined is equivalent $d \in \mathbb{R}^n_{++}$, $1^{\mathsf{T}} d = 1$. We can have gradient projection, $(I - \frac{1}{n} 1 1^{\mathsf{T}}) \nabla_d \mathsf{cond}(H_{u(d)}^{-1})$, then ensure $d \in \mathbb{R}^n_{++}$

via line-search during updates. This comes at a cost: P1's objective is non-convex (although quasiconfex), and methods like line search are needed to determine the step-size during an optimization. The gradient of the condition number, by Lemma A.2.2, can be expressed as:

$$\nabla_d \,\mathsf{cond}(H_{u(d)}^{-1}) = \nabla_d \frac{\lambda_{\max}(H_{u(d)}^{-1})}{\lambda_{\min}(H_{u(d)}^{-1})} = \frac{\lambda_{[1,d]}}{\lambda_{[n,d]}} \cdot D_d^{-1} \left(\left| v_{[1,d]} \right|^2 - \left| v_{[n,d]} \right|^2 \right), \qquad \text{[A.10]}$$

where $\lambda_{[1,d]}$, $\lambda_{[n,d]}$ are the largest, and smallest eigenvalue of $H_{u(d)}^{-1}$ respectively.

A.3.2 Spectral Radius Minimization

We formulate an unconstrained spectral radius minimization problem:

$$\arg\min_{d \in \mathbb{R}^n} \| I - H^{-1/2} D_d H^{-1/2} \|_\infty, \qquad \text{[P2]}$$

which is convex, as $H^{-1/2} D_d H^{-1/2}$ is linear in d.

Lemma A.3.1. *A sufficient necessary set of conditions for $d^\star$ to be a minimizer for the convex problem* [P2] *includes:*

$$\lambda_{\max}(H^{-1/2} D_{d^\star} H^{-1/2}) - 1 = 1 - \lambda_{\min}(H^{-1/2} D_{d^\star} H^{-1/2}), \qquad \text{[A.11]}$$

$$\exists \mu \geq 0, \;\; s.t., \quad 0 \in \partial_d \, \lambda_{\max}(H^{-1/2} D_d H^{-1/2}) + \mu \partial_d (-\lambda_{\min}(H^{-1/2} D_d H^{-1/2})), \qquad \text{[A.12]}$$

Proof. We reformulate [P2] as the following convex problem:

$$\begin{aligned} \arg\min_{c,d} \quad & c \\ \text{s.t.} \quad & c \geq 0 \\ & \lambda_{\max}(H^{-1/2} D_d H^{-1/2}) - 1 \leq c \\ & 1 - \lambda_{\min}(H^{-1/2} D_d H^{-1/2}) \leq c. \end{aligned}$$

The Lagrangian function of this problem is:

$$L(c, d, \mu_1, \mu_2, \mu_3) = c - \mu_1 c + \mu_2(\lambda_{\max}(H^{-1/2} D_d H^{-1/2}) - (1 + c)) \\ + \mu_3((1 - c) - \lambda_{\max}(H^{-1/2} D_d H^{-1/2})). \qquad [A.13]$$

The corresponding KKT conditions are:

$$\begin{aligned} c^\star \geq 0, \quad \mu_1^\star, \mu_2^\star, \mu_3^\star \geq 0 & \qquad \text{(primal, dual feasibility)} \\ \mu_1^\star c^\star &= \mu_2^\star(\lambda_{\max}(H^{-1/2} D_{d^\star} H^{-1/2}) - (1 + c^\star)) \\ &= \mu_3^\star((1 - c^\star) - \lambda_{\max}(H^{-1/2} D_{d^\star} H^{-1/2})) = 0 \qquad \text{(complementary slackness)} \\ 0 &= 1 - \mu_1^\star - \mu_2^\star - \mu_3^\star \\ 0 &\in \mu_2^\star \partial_d \lambda_{\max}(H^{-1/2} D_{d^\star} H^{-1/2}) \\ &\quad + \mu_3^\star \partial_d(-\lambda_{\min}(H^{-1/2} D_{d^\star} H^{-1/2})), \qquad \text{(stationarity)} \end{aligned}$$

and the stationarity condition implies [A.12]. Then, for [A.11], it is verifiable that:

$$0 \notin \partial_d \lambda_{\max}(H^{-1/2} D_{d^\star} H^{-1/2}), \quad 0 \notin \partial_d(-\lambda_{\min}(H^{-1/2} D_{d^\star} H^{-1/2})),$$

for any invertible H. So the stationarity also implies that $\mu_2^\star, \mu_3^\star > 0$, which when combined with the complementary slackness, indicates [A.11]. □

Theorem A.3.2. *If d minimizes* [P2]*, then $\frac{1}{\lambda_{\min}(H^{-1/2} D_d H^{-1/2})} \cdot d$ minimizes* [P0]*, and vice versa.*

Proof. Let $\hat{d} = \frac{1}{\lambda_{\min}(H^{-1/2} D_d H^{-1/2})} \cdot d$, we have $\lambda_{\min}(H^{-1/2} D_{\hat{d}} H^{-1/2}) = 1$, which meets the complementary slackness, and primal feasibility in Lemma A.2.3. For the stationarity, due to the 0-degree positive homogeneity of the derivatives of eigenvalues with respect to d, [A.12] is equivalent to the stationarity in A.2.3. □

A.3.2.1 Reverse problem of [P2]

We define vector inverse $(\cdot)^{-1}$:$\mathbb{R}^n \mapsto \mathbb{R}^n$, such that, for $i = 1 \dots n$, the i-th element $(d^{-1})_{[i]} = (d_{[i]})^{-1}$, and, if $d_{[i]} \neq 0$; $(d^{-1})_{[i]} = 0$, if $d_{[i]} = 0$. For convex problem:

$$\arg\min_{b \in \mathbb{R}^n} \|I - H^{1/2} D_b H^{1/2}\|_\infty, \qquad \text{[P3]}$$

we have the following theorem.

Theorem A.3.3. *For invertible H, if b minimizes* [P3]*, with $\lambda_1 \coloneqq \lambda_{\max}(H^{1/2} D_b H^{1/2})$, and $\lambda_n \coloneqq \lambda_{\min}(H^{1/2} D_b H^{1/2})$; then, $d \coloneqq \frac{2\lambda_1\lambda_n}{\lambda_1+\lambda_n} b^{-1}$ minimizes* [P2]*, and vice versa.*

Proof. This theorem is straightforward:

$$\lambda_{\max}(H^{-1/2} D_d H^{-1/2}) = \frac{1}{\lambda_n} \cdot \frac{2\lambda_1\lambda_n}{\lambda_1 + \lambda_n}, \quad \lambda_{\min}(H^{-1/2} D_d H^{-1/2}) = \frac{1}{\lambda_1} \cdot \frac{2\lambda_1\lambda_n}{\lambda_1 + \lambda_n},$$

so, d satisfies the [A.11]. Moreover, $H^{-1/2} D_d H^{-1/2}$ share the same eigenvectors of $H^{1/2} D_b H^{1/2}$, but in the opposite order, since the latter is a scaled inversion of the former, and the inversion flips the order. Therefore, by Lemma A.0.5, we have:

$$\partial_b \lambda_{\max}(H^{1/2} D_b H^{1/2}) = \mu_1 \partial_d \lambda_{\min}(H^{-1/2} D_d H^{-1/2}),$$
$$\partial_b \lambda_{\min}(H^{1/2} D_b H^{1/2}) = \mu_2 \partial_d \lambda_{\max}(H^{-1/2} D_d H^{-1/2}),$$

for some μ_1, $\mu_2 \geq 0$; and this implies d satisfies the KKT stationarity for [P2]. □

A.4 Discussion

The simple geometric multiplicity assumption is in general unrealisitic to be based on for developing efficient algorithms to solve the optimization problem. The non-uniqueness of subgradients complicates vanilla line-search based optimizations.

An illustrative convex problem of the complication is:

$$\underset{x\in\mathbb{R}^2_+}{\arg\min} \quad \max(x_{[1]}, x_{[2]}),$$

where $x_{[1]}$, and $x_{[2]}$ are the first and second element of x, respectively. Consider an initilization, x^0, such that $x^0_{[1]} > x^0_{[2]} > 0$. Without analytical prior knowledge of, max, line-search fails like coordinate descent: stucking at an x^1 that is near the crease of the cost landscape where $x^1_{[1]}$ is arbitrarily close, but greater than $x^1_{[2]}$. This example resembles the majorization problem:

$$\underset{d\in\mathcal{D}_H}{\arg\min} \quad \max(\lambda_1(H^{-1}_{v(d)}), \ldots, \lambda_n(H^{-1}_{v(d)})). \tag{A.14}$$

As a consequence, practical algorithm for the majorization problem needs to be more sophisticated than a vanilla line-search. Our adventure cuts out here. Below are some comments on future works on this problem.

Recognizing that the difficulty is the non-smoothness, an intuitive direction is to smooth the $\lambda_{\max}$ objective[3]. In particular, D'Aspremont [12] investigated replacing the maximum operation in determining the largest eigenvalue with softmax, which, when applied to [A.14], gives:

$$\underset{d\in\mathcal{D}_H}{\arg\min} \quad \mu \log\left(\sum_{i=1}^{n} e^{\lambda_i(H^{-1}_{v(d)})/\mu}\right), \tag{A.15}$$

where $\mu > 0$ tunes the smoothness: as $\mu \to 0$, the softmax approaches $\lambda_{\max}$. Nevertheless, the new loss requires knowing all λ_i, and is not directly applicable for large scale problems. To cap at a $k \ll n$ in softmax, we need to study its impact on convergence and optimality (e.g., boundness of the solution for [A.15] to the solution

[3]One might attempt instead smoothing the derivatives using momentum terms like as Nesterov's accelerated gradient descent does, which utilizes the Lipschitz constants. However, second order knowledge of A.14 requires pseudo-inverting H according to Chapter 8.11 of [45].

for [A.14] as a function of μ). In [13], D'Aspremont further investigated stochastic smoothing, which avoids full eigen-decompositions, and may be useful here.

Another general approach for handling non-smoothness is the bundle method [62], which systematically update a local (linear, easy to solve) approximation of the original problem. Helmberg et al. extended this approach to semidefinite programing in their spectral bundle method works [29–31]. Later, Kangal et al. unified the approach with subspace methods, and provided convergence rate analysis in [32]. It is of interest to also try applying them on the majorization problem.

Finally, to solve the diagonal majorizer of certain applications, e.g., MRI reconstruction, we can exploit their system matrix structures to efficiently compute the eigen-pairs, and the derivatives of d. For instance, let P be a permutation matrix that reverts the order of elements in a vector, (e.g., $P[1,2,3]^{\mathsf{T}} = [3,2,1]^{\mathsf{T}}$). When the matrix to be majorized, H, is Hermitian and Toeplitz, let q be one of its eigenvector that corresponds to the eigenvalue λ, then $\mathsf{conj}(Pq)$, is also an eigenvector corresponding to λ; and hence we know the optimal $d^\star$ for this H must be symmetric, i.e., $d^\star = Pd^\star$.

BIBLIOGRAPHY

[1] Michael J. Allison, Sathish Ramani, and Jeffrey A. Fessler. Accelerated Regularized Estimation of MR Coil Sensitivities Using Augmented Lagrangian Methods. *IEEE Transactions on Medical Imaging*, 32(3):556–564, Mar 2013.

[2] Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, Jan 2009.

[3] Jon Louis Bentley. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM*, 18(9):509–517, Sep 1975.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Zhipeng Cao, Manus J. Donahue, Jun Ma, and William A. Grissom. Joint design of large-tip-angle parallel RF pulses and blipped gradient trajectories. *Magnetic Resonance in Medicine*, 75(3):1198–1208, 2016.

[6] Alex Cerjanic, Joseph L. Holtrop, Giang Chau Ngo, Brent Leback, Galen Arnold, Mark Van Moer, Genevieve LaBelle, Jeffrey A. Fessler, and Bradley P. Sutton. PowerGrid: A open source library for accelerated iterative magnetic resonance image reconstruction. In *ISMRM*, page 0525, 2016.

[7] Alan Chu and Douglas C. Noll. Coil compression in simultaneous multislice functional MRI with concentric ring slice-GRAPPA and SENSE. *Magnetic Resonance in Medicine*, 76(4):1196–1209, Oct 2016.

[8] Frank H Clarke. Nonsmooth analysis and optimization. In *Proceedings of the international congress of mathematicians*, volume 5, pages 847–853, 1983.

[9] M. A. Cloos, N. Boulant, M. Luong, G. Ferrand, E. Giacomini, Denis Le Bihan, and A. Amadon. kT-points: Short three-dimensional tailored RF pulses for flip-angle homogenization over an extended volume. *Magnetic Resonance in Medicine*, 67(1):72–80, Jan 2012.

[10] Steven M. Conolly, Dwight G. Nishimura, and Albert Macovski. Optimal Control Solutions to the Magnetic Resonance Selective Excitation Problem. *IEEE Transactions on Medical Imaging*, 5(2):106–115, Jun 1986.

[11] Steven M. Conolly, Dwight G. Nishimura, Albert Macovski, and Gary Glover. Variable-rate selective excitation. *Journal of Magnetic Resonance*, 78(3):440 – 458, 1988.

[12] Alexandre D'Aspremont. Smooth Optimization with Approximate Gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, Jan 2008.

[13] Alexandre D'Aspremont and Noureddine El Karoui. A Stochastic Smoothing Algorithm for Semidefinite Programming. *SIAM Journal on Optimization*, 24(3):1138–1177, Jan 2014.

[14] Mathias Davids, Lothar R. Schad, Lawrence L. Wald, and Bastien Guérin. Fast three-dimensional inner volume excitations using parallel transmission and optimized k-space trajectories. *Magnetic Resonance in Medicine*, 76(4):1170–1182, Oct 2016.

[15] Weiran Deng, Benjamin Zahneisen, and V. Andrew Stenger. Rotated stack-of-spirals partial acquisition for rapid volumetric parallel MRI. *Magnetic Resonance in Medicine*, 76(1):127–135, Jul 2016.

[16] David A Feinberg and Essa Yacoub. The rapid development of high speed, resolution and precision in fMRI. *Neuroimage*, 62(2):720–725, Aug 2012.

[17] Jeffrey A. Fessler and Bradley P. Sutton. Nonuniform fast fourier transforms using min-max interpolation. *IEEE Transactions on Signal Processing*, 51(2):560–574, Feb 2003.

[18] Andreas Griewank and Andrea Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics, Jan 2008.

[19] William A. Grissom, Mohammad-Mehdi Khalighi, Laura I. Sacolick, Brian K. Rutt, and Mika W. Vogel. Small-tip-angle spokes pulse design using interleaved greedy and local optimization methods. *Magnetic Resonance in Medicine*, 68(5):1553–1562, Nov 2012.

[20] William A. Grissom, Kawin Setsompop, Samuel A. Hurley, Jeffrey Tsao, Julia V. Velikina, and Alexey A. Samsonov. Advancing RF pulse design using an open-competition format: Report from the 2015 ISMRM challenge. *Magnetic Resonance in Medicine*, 78(4):1352–1361, 2017.

[21] William A. Grissom, Dan Xu, A.B. Kerr, Jeffrey A. Fessler, and Douglas C. Noll. Fast Large-Tip-Angle Multidimensional and Parallel RF Pulse Design in MRI. *IEEE Transactions on Medical Imaging*, 28(10):1548–1559, Oct 2009.

[22] Mark A. Griswold, Robin M. Heidemann, and Peter M. Jakob. Direct Parallel Imaging Reconstruction of Radially Sampled Data Using GRAPPA with Relative Shifts. In *ISMRM*, volume 11, page 2349, 2003.

[23] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic resonance in medicine*, 47(6):1202–1210, Jun 2002.

[24] Paul T. Gurney, Brian Andrew Hargreaves, and Dwight G. Nishimura. Design and analysis of a practical 3D cones trajectory. *Magnetic Resonance in Medicine*, 55(3):575–582, Mar 2006.

[25] Justin P. Haldar. Low-Rank Modeling of Local k-Space Neighborhoods (LORAKS) for Constrained MRI. *IEEE Transactions on Medical Imaging*, 33(3):668–681, Mar 2014.

[26] C. J. Hardy, P. A. Bottomley, M. O'Donnell, and P. B. Roemer. Optimization of two-dimensional spatially selective NMR pulses by simulated annealing. *Journal of Magnetic Resonance*, 77(2):233–250, 1988.

[27] Brian A. Hargreaves, Charles H. Cunningham, Dwight G. Nishimura, and Steven M. Conolly. Variable-rate selective excitation for rapid MRI sequences. *Magnetic Resonance in Medicine*, 52(3):590–597, 2004.

[28] Robin M. Heidemann, Mark A. Griswold, Nicole Seiberlich, Gunnar Krüger, Stephan A.R. Kannengiesser, Berthold Kiefer, Graham Wiggins, Lawrence L. Wald, and Peter M. Jakob. Direct parallel image reconstructions for spiral trajectories using GRAPPA. *Magnetic Resonance in Medicine*, 56(2):317–326, Aug 2006.

[29] C. Helmberg and K.C. Kiwiel. A spectral bundle method with bounds. *Mathematical Programming*, 93(2):173–194, Dec 2002.

[30] C. Helmberg, M.L. Overton, and F. Rendl. The spectral bundle method with second-order information. *Optimization Methods and Software*, 29(4):855–876, Jul 2014.

[31] C Helmberg and F Rendl. A Spectral Bundle Method for Semidefinite Programming. *SIAM Journal on Optimization*, 10(3):673–696, Jan 2000.

[32] Fatih Kangal, Karl Meerbergen, Emre Mengi, and Wim Michiels. A Subspace Method for Large-Scale Eigenvalue Optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(1):48–82, jan 2018.

[33] K. J. Layton, S. Kroboth, F. Jia, S. Littin, H. Yu, J. Leupold, J-F. Nielsen, T. Stöcker, and M. Zaitsev. Pulseq: A rapid and hardware-independent pulse sequence prototyping framework. *Mag. Res. Med.*, 77(4):1544–52, April 2017.

[34] Chunlei Liu, Roland Bammer, and Michael E. Moseley. Parallel imaging reconstruction for arbitrary trajectories using k-space sparse matrices (kSPA). *Magnetic Resonance in Medicine*, 58(6):1171–1181, Dec 2007.

[35] Tianrui Luo, Jon-fredrik Nielsen, and Douglas C. Noll. Improved aliasing suppression in steady-state, parallel imaging using inner volume excitation. In *ISMRM*, page 3221, 2016.

[36] Tianrui Luo, Douglas C. Noll, Jeffrey A. Fessler, and Jon-fredrik Nielsen. A Fast and General Non-Cartesian GRAPPA Reconstruction Method. In *ISMRM*, page 2821, 2018.

[37] Tianrui Luo, Douglas C Noll, Jeffrey A Fessler, and Jon-Fredrik Nielsen. A GRAPPA algorithm for arbitrary 2D/3D non-Cartesian sampling trajectories with rapid calibration. *Magnetic Resonance in Medicine*, (February):mrm.27801, May 2019.

[38] Tianrui Luo, Douglas C. Noll, Jeffrey A. Fessler, and Jon-Fredrik Nielsen. Joint design of rf and gradient waveforms via auto-differentiation for 3d tailored excitation in mri, 2020. arXiv 2008.10594.

[39] Tianrui Luo, Douglas C. Noll, and Jon-fredrik Nielsen. 3D Inner Volume Imaging with 3D Tailored Outer Volume Suppression RF Pulses. In *ISMRM*, page 4631, 2019.

[40] Tianrui Luo, Douglas C. Noll, and Jon-fredrik Nielsen. 3D Inner Volume bSSFP functional MRI: preliminary results. In *ISMRM*, page 2889, 2020.

[41] Michael Lustig, David L. Donoho, J.M. Santos, and John M. Pauly. Compressed Sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72–82, Mar 2008.

[42] Michael Lustig, Seung-Jean Kim, and John M. Pauly. A fast method for designing time-optimal gradient waveforms for arbitrary k-space trajectories. *IEEE Transactions on Medical Imaging*, 27(6):866–873, Jun 2008.

[43] Michael Lustig and John M. Pauly. SPIRiT: Iterative self-consistent parallel imaging reconstruction from arbitrary k-space. *Magnetic Resonance in Medicine*, 64(2):n/a–n/a, 2010.

[44] Chao Ma, Dan Xu, Kevin F. King, and Zhi-Pei Liang. Joint design of spoke trajectories and RF pulses for parallel excitation. *Magnetic Resonance in Medicine*, 65(4):973–985, Apr 2011.

[45] Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 1988.

[46] Shaihan J. Malik, Shiva Keihaninejad, Alexander Hammers, and Joseph V. Hajnal. Tailored excitation in 3D with spiral nonselective (SPINS) RF pulses. *Magnetic Resonance in Medicine*, 67(5):1303–1315, May 2012.

[47] Madison G. McGaffin and Jeffrey A. Fessler. Algorithmic design of majorizers for large-scale inverse problems, 2015. arXiv 1508.02958.

[48] Karla L. Miller, Brian A. Hargreaves, Jongho Lee, David Ress, R. Christopher deCharms, and John M. Pauly. Functional brain imaging using a blood oxygenation sensitive steady state. *Magnetic Resonance in Medicine*, 50(4):675–683, 2003.

[49] Dimitris Mitsouras, Robert V. Mulkern, and Frank J. Rybicki. Strategies for inner volume 3D fast spin echo magnetic resonance imaging using nonselective refocusing radio frequency pulsesa). *Medical Physics*, 33(1):173–186, 2005.

[50] Jon-Fredrik Nielsen and Douglas C. Noll. TOPPE: A framework for rapid prototyping of MR pulse sequences. *Magnetic Resonance in Medicine*, 79(6):3128–3134, 2018.

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.

[52] John Pauly, Dwight Nishimura, and Albert Macovski. A k-space analysis of small-tip-angle excitation. *Journal of Magnetic Resonance*, 81(1):43–56, Jan 1989.

[53] S J Peltier and D C Noll. Physiological noise in multi-shot functional MRI. In *ISMRM*, page 118, 2002.

[54] Scott James Peltier. *Characterization and compensation of systematic noise in functional magnetic resonance imaging.* PhD thesis, University of Michigan, Ann Arbor, 2003.

[55] Jonathan R. Polimeni, Himanshu Bhat, Thomas Witzel, Thomas Benner, Thorsten Feiweier, Souheil J. Inati, Ville Renvall, Keith Heberlein, and Lawrence L. Wald. Reducing sensitivity losses due to respiration and motion in accelerated echo planar imaging by reordering the autocalibration data acquisition. *Magnetic Resonance in Medicine*, 75(2):665–679, Feb 2016.

[56] Klaas P. Pruessmann, Markus Weiger, Peter Börnert, and Peter Boesiger. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magnetic Resonance in Medicine*, 46(4):638–651, Oct 2001.

[57] Klaas P. Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5):952–962, Nov 1999.

[58] V. Rasche, R. Proksa, R. Sinkus, P. Bornert, and H. Eggers. Resampling of data between arbitrary grids using convolution interpolation. *IEEE Transactions on Medical Imaging*, 18(5):385–392, May 1999.

[59] Philip M Robson, Aaron K Grant, Ananth J Madhuranthakam, Riccardo Lattanzi, Daniel K Sodickson, and Charles A McKenzie. Comprehensive quantification of signal-to-noise ratio and g-factor for image-based and k-space-based

parallel imaging reconstructions. *Magnetic resonance in medicine*, 60(4):895–907, Oct 2008.

[60] Klaus Scheffler, Erich Seifritz, Deniz Bilecen, Ramesh Venkatesan, Jürgen Hennig, Michael Deimling, and E. Mark Haacke. Detection of BOLD changes by means of a frequency-sensitive trueFISP technique: preliminary results. *NMR in Biomedicine*, 14(7-8):490–496, 2001.

[61] Johannes T. Schneider, Raffi Kalayciyan, Martin Haas, Sarah R. Herrmann, Wolfgang Ruhm, Jürgen Hennig, and Peter Ullmann. Inner-volume imaging in vivo using three-dimensional parallel spatially selective excitation. *Magnetic Resonance in Medicine*, 69(5):1367–1378, May 2013.

[62] Helga Schramm and Jochem Zowe. A Version of the Bundle Idea for Minimizing a Nonsmooth Function: Conceptual Idea, Convergence Analysis, Numerical Results. *SIAM Journal on Optimization*, 2(1):121–152, Feb 1992.

[63] Nicole Seiberlich, Felix A. Breuer, Martin Blaimer, Kestutis Barkauskas, Peter M. Jakob, and Mark A. Griswold. Non-Cartesian data reconstruction using GRAPPA operator gridding (GROG). *Magnetic Resonance in Medicine*, 58(6):1257–1265, Dec 2007.

[64] Nicole Seiberlich, Felix A. Breuer, Philipp Ehses, Hisamoto Moriguchi, Martin Blaimer, Peter M. Jakob, and Mark A. Griswold. Using the GRAPPA operator and the generalized sampling theorem to reconstruct undersampled non-Cartesian data. *Magnetic Resonance in Medicine*, 61(3):705–715, Mar 2009.

[65] Nicole Seiberlich, Gregory Lee, Philipp Ehses, Jeffrey L. Duerk, Robert Gilkeson, and Mark Griswold. Improved temporal resolution in cardiac imaging using through-time spiral GRAPPA. *Magnetic Resonance in Medicine*, 66(6):1682–1688, Dec 2011.

[66] Kawin Setsompop, Borjan A. Gagoski, Jonathan R. Polimeni, Thomas Witzel, Van J. Wedeen, and Lawrence L. Wald. Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, 67(5):1210–1224, May 2012.

[67] Kawin Setsompop, Lawrence L. Wald, Vijayanand Alagappan, Borjan A. Gagoski, and Elfar Adalsteinsson. Magnitude least squares optimization for parallel radio frequency excitation design demonstrated at 7 Tesla with eight channels. *Magnetic Resonance in Medicine*, 59(4):908–915, Apr 2008.

[68] F. S. Sisser. Elimination of bounds in optimization problems by transforming variables. *Mathematical Programming*, 20(1):110–21, Dec 1981.

[69] James A Sorenson and X Wang. ROC methods for evaluation of fMRI techniques. *Magnetic Resonance in Medicine*, 36(5):737–744, 1996.

[70] Rüdiger Stirnberg, Willem Huijbers, Daniel Brenner, Benedikt A. Poser, Monique Breteler, and Tony Stöcker. Rapid whole-brain resting-state fMRI at 3 T: Efficiency-optimized three-dimensional EPI versus repetition time-matched simultaneous-multi-slice EPI. *NeuroImage*, 163:81–92, Dec 2017.

[71] Hao Sun. *Topics in steady-state MRI sequences and RF pulse optimization*. PhD thesis, University of Michigan, Ann Arbor, 2015.

[72] Hao Sun, Jeffrey A. Fessler, Douglas C. Noll, and Jon-Fredrik Nielsen. Joint Design of Excitation k-Space Trajectory and RF Pulse for Small-Tip 3D Tailored Excitation in MRI. *IEEE Transactions on Medical Imaging*, 35(2):468–479, Feb 2016.

[73] Hao Sun, Jeffrey A. Fessler, Douglas C. Noll, and Jon-Fredrik Nielsen. Rapid inner-volume imaging in the steady-state with 3D selective excitation and small-tip fast recovery imaging. *Magnetic Resonance in Medicine*, 76(4):1217–1223, Oct 2016.

[74] S. Lalith Talagala, Joelle E. Sarlls, Siyuan Liu, and Souheil J. Inati. Improvement of temporal signal-to-noise ratio of GRAPPA accelerated echo planar imaging using a FLASH based calibration scan. *Magnetic Resonance in Medicine*, 75(6):2362–2371, Jun 2016.

[75] Ye Tian, Kay Condie Erb, Ganesh Adluru, Devavrat Likhite, Apoorva Pedgaonkar, Michael Blatt, Srikant Kamesh Iyer, John Roberts, and Edward DiBella. Technical Note: Evaluation of pre-reconstruction interpolation methods for iterative reconstruction of radial k-space data. *Medical Physics*, 44(8):4025–4034, Aug 2017.

[76] Shao Tingting, Xia Ling, Tao Guisheng, Chi Jieru, Liu Feng, and Stuart Crozier. Advanced Three-Dimensional Tailored RF Pulse Design in Volume Selective Parallel Excitation. *IEEE Transactions on Medical Imaging*, 31(5):997–1007, May 2012.

[77] B.J. Wilm, J. Svensson, A. Henning, K.P. Pruessmann, P. Boesiger, and S.S. Kollias. Reduced field-of-view MRI using outer volume suppression for spinal cord diffusion imaging. *Magnetic Resonance in Medicine*, 57(3):625–630, 2007.

[78] Ernest N. Yeh, Charles A. McKenzie, Michael A. Ohliger, and Daniel K. Sodickson. Parallel magnetic resonance imaging with adaptive radius in k-space (PARS): Constrained image reconstruction using k-space locality in radiofrequency coil encoded data. *Magnetic Resonance in Medicine*, 53(6):1383–1392, Jun 2005.

[79] Chun-yu Yip, Jeffrey A. Fessler, and Douglas C. Noll. Iterative RF pulse design for multidimensional, small-tip-angle selective excitation. *Magnetic Resonance in Medicine*, 54(4):908–917, Oct 2005.

[80] Chun-Yu Yip, William A. Grissom, Jeffrey A. Fessler, and Douglas C. Noll. Joint design of trajectory and RF pulses for parallel excitation. *Magnetic Resonance in Medicine*, 58(3):598–604, Sep 2007.

[81] Daehyun Yoon, Jeffrey A. Fessler, Anna C. Gilbert, and Douglas C. Noll. Fast joint design method for parallel excitation radiofrequency pulse and gradient waveforms considering off-resonance. *Magnetic Resonance in Medicine*, 68(1):278–285, Jul 2012.

[82] A.C. Zelinski, Lawrence L. Wald, Kawin Setsompop, V.K. Goyal, and Elfar Adalsteinsson. Sparsity-Enforced Slice-Selective MRI RF Excitation Pulse Design. *IEEE Transactions on Medical Imaging*, 27(9):1213–1229, Sep 2008.

[83] Tao Zhang, John M. Pauly, Shreyas S. Vasanawala, and Michael Lustig. Coil compression for accelerated imaging with Cartesian sampling. *Magnetic Resonance in Medicine*, 69(2):571–582, 2013.

[84] Ziwu Zhou, Fei Han, Lirong Yan, Danny J.J. Wang, and Peng Hu. Golden-ratio rotated stack-of-stars acquisition for improved volumetric MRI. *Magnetic Resonance in Medicine*, 78(6):2290–2298, Dec 2017.