Iteratively Reweighted ℓ_1 Approaches to Sparse Composite Regularization

Rizwan Ahmad and Philip Schniter, *Fellow, IEEE*

***Abstract*—Motivated by the observation that a given signal $\boldsymbol{x}$ admits sparse representations in multiple dictionaries $\boldsymbol{\Psi}_d$ but with varying levels of sparsity across dictionaries, we propose two new algorithms for the reconstruction of (approximately) sparse signals from noisy linear measurements. Our first algorithm, Co-L1, extends the well-known lasso algorithm from the L1 regularizer $\|\boldsymbol{\Psi x}\|_1$ to composite regularizers of the form $\sum_d \boldsymbol{\lambda}_d \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1$ while self-adjusting the regularization weights $\boldsymbol{\lambda}_d$. Our second algorithm, Co-IRW-L1, extends the well-known iteratively reweighted L1 algorithm to the same family of composite regularizers. We provide several interpretations of both algorithms: 1) majorization-minimization (MM) applied to a non-convex log-sum-type penalty; 2) MM applied to an approximate ℓ_0-type penalty; 3) MM applied to Bayesian MAP inference under a particular hierarchical prior; and 4) variational expectation maximization (VEM) under a particular prior with deterministic unknown parameters. A detailed numerical study suggests that our proposed algorithms yield significantly improved recovery SNR when compared to their noncomposite L1 and IRW-L1 counterparts.**

***Index Terms*—Bayesian methods, composite regularization, iterative reweighting algorithms, majorization minimization, sparse optimization, variational inference.**

I. Introduction

We consider the problem of recovering the signal (or image) $\boldsymbol{x} \in \mathbb{C}^N$ from noisy linear measurements of the form

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{w} \in \mathbb{C}^M, \tag{1}$$

where $\boldsymbol{\Phi} \in \mathbb{C}^{M \times N}$ is a known measurement operator and $\boldsymbol{w} \in \mathbb{C}^M$ is additive noise. Such problems arise in imaging, machine learning, radar, communications, speech, and many other applications. We are particularly interested in the case that $M \ll N$, where $\boldsymbol{x}$ cannot be uniquely determined from the measurements $\boldsymbol{y}$, even in the absence of noise. This latter situation arises in many of the aforementioned applications, as well as in broad area of signal recovery methods associated with *compressive sensing* (CS) [1].

Manuscript received April 20, 2015; revised September 24, 2015; accepted September 25, 2015. Date of publication October 01, 2015; date of current version December 08, 2015. This work was supported in part by NSF under Grant CCF-1218754 and Grant CCF-1018368. Portions of this work were presented at the 2015 ISMRM Annual Meeting and Exhibition. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jong Chul Ye.

The authors are with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: rizwan.ahmad@osumc.edu; schniter@ece.osu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCI.2015.2485078

A. Regularized ℓ_2 Minimization

By incorporating (partial) prior knowledge about the signal and noise power, it may be possible to accurately recover $\boldsymbol{x}$ from $M \ll N$ measurements $\boldsymbol{y}$. In this work, we consider signal recovery based on optimization problems of the form

$$\arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 + R(\boldsymbol{x}) \tag{2}$$

where γ is a tuning parameter that reflects knowledge of the noise level and $R(\boldsymbol{x})$ is a penalty, or regularization, that reflects prior knowledge about the signal $\boldsymbol{x}$ [2]. We briefly summarize several common instances of $R(\boldsymbol{x})$ below.

1) If $\boldsymbol{x}$ is known to be *sparse* (i.e., contains sufficiently few non-zero coefficients) or approximately sparse, then one would ideally like to use the ℓ_0 penalty (i.e., counting "norm") $R(\boldsymbol{x}) = \|\boldsymbol{x}\|_0 \triangleq \sum_{n=1}^N 1_{|x_n|>0}$, where $1_{\{\cdot\}}$ is the indicator function. However, since this choice makes (2) NP-hard, it is not often used in practice.
2) The ℓ_1 penalty, $R(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{n=1}^N |x_n|$, is a well-known surrogate to the ℓ_0 penalty that renders (2) convex, and thus amenable to polynomial-time solution. In this case, (2) is known as the *basis pursuit denoising* [3] or *lasso* [4] problem, which is commonly used in *synthesis-based* CS [1].
3) Various non-convex surrogates for the ℓ_0 penalty have also been considered, such as the ℓ_p penalty $R(\boldsymbol{x}) = \|\boldsymbol{x}\|_p^p = \sum_{n=1}^N |x_n|^p$ with $p \in (0,1)$ and the log-sum penalty $R(\boldsymbol{x}) = \sum_{n=1}^N \log(\epsilon + |x_n|)$ with $\epsilon \geq 0$. Although (2) becomes difficult to solve, it can be tractably approximated. See [2] for a more complete discussion.
4) The choice $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_1$, with known matrix $\boldsymbol{\Psi} \in \mathbb{C}^{L \times N}$, leads to *analysis-based* CS [5] and the *generalized lasso* [6]. Penalties of this form are appropriate when prior knowledge suggests that the transform coefficients $\boldsymbol{\Psi x}$ are (approximately) sparse, as opposed to the signal $\boldsymbol{x}$ itself being sparse. When $\boldsymbol{\Psi}$ is a finite-difference operator, $\|\boldsymbol{\Psi x}\|_1$ yields anisotropic *total variation* regularization [7].
5) Non-convex penalties can also be placed on the transform coefficients $\boldsymbol{\Psi x}$, leading to, e.g., $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_p^p = \sum_{l=1}^L |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}|^p$ with $p \in (0,1)$ or $R(\boldsymbol{x}) = \sum_{l=1}^L \log(\epsilon + |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}|)$ with $\epsilon \geq 0$.

A popular approach to solve (2) with a non-convex penalty $R(\boldsymbol{x})$ is through *iteratively reweighted* ℓ_1 (IRW-L1)[1] [9]. There, (2) with fixed non-convex $R(\boldsymbol{x})$ is approximated by solving a sequence of convex problems

[1] Iteratively reweighted ℓ_2 is a popular alternative, e.g., [8]–[12].

$$\boldsymbol{x}^{(t)} = \arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + R^{(t)}(\boldsymbol{x}), \tag{3}$$

where, at iteration t, the penalty $R^{(t)}(\boldsymbol{x}) = \sum_{n=1}^{N} w_n^{(t)}|x_n|$ with each weight $w_n^{(t)}$ set based on the previous estimate $x_n^{(t-1)}$. Constrained formulations of IRW-L1 based on "$\boldsymbol{x}^{(t)} = \arg\min_{\boldsymbol{x}} R^{(t)}(\boldsymbol{x})$ s.t. $\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2 \le \delta$," have also been considered, such as in [12]–[14]. Many of the papers cited above show empirical results where the performance of IRW-L1 surpasses that of standard ℓ_1.

B. Sparsity-Inducing Composite Regularizers

In this work, we focus on sparsity-inducing *composite* regularizers of the form

$$R_1^D(\boldsymbol{x};\boldsymbol{\lambda}) \triangleq \sum_{d=1}^{D} \lambda_d \|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1, \tag{4}$$

where each $\boldsymbol{\Psi}_d \in \mathbb{C}^{L_d\times N}$ is a known analysis operator and $\lambda_d \ge 0$ is a corresponding regularization weight. Our goal is to recover the signal $\boldsymbol{x}$ from measurements (1) by optimizing (2) with the composite regularizer (4). Doing so requires an optimization of the weights $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_D]^{\mathsf{T}}$ in (4). We are also interested in iteratively re-weighted extensions of this problem that, at iteration t, use composite regularizers of the form[2]

$$R^{(t)}(\boldsymbol{x}) = \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)}\boldsymbol{\Psi}_d\boldsymbol{x}\|_1, \tag{5}$$

where $\boldsymbol{W}_d^{(t)}$ are diagonal matrices. This latter approach requires the optimization of both $\lambda_d^{(t)}$ and $\boldsymbol{W}_d^{(t)}$ for all d.

As a motivating example, suppose that $\{\boldsymbol{\Psi}_d\}$ is a collection of orthonormal bases that includes, e.g., spikes, sines, and various wavelet bases. The signal $\boldsymbol{x}$ may be sparse in some of these bases, but not all. Thus, we would like to adjust each λ_d in (4) to appropriately weight the contribution from each basis. But it is not clear how to do this, especially since $\boldsymbol{x}$ is unknown. As another example, suppose that $\boldsymbol{x}$ contains a (rasterized) sequence of images and that $\|\boldsymbol{\Psi}_1\boldsymbol{x}\|_1$ measures temporal total-variation while $\|\boldsymbol{\Psi}_2\boldsymbol{x}\|_1$ measures spatial total-variation. Intuitively, we would like to weight these two regularizations differently, depending on whether the image pixels vary more in the temporal or spatial dimensions. But it is not clear how to do this, especially since $\boldsymbol{x}$ is unknown.

C. Contributions

In this work, we propose novel iteratively reweighted approaches to sparse reconstruction based on composite regularizations of the form (4)–(5) with automatic tuning of the regularization weights $\boldsymbol{\lambda}$ and $\boldsymbol{W}_d$. For each of our proposed algorithms, we will provide four interpretations:

1) MM applied to a non-convex log-sum-type penalty,
2) MM applied to an approximate ℓ_0-type penalty,
3) MM applied to Bayesian MAP inference based on Gamma and Jeffrey's hyperpriors [15], [16], and
4) variational expectation maximization (VEM) [17], [18] applied to a Laplacian or generalized-Pareto prior with deterministic unknown parameters.

We show that the MM interpretation guarantees convergence in the sense of satisfying an asymptotic stationary point condition [19]. Moreover, we establish connections between our proposed approaches and existing IRW-L1 algorithms, and we provide novel VEM-based and Bayesian MAP interpretations of those existing algorithms.

Finally, through the detailed numerical study in Sec. IV, we establish that our proposed algorithms yield significant gains in recovery accuracy relative to existing methods with only modest increases in runtime. In particular, when $\{\boldsymbol{\Psi}_d\}$ are chosen so that the sparsity of $\boldsymbol{\Psi}_d\boldsymbol{x}$ varies with d, this structure can be exploited for improved recovery. The more disparate the sparsity, the greater the improvement.

D. Related Work

As discussed above, the generalized lasso [6] is one of the most common approaches to L1-regularized analysis-CS [5], i.e., the optimization (2) under the regularizer $R(\boldsymbol{x}) = \|\boldsymbol{\Psi}\boldsymbol{x}\|_1$. The Co-L1 algorithm that we present in Sec. II can be interpreted as a generalization of this L1 method to *composite* regularizers of the form (4). Meanwhile, the iteratively reweighted extension of the generalized lasso, IRW-L1 [9], often yields significantly better reconstruction accuracy with a modest increase in complexity (e.g., [13], [14]). The Co-IRW-L1 algorithm that we present in Sec. III can be interpreted as a generalization of this IRW-L1 method to *composite* regularizers of the form (5). The existing non-composite L1 and IRW-L1 approaches essentially place an identical weight $\lambda_d = 1$ on every term in (4)–(5), and thus make no attempt to leverage differences in the sparsity of the transform coefficients $\boldsymbol{\Psi}_d\boldsymbol{x}$ across the sub-dictionary index d. However, the numerical results that we present in Sec. IV suggest that there can be significant advantages to optimizing λ_d, which is precisely what our methods do.

The problem of optimizing the weights λ_d of composite regularizers $R(\boldsymbol{x};\boldsymbol{\lambda}) = \sum_d \lambda_d R_d(\boldsymbol{x})$ is a long-standing problem with a rich literature (see, e.g., the recent book [20]). However, the vast majority of that literature focuses on the Tikhonov case where $R_d(\boldsymbol{x})$ are quadratic (see, e.g., [21]–[24]). One notable exception is [25], which assumes continuously differentiable $R_d(\boldsymbol{x})$ and thus does not cover our composite ℓ_1 prior (4). Another notable exception is [26], which assumes i) the availability of a noiseless training example of $\boldsymbol{x}$ to help tune the L1 regularization weights $\boldsymbol{\lambda}$ in (4), and ii) the trivial measurement matrix $\boldsymbol{\Phi} = \boldsymbol{I}$. In contrast, our proposed methods operate without any training and support generic measurement matrices $\boldsymbol{\Phi}$.

In the special case that each $\boldsymbol{\Psi}_d$ is composed of a subset of rows from the $N \times N$ identity matrix, the regularizers (4)–(5) can induce *group* sparsity in the recovery of $\boldsymbol{x}$, in that certain sub-vectors $\boldsymbol{x}_d \triangleq \boldsymbol{\Psi}_d\boldsymbol{x}$ of $\boldsymbol{x}$ are driven to zero while others are not. The paper [27] develops an IRW-L1-based approach to group-sparse signal recovery for equal-sized non-overlapping groups that can be considered as a special case of the Co-L1 algorithm that we develop in Sec. II. However, our

[2]Although (5) is over-parameterized, the form of (5) is convenient for algorithm development.

approach is more general in that it handles possibly non-equal and/or overlapping groups, not to mention sparsity in a generic set of sub-dictionaries $\boldsymbol{\Psi}_d$. Recently, Bayesian MAP group-sparse recovery was considered in [28]. However, the technique described there uses Gaussian scale mixtures or, equivalently, weighted-L2 regularizers $R(\boldsymbol{x};\boldsymbol{\lambda}) = \sum_d \lambda_d \|\boldsymbol{x}_d\|_2$, while our methods use weighted-ℓ_1 regularizers (4)–(5).

E. Notation

We use boldface capital letters like $\boldsymbol{\Psi}$ for matrices, boldface small letters like $\boldsymbol{x}$ for vectors, and $(\cdot)^{\mathsf{T}}$ for transposition. We use $\|\boldsymbol{x}\|_p = (\sum_n |x_n|^p)^{1/p}$ for the ℓ_p norm of $\boldsymbol{x}$, with x_n representing the n^{th} coefficient in $\boldsymbol{x}$ and $p > 0$. We then use $\|\boldsymbol{x}\|_0 = \lim_{p\to 0} \sum_n |x_n|^p$ [12] when referring to the ℓ_0 quasi-norm, which counts the number of nonzero coefficients in $\boldsymbol{x}$. We define the "mixed $\ell_{p,0}$ quasi-norm" with $p > 0$ as[3] $\lim_{q\to 0} \sum_d (\sum_l |x_{d,l}|^p)^q$, and the "mixed $\ell_{0,0}$ quasi-norm" as $\lim_{p,q\to 0} \sum_d (\sum_l |x_{d,l}|^p)^q$. We use $\nabla g(\boldsymbol{x})$ for the gradient of a functional $g(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and 1_A for the indicator function that returns the value 1 when A is true and 0 when A is false. We use $p(\boldsymbol{x};\boldsymbol{\lambda})$ for the pdf of random vector $\boldsymbol{x}$ under deterministic parameters $\boldsymbol{\lambda}$, and $p(\boldsymbol{x}|\boldsymbol{\lambda})$ for the pdf of $\boldsymbol{x}$ conditioned on the random vector $\boldsymbol{\lambda}$. We use $D_{\mathsf{KL}}(q\|p)$ to denote the Kullback-Leibler (KL) divergence of pdf p from pdf q, and we use $\mathbb{R}$ and $\mathbb{C}$ to denote the real and complex fields, respectively.

II. The Co-L1 Algorithm

We first propose the Composite-L1 (Co-L1) algorithm, which is summarized in Algorithm 1. There, L_d denotes the number of rows in $\boldsymbol{\Psi}_d$.

Algorithm 1. The Co-L1 Algorithm

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma > 0$, $\epsilon \geq 0$
2: if $\boldsymbol{\Psi}_d\boldsymbol{x} \in \mathbb{R}^{L_d}$, use $C_d = 1$; if $\boldsymbol{\Psi}_d\boldsymbol{x} \in \mathbb{C}^{L_d}$, use $C_d = 2$.
3: initialization: $\lambda_d^{(1)} = 1\ \forall d$
4: for $t = 1, 2, 3, \ldots$
5: $\quad \boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^D \lambda_d^{(t)} \|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1$
6: $\quad \lambda_d^{(t+1)} \leftarrow \dfrac{C_d L_d}{\epsilon + \|\boldsymbol{\Psi}_d \boldsymbol{x}^{(t)}\|_1}, d = 1, \ldots, D$
7: end
8: output: $\boldsymbol{x}^{(t)}$

The main computational step of Co-L1 is the L2+L1 minimization in line 5, which can be recognized as (2) under the composite regularizer R_1^D from (4). This is a convex optimization problem that can be readily solved by existing techniques (e.g., ADMM [30], [31], Douglas-Rachford splitting [32], MFISTA [33], NESTA-UP [34], GAMP [35], etc.), the specific choice of which is immaterial to this paper.

Note that Co-L1 requires the user to set a small regularization term $\epsilon \geq 0$ whose role is to prevent the denominator in line 6 from reaching zero. For typical choices of $\boldsymbol{\Psi}_d$ and γ, the vector $\boldsymbol{\Psi}_d\boldsymbol{x}^{(t)}$ will almost never be exactly zero, in which case it suffices to set $\epsilon = 0$. Also, Co-L1 requires the user to set the measurement fidelity weight γ. With additive white Gaussian noise (AWGN) of variance $\sigma^2 > 0$, the Bayesian MAP interpretation discussed in Sec. II-D suggests setting $\gamma = \frac{1}{2\sigma^2}$ for real-valued AWGN or $\gamma = \frac{1}{\sigma^2}$ for circular complex-valued AWGN. These are, in fact, the settings that we used for all numerical results in Sec. IV.

Note line 5 of Algorithm 1 can be equivalently restated as

$$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \sum_{d=1}^D \lambda_d^{(t)} \|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1 \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta. \quad (6)$$

By equivalent, we mean that, for any $\delta > 0$, there exists a γ for which the solutions of line 5 and (6) are identical [36]. A version of this manuscript that focuses on the constrained case can be found at [37]. Numerical experiments therein show that the performance of Co-L1 using (6) with the hand-tuned value $\delta = 0.8\sqrt{M\sigma^2}$ is very similar to that of Algorithm 1 with γ chosen as described above.

Co-L1's update of the weights $\boldsymbol{\lambda}$, defined by line 6 of Algorithm 1, can be interpreted in various ways, as we detail below. For ease of explanation, we first consider the case where $\boldsymbol{\Psi}_d\boldsymbol{x}$ is real-valued $\forall d$, and later discuss the complex-valued case in Sec. II-F.

Theorem 1 (Co-L1): The Co-L1 algorithm in Algorithm 1 has the following interpretations:

1) MM applied to (2) under the log-sum penalty

$$R_{\text{ls}}^D(\boldsymbol{x};\epsilon) \triangleq \sum_{d=1}^D L_d \log(\epsilon + \|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1), \quad (7)$$

2) as $\epsilon \to 0$, MM applied to (2) under the weighted $\ell_{1,0}$ [29] penalty

$$R_{10}^D(\boldsymbol{x}) \triangleq \sum_{d=1}^D L_d\, 1_{\|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1 > 0}, \quad (8)$$

3) MM applied to Bayesian MAP estimation under an additive white Gaussian noise (AWGN) likelihood and the hierarchical prior

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{d=1}^D \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d \|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1\right) \quad (9)$$

$$\boldsymbol{\lambda} \sim \text{i.i.d. } \Gamma(0, \epsilon^{-1}) \quad (10)$$

where $\boldsymbol{z}_d \triangleq \boldsymbol{\Psi}_d\boldsymbol{x} \in \mathbb{R}^{L_d}$ is i.i.d. Laplacian given λ_d, and λ_d is Gamma distributed with scale parameter ϵ^{-1} and shape parameter zero, which becomes Jeffrey's non-informative hyperprior $p(\lambda_d) \propto 1_{\lambda_d > 0}/\lambda_d$ when $\epsilon = 0$,

4) variational EM under an AWGN likelihood and the prior

$$p(\boldsymbol{x};\boldsymbol{\lambda}) \propto \prod_{d=1}^D \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d(\|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1 + \epsilon)\right), \quad (11)$$

which, when $\epsilon = 0$, is i.i.d. Laplacian on $\boldsymbol{z}_d = \boldsymbol{\Psi}_d\boldsymbol{x} \in \mathbb{R}^{L_d}$ with deterministic scale parameter $\lambda_d > 0$.

[3] Our $\ell_{p,0}$ and $\ell_{0,0}$ definitions are motivated by the standard $\ell_{p,q}$ mixed norm definition (for $p, q > 0$), which is $(\sum_d (\sum_l |x_{d,l}|^p)^{q/p})^{1/q}$ [29].

Proof: See Sections II-A to II-E below. ■

Importantly, the MM interpretation implies convergence (in the sense of an asymptotic stationary point condition) when $\epsilon > 0$, as detailed in Sec. II-B.

A. Log-Sum MM Interpretation of Co-L1

Consider the optimization problem

$$\arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + R_{\text{ls}}^D(\boldsymbol{x};\epsilon) \tag{12}$$

with R_{ls}^D from (7). Inspired by [13, § 2.3], we write (12) as

$$\begin{aligned}\arg\min_{\boldsymbol{x},\boldsymbol{u}} &\gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2 + \sum_{d=1}^{D} L_d \log\left(\epsilon + \sum_{l=1}^{L_d} u_{d,l}\right)\\ &\text{s.t. } |\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}| \le u_{d,l}\ \forall d,l,\end{aligned} \tag{13}$$

where $\boldsymbol{\psi}_{d,l}^{\mathsf{T}}$ is the lth row of $\boldsymbol{\Psi}_d$. Problem (13) is of the form

$$\arg\min_{\boldsymbol{v}} g(\boldsymbol{v}) \text{ s.t. } \boldsymbol{v}\in\mathcal{C}, \tag{14}$$

where $\boldsymbol{v} = [\boldsymbol{u}^{\mathsf{T}}, \boldsymbol{x}^{\mathsf{T}}]^{\mathsf{T}}$, $\mathcal{C}$ is a convex set,

$$g(\boldsymbol{v}) = \gamma\|\boldsymbol{y} - [\boldsymbol{0} \quad \boldsymbol{\Phi}]\boldsymbol{v}\|_2^2 + \sum_{d=1}^{D} L_d \log\left(\epsilon + \sum_{k\in\mathcal{K}_d} v_k\right) \tag{15}$$

is a non-convex penalty, and the set $\mathcal{K}_d \triangleq \{k : \sum_{i=1}^{d-1} L_i < k \le \sum_{i=1}^{d} L_i\}$ contains the indices k such that $v_k \in \{u_{d,l}\}_{l=1}^{L_d}$.

Since $g(\boldsymbol{v})$ is the sum of convex and concave terms, i.e., a "difference of convex" (DC) functions, (14) can be recognized as a DC program [38]. Majorization-minimization (MM) [19], [39] is a popular method to attack non-convex problems of this form. In particular, MM iterates the following two steps: (i) construct a surrogate $g(\boldsymbol{v};\boldsymbol{v}^{(t)})$ that majorizes $g(\boldsymbol{v})$ at $\boldsymbol{v}^{(t)}$, and (ii) update $\boldsymbol{v}^{(t+1)} = \arg\min_{\boldsymbol{v}\in\mathcal{C}} g(\boldsymbol{v};\boldsymbol{v}^{(t)})$. By "majorize," we mean that $g(\boldsymbol{v};\boldsymbol{v}^{(t)}) \ge g(\boldsymbol{v})$ for all $\boldsymbol{v}$ with equality when $\boldsymbol{v} = \boldsymbol{v}^{(t)}$.

Due to the DC form of $g(\boldsymbol{v})$ in (15), a majorizing surrogate can be constructed by linearizing the concave term about its tangent at $\boldsymbol{v}^{(t)}$. In particular, say $g(\boldsymbol{v}) = g_1(\boldsymbol{v}) + g_2(\boldsymbol{v})$, where g_1 is the convex (quadratic) term and g_2 is the concave (log-sum) term, and say ∇g_2 is the gradient of g_2 w.r.t. $\boldsymbol{v}$. Then

$$g(\boldsymbol{v};\boldsymbol{v}^{(t)}) \triangleq g_1(\boldsymbol{v}) + g_2(\boldsymbol{v}^{(t)}) + \nabla g_2(\boldsymbol{v}^{(t)})^{\mathsf{T}}[\boldsymbol{v}-\boldsymbol{v}^{(t)}] \tag{16}$$

majorizes $g(\boldsymbol{v})$ at $\boldsymbol{v}^{(t)}$, and so the MM iterations become

$$\boldsymbol{v}^{(t+1)} = \arg\min_{\boldsymbol{v}\in\mathcal{C}} g_1(\boldsymbol{v}) + \nabla g_2(\boldsymbol{v}^{(t)})^{\mathsf{T}}\boldsymbol{v} \tag{17}$$

after neglecting the $\boldsymbol{v}$-invariant terms.

Examining the log-sum term in (15), we see that

$$[\nabla g_2(\boldsymbol{v}^{(t)})]_k = \begin{cases} \dfrac{L_{d(k)}}{\epsilon + \sum_{i\in\mathcal{K}_{d(k)}} v_i^{(t)}} & \text{if } d(k)\neq 0 \\ 0 & \text{else,}\end{cases} \tag{18}$$

where $d(k)$ is the index $d\in\{1,\dots,D\}$ of the set $\mathcal{K}_d$ containing k, or 0 if no such set exists. Thus MM prescribes

$$\boldsymbol{v}^{(t+1)} = \arg\min_{\boldsymbol{v}\in\mathcal{C}} \gamma\|\boldsymbol{y} - [\boldsymbol{0}\ \boldsymbol{\Phi}]\boldsymbol{v}\|_2^2 + \sum_{d=1}^{D}\sum_{k\in\mathcal{K}_d} \frac{L_d v_k}{\epsilon + \sum_{i\in\mathcal{K}_d} v_i^{(t)}}, \tag{19}$$

or equivalently

$$\boldsymbol{x}^{(t+1)} = \arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \frac{L_d \sum_{l=1}^{L_d} |\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon + \sum_{l=1}^{L_d} |\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}^{(t)}|} \tag{20}$$

$$= \arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t+1)}\|\boldsymbol{\Psi}_d\boldsymbol{x}\|_1 \tag{21}$$

for

$$\lambda_d^{(t+1)} = \frac{L_d}{\epsilon + \|\boldsymbol{\Psi}_d\boldsymbol{x}^{(t)}\|_1}, \tag{22}$$

which coincides with Algorithm 1. This establishes Part 1 of Theorem 1.

B. Convergence of Co-L1

The paper [19] studies the convergence of MM, and includes a special discussion of the application of MM to DC programming. In the language of our Sec. II-A, [19] establishes that, when g_2 is differentiable with a Lipschitz continuous gradient, the MM sequence $\{\boldsymbol{v}^{(t)}\}_{t\ge1}$ satisfies an asymptotic stationary point (ASP) condition. Although this falls short of establishing convergence to a local minimum (which is difficult for generic non-convex problems), the ASP condition is based on a classical necessary condition for a local minimum. In particular, using $\nabla g(\boldsymbol{v};\boldsymbol{d})$ to denote the directional derivative of g at $\boldsymbol{v}$ in the direction $\boldsymbol{d}$, it is known [40] that $\boldsymbol{v}_\star$ locally minimizes g over $\mathcal{C}$ only if $\nabla g(\boldsymbol{v}_\star;\boldsymbol{v}-\boldsymbol{v}_\star)\ge 0$ for all $\boldsymbol{v}\in\mathcal{C}$. Thus, in [19], it is said that $\{\boldsymbol{v}^{(t)}\}_{t\ge1}$ satisfies an ASC condition if

$$\liminf_{t\to+\infty}\ \inf_{\boldsymbol{v}\in\mathcal{C}} \frac{\nabla g(\boldsymbol{v}^{(t)};\boldsymbol{v}-\boldsymbol{v}^{(t)})}{\|\boldsymbol{v}-\boldsymbol{v}^{(t)}\|_2} \ge 0. \tag{23}$$

In our case, g_2 from (15) is indeed differentiable, with gradient ∇g_2 given by (18). Moreover, Appendix A shows that this gradient is Lipschitz continuous when $\epsilon > 0$. Thus, the sequence of estimates produced by Algorithm 1 satisfies the ASP condition (23).

C. Approximate $\ell_{1,0}$ Interpretation of Co-L1

In the limit of $\epsilon\to 0$, the log-sum minimization

$$\arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{n=1}^{N}\log(\epsilon+|x_n|) \tag{24}$$

for $\gamma > 0$ is known [12] to be equivalent to ℓ_0 minimization

$$\arg\min_{\boldsymbol{x}} \gamma'\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \|\boldsymbol{x}\|_0 \tag{25}$$

for some $\gamma' > 0$. (See Appendix B for a proof.) This equivalence can be seen intuitively as follows. As $\epsilon \to 0$, the contribution to the regularization term $\sum_{n=1}^{N} \log(\epsilon + |x_n|)$ from each non-zero x_n remains finite, while that from each zero-valued x_n approaches $-\infty$. Since we are interested in minimizing the regularization term, we get a huge reward for each zero-valued x_n, or—equivalently—a huge penalty for each non-zero x_n.

To arrive at an ℓ_0 interpretation of the Co-L1 algorithm, we consider the corresponding optimization problem (12) in the limit that $\epsilon \to 0$. There we see that the regularization term $R_{\text{ls}}^{D}(\boldsymbol{x}; 0)$ from (7) yields L_d huge rewards when $\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 = 0$, or equivalently L_d huge penalties when $\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \neq 0$, for each $d \in \{1, \ldots, D\}$. Thus, we can interpret Co-L1 as attempting to solve the optimization problem (8), which is a weighted version of the "$\ell_{p,q}$ mixed norm" problem from [29] for $p = 1$ and $q \to 0$. This establishes Part 2 of Theorem 1.

D. Bayesian MAP Interpretation of Co-L1

The MAP estimate [41] of $\boldsymbol{x}$ from $\boldsymbol{y}$ is

$$\boldsymbol{x}_{\mathsf{MAP}} \triangleq \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \{-\log p(\boldsymbol{x}|\boldsymbol{y})\} \tag{26}$$

$$= \arg\min_{\boldsymbol{x}} \{-\log p(\boldsymbol{x}) - \log p(\boldsymbol{y}|\boldsymbol{x})\} \tag{27}$$

$$= \arg\min_{\boldsymbol{x}} \left\{-\log p(\boldsymbol{x}) + \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2\right\}, \tag{28}$$

where (26) used the monotonicity of $\log$, (27) used Bayes rule, and (28) used the AWGN likelihood. Note that, for real-valued AWGN with σ^2 variance, $\gamma = \frac{1}{2\sigma^2}$, while for circular complex-valued AWGN with σ^2 variance, $\gamma = \frac{1}{\sigma^2}$.

Next, we derive the $-\log p(\boldsymbol{x})$ term in (28) that results from the hierarchical prior (9)–(10). Recall that, with shape parameter κ and scale parameter θ, the Gamma pdf [42] is $\Gamma(\lambda_d; \kappa, \theta) = 1_{\lambda_d > 0} \lambda_d^{\kappa-1} \theta^{-\kappa} \exp(-\lambda_d/\theta)/\Gamma(\kappa)$, where $\Gamma(\kappa)$ is the Gamma function. Since $\Gamma(\lambda_d; \kappa, \theta) \propto 1_{\lambda_d > 0} \lambda_d^{\kappa-1} \exp(-\lambda_d/\theta)$, we note that $\Gamma(\lambda_d; 0, \infty) \propto 1_{\lambda_d > 0}/\lambda_d$, which is Jeffrey's non-informative hyperprior [15], [42] for the Laplace scale parameter λ_d. Then, according to (9)–(10), the prior equals

$$p(\boldsymbol{x}) = \int_{\mathbb{R}^D} p(\boldsymbol{x}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{\lambda} \tag{29}$$

$$\propto \prod_{d=1}^{D} \int_0^\infty \left(\frac{\lambda_d}{2}\right)^{L_d} \exp(-\lambda_d \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1) \frac{\exp(-\lambda_d \epsilon)}{\lambda_d} \, \mathrm{d}\lambda_d \tag{30}$$

$$= \prod_{d=1}^{D} \frac{(L_d - 1)!}{(2(\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 + \epsilon))^{L_d}} \tag{31}$$

which implies that

$$-\log p(\boldsymbol{x}) = \text{const} + \sum_{d=1}^{D} L_d \log\left(\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 + \epsilon\right). \tag{32}$$

Equations (28), (32), and (7) imply

$$\boldsymbol{x}_{\mathsf{MAP}} = \arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + R_{\text{ls}}^{D}(\boldsymbol{x}; 0). \tag{33}$$

Finally, applying the MM algorithm to this optimization problem (as detailed in Sec. II-A), we arrive at Algorithm 1. We note that [16] proposed to use Gamma and Jeffrey's hyperpriors with MM for total-variation image deblurring, although their algorithm is not of the IRW-L1 form. This establishes Part 3 of Theorem 1.

E. Variational EM Interpretation of Co-L1

The variational expectation-maximization (VEM) algorithm [17], [18] is an iterative approach to maximum-likelihood (ML) estimation that generalizes the EM algorithm from [43]. We now provide a brief review of the VEM algorithm and describe how it can be applied to estimate $\boldsymbol{\lambda}$ in (11).

First, note that the log-likelihood can be written as

$$\log p(\boldsymbol{y}; \boldsymbol{\lambda}) = \int q(\boldsymbol{x}) \log p(\boldsymbol{y}; \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{x} \tag{34}$$

$$= \int q(\boldsymbol{x}) \log \left[\frac{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda})}{q(\boldsymbol{x})} \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\lambda})}\right] \mathrm{d}\boldsymbol{x} \tag{35}$$

$$= \underbrace{\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda})}{q(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}}_{\triangleq F\left(q(\boldsymbol{x}); \boldsymbol{\lambda}\right)} + \underbrace{\int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\lambda})} \, \mathrm{d}\boldsymbol{x}}_{\triangleq D_{\mathsf{KL}}\left(q(\boldsymbol{x}) \| p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\lambda})\right)}, \tag{36}$$

for an arbitrary pdf $q(\boldsymbol{x})$, where $D_{\mathsf{KL}}(q\|p)$ denotes the KL divergence of p from q. Because $D_{\mathsf{KL}}(q\|p) \geq 0$ for any q and p, we see that $F(q(\boldsymbol{x}); \boldsymbol{\lambda})$ is a lower bound on $\log p(\boldsymbol{y}; \boldsymbol{\lambda})$. The EM algorithm performs ML estimation by iterating

$$q^{(t)}(\boldsymbol{x}) = \arg\min_{q} D_{\mathsf{KL}}\left(q(\boldsymbol{x}) \middle\| p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\lambda}^{(t)})\right) \tag{37}$$

$$\boldsymbol{\lambda}^{(t+1)} = \arg\max_{\boldsymbol{\lambda}} F(q^{(t)}(\boldsymbol{x}); \boldsymbol{\lambda}), \tag{38}$$

where the "E" step (37) tightens the lower bound and the "M" step (38) maximizes the lower bound.

The EM algorithm places no constraints on $q(\boldsymbol{x})$, in which case the solution to (37) is simply $q^{(t)}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\lambda}^{(t)})$, i.e., the posterior pdf of $\boldsymbol{x}$ under $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. In many applications, however, this posterior is too difficult to compute and/or use in (38). To circumvent this problem, the VEM algorithm constrains $q(\boldsymbol{x})$ to some family of distributions $\mathcal{Q}$ that makes (37)–(38) tractable.

For our application of the VEM algorithm, we constrain to distributions of the form

$$q(\boldsymbol{x}) \propto \lim_{T \to 0} \exp\left(\tfrac{1}{T} \log p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\lambda})\right), \tag{39}$$

which has the effect of concentrating the mass in $q(\boldsymbol{x})$ at its mode. Plugging this $q(\boldsymbol{x})$ and $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}) = p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}; \boldsymbol{\lambda})$ into (36), we see that the M step (38) reduces to

$$\boldsymbol{\lambda}^{(t+1)} = \arg\max_{\boldsymbol{\lambda}} \log p(\boldsymbol{x}; \boldsymbol{\lambda}) \Big|_{\boldsymbol{x} = \boldsymbol{x}_{\mathsf{MAP}}^{(t)}} \tag{40}$$

$$\text{for } \boldsymbol{x}_{\mathsf{MAP}}^{(t)} \triangleq \arg\max_{\boldsymbol{x}} p\left(\boldsymbol{x} \middle| \boldsymbol{y}; \boldsymbol{\lambda}^{(t)}\right), \tag{41}$$

where (41) be interpreted as the E step. For the particular $p(\boldsymbol{x};\boldsymbol{\lambda})$ in (11), we have that

$$\log p(\boldsymbol{x};\boldsymbol{\lambda}) = \text{const} + \sum_{d=1}^{D} \left[L_d \log(\lambda_d) - \lambda_d(\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 + \epsilon) \right], \tag{42}$$

and by zeroing the gradient w.r.t. $\boldsymbol{\lambda}$, we find that (40) becomes

$$\lambda_d^{(t+1)} = \frac{L_d}{\|\boldsymbol{\Psi}_d \boldsymbol{x}_{\mathsf{MAP}}^{(t)}\|_1 + \epsilon}, \qquad d = 1, \ldots, D. \tag{43}$$

Meanwhile, from (28) and (11), we find that (41) becomes

$$\boldsymbol{x}_{\mathsf{MAP}}^{(t)} = \arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1. \tag{44}$$

In conclusion, our VEM algorithm iterates the steps (43)–(44), which match the steps in Algorithm 1. This establishes Part 4 of Theorem 1.

F. Co-L1 for Complex-Valued $\boldsymbol{\Psi}_d \boldsymbol{x}$

In Theorem 1 and Sections II-A-II-E, real-valued analysis outputs $\boldsymbol{\Psi}_d \boldsymbol{x}$ were assumed for ease of explanation. We now extend the previous results to the case of complex-valued $\boldsymbol{\Psi}_d \boldsymbol{x}$. For this, we focus on the VEM interpretation (recall Part 4 of Theorem 1), noting that a similar justification can be made based on the Bayesian MAP interpretation. In particular, we assume an AWGN likelihood and a complex-valued extension of the prior (11):

$$p(\boldsymbol{x};\boldsymbol{\lambda}) \propto \prod_{d=1}^{D} \left(\frac{\lambda_d}{2\pi} \right)^{2L_d} \exp\left(-\lambda_d(\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 + \epsilon) \right), \tag{45}$$

which, when $\epsilon = 0$, is i.i.d. complex-valued Laplacian on $\boldsymbol{z}_d = \boldsymbol{\Psi}_d \boldsymbol{x} \in \mathbb{C}^{L_d}$ with deterministic scale parameter $\lambda_d > 0$. To show this, we follow the steps in Sec. II-E up to the log-prior in (42), which now becomes

$$\log p(\boldsymbol{x};\boldsymbol{\lambda}) = \text{const} + \sum_{d=1}^{D} \left[2L_d \log(\lambda_d) - \lambda_d(\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 + \epsilon) \right]. \tag{46}$$

Zeroing the gradient w.r.t. $\boldsymbol{\lambda}$, we find that the VEM update in (40) becomes

$$\lambda_d^{(t+1)} = \frac{2L_d}{\|\boldsymbol{\Psi}_d \boldsymbol{x}_{\mathsf{MAP}}^{(t)}\|_1 + \epsilon}, \qquad d = 1, \ldots, D, \tag{47}$$

which is twice as large as the real-valued case in (43).

G. New Interpretations of the IRW-L1 Algorithm

The proposed Co-L1 algorithm is related to the analysis-CS formulation of the well-known IRW-L1 algorithm [9]. For clarity, and for later use in Sec. III, we summarize this latter algorithm in Algorithm 2, and note that the synthesis-CS formulation follows from the special case that $\boldsymbol{\Psi} = \boldsymbol{I}$.

Algorithm 2. The IRW-L1 Algorithm

input: $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L]^{\mathsf{T}}$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma \geq 0$, $\epsilon \geq 0$
initialization: $\boldsymbol{W}^{(1)} = \boldsymbol{I}$
for $t = 1, 2, 3, \ldots$
  $\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \|\boldsymbol{W}^{(t)} \boldsymbol{\Psi} \boldsymbol{x}\|_1$
  $\boldsymbol{W}^{(t+1)} \leftarrow \text{diag}\left\{ \frac{1}{\epsilon + |\boldsymbol{\psi}_1^{\mathsf{T}} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon + |\boldsymbol{\psi}_L^{\mathsf{T}} \boldsymbol{x}^{(t)}|} \right\}$
end
output: $\boldsymbol{x}^{(t)}$

Comparing Algorithm 2 to Algorithm 1, we see that IRW-L1 coincides with real-valued Co-L1 in the case that every sub-dictionary $\boldsymbol{\Psi}_d$ has dimension one, i.e., $C_d = 1 = L_d \ \ \forall d$ and $D = L$, where $L \triangleq \sum_{d=1}^{D} L_d$ denotes the total number of analysis coefficients. Thus, the Co-L1 interpretations from Theorem 1 can be directly translated to IRW-L1 as follows.

Corollary 2 (IRW-L1): The IRW-L1 algorithm from Algorithm 2 has the following interpretations:

1) MM applied to (2) under the log-sum penalty

$$R_{\text{ls}}^{L}(\boldsymbol{x};\epsilon) = \sum_{l=1}^{L} \log(\epsilon + |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}|), \tag{48}$$

recalling the definition of R_{ls}^{L} from (7),

2) as $\epsilon \to 0$, MM applied to (2) under the ℓ_0 penalty

$$R_0^L(\boldsymbol{x}) \triangleq \sum_{l=1}^{L} 1_{|\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}| > 0}, \tag{49}$$

3) MM applied to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{l=1}^{L} \frac{\lambda_l}{2} \exp\left(-\lambda_l |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}| \right) \tag{50}$$

$$\boldsymbol{\lambda} \sim \text{i.i.d.} \quad \Gamma(0, \epsilon^{-1}) \tag{51}$$

where $z_l = \boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}$ is Laplacian given λ_l, and λ_l is Gamma distributed with scale parameter ϵ^{-1} and shape parameter zero, which becomes Jeffrey's non-informative hyper-prior $p(\lambda_l) \propto 1_{\lambda_l > 0}/\lambda_l$ when $\epsilon = 0$.

4) variational EM under an AWGN likelihood and the prior

$$p(\boldsymbol{x};\boldsymbol{\lambda}) \propto \prod_{l=1}^{L} \frac{\lambda_l}{2} \exp\left(-\lambda_l(|\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}| + \epsilon) \right). \tag{52}$$

which, when $\epsilon = 0$, is independent Laplacian on $\boldsymbol{z} = \boldsymbol{\Psi}\boldsymbol{x} \in \mathbb{R}^L$ under the positive deterministic scale parameters in $\boldsymbol{\lambda}$.

While Part 1 and Part 2 of Corollary 2 were established for the ℓ_2-constrained synthesis-CS formulation of IRW-L1 in [13], we believe that Part 3 and Part 4 are novel interpretations of IRW-L1.

III. The Co-IRW-L1 Algorithm

We now propose the Co-IRW-L1-$\boldsymbol{\epsilon}$ algorithm, which is summarized in Algorithm 3. Co-IRW-L1-$\boldsymbol{\epsilon}$ can be thought

of as a hybrid of the Co-L1 and IRW-L1 approaches from Algorithms 1 and 2, respectively. Like with Co-L1, the Co-IRW-L1-$\boldsymbol{\epsilon}$ algorithm uses sub-dictionary dependent weights λ_d that are updated at each iteration t using a sparsity metric on $\boldsymbol{\Psi}_d \boldsymbol{x}^{(t)}$. But, like with IRW-L1, the Co-IRW-L1-$\boldsymbol{\epsilon}$ algorithm also uses diagonal weight matrices $\boldsymbol{W}_d^{(t)}$ that are updated at each iteration. As with both Co-L1 and IRW-L1, the computational burden of Co-IRW-L1-$\boldsymbol{\epsilon}$ is dominated by the L2+L1 minimization problem in line 4 of Algorithm 3, which is readily solved by existing techniques like MFISTA.

Algorithm 3. The Real-Valued Co-IRW-L1-$\boldsymbol{\epsilon}$ Algorithm

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma > 0$, $\epsilon_d > 0 \; \forall d$, $\varepsilon \geq 0$,
2: initialization: $\lambda_d^{(1)} = 1 \; \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I} \; \forall d$
3: for $t = 1, 2, 3, \ldots$
4: $\quad \boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1$
5: $\quad \lambda_d^{(t+1)} \leftarrow \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log\left(1 + \varepsilon + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}{\epsilon_d}\right) \right]^{-1} + 1, \quad \forall d = 1, \ldots, D$
6: $\quad \boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag}\left\{ \frac{1}{\epsilon_d(1+\varepsilon) + |\boldsymbol{\psi}_{d,1}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d(1+\varepsilon) + |\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}} \boldsymbol{x}^{(t)}|} \right\}, \quad \forall d$
7: end
8: output: $\boldsymbol{x}^{(t)}$

The Co-IRW-L1-$\boldsymbol{\epsilon}$ algorithm can be interpreted in various ways, as we detail below. For clarity, we first consider fixed regularization parameters $\boldsymbol{\epsilon} \triangleq [\epsilon_1, \ldots, \epsilon_D]^{\mathsf{T}}$ and later, in Sec. III-F, we describe how they can be adapted at each iteration, leading to the Co-IRW-L1 algorithm. Also, to simplify the development, we first consider the real-valued case and discuss the complex-valued case later, in Sec. III-G.

Theorem 3 (Co-IRW-L1-$\boldsymbol{\epsilon}$): The real-valued Co-IRW-L1-$\boldsymbol{\epsilon}$ algorithm in Algorithm 3 has the following interpretations:

1) MM applied to (2) under the log-sum-log penalty

$$R_{\text{lsl}}(\boldsymbol{x}; \boldsymbol{\epsilon}, \varepsilon) \triangleq \sum_{d=1}^{D} \sum_{l=1}^{L_d} \log\left[\left(\epsilon_d(1+\varepsilon) + |\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}|\right) \times \sum_{i=1}^{L_d} \log\left(1 + \varepsilon + \frac{|\boldsymbol{\psi}_{d,i}^{\mathsf{T}} \boldsymbol{x}|}{\epsilon_d}\right)\right], \tag{53}$$

2) as $\varepsilon \to 0$ and $\epsilon_d \to 0 \quad \forall d$, MM applied to (2) under the $\ell_0 + \ell_{0,0}$ penalty

$$R_{0,00}^{D}(\boldsymbol{x}) \triangleq \|\boldsymbol{\Psi}\boldsymbol{x}\|_0 + \sum_{d=1}^{D} L_d \, 1_{\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_0 > 0}, \tag{54}$$

3) MM applied to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(\boldsymbol{x}|\boldsymbol{\lambda}; \boldsymbol{\epsilon}) \propto \prod_{d=1}^{D} \prod_{l=1}^{L_d} \frac{\lambda_d}{2\epsilon_d} \left(1 + \varepsilon + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}|}{\epsilon_d}\right)^{-(\lambda_d + 1)} \tag{55}$$

$$p(\boldsymbol{\lambda}) = \prod_{d=1}^{D} p(\lambda_d), \quad p(\lambda_d) \propto \begin{cases} \frac{1}{\lambda_d} & \lambda_d > 0 \\ 0 & \text{else} \end{cases}, \tag{56}$$

where, when $\varepsilon = 0$, the variables $\boldsymbol{z}_d = \boldsymbol{\Psi}_d \boldsymbol{x} \in \mathbb{R}^{L_d}$ are i.i.d. generalized-Pareto [44] given λ_d, and $p(\lambda_d)$ is Jeffrey's non-informative hyperprior [15], [42] for the random shape parameter λ_d.

4) variational EM under an AWGN likelihood and the prior

$$p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon}) \propto \prod_{d=1}^{D} \prod_{l=1}^{L_d} \frac{\lambda_d - 1}{2\epsilon_d} \left(1 + \varepsilon + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}|}{\epsilon_d}\right)^{-\lambda_d} \tag{57}$$

where, when $\varepsilon = 0$, the variables $\boldsymbol{z}_d = \boldsymbol{\Psi}_d \boldsymbol{x} \in \mathbb{R}^{L_d}$ are i.i.d. generalized-Pareto with deterministic shape parameter $\lambda_d > 1$ and scale parameter $\epsilon_d > 0$.

Proof: See Sections III-A to III-E below. ■

As with Co-L1, the MM interpretation implies convergence (in the sense of an asymptotic stationary point condition) when $\varepsilon > 0$, as detailed in Sec. III-B.

A. Log-Sum-Log MM Interpretation of Co-IRW-L1-$\boldsymbol{\epsilon}$

Consider the optimization problem

$$\arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + R_{\text{lsl}}(\boldsymbol{x}; \boldsymbol{\epsilon}, \varepsilon) \tag{58}$$

with R_{lsl} defined in (53). We attack this optimization problem using the MM approach detailed in Sec. II-A. The difference is that now the function g_2 is defined as

$$g_2(\boldsymbol{v}) = \sum_{d=1}^{D} \sum_{k \in \mathcal{K}_d} \log\left[(\epsilon_d(1+\varepsilon) + v_k) \sum_{i \in \mathcal{K}_d} \log\left(1 + \varepsilon + \frac{v_i}{\epsilon_d}\right)\right] \tag{59}$$

$$= \sum_{d=1}^{D} \left[L_d \log \sum_{i \in \mathcal{K}_d} \log\left(1 + \varepsilon + \frac{v_i}{\epsilon_d}\right) + \sum_{k \in \mathcal{K}_d} \log\left(\epsilon_d(1+\varepsilon) + v_k\right)\right], \tag{60}$$

which has a gradient of

$$[\nabla g_2(\boldsymbol{v}^{(t)})]_k \tag{61}$$

$$= \left(\frac{L_{d(k)}}{\sum_{i \in \mathcal{K}_{d(k)}} \log\left(1 + \varepsilon + \frac{v_i^{(t)}}{\epsilon_{d(k)}}\right)} + 1 \right) \frac{1}{\epsilon_{d(k)}(1+\varepsilon) + v_k^{(t)}} \tag{62}$$

when $d(k) \neq 0$ and otherwise $[\nabla g_2(\boldsymbol{v}^{(t)})]_k = 0$. Thus, recalling (17), MM prescribes

$$\boldsymbol{v}^{(t+1)} = \arg\min_{\boldsymbol{v}\in\mathcal{C}} \sum_{d=1}^{D} \sum_{k\in\mathcal{K}_d} \left(\frac{L_d}{\sum_{i\in\mathcal{K}_d} \log\left(1+\varepsilon+\frac{v_i^{(t)}}{\epsilon_d}\right)} + 1 \right) \times \left(\frac{v_k}{\epsilon_d(1+\varepsilon)+v_k^{(t)}} \right) + \gamma\|\boldsymbol{y} - [\boldsymbol{0} \quad \boldsymbol{\Phi}]\boldsymbol{v}\|_2^2, \quad (63)$$

or equivalently

$$\boldsymbol{x}^{(t+1)} = \arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} \sum_{l=1}^{L_d} \lambda_d^{(t+1)} \left(\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}^{(t)}|} \right) + \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 \quad (64)$$

for

$$\lambda_d^{(t+1)} = \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}^{(t)}|}{\epsilon_d}\right) \right]^{-1} + 1, \quad (65)$$

which coincides with Algorithm 3. This establishes Part 1 of Theorem 3.

B. Convergence of Co-IRW-L1-$\boldsymbol{\epsilon}$

The convergence of Co-IRW-L1-$\boldsymbol{\epsilon}$ (in the sense of an asymptotic stationary point condition) for $\varepsilon > 0$ can be shown using the same procedure as in Sec. II-B. To do this, we only need to verify that the gradient ∇g_2 in (61) is Lipschitz continuous when $\varepsilon > 0$, which we do in Appendix C.

C. Approximate $\ell_0 + \ell_{0,0}$ Interpretation of Co-IRW-L1-$\boldsymbol{\epsilon}$

Recalling the discussion in Sec. II-C, we now consider the behavior of the $R_{\text{lsl}}(\boldsymbol{x};\boldsymbol{\epsilon},\varepsilon)$ regularizer in (53) as $\varepsilon \to 0$ and $\epsilon_d \to 0 \quad \forall d$. For this, it helps to decouple (53) into two terms:

$$R_{\text{lsl}}(\boldsymbol{x};\boldsymbol{\epsilon},\varepsilon) = \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left(\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|\right) + \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left[\sum_{i=1}^{L_d} \log\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,i}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)\right]. \quad (66)$$

As $\epsilon_d \to 0 \quad \forall d$, the first term in (66) contributes an infinite valued "reward" for each pair (d,l) such that $|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}| = 0$, or a finite valued cost otherwise. As for the second term, we see that $\lim_{\varepsilon\to 0,\epsilon_d\to 0} \sum_{i=1}^{L_d} \log\left(1+\varepsilon+|\boldsymbol{\psi}_{d,i}^{\mathsf{T}}\boldsymbol{x}|/\epsilon_d\right) = 0$ if and only if $|\boldsymbol{\psi}_{d,i}^{\mathsf{T}}\boldsymbol{x}| = 0 \quad \forall i \in \{1,\dots,L_d\}$, i.e., if and only if $\|\boldsymbol{\Psi}_d\boldsymbol{x}\|_0 = 0$. And when $\|\boldsymbol{\Psi}_d\boldsymbol{x}\|_0 = 0$, the second term in (66) contributes L_d infinite valued rewards. In summary, as $\varepsilon \to 0$ and $\epsilon_d \to 0 \quad \forall d$, the first term in (66) behaves like $\|\boldsymbol{\Psi}\boldsymbol{x}\|_0$ and the second term like the weighted $\ell_{0,0}$ quasi-norm $\sum_{d=1}^{D} L_d 1_{\|\boldsymbol{\Psi}_d\boldsymbol{x}\|_0>0}$, as stated in (54). This establishes Part 2 of Theorem 3.

D. Bayesian MAP Interpretation of Co-IRW-L1-$\boldsymbol{\epsilon}$

To show that Co-IRW-L1-$\boldsymbol{\epsilon}$ can be interpreted as Bayesian MAP estimation under the hierarchical prior (55)–(56), we first compute the prior $p(\boldsymbol{x})$. To start,

$$p(\boldsymbol{x}) = \int_{\mathbb{R}^D} p(\boldsymbol{\lambda})p(\boldsymbol{x}|\boldsymbol{\lambda})\,\mathrm{d}\,\boldsymbol{\lambda} \quad (67)$$

$$\propto \prod_{d=1}^{D} \int_0^\infty \frac{1}{\lambda_d} \prod_{l=1}^{L_d} \frac{\lambda_d}{2\epsilon_d} \left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)^{-(\lambda_d+1)} \mathrm{d}\,\lambda_d. \quad (68)$$

Writing $(1+\varepsilon+|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|/\epsilon_d)^{-(\lambda_d+1)} = \exp(-(\lambda_d+1)Q_{d,l})$ for $Q_{d,l} \triangleq \log(1+\varepsilon+|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|/\epsilon_d)$, we get

$$p(\boldsymbol{x}) \propto \prod_{d=1}^{D} \frac{1}{(2\epsilon_d)^{L_d}} \int_0^\infty \lambda_d^{L_d-1} e^{-(\lambda_d+1)\sum_{l=1}^{L_d} Q_{d,l}}\,\mathrm{d}\,\lambda_d. \quad (69)$$

Defining $Q_d \triangleq \sum_{l=1}^{L_d} Q_{d,l}$ and changing the variable of integration to $\tau_d \triangleq \lambda_d Q_d$, we find

$$p(\boldsymbol{x}) \propto \prod_{d=1}^{D} \frac{e^{-Q_d}}{(2\epsilon_d Q_d)^{L_d}} \underbrace{\int_0^\infty \tau_d^{L_d-1} e^{-\tau_d}\,\mathrm{d}\,\tau_d}_{(L_d-1)!} \quad (70)$$

$$\propto \prod_{d=1}^{D} \left[\frac{1}{\epsilon_d \sum_{i=1}^{L_d} \log(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,i}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d})} \right]^{L_d} \times \prod_{l=1}^{L_d} \frac{1}{1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}} \quad (71)$$

$$= \prod_{d=1}^{D}\prod_{l=1}^{L_d} \left[\left(\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|\right) \times \sum_{i=1}^{L_d} \log\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,i}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right) \right]^{-1}, \quad (72)$$

which implies that

$$-\log p(\boldsymbol{x}) = \text{const} + R_{\text{lsl}}(\boldsymbol{x};\boldsymbol{\epsilon},\varepsilon) \quad (73)$$

for $R_{\text{lsl}}(\boldsymbol{x};\boldsymbol{\epsilon},\varepsilon)$ defined in (53).

Plugging (73) into (28), we see that

$$\boldsymbol{x}_{\mathsf{MAP}} = \arg\min_{\boldsymbol{x}} \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + R_{\text{lsl}}(\boldsymbol{x};\boldsymbol{\epsilon},\varepsilon), \quad (74)$$

which is equivalent to the optimization problem in (58). We showed in Sec. III-A that, by applying the MM algorithm to (58), we arrive at Algorithm 3. This establishes Part 3 of Theorem 3.

E. Variational EM Interpretation of Co-IRW-L1-$\boldsymbol{\epsilon}$

To justify the variational EM (VEM) interpretation of Co-IRW-L1-$\boldsymbol{\epsilon}$, we closely follow the approach used for Co-L1 in

Sec. II-E. The main difference is that now the prior takes the form of $p(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\epsilon})$ from (57). Thus, (42) becomes

$$\log p(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\epsilon}) = \sum_{d=1}^{D}\sum_{l=1}^{L_d}\left[\log\left(\frac{\lambda_d-1}{\epsilon_d}\right)-\lambda_d\log\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)\right] + \text{const} \tag{75}$$

and by zeroing the gradient w.r.t. $\boldsymbol{\lambda}$ we see that the M step (43) becomes

$$\frac{1}{\lambda_d^{(t+1)}-1} = \frac{1}{L_d}\log\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}_{\mathsf{MAP}}^{(t)}|}{\epsilon_d}\right),\quad d=1,\dots,D, \tag{76}$$

where again $\boldsymbol{x}_{\mathsf{MAP}}^{(t)}$ denotes the MAP estimate of $\boldsymbol{x}$ under $\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(t)}$. From (28) and (57), we see that

$$\boldsymbol{x}_{\mathsf{MAP}}^{(t)} = \arg\min_{\boldsymbol{x}}\sum_{d=1}^{D}\lambda_d^{(t)}\sum_{l=1}^{L_d}\log\left(|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|+\epsilon_d(1+\varepsilon)\right) + \gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2, \tag{77}$$

which (for $\varepsilon=0$) is a $\boldsymbol{\lambda}^{(t)}$-weighted version of the IRW-L1 log-sum optimization problem (recall Part 1 of Corollary 2). To solve (77), we apply MM. With a small modification of the MM derivation from Sec. II-A, we obtain the 2-step iteration

$$\boldsymbol{x}_{\mathsf{MAP}}^{(i)} = \arg\min_{\boldsymbol{x}}\gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2+\sum_{d=1}^{D}\lambda_d^{(t)}\|\boldsymbol{W}_d^{(i)}\boldsymbol{\Psi}_d\boldsymbol{x}\|_1 \tag{78}$$

$$\boldsymbol{W}_d^{(i+1)} = \text{diag}\left\{\frac{1}{\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,1}^{\mathsf{T}}\boldsymbol{x}^{(i)}|},\cdots,\frac{1}{\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}}\boldsymbol{x}^{(i)}|}\right\}. \tag{79}$$

By using only a single MM iteration per VEM iteration, the MM index "i" can be rewritten as the VEM index "t," in which case the VEM algorithm becomes

$$\boldsymbol{x}^{(t)} = \arg\min_{\boldsymbol{x}}\gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2+\sum_{d=1}^{D}\lambda_d^{(t)}\|\boldsymbol{W}_d^{(t)}\boldsymbol{\Psi}_d\boldsymbol{x}\|_1 \tag{80}$$

$$\boldsymbol{W}_d^{(t+1)} = \text{diag}\left\{\frac{1}{\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,1}^{\mathsf{T}}\boldsymbol{x}^{(t)}|},\dots,\frac{1}{\epsilon_d(1+\varepsilon)+|\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}}\boldsymbol{x}^{(t)}|}\right\},\forall d \tag{81}$$

$$\lambda_d^{(t+1)} = \left[\frac{1}{L_d}\log\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}^{(t)}|}{\epsilon_d}\right)\right]^{-1}+1,\quad \forall d, \tag{82}$$

which matches the steps in Algorithm 3. This establishes Part 4 of Theorem 3.

F. Co-IRW-L1

Until now, we have considered the Co-IRW-L1-$\boldsymbol{\epsilon}$ parameters $\boldsymbol{\epsilon}=[\epsilon_1,\dots,\epsilon_D]^{\mathsf{T}}$ to be fixed and known. But it is not clear how to set these parameters in practice. Thus, in this section, we describe an extension of Co-IRW-L1-$\boldsymbol{\epsilon}$ that adapts the $\boldsymbol{\epsilon}$ vector at every iteration. The resulting procedure, which we will refer to as Co-IRW-L1, is summarized in Algorithm 4.

Algorithm 4. The Co-IRW-L1 Algorithm

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma>0$, $\varepsilon\geq 0$
2: if $\boldsymbol{\Psi}\boldsymbol{x}\in\mathbb{R}^L$, use $\Lambda=(1,\infty)$ and $\log p(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\epsilon})$ from (75);
if $\boldsymbol{\Psi}\boldsymbol{x}\in\mathbb{C}^L$, use $\Lambda=(2,\infty)$ and $\log p(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\epsilon})$ from (84).
3: initialization: $\lambda_d^{(1)}=1\ \forall d$, $\boldsymbol{W}_d^{(1)}=\boldsymbol{I}\ \forall d$
4: for $t=1,2,3,\dots$
5: $\boldsymbol{x}^{(t)}\leftarrow\arg\min_{\boldsymbol{x}}\gamma\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2+\sum_{d=1}^{D}\lambda_d^{(t)}\|\boldsymbol{W}_d^{(t)}\boldsymbol{\Psi}_d\boldsymbol{x}\|_1$
6: $(\lambda_d^{(t+1)},\epsilon_d^{(t+1)})\leftarrow\arg\max_{\lambda_d\in\Lambda,\epsilon_d>0}\log p(\boldsymbol{x}^{(t)};\boldsymbol{\lambda},\boldsymbol{\epsilon}),\ d=1,\dots,D$
7: $\boldsymbol{W}_d^{(t+1)}\leftarrow\text{diag}\left\{\frac{1}{\epsilon_d^{(t+1)}(1+\varepsilon)+|\boldsymbol{\psi}_{d,1}^{\mathsf{T}}\boldsymbol{x}^{(t)}|},\cdots,\frac{1}{\epsilon_d^{(t+1)}(1+\varepsilon)+|\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}}\boldsymbol{x}^{(t)}|}\right\},\quad\forall d$
8: end
9: output: $\boldsymbol{x}^{(t)}$

Although there does not appear to be a closed-form solution to the joint maximization problem in line 6 of Algorithm 4, it is over two real parameters and thus can be solved numerically without a significant computational burden.

Algorithm 4 can be interpreted as a generalization of the VEM approach to Co-IRW-L1-$\boldsymbol{\epsilon}$ that is summarized in Part 4 of Theorem 3 and detailed in Sec. III-E. Whereas Co-IRW-L1-$\boldsymbol{\epsilon}$ used VEM to estimate the $\boldsymbol{\lambda}$ parameters in the prior (57) for a fixed value of $\boldsymbol{\epsilon}$, Co-IRW-L1 uses VEM to *jointly* estimate $(\boldsymbol{\lambda},\boldsymbol{\epsilon})$ in (57). Thus, Co-IRW-L1 can be derived by repeating the steps in Sec. III-E, except that now the maximization of $\log p(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\epsilon})$ in (75) is performed jointly over $(\boldsymbol{\lambda},\boldsymbol{\epsilon})$, as reflected by line 6 of Algorithm 4.

G. Co-IRW-L1 for Complex-Valued $\boldsymbol{\Psi}_d\boldsymbol{x}$

In Sections III-A-III-F, the analysis outputs $\boldsymbol{\Psi}_d\boldsymbol{x}$ were assumed to be real-valued. We now extend the previous results to the case of complex-valued $\boldsymbol{\Psi}_d\boldsymbol{x}$. For this, we focus on the Co-IRW-L1 algorithm, since Co-IRW-L1-$\boldsymbol{\epsilon}$ follows as the special case where $\boldsymbol{\epsilon}$ is fixed at a user-supplied value.

Recalling that Co-IRW-L1 was constructed by generalizing the VEM interpretation of Co-IRW-L1-$\boldsymbol{\epsilon}$, we reconsider this VEM interpretation for the case of complex-valued $\boldsymbol{\Psi}_d\boldsymbol{x}$. In particular, we assume an AWGN likelihood and the following complex-valued extension of the prior (57):

$$p(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\epsilon})\propto\prod_{d=1}^{D}\prod_{l=1}^{L_d}\frac{(\lambda_d-1)(\lambda_d-2)}{2\pi\epsilon_d^2}\left(1+\varepsilon+\frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)^{-\lambda_d} \tag{83}$$

which (for $\varepsilon = 0$) is i.i.d. generalized-Pareto on $\boldsymbol{z}_d = \boldsymbol{\Psi}_d \boldsymbol{x} \in \mathbb{C}^{L_d}$ with deterministic shape parameter $\lambda_d > 2$ and deterministic scale parameter $\epsilon_d > 0$. In this case, the log-prior (75) changes to

$$\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon}) = \text{const} + \sum_{d=1}^{D} \sum_{l=1}^{L_d} \left[\log \left(\frac{(\lambda_d - 1)(\lambda_d - 2)}{\epsilon_d^2} \right) - \lambda_d \log \left(1 + \varepsilon + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}|}{\epsilon_d} \right) \right] \tag{84}$$

which is then maximized over $(\boldsymbol{\lambda}, \boldsymbol{\epsilon})$ in line 6 of Algorithm 4.

IV. Numerical Results

We now present results from a numerical study into the performance of the proposed Co-L1 and Co-IRW-L1 methods, given as Algorithm 1 and Algorithm 4, respectively. Three experiments are discussed below, all of which focus on the problem of recovering an N-pixel image (or image sequence) $\boldsymbol{x}$ from M-sample noisy compressed measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{w}$, with $M \ll N$. In the first experiment, we recover synthetic 2D finite-difference signals; in the second experiment, we recover the Shepp-Logan phantom and the Cameraman image; and in the third experiment, we recover dynamic MRI sequences, also known as "cines."

As discussed in Sec. I-D, Co-L1 can be considered as the composite extension of the standard L1-regularized approach to analysis CS, i.e., (2) under the non-composite L1 regularizer $R(\boldsymbol{x}) = \|\boldsymbol{\Psi}\boldsymbol{x}\|_1$. Similarly, Co-IRW-L1 can be considered as the composite extension of the standard IRW approach to the same problem. Thus, we compare our proposed composite methods against these two non-composite methods, referring to them simply as "L1" and "IRW-L1" in the sequel.

A. Experimental Setup

For the dynamic MRI experiment, we constructed $\boldsymbol{\Phi}$ using randomly sub-sampled Fourier measurements at each time instant with a varying sampling pattern across time. More details are given in Sec. IV-D. For the other experiments, we used a "spread spectrum" operator [45] of the form $\boldsymbol{\Phi} = \boldsymbol{D}\boldsymbol{F}\boldsymbol{C}$, where $\boldsymbol{C} \in \mathbb{R}^{N \times N}$ is diagonal matrix with i.i.d equiprobable ± 1 entries, $\boldsymbol{F} \in \mathbb{C}^{N \times N}$ is the discrete Fourier transform (DFT), and $\boldsymbol{d} \in \mathbb{R}^{M \times N}$ is a row-selection operator that selects M rows of $\boldsymbol{F}\boldsymbol{C} \in \mathbb{C}^{N \times N}$ uniformly at random.

In all cases, the noise $\boldsymbol{w}$ was zero-mean, white, and circular Gaussian (i.e., independent real and imaginary components of equal variance). Denoting the noise variance by σ^2, we define the measurement signal-to-noise ratio (SNR) as $\|\boldsymbol{y}\|_2^2/(M\sigma^2)$ and the recovery SNR of signal estimate $\hat{\boldsymbol{x}}$ as $\|\boldsymbol{x}\|_2^2/\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$.

Note that, when $\boldsymbol{x}$ is real-valued, the measurements $\boldsymbol{y}$ will be complex-valued due to the construction of $\boldsymbol{\Phi}$. Thus, to allow the use of real-valued L1 solvers, we split each complex-valued element of $\boldsymbol{y}$ (and the corresponding rows of $\boldsymbol{\Phi}$ and $\boldsymbol{w}$) into real and imaginary components, resulting in a real-only model. However, to avoid possible redundancy issues caused by the conjugate symmetry of the noiseless Fourier measurements $\boldsymbol{F}\boldsymbol{C}\boldsymbol{x}$, we ensured that $\boldsymbol{D}$ selected at most one sample from each complex-conjugate pair.

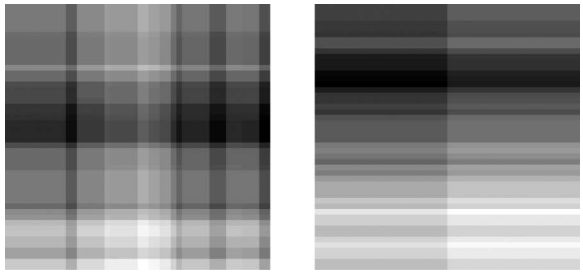

Fig. 1. Examples of the 2D finite-difference signal $\boldsymbol{X}$ used in the first experiment. On the left is a realization generated under a transition ratio of $\alpha = 14/14 = 1$, and on the right is a realization generated under $\alpha = 27/1 = 27$.

We used MFISTA [33] to implement the L2+L1 optimization needed for all methods. The maximum number of outer, reweighting iterations for Co-L1 and Co-IRW-L1 was set to 16, while the maximum number of inner MFISTA iterations was set at 60, with early termination if $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2/\|\boldsymbol{x}^{(t)}\|_2 < 1 \times 10^{-6}$. In all experiments, we used $\gamma = 1/\sigma^2$ (as motivated before (6)) and $\epsilon = 0 = \varepsilon$.

B. Synthetic 2D Finite-Difference Signals

Our first experiment aims to answer the following question. If we know that the sparsity of $\boldsymbol{\Psi}_1\boldsymbol{x}$ differs from the sparsity of $\boldsymbol{\Psi}_2\boldsymbol{x}$, then can we exploit this knowledge for signal recovery, even if we don't know *how* the sparsities are different? This is precisely the goal of composite regularizations like (4).

To investigate this question, we constructed 2D signals with finite-difference structure in both the vertical and horizontal domains. In particular, we constructed $\boldsymbol{x} = \boldsymbol{x}_1\boldsymbol{1}^{\mathsf{T}} + \boldsymbol{1}\boldsymbol{x}_2^{\mathsf{T}}$, where both $\boldsymbol{x}_1 \in \mathbb{R}^{48}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{48}$ are finite-difference signals and $\boldsymbol{1} \in \mathbb{R}^{48}$ contains only ones. The locations of the transitions in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ were selected uniformly at random and the amplitudes of the transitions were drawn i.i.d. zero-mean Gaussian. The total number of transitions in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ was fixed at 28, but the ratio of the number of transitions in $\boldsymbol{x}_1$ to the number in $\boldsymbol{x}_2$, denoted by α, was varied from 1 to 27. The case $\alpha = 1$ corresponds to $\boldsymbol{x}$ having 14 vertical transitions and 14 horizontal transitions, while the case $\alpha = 27$ corresponds to $\boldsymbol{x}$ having 27 vertical transitions and a single horizontal transition. (See Fig. 1 for examples.) Finally, the signal $\boldsymbol{x} \in \mathbb{R}^N$ appearing in our model (1) was created by vectorizing $\boldsymbol{x}$, yielding a total of $N = 48^2 = 2304$ pixels.

Given $\boldsymbol{x}$, noisy observations $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{w}$ were generated using the random "spread spectrum" measurement operator $\boldsymbol{\Phi}$ described earlier at a sampling ratio of $M/N = 0.25$, with additive white Gaussian noise (AWGN) $\boldsymbol{w}$ scaled to achieve a measurement SNR of 40 dB. All recovery algorithms used vertical and horizontal finite-difference operators $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, respectively, with $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1^{\mathsf{T}}, \boldsymbol{\Psi}_2^{\mathsf{T}}]^{\mathsf{T}}$ in the non-composite case.

Figure 2 shows recovery SNR versus α for the non-composite L1 and IRW-L1 techniques and our proposed Co-L1 and Co-IRW-L1 techniques. Each SNR in the figure represents the median value from 25 trials, each using an independent realization of the triple $(\boldsymbol{\Phi}, \boldsymbol{x}, \boldsymbol{w})$. The figure shows that the

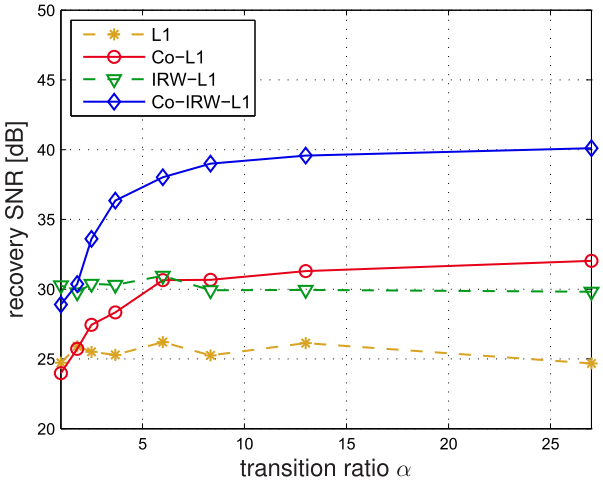

Fig. 2. Recovery SNR versus transition ratio α for the first experiment, which used 2D finite-difference signals, spread-spectrum measurements at $M/N = 0.25$, AWGN at 40 dB, and finite-difference operators for $\boldsymbol{\Psi}_d$. Each recovery SNR represents the median value from 25 independent trials.

Fig. 3. Left: the real-valued cropped Cameraman image of size $N = 96 \times 104$. Right: the complex-valued Shepp-Logan phantom of size $N = 96 \times 96$. For the Shepp-Logan phantom, the real and imaginary parts of $\boldsymbol{x}$ were identical, and only the real part is shown here.

recovery SNR of both L1 and IRW-L1 is roughly invariant to the transition ratio α, which makes sense because the overall sparsity of $\boldsymbol{\Psi x}$ is fixed at 28 transitions by construction. In contrast, the recovery SNRs of Co-L1 and Co-IRW-L1 vary with α, with higher values of α yielding a more structured signal and thus higher recovery SNR when this structure is properly exploited.

C. Cameraman and Shepp-Logan Recovery

For our second experiment, we investigate algorithm performance versus sampling ratio M/N when recovering the well-known Shepp-Logan phantom and Cameraman images. In particular, we used the $N = 96 \times 104$ cropped real-valued Cameraman image and the $N = 96 \times 96$ complex-valued Shepp-Logan phantom shown in Fig. 3, and we constructed compressed noisy measurements $\boldsymbol{y}$ using spread-spectrum $\boldsymbol{\Phi}$ and AWGN $\boldsymbol{w}$ at a measurement SNR of 30 dB in the Cameraman case and 40 dB in the Shepp-Logan case.

For the Cameraman image, we constructed the analysis operator $\boldsymbol{\Psi} \in \mathbb{R}^{8N \times N}$ by concatenating undecimated db1 and db2 2D wavelet transforms (UWT-db1-db2) with one level of decomposition. For the Shepp-Logan phantom image, we constructed the analysis operator $\boldsymbol{\Psi} \in \mathbb{R}^{4N \times N}$ from the undecimated db1 2D wavelet transform (UWT-db1) with one level of decomposition. The Co-L1 and Co-IRW-L1 algorithms treated each of the sub-bands of the wavelet transform as a separate sub-dictionary $\boldsymbol{\Psi}_d$ in their composite regularizers.

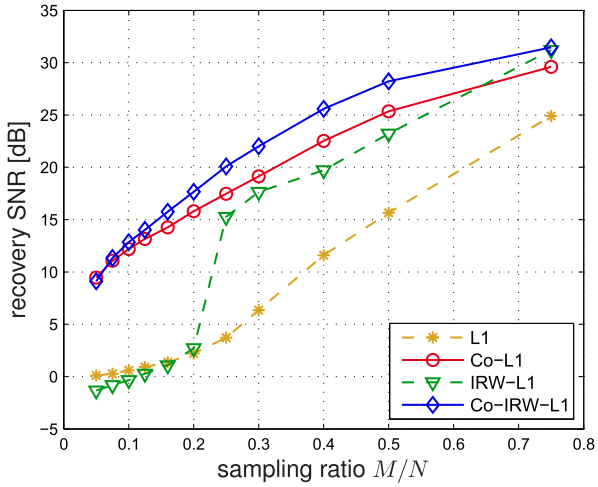

Fig. 4. Recovery SNR versus sampling ratio M/N for the cropped Cameraman image. Measurements were constructed using a spread-spectrum operator and AWGN at 30 dB SNR, and recovery used UWT-db1-db2 at one level of decomposition. Each SNR value represents the median value from 7 independent trials.

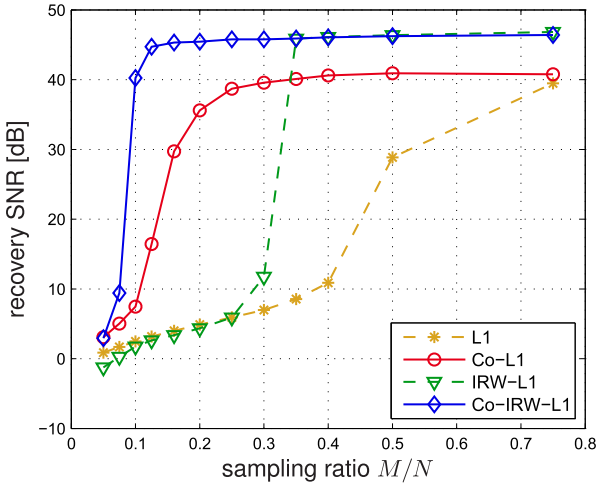

Fig. 5. Recovery SNR versus sampling ratio M/N for the Shepp-Logan phantom. Measurements were constructed using a spread-spectrum operator and AWGN at 40 dB SNR, and recovery used UWT-db1 at one level of decomposition. Each recovery SNR represents the median value from 7 independent trials.

Fig. 4 shows recovery SNR versus sampling ratio M/N for the Cameraman image, while Fig. 5 shows the same for the Shepp-Logan phantom. Each recovery SNR represents the median value from 7 independent realizations of $(\boldsymbol{\Phi}, \boldsymbol{w})$. Both figures show that Co-L1 and Co-IRW-L1 outperform their non-composite counterparts, especially at low sampling ratios; the gap between Co-IRW-L1 and and IRW-L1 closes at $M/N \geq 0.35$ for the Shepp-Logan phantom.

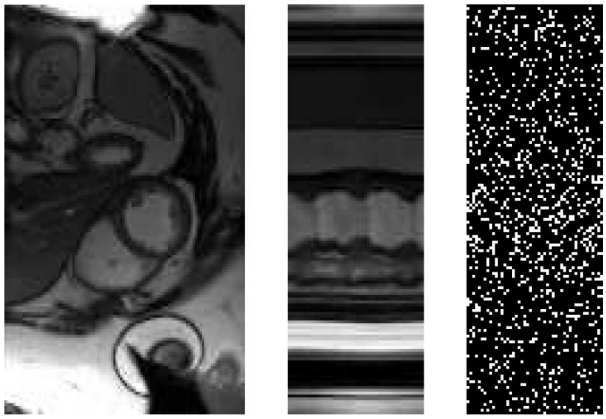

Fig. 6. Left: A 144×85 spatial slice from the $144 \times 85 \times 48$ dMRI dataset. Middle: The 144×48 spatio-temporal slice used for the dMRI experiment. Right: a realization of the variable-density k-space sampling pattern, versus time, at $M/N = 0.30$.

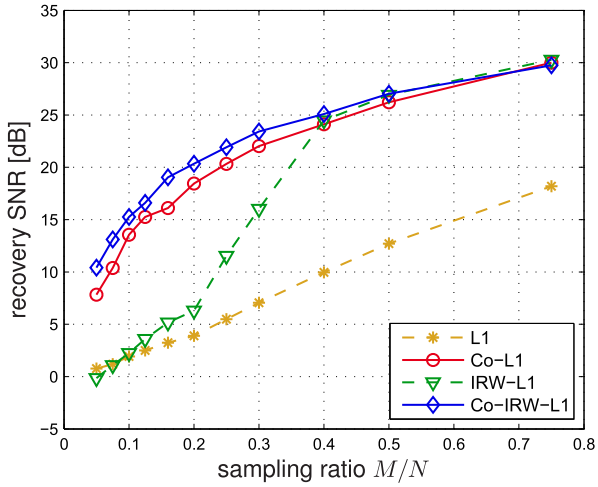

Fig. 7. Recovery SNR versus sampling ratio M/N for the dMRI experiment. Each SNR value represents the median value from 7 independent trials. Measurements were constructed using variable-density sub-sampled Fourier operator and AWGN at 30 dB measurement SNR, and recovery used a concatenation of db1-db3 orthogonal 2D wavelet bases at two levels of decomposition.

D. Dynamic MRI

For our third experiment, we investigate a simplified version of the "dynamic MRI" (dMRI) problem. In dMRI, one attempts to recover a sequence of MRI images, known as an MRI cine, from highly under-sampled "k-t-domain" measurements $\{\boldsymbol{y}_t\}_{t=1}^T$ constructed as

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_t \boldsymbol{x}_t + \boldsymbol{w}_t, \tag{85}$$

where $\boldsymbol{x}_t \in \mathbb{R}^{N_1 N_2}$ is a vectorized $(N_1 \times N_2)$-pixel image at time t, $\boldsymbol{\Phi}_t \in \mathbb{R}^{M_1 \times N_1 N_2}$ is a sub-sampled Fourier operator at time t, and $\boldsymbol{w}_t \in \mathbb{R}^{M_1}$ is AWGN. This real-valued $\boldsymbol{\Phi}_t$ is constructed from the complex-valued $N_1 N_2 \times N_1 N_2$ 2D DFT matrix by randomly selecting $0.5M_1$ rows and then splitting each of those rows into its real and imaginary components. Here, it is usually advantageous to vary the sampling pattern with time and to sample more densely at low frequencies, where most of the signal energy lies (e.g., [46]). Putting (85) into the form of our measurement model (1), we get

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{\Phi}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Phi}_T \end{bmatrix}}_{\boldsymbol{\Phi}} \underbrace{\begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_T \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_T \end{bmatrix}}_{\boldsymbol{w}}, \tag{86}$$

with total measurement dimension $M = M_1 T$ and total signal dimension $N = N_1 N_2 T$.

As ground truth, we used a high-quality dMRI cardiac cine $\boldsymbol{x}$ of dimensions $N_1 = 144$, $N_2 = 85$, and $T = 48$. The left pane in Fig. 6 shows a 144×85 image from this cine extracted at a single time t, while the middle pane shows a 144×48 spatio-temporal profile from this cine extracted at a single horizontal location. This middle pane shows that the temporal dimension is much more structured than the spatial dimension, suggesting that there may be an advantage to weighting the spatial and temporal dimensions differently in a composite regularizer.

To test this hypothesis, we constructed an experiment where the goal was to recover the 144×48 spatio-temporal profile shown in the middle pane of Fig. 6, as opposed to the full 3D cine, from subsampled k-t-domain measurements. For this purpose, we constructed measurements $\{\boldsymbol{y}\}_{t=1}^T$ as described above, but with $N_2 = 1$ (and thus a 1D DFT), and used a variable density random sampling method. The right pane of Fig. 6 shows a typical realization of the sampling pattern versus time. Finally, we selected the AWGN variance that yielded measurement SNR $= 30$ dB.

For the non-composite L1 and IRW-L1 algorithms, we constructed the analysis operator $\boldsymbol{\Psi} \in \mathbb{R}^{3N \times N}$ from a vertical concatenation of the db1-db3 orthogonal 2D discrete wavelet bases, each with two levels of decomposition. For the Co-L1 and Co-IRW-L1 algorithms, we assigned each of the 21 sub-bands in $\boldsymbol{\Psi}$ to a separate sub-dictionary $\boldsymbol{\Psi}_d \in \mathbb{R}^{L_d \times N}$. Note that the sub-dictionary size L_d decreases with the level in the decomposition. By weighting certain sub-dictionaries differently than others, the composite regularizers can exploit differences in spatial versus temporal structure.

Fig. 7 shows recovery SNR versus sampling ratio M/N for the four algorithms under test. Each reported SNR represents the median SNR from 7 independent realizations of $(\boldsymbol{\Phi}, \boldsymbol{w})$. The figure shows that Co-L1 outperforms its non-composite counterparts at all tested values of M/N, while Co-IRW-L1 outperforms its noncomposite counterpart for $M/N \leq 0.4$. Although not shown here, we obtained similar results with other cine datasets and with an UWT-db1-based analysis operator.

For qualitative comparison, Fig. 8 shows the spatio-temporal profile recovered by each of the four algorithms under test at $M/N = 0.3$ for a typical realization of $(\boldsymbol{\Phi}, \boldsymbol{w})$. Compared to the ground-truth profile shown in the middle pane of Fig. 6, the profiles recovered by L1 and IRW-L1 show visible artifacts that appear as vertical streaks. In contrast, the profiles recovered by Co-L1 and Co-IRW-L1 preserve most of the features present in the ground-truth profile.

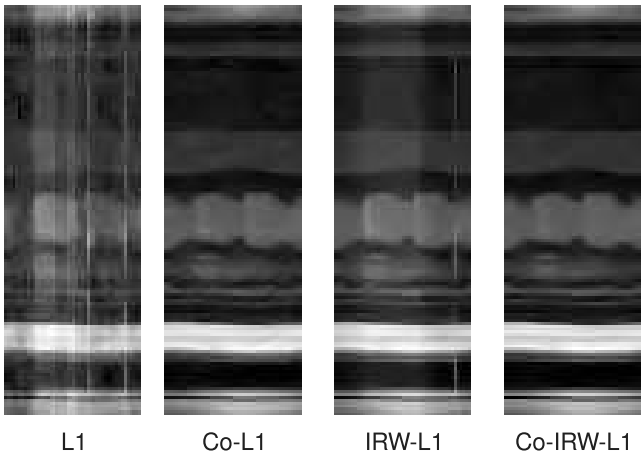

Fig. 8. Recovered dMRI spatio-temporal profiles at $M/N = 0.30$.

TABLE I
COMPUTATION TIMES (IN SECONDS) FOR THE PRESENTED EXPERIMENTAL STUDIES. THE TIMES ARE AVERAGED OVER TRIAL RUNS AND DIFFERENT SAMPLING RATIOS

	Shepp-Logan	Cameraman	MRI
L1	8.12	9.88	22.0
Co-L1	8.83	12.8	21.7
IRW-L1	7.95	12.7	24.1
Co-IRW-L1	9.29	16.9	29.6

E. Algorithm Runtime

Table I reports the average runtimes of the L1, Co-L1, IRW-L1, and Co-IRW-L1 algorithms for the experiments in Sections IV-C and IV-D. There we see that the runtime of Co-L1 was $1.29\times$ that of L1 for the worst case, and the runtime of Co-IRW-L1 was $1.33\times$ that of IRW-L1 for the worst case.

F. Choice of Dictionary

In our last experiment, we investigate the performance of Co-IRW-L1 versus choice of $\{\boldsymbol{\Psi}_d\}$. For this, we constructed $\{\boldsymbol{\Psi}_d\}$ using a concatenation of either undecimated or orthogonal 2D Daubechies wavelet transforms, and we varied both the number of transforms in the concatenation as well as the number of levels in the wavelet decomposition. We then attempted to recover the Cameraman image from spread-spectrum measurements at $M/N = 0.4$ in AWGN at 30 dB SNR. As usual, the Co-IRW-L1 algorithm treated each wavelet sub-band as a separate sub-dictionary.

The recovery SNR for various choices of $\boldsymbol{\Psi}$ is shown in Fig. 9. For the case of orthogonal wavelet transforms (OWT), a significant performance improvement was observed in going from one to two transforms, regardless of the wavelet decomposition level. However, a slight performance degradation was observed when concatenating more than two OTWs. Moreover, the effect of varying the level of decomposition was mild unless no concatenation (i.e., db1) was used. For the undecimated wavelet transform (UWT) case, the recovery SNR was essentially invariant to both the level of decomposition and the number of concatenated transforms, with only a slight degradation when five transforms were concatenated. Overall, the UWT performed significantly better than the OWT. Similar trends were observed for the Co-L1 algorithm in experiments not shown here.

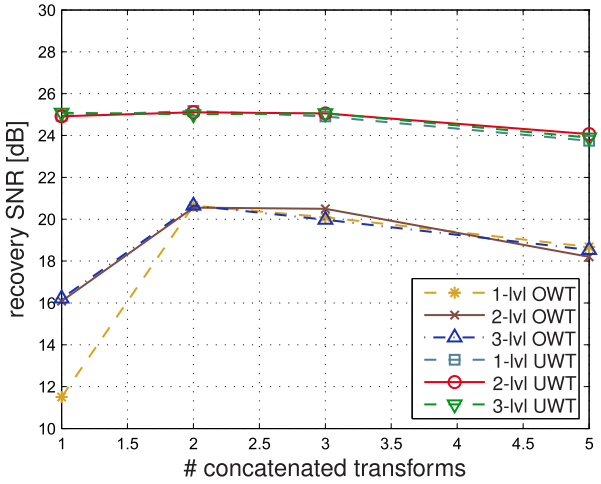

Fig. 9. Co-IRW-L1 recovery SNR for different choices of $\boldsymbol{\Psi}_d$. Measurements were constructed from the cropped cameraman image using a spread-spectrum operator, AWGN at 30 dB SNR, and sampling ratio $M/N = 0.40$. Here, OWT represents a concatenation of 2D orthogonal Daubechies wavelet transforms, UWT represents a concatenation of 2D undecimated Daubechies wavelet transforms, and "lvl" denotes the level of decomposition. Each SNR value represents the median value from 3 independent trials.

V. CONCLUSIONS

Motivated by the observation that a given signal $\boldsymbol{x}$ admits sparse representations in multiple dictionaries $\boldsymbol{\Psi}_d$ but with varying levels of sparsity across dictionaries, we proposed two new algorithms for the reconstruction of (approximately) sparse signals from noisy linear measurements. Our first algorithm, Co-L1, extends the well-known lasso algorithm [3], [4], [6] from the L1 penalty $\|\boldsymbol{\Psi}\boldsymbol{x}\|_1$ to composite L1 penalties of the form (4) while self-adjusting the regularization weights λ_d. Our second algorithm, Co-IRW-L1, extends the well-known IRW-L1 algorithm [9], [12], [13] to the same family of composite penalties while self-adjusting the regularization weights λ_d and the regularization parameters ϵ_d.

We provided several interpretations of both algorithms: i) majorization-minimization (MM) applied to a non-convex log-sum-type penalty, ii) MM applied to an approximate ℓ_0-type penalty, iii) MM applied to Bayesian MAP inference under a particular hierarchical prior, and iv) variational expectation-maximization (VEM) under a particular prior with deterministic unknown parameters. Also, we leveraged the MM interpretation to establish convergence in the form of an asymptotic stationary point condition [19]. Furthermore, we noted that the Bayesian MAP and VEM viewpoints yield novel interpretations of the original IRW-L1 algorithm. Finally, we present a detailed numerical study that suggests that our proposed algorithms yield significantly improved recovery SNR when compared to their non-composite L1 and IRW-L1 counterparts with a modest (e.g., $1.3\times$) increase in runtime.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable feedback.

APPENDIX A
LIPSCHITZ CONTINUITY OF CO-L1 GRADIENT

In this appendix, we establish the Lipschitz continuity of ∇g_2 from (18) in the case that $\epsilon > 0$. We first recall that, for ∇g_2 to be Lipschitz continuous over the domain $\boldsymbol{v} \in \mathcal{C}$, there must exist some constant β such that, for all $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{C}$,

$$\|\nabla g_2(\boldsymbol{v}) - \nabla g_2(\boldsymbol{v}')\|_2^2 \leq \beta \|\boldsymbol{v} - \boldsymbol{v}'\|_2^2 \tag{87}$$

From (18), we have

$$\|\nabla g_2(\boldsymbol{v}) - \nabla g_2(\boldsymbol{v}')\|_2^2$$
$$= \sum_{k=1}^{L} \left(\frac{L_{d(k)}}{\epsilon + \sum_{i \in \mathcal{K}_{d(k)}} v_i} - \frac{L_{d(k)}}{\epsilon + \sum_{i \in \mathcal{K}_{d(k)}} v'_i} \right)^2 \tag{88}$$
$$= \sum_{k=1}^{L} \frac{L_{d(k)}^2 \left[\sum_{i \in \mathcal{K}_{d(k)}} (v'_i - v_i) \right]^2}{\left(\epsilon + \sum_{i \in \mathcal{K}_{d(k)}} v_i \right)^2 \left(\epsilon + \sum_{i \in \mathcal{K}_{d(k)}} v'_i \right)^2} \tag{89}$$
$$= \sum_{d=1}^{D} \sum_{l=1}^{L_d} \frac{L_d^2 \left[\sum_{i=1}^{L_d} (u'_{d,i} - u_{d,i}) \right]^2}{\left(\epsilon + \sum_{i \in \mathcal{K}_{d(k)}} v_i \right)^2 \left(\epsilon + \sum_{i \in \mathcal{K}_{d(k)}} v'_i \right)^2}. \tag{90}$$

We can then upper bound the latter as follows.

$$\|\nabla g_2(\boldsymbol{v}) - \nabla g_2(\boldsymbol{v}')\|_2^2 \leq \sum_{d=1}^{D} \sum_{l=1}^{L_d} \frac{L_d^2}{\epsilon^4} \left[\sum_{i=1}^{L_d} (u'_{d,i} - u_{d,i}) \right]^2 \tag{91}$$
$$\leq \sum_{d=1}^{D} \frac{L_d^3}{\epsilon^4} \left[\sum_{i=1}^{L_d} |u'_{d,i} - u_{d,i}| \right]^2 \tag{92}$$
$$\leq \sum_{d=1}^{D} \frac{L_d^4}{\epsilon^4} \sum_{i=1}^{L_d} (u'_{d,i} - u_{d,i})^2 \tag{93}$$
$$\leq \frac{L_{\max}^4}{\epsilon^4} \sum_{k=1}^{L} (v'_k - v_k)^2 \tag{94}$$
$$\leq \frac{L_{\max}^4}{\epsilon^4} \sum_{k=1}^{L+N} (v'_k - v_k)^2 \tag{95}$$
$$= \frac{L_{\max}^4}{\epsilon^4} \|\boldsymbol{v} - \boldsymbol{v}'\|_2^2, \tag{96}$$

where (91) follows from the fact that $u_{d,l} \geq 0 \quad \forall d, l$ (according to (13)), (93) follows from the fact that $\|\boldsymbol{x}\|_1 \leq \sqrt{N}\|\boldsymbol{x}\|_2$ for $\boldsymbol{x} \in \mathbb{C}^N$, and (94) uses $L_{\max} \triangleq \max_d L_d$. Comparing (96) to (87), we see that ∇g_2 from (18) is Lipschitz continuous.

APPENDIX B
EQUIVALENCE OF LOG-SUM AND ℓ_0 MINIMIZATION

In this appendix, we establish that the log-sum optimization (24) becomes equivalent to the ℓ_0 optimization (25) as $\epsilon \to 0$. We first note that, for any $\epsilon > 0$,

$$\frac{1}{\log(1/\epsilon)} \sum_{n=1}^{N} \log(\epsilon + |x_n|) \tag{97}$$
$$= \frac{1}{\log(1/\epsilon)} \left[\sum_{n:\, x_n = 0} \log(\epsilon) + \sum_{n:\, x_n \neq 0} \log(\epsilon + |x_n|) \right] \tag{98}$$
$$= \|\boldsymbol{x}\|_0 - N + \frac{\sum_{n:\, x_n \neq 0} \log(\epsilon + |x_n|)}{\log(1/\epsilon)}, \tag{99}$$

where $\|\boldsymbol{x}\|_0$ is defined as the counting norm, i.e., $\|\boldsymbol{x}\|_0 \triangleq |\{x_n : x_n \neq 0\}|$. Applying this result to the objective function in (24), we have

$$\gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{n=1}^{N} \log(\epsilon + |x_n|)$$
$$\propto \underbrace{\frac{\gamma}{\log(1/\epsilon)}}_{\triangleq\, \gamma'} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \|\boldsymbol{x}\|_0 - N + \frac{\sum_{n:\, x_n \neq 0} \log(\epsilon + |x_n|)}{\log(1/\epsilon)}. \tag{100}$$

Clearly the global scaling and offset by N in (100) are inconsequential to the minimization in (24). Furthermore, by making $\epsilon > 0$ arbitrarily small, we can make the last term in (100) arbitrarily small[4] and thus negligible compared to the other terms. It is in this sense that we say that (24) is equivalent to (25) as $\epsilon \to 0$.

APPENDIX C
LIPSCHITZ CONTINUITY OF CO-IRW-L1-$\boldsymbol{\epsilon}$ GRADIENT

In this appendix, we establish the Lipschitz continuity of ∇g_2 from (61) in the case that $\varepsilon > 0$, recalling the Lipschitz definition (87). To ease the exposition, we focus on the $L = 1$ case, noting that a similar (but more tedious) technique can be applied to the general case.

From the $L = 1$ case of (61), we have

$$|\nabla g_2(v) - \nabla g_2(v')|^2$$
$$= \left[\left(\frac{1}{\log(1 + \varepsilon + \frac{v}{\epsilon_1})} + 1 \right) \frac{1}{\epsilon_1(1+\varepsilon) + v} - \left(\frac{1}{\log(1 + \varepsilon + \frac{v'}{\epsilon_1})} + 1 \right) \frac{1}{\epsilon_1(1+\varepsilon) + v'} \right]^2 \tag{101}$$
$$= [A + B]^2 \tag{102}$$
$$\leq [|A| + |B|]^2 \leq 2\left[A^2 + B^2\right], \tag{103}$$

since $\|\boldsymbol{x}\|_1 \leq \sqrt{N}\|\boldsymbol{x}\|_2$ for $\boldsymbol{x} \in \mathbb{C}^N$, and where

[4]Note that, as $\epsilon \to 0$, the numerator of the last term in (100) converges to the finite value $\sum_{n:\, x_n \neq 0} \log(|x_n|)$ while the denominator grows to $+\infty$.

$$A \triangleq \frac{1}{\epsilon_1(1+\varepsilon)+v} - \frac{1}{\epsilon_1(1+\varepsilon)+v'} \tag{104}$$

$$B \triangleq \frac{1}{(\epsilon_1(1+\varepsilon)+v)\log(1+\varepsilon+\frac{v}{\epsilon_1})} - \frac{1}{(\epsilon_1(1+\varepsilon)+v')\log(1+\varepsilon+\frac{v'}{\epsilon_1})}. \tag{105}$$

Examining A^2, we find that

$$A^2 = \left(\frac{1}{\epsilon_1(1+\varepsilon)+v} - \frac{1}{\epsilon_1(1+\varepsilon)+v'}\right)^2 \tag{106}$$

$$= \left(\frac{\epsilon_1(1+\varepsilon)+v'-[\epsilon_1(1+\varepsilon)+v]}{[\epsilon_1(1+\varepsilon)+v][\epsilon_1(1+\varepsilon)+v']}\right)^2 \tag{107}$$

$$\leq (v'-v)^2/\epsilon_1^4 \tag{108}$$

since $\epsilon_1, \varepsilon > 0$ and $v, v' \geq 0$. Next, we write B^2 as

$$B^2 = \frac{1}{\epsilon_1^2}\left(\frac{1}{\alpha\log(\alpha)} - \frac{1}{\alpha'\log(\alpha')}\right)^2 \tag{109}$$

$$= \frac{1}{\epsilon_1^2}\left(\frac{\alpha'\log(\alpha')-\alpha\log(\alpha)}{\alpha\log(\alpha)\alpha'\log(\alpha')}\right)^2 \tag{110}$$

with $\alpha \triangleq 1+\varepsilon+\frac{v}{\epsilon_1}$ and $\alpha' \triangleq 1+\varepsilon+\frac{v'}{\epsilon_1}$, and realize

$$\alpha'\log(\alpha') - \alpha\log(\alpha)$$

$$= (\alpha + \frac{v'-v}{\epsilon_1})\log(\alpha') - \alpha\log(\alpha) \tag{111}$$

$$= \alpha\log(\alpha') - \alpha\log(\alpha) + \frac{v'-v}{\epsilon_1}\log(\alpha') \tag{112}$$

which implies that

$$B^2 = \frac{1}{\epsilon_1^2}\left(\underbrace{\frac{1}{\alpha'\log(\alpha)} - \frac{1}{\alpha'\log(\alpha')}}_{\triangleq B_1} + \underbrace{\frac{(v'-v)/\epsilon_1}{\alpha\alpha'\log(\alpha)}}_{\triangleq B_2}\right)^2 \tag{113}$$

$$\leq \frac{[|B_1|+|B_2|]^2}{\epsilon_1^2} \leq \frac{2\left[B_1^2+B_2^2\right]}{\epsilon_1^2}. \tag{114}$$

Examining B_1^2 we find

$$B_1^2 = \frac{1}{\alpha'^2}\left(\frac{1}{\log(\alpha)} - \frac{1}{\log(\alpha')}\right)^2 \tag{115}$$

$$= \frac{1}{\alpha'^2}\left(\frac{\log(\alpha')-\log(\alpha)}{\log(\alpha)\log(\alpha')}\right)^2 \tag{116}$$

$$= \frac{1}{\alpha'^2}\frac{\log(\alpha'/\alpha)^2}{\log(\alpha)^2\log(\alpha')^2}. \tag{117}$$

Because $\epsilon_1, \varepsilon > 0$ and $v, v' \geq 0$, we have that $\alpha, \alpha' > 1$ and $\log(\alpha)^2 \geq \log(1+\varepsilon)$ and $\log(\alpha')^2 \geq \log(1+\varepsilon)$, so that

$$B_1^2 \leq \frac{\log(\alpha'/\alpha)^2}{\log(1+\varepsilon)^4}. \tag{118}$$

Moreover,

$$\log(\alpha'/\alpha)^2 = \log\left(\frac{\alpha+\frac{v'-v}{\epsilon_1}}{\alpha}\right)^2 \tag{119}$$

$$= \log\left(1+\frac{v'-v}{\epsilon_1\alpha}\right)^2 \tag{120}$$

$$\leq \max\left\{\left(\frac{v'-v}{\epsilon_1\alpha}\right)^2, \left(\frac{v'-v}{\epsilon_1\alpha+v'-v}\right)^2\right\} \tag{121}$$

$$= \frac{(v'-v)^2}{\epsilon_1^2}\max\left\{\frac{1}{\alpha^2}, \frac{1}{(\alpha')^2}\right\} \tag{122}$$

$$\leq \frac{(v'-v)^2}{\epsilon_1^2}, \tag{123}$$

where (121) used the property that $\frac{x}{1+x} \leq \log(1+x) \leq x$ for $x > -1$, and (123) used $\alpha, \alpha' > 1$. Finally, we have

$$B_2^2 = \frac{(v'-v)^2}{\epsilon_1^2(\alpha\alpha')^2\log(1+\varepsilon+v/\epsilon_1)^2} \tag{124}$$

$$\leq \frac{(v'-v)^2}{\epsilon_1^2\log(1+\varepsilon)^2} \tag{125}$$

where the latter step used $\alpha, \alpha' > 1$ and $1+\varepsilon > 0$ and $v/\epsilon_1 \geq 0$. Putting together (103), (108), (114), (118), (123) and (125), we see that there exists $\beta > 0$ such that

$$|\nabla g_2(v) - \nabla g_2(v')|^2 \leq \beta(v'-v)^2 \quad \forall (v', v) \in \mathcal{C}, \tag{126}$$

implying that ∇g_2 is Lipschitz continuous.

REFERENCES

[1] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Prob.*, vol. 23, pp. 947–968, 2007.

[6] R. J. Tibshirani, "Solution path of the generalized lasso," *Ann. Stat.*, vol. 39, no. 3, pp. 1335–1371, 2011.

[7] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, 1992.

[8] M. A. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, Sep. 2003.

[9] M. A. T. Figueiredo and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.

[10] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, USA, Apr. 2008, pp. 3869–3872.

[11] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[12] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, Apr. 2010.

[13] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. App.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
[14] R. E. Carrillo, J. D. McEwen, D. Van De Ville, J.-P. Thiran, and Y. Wiaux, "Sparsity averaging for compressive imaging," *IEEE Signal Process. Lett.*, vol. 20, no. 6, pp. 591–594, Jun. 2013.
[15] M. A. T. Figueiredo and R. D. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffreys' noninformative prior," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1322–1331, Sep. 2001.
[16] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Adaptive total variation image deblurring: A majorization-minimization approach," *Signal Process.*, vol. 89, no. 9, pp. 1683–1693, 2009.
[17] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA, USA: MIT Press, 1998, pp. 355–368.
[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2007.
[19] J. Mairal, "Optimization with first-order surrogate functions," in *Proc. Int. Conf. Mach. Learn.*, 2013, vol. 28, pp. 783–791.
[20] S. Lu and S. V. Pereverzev, *Regularization Theory for Ill-Posed Problems*. Berlin, Germany: Walter de Gruyter, 2013.
[21] C. Brezinski, M. Redivo-Zaglia, G. Rodriguez, and S. Seatzu, "Multi-parameter regularization techniques for ill-conditioned linear systems," *Numer. Math.*, vol. 94, no. 2, pp. 203–228, 2003.
[22] P. Xu, Y. Fukuda, and Y. Liu, "Multiple parameter regularization: Numerical solutions and applications to the determination of geopotential from precise satellite orbits," *J. Geod.*, vol. 80, no. 1, pp. 17–27, 2006.
[23] S. Gazzola and P. Novati, "Multi-parameter Arnoldi-Tikhonov methods," *Electron. Trans. Numer. Anal.*, vol. 40, pp. 452–475, 2013.
[24] M. Fornasier, V. Naumova, and S. V. Pereverzyev, "Multi-parameter regularization techniques for ill-conditioned linear systems," *SIAM J. Numer. Anal.*, vol. 52, no. 4, pp. 1770–1794, 2014.
[25] M. Belge, M. E. Kilmer, and E. L. Miller, "Efficient determination of multiple regularization parameters in a generalized L-curve framework," *Inverse Prob.*, vol. 18, no. 4, pp. 1161–1183, 2002.
[26] K. Kunisch and T. Pock, "A bilevel optimization approach for parameter learning in variational models," *SIAM J. Imag. Sci.*, vol. 6, no. 2, pp. 938–983, 2013.
[27] A. Rakotomamonjy, "Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms," *Signal Process.*, vol. 91, pp. 1505–1526, 2011.
[28] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2906–2921, Jun. 2014.
[29] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 2, pp. 303–324, 2009.
[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[31] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Sep. 2010.
[32] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 6564–574, Dec. 2007.
[33] Z. Tan, Y. Eldar, A. Beck, and A. Nehorai, "Smoothing and decomposition for analysis sparse recovery," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1762–1774, Apr. 2014.
[34] S. Becker, J. Bobin, and E. J. Candès, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 1–39, 2011.
[35] M. Borgerding, P. Schniter, J. Vila, and S. Rangan, "Generalized approximate message passing for cosparse analysis compressive sensing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 3756–3760 (see also arXiv:1312.3698).
[36] D. Lorenz and N. Worliczek, "Necessary conditions for variational regularization schemes," *Inverse Prob.*, vol. 29, no. 7, p. 075016, 2013.
[37] R. Ahmad and P. Schniter, "Iteratively reweighted ℓ_1 approaches to ℓ_2-constrained sparse composite regularization," arXiv:1504.05110v2, Aug. 2015.
[38] R. Horst and N. Thoai, "DC programming: Overview," *J. Optim. Theory. Appl.*, vol. 103, no. 1, pp. 1–43, 1999.
[39] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.
[40] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*. New York, NY, USA: Springer, 2006.
[41] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York, NY, USA: Springer, 1994.
[42] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Berlin, Germany: Springer-Verlag, 1985.
[43] A. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, pp. 1–17, 1977.
[44] V. Cevher, "Learning with compressible priors," in *Proc. Neural Inf. Process. Syst. Conf.*, Vancouver, BC, Canada, Dec. 2009, pp. 261–269.
[45] G. Puy, P. Vandergheynst, R. Gribonval, and Y. Wiaux, "Universal and efficient compressed sensing by spread spectrum and application to realistic Fourier imaging techniques," *EURASIP J. Appl. Signal Process.*, vol. 2012, no. 6, pp. 1–13, 2012.
[46] R. Ahmad, H. Xue, S. Giri, Y. Ding, J. Craft, and O. P. Simonetti, "Variable density incoherent spatiotemporal acquisition (VISTA) for highly accelerated cardiac MRI," *Magn. Reson. Med.*, 2014 [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/mrm.25507/abstract

Rizwan Ahmad received the B.S. degree (Hons.) in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2000, and the M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, the Ohio State University, Columbus, OH, USA, in 2004 and 2007, respectively. Since 2014, he has been with the Department of Electrical and Computer Engineering, the Ohio State University, where he is currently working as a Research Assistant Professor. His research interests include biomedical imaging, signal processing, tomographic reconstruction, EPR imaging, and cardiac magnetic resonance imaging.

Philip Schniter (F'14) received the B.S. and M.S. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992 and 1993, respectively, and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 2000. From 1993 to 1996, he was with Tektronix Inc., Beaverton, OR, USA, as a Systems Engineer. After receiving the Ph.D. degree, he joined with the Department of Electrical and Computer Engineering, the Ohio State University, Columbus, OH, USA where he is currently a Professor and a Member of the Information Processing Systems (IPS) Laboratory. From 2008 to 2009, he was a Visiting Professor with Eurecom, Sophia Antipolis, France, and Supélec, Gif-sur-Yvette, France. His research interests include statistical signal processing, wireless communications and networks, and machine learning. In 2003, he was the recipient of the National Science Foundation CAREER Award.