Chapter 1
Recovering Signals with Unknown Sparsity in Multiple Dictionaries

Rizwan Ahmad and Philip Schniter

Abstract Motivated by the observation that a given signal x may admit sparse representations in multiple dictionaries Ψ_d, but with varying levels of sparsity across dictionaries, we propose two new algorithms for signal reconstruction from noisy linear measurements. Our first algorithm, extends the well-known basis-pursuit-denoising algorithm from the L1 regularizer $\|\Psi x\|_1$ to composite regularizers of the form $\sum_d \lambda_d \|\Psi_d x\|_1$ while self-adjusting the regularization weights λ_d. Our second algorithm extends the well-known iteratively reweighted L1 algorithm to the same family of composite regularizers. For each algorithm, we provide several interpretations: 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.A detailed numerical study suggests that, when compared to their non-composite counterparts, our composite algorithms yield significantly improvements in accuracy with only modest increases in computational complexity.

1.1 Introduction

Consider the problem of recovering a signal or image $x \in \mathbb{C}^n$ from noisy linear measurements

$$y = \Phi x + e \in \mathbb{C}^m, \tag{1.1}$$

Rizwan Ahmad
Dept. of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA,
e-mail: ahmad.46@osu.edu

Philip Schniter
Dept. of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA,
e-mail: schniter.1@osu.edu

where the measurement operator $\Phi \in \mathbb{C}^{m \times n}$ is known and $e \in \mathbb{C}^m$ is additive noise. Such problems arise in imaging, communications, speech, radar, machine learning, and many other applications. We are particularly interested in the case where $m \ll n$, under which x cannot be uniquely determined from the measurements y, even in the absence of noise. This latter situation arises in many of the applications mentioned earlier, and it has recently been popularized under the framework of *compressive sensing* (CS) [12, 22, 27].

1.1.1 ℓ_2-Constrained Regularization

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

$$\arg\min_x R(x) \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \varepsilon, \tag{1.2}$$

where $\varepsilon \geq 0$ a data-fidelity tolerance that reflects prior knowledge of the noise power and $R(x)$ is a penalty, or regularization, that reflects prior knowledge about the signal x [35]. We briefly summarize several common instances of $R(x)$ below.

1. If 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") $R(x) = \|x\|_0 \triangleq |\operatorname{supp}(x)|$. However, since this choice makes (1.2) NP-hard, it is rarely considered in practice.
2. The ℓ_1 penalty, $R(x) = \|x\|_1 = \sum_{j=1}^n |x_j|$, is a commonly used surrogate to ℓ_0 that makes (1.2) convex and thus solvable in polynomial time. Under this penalty, (1.2) is known as *basis pursuit denoising* [17] or as the *lasso* [44]. It is commonly used in *synthesis-based* CS [12, 22, 27].
3. Non-convex surrogates to the ℓ_0 penalty have also been proposed. Well-known varieties include the ℓ_p penalty $R(x) = \|x\|_p^p = \sum_{j=1}^n |x_j|^p$ with $p \in (0,1)$, and the log-sum penalty $R(x) = \sum_{j=1}^n \log(\delta + |x_j|)$ with $\delta \geq 0$. Although (1.2) becomes difficult to solve exactly in a guaranteed manner, it can be approximated, leading to excellent empirical performance. Further details will be given below.
4. The choice $R(x) = \|\Psi x\|_1$, with known matrix $\Psi \in \mathbb{C}^{L \times n}$, is familiar from *analysis-based* CS [21]. Penalties of this form are appropriate when prior knowledge suggests that the transform coefficients Ψx are (approximately) sparse, as opposed to the signal x itself being sparse. In this case, (1.2) can be solved by the *generalized lasso* [45]. When Ψ is a finite-difference operator, $\|\Psi x\|_1$ yields anisotropic *total variation* regularization [42].
5. Non-convex penalties can also be placed on the transform coefficients Ψx, leading to, e.g., $R(x) = \|\Psi x\|_p^p = \sum_{l=1}^L |\psi_l^T x|^p$ with $p \in (0,1)$ or $R(x) = \sum_{l=1}^L \log(\delta + |\psi_l^T x|)$ with $\delta \geq 0$, where ψ_l^T denotes the lth row of Ψ.

With a non-convex penalty $R(x)$, a popular approach to solving (1.2) is through *iteratively reweighted* $\ell_!$ (IRW-L1) [13, 46]. There, (1.2) with a fixed non-convex $R(x)$ is approximated by a sequence of convex problems

$$x^{(t)} = \arg\min_x R^{(t)}(x) \;\text{ s.t. }\; \|y - \Phi x\|_2^2 \le \varepsilon \tag{1.3}$$

with $R^{(t)}(x) = \sum_{j=1}^n w_j^{(t)} |x_j|$ a weighted ℓ_1 norm, where the weights $w^{(t)}$ are computed from the previous estimate $x^{(t-1)}$. When $w_j^{(t)} = (\delta + |x_j^{(t-1)}|)^{-1}$ for a small constant $\delta \ge 0$, the IRW-L1 algorithm can be interpreted [13] as a majorization-minimization (MM) [29] approach to (1.2) under the *log-sum* penalty $R(x) = \sum_{j=1}^n \log(\delta + |x_j|)$, which can be considered as a non-convex surrogate to the ℓ_0 penalty. Various empirical and theoretical studies [13, 46, 30] of this latter case have shown performance surpassing that of the ℓ_1 penalty. Unconstrained formulations of IRW-L1 based on "$\arg\min_x R^{(t)}(x) + \gamma\|y - \Phi x\|_2^2$" have also been considered, such as in the seminal work [25]. Likewise, constrained and unconstrained versions of iteratively reweighted ℓ_2 were considered in [23, 25, 16, 19, 46]. See [35] for further discussion.

1.1.2 Sparsity-Inducing Composite Regularizers

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

$$R_1(x) \triangleq \sum_{d=1}^D \lambda_d \|\Psi_d x\|_1, \tag{1.4}$$

where each $\Psi_d \in \mathbb{C}^{L_d \times n}$ is a known analysis operator and $\lambda_d \ge 0$ is its regularization weight. Our goal is to recover the signal x from measurements (1.1) using a constrained optimization (1.2) under the composite regularizer (1.4). Doing so requires an optimization of the weights $\lambda \triangleq [\lambda_1, \dots, \lambda_D]^T$ in (1.4). We are also interested in iteratively re-weighted extensions of this problem that, at iteration t, use composite regularizers of the form[1]

$$R^{(t)}(x) = \sum_{d=1}^D \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1, \tag{1.5}$$

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

As a motivating example, suppose that $\{\Psi_d\}$ is a collection of orthonormal bases that includes, e.g., spikes, sines, and various wavelet bases. The signal x may be sparse in some of these bases, but not all. Thus, we would like to adjust each λ_d

[1] Although (1.5) is over-parameterized, the form of (1.5) is convenient for algorithm development.

in (1.4) to appropriately weight the contribution from each basis. But it is not clear how to do this when x is unknown. As another example, suppose that x contains a (rasterized) sequence of images and that $\|\Psi_1 x\|_1$ measures temporal total-variation while $\|\Psi_2 x\|_1$ measures spatial total-variation. Intuitively, we would like to weight these two regularizations differently, depending on whether the image varies more in the temporal or spatial dimensions. But it is not clear how to do this when x is unknown.

1.1.3 Contributions

In this work, we propose novel iteratively reweighted approaches to sparse reconstruction based on composite regularizations of the form (1.4)-(1.5) with automatic tuning of the regularization weights λ and 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 [7, 24, 37], and
4. variational expectation maximization (VEM) [36, 8] 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 [34]. 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. 1.4, 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 $\{\Psi_d\}$ are chosen so that the sparsity of $\Psi_d x$ varies with d, this structure can be exploited for improved recovery. The more disparate the sparsity, the greater the improvement.

1.1.4 Related Work

As discussed above, the generalized lasso [45] is one of the most common approaches to L1-regularized analysis-CS [21], i.e., the optimization (1.2) under the regularizer $R(x) = \|\Psi x\|_1$. The Co-L1 algorithm that we present in Sec. 1.2 can be interpreted as a generalization of this L1 method to *composite* regularizers of the form (1.4). Meanwhile, the iteratively reweighted extension of the generalized lasso, IRW-L1 [13], 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. 1.3 can then be interpreted as a generalization of this IRW-L1 method to *composite* regularizers of the form (1.5). The existing non-composite L1 and IRW-L1 approaches essentially place an identical weight $\lambda_d = 1$ on every term in (1.4)-(1.5), and thus make no attempt to leverage differences in the sparsity of the transform coefficients $\Psi_d x$ across the sub-dictionary index d. However, the numerical results that we present in Sec. 1.4 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 $R(x;\lambda) = \sum_d \lambda_d R_d(x)$ is a long-standing problem with a rich literature (see, e.g., the recent book [33]). However, the vast majority of that literature focuses on the Tikhonov case where $R_d(x)$ are quadratic (see, e.g., [11, 47, 28, 26]). One notable exception is [6], which assumes continuously differentiable $R_d(x)$ and thus does not cover our composite ℓ_1 prior (1.4). Another notable exception is [32], which assumes i) the availability of a noiseless training example of x to help tune the L1 regularization weights λ in (1.4), and ii) the trivial measurement matrix $\Phi = I$. In contrast, our proposed methods operate without any training and support generic measurement matrices Φ.

In the special case that each Ψ_d is composed of a subset of rows from the $n \times n$ identity matrix, the regularizers (1.4)-(1.5) can induce *group* sparsity in the recovery of x, in that certain sub-vectors $x_d \triangleq \Psi_d x$ of x are driven to zero while others are not. The paper [40] 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. 1.2. However, our 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 Ψ_d. Recently, Bayesian MAP group-sparse recovery was considered in [4]. However, the technique described there uses Gaussian scale mixtures or, equivalently, weighted-ℓ_2 regularizers $R(x;\lambda) = \sum_d \lambda_d \|x_d\|_2$, while our methods use weighted-ℓ_1 regularizers (1.4)-(1.5).

A recent work [2] considered the *unconstrained* version of the problem considered in this chapter, where the aim is to solve a non-convex optimization problem of the form

$$\arg\min_x R(x) + \gamma \|y - \Phi x\|_2, \tag{1.6}$$

for some $\gamma > 0$, through a sequence of convex problems

$$x^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1 + \gamma \|y - \Phi x\|_2^2, \tag{1.7}$$

where $\{\lambda_d^{(t)}, W_d^{(t)}\}_{d=1}^{D}$ are set using $x^{(t-1)}$. Although the unconstrained case bears some similarity to the *constrained* case considered in this chapter, each case leads to a distinct set of algorithms, interpretations, and analyses.

1.1.5 Notation

We use capital letters like Ψ for matrices, small letters like x for vectors, and $(\cdot)^T$ for transposition. We use $\|x\|_p \triangleq (\sum_j |x_j|^p)^{1/p}$ for the ℓ_p norm of vector x, with x_j representing the nth coefficient in x. When referring to the "mixed $\ell_{p,q}$ norm" of a matrix X, we mean $(\sum_d(\sum_l |x_{d,l}|^p)^{q/p})^{1/q}$ as in [31], where $x_{d,l}$ is the dth row and lth column of X. We adopt the index-set abbreviation $[D] \triangleq \{1,\ldots,D\}$ and use I to denote the identity matrix. We use $\nabla g(x)$ for the gradient of a functional $g(x)$ with respect to 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 $p(x;\lambda)$ for the pdf of random vector x under deterministic parameters λ, and $p(x|\lambda)$ for the pdf of x conditioned on the random vector λ. We use $D_{\mathsf{KL}}(q\|p)$ to denote the Kullback-Leibler (KL) divergence of the pdf p from the pdf q, and we use $\mathbb{R}$ and $\mathbb{C}$ to denote the real and complex fields, respectively.

1.2 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 Ψ_d.

Algorithm 1 The Co-L1 Algorithm

```
1: input: $\{\Psi_d\}_{d=1}^D$, $\Phi$, $y$, $\varepsilon \geq 0$, $\delta \geq 0$
2: initialization: $\lambda_d^{(1)} = 1\ \forall d$
3: for $t = 1,2,3,\ldots$
4:     $x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^D \lambda_d^{(t)} \|\Psi_d x\|_1$ s.t. $\|y - \Phi x\|_2 \leq \varepsilon$
5:     $\lambda_d^{(t+1)} \leftarrow \frac{L_d}{\delta + \|\Psi_d x^{(t)}\|_1}$, $d \in [D]$
6: end
7: output: $x^{(t)}$
```

The main computational step of Co-L1 is the constrained ℓ_1 minimization in line 4, which can be recognized as (1.2) under the composite regularizer R_1 from (1.4). This is a convex optimization problem that can be readily solved by existing techniques (e.g., Douglas-Rachford splitting [18], ADMM [10, 1], NESTA-UP [5], MFISTA via smoothing and decomposition [43], etc.), the specific choice of which is immaterial to this paper.

Note that Co-L1 requires the user to set a small regularization term $\delta \geq 0$ whose role is to prevent the denominator in line 5 from reaching zero. For typical choices of the analysis operators Ψ_d and ε, the vector $\Psi_d x^{(t)}$ will almost never be exactly zero, in which case it suffices to set $\delta = 0$. Also, Co-L1 requires the user to set the measurement fidelity constraint $\varepsilon \geq 0$. For additive white Gaussian noise (AWGN) of variance $\sigma^2 > 0$, the choice $\varepsilon = 0.8\sqrt{\sigma^2 m}$ works empirically well, and we used this setting for all numerical results in Sec. 1.4.

Co-L1's update of the weights λ, defined by line 5 of Algorithm 1, can be interpreted in various ways, as we detail below. For ease of explanation, we first consider the case where the signal x is real-valued, and later discuss the complex-valued case in Sec. 1.2.6. As we will see, the steps in Algorithm 1 apply to both real- and complex-valued x.

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

1. *MM applied to (1.2) under the log-sum penalty*

$$R_{ls}^{D}(x;\delta) \triangleq \sum_{d=1}^{D} L_d \log(\delta + \|\Psi_d x\|_1), \tag{1.8}$$

2. *as $\delta \to 0$, an approximate solution to the weighted $\ell_{1,0}$ [31] problem*

$$\arg\min_{x} \sum_{d=1}^{D} L_d \, 1_{\|\Psi_d x\|_1 > 0} \;\; s.t. \;\; \|y - \Phi x\|_2 \leq \varepsilon, \tag{1.9}$$

3. *for $\varepsilon = 0$, MM applied to Bayesian MAP estimation under a noiseless likelihood and the hierarchical prior*

$$p(x|\lambda) = \prod_{d=1}^{D} \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d \|\Psi_d x\|_1\right) \tag{1.10}$$

$$\lambda \sim i.i.d.\ \Gamma(0, \delta^{-1}) \tag{1.11}$$

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

4. *for $\varepsilon = 0$, variational EM under a noiseless likelihood and the prior*

$$p(x;\lambda) \propto \prod_{d=1}^{D} \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d(\|\Psi_d x\|_1 + \delta)\right), \tag{1.12}$$

which, when $\delta = 0$, is i.i.d. Laplacian on $z_d = \Psi_d x \in \mathbb{R}^{L_d}$ with deterministic scale parameter $\lambda_d > 0$.

Proof. See Sections 1.2.1 to 1.2.5 below.

Importantly, the MM interpretation implies convergence (in the sense of an asymptotic stationary point condition) when $\delta > 0$, as detailed in Sec. 1.2.2.

1.2.1 Log-Sum MM Interpretation of Co-L1

Consider the optimization problem

$$\arg\min_x R_{\mathsf{ls}}^D(x;\delta) \;\text{ s.t. }\; \|y-\Phi x\|_2 \leq \varepsilon \tag{1.13}$$

with R_{ls}^D from (1.8). Inspired by [13, §2.3], we write (1.13) as

$$\arg\min_{x,u} \sum_{d=1}^D L_d \log\left(\delta + \sum_{l=1}^{L_d} u_{d,l}\right) \;\text{ s.t. }\; \begin{cases} \|y-\Phi x\|_2 \leq \varepsilon \\ |\psi_{d,l}^T x| \leq u_{d,l} \;\forall d,l, \end{cases} \tag{1.14}$$

where $\psi_{d,l}^T$ is the lth row of Ψ_d. Problem (1.14) is of the form

$$\arg\min_v g(v) \;\text{ s.t. }\; v \in C, \tag{1.15}$$

where $v=[u^T,x^T]^T$, C is a convex set,

$$g(v) = \sum_{d=1}^D L_d \log\left(\delta + \sum_{k\in K_d} v_k\right) \tag{1.16}$$

is a concave penalty, and the set $K_d \triangleq \{k : \sum_{d'=1}^{d-1} L_{d'} < k \leq \sum_{d'=1}^{d} L_{d'}\}$ contains the indices k such that $v_k \in \{u_{d,l}\}_{l=1}^{L_d}$.

Majorization-minimization (MM) [29, 34] is a popular method to attack non-convex problems of this form. In particular, MM iterates the following two steps: (i) construct a surrogate $g(v;v^{(t)})$ that majorizes $g(v)$ at $v^{(t)}$, and (ii) update $v^{(t+1)} = \arg\min_{v\in C} g(v;v^{(t)})$. By "majorize," we mean that $g(v;v^{(t)}) \geq g(v)$ for all v with equality when $v=v^{(t)}$.

Due to the concavity of our g, we can construct a majorizing surrogate using the tangent of g at $v^{(t)}$. In particular, let ∇g denote the gradient of g w.r.t. v. Then

$$g(v;v^{(t)}) = g(v^{(t)}) + \nabla g(v^{(t)})^T [v - v^{(t)}] \tag{1.17}$$

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

$$v^{(t+1)} = \arg\min_{v\in C} \nabla g(v^{(t)})^T v \tag{1.18}$$

after neglecting the v-invariant terms. From (1.16), we see that

$$[\nabla g(v^{(t)})]_k = \begin{cases} \dfrac{L_{d(k)}}{\delta + \sum_{i\in K_{d(k)}} v_i^{(t)}} & \text{if } d(k) \neq 0 \\ 0 & \text{else,} \end{cases} \tag{1.19}$$

where $d(k)$ is the index $d \in [D]$ of the set K_d containing k, or 0 if no such set exists. Thus MM prescribes

$$v^{(t+1)} = \arg\min_{v \in C} \sum_{d=1}^{D} \sum_{k \in K_d} \frac{L_d v_k}{\delta + \sum_{i \in K_d} v_i^{(t)}}, \tag{1.20}$$

or equivalently

$$x^{(t+1)} = \arg\min_{x} \sum_{d=1}^{D} \frac{L_d \sum_{l=1}^{L_d} |\psi_{d,l}^T x|}{\delta + \sum_{l=1}^{L_d} |\psi_{d,l}^T x^{(t)}|} \text{ s.t. } \|y - \Phi x\|_2 \le \varepsilon \tag{1.21}$$

$$= \arg\min_{x} \sum_{d=1}^{D} \lambda_d^{(t+1)} \|\Psi_d x\|_1 \text{ s.t. } \|y - \Phi x\|_2 \le \varepsilon \tag{1.22}$$

for

$$\lambda_d^{(t+1)} = \frac{L_d}{\delta + \|\Psi_d x^{(t)}\|_1}, \tag{1.23}$$

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

1.2.2 Convergence of Co-L1

The recent paper [34] studies the convergence of MM algorithms. In particular, it establishes that when the optimization objective $g(v)$ is differentiable in $v \in C$ with a Lipschitz continuous gradient, the MM sequence $\{v^{(t)}\}_{t \ge 1}$ satisfies an asymptotic stationary point (ASP) condition. Although it falls short of establishing convergence to a local minimum (which is very difficult for general non-convex optimization problems), the ASP condition is based on a classical necessary condition for a local minimum. In particular, using $\nabla g(v;d)$ to denote the directional derivative of g at v in the direction d, it is known [9] that $v_\star$ locally minimizes g over C only if $\nabla g(v_\star; v - v_\star) \ge 0$ for all $v \in C$. Thus, in [34], it is said that $\{v^{(t)}\}_{t \ge 1}$ satisfies an ASC condition if

$$\liminf_{t \to +\infty} \inf_{v \in C} \frac{\nabla g(v^{(t)}; v - v^{(t)})}{\|v - v^{(t)}\|_2} \ge 0. \tag{1.24}$$

In our case, g from (1.16) is indeed differentiable, with gradient given in (1.19). Moreover, [2, App. A] shows that the gradient is Lipschitz continuous when $\delta > 0$. Thus, the sequence of estimates produced by Algorithm 1 satisfies the ASP condition (1.24).

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

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

$$\arg\min_x \sum_{j=1}^{n} \log(\delta + |x_j|) \text{ s.t. } \|y - \Phi x\|_2 \le \varepsilon \tag{1.25}$$

is known [41] to be equivalent to ℓ_0 minimization

$$\arg\min_x \|x\|_0 \text{ s.t. } \|y - \Phi x\|_2 \le \varepsilon. \tag{1.26}$$

(See [2, App. B] for a proof.) This equivalence can be seen intuitively as follows. As $\delta \to 0$, the contribution to the regularization term $\sum_{j=1}^{n} \log(\delta + |x_j|)$ from each non-zero x_j remains finite, while that from each zero-valued x_j approaches $-\infty$. Since we are interested in minimizing the regularization term, we get a huge reward for each zero-valued x_j, or—equivalently—a huge penalty for each non-zero x_j.

To arrive at an ℓ_0 interpretation of the Co-L1 algorithm, we consider the corresponding optimization problem (1.13) in the limit that $\delta \to 0$. There we see that the regularization term $R^D_{\mathsf{ls}}(x;0)$ from (1.8) yields L_d huge rewards when $\|\Psi_d x\|_1 = 0$, or equivalently L_d huge penalties when $\|\Psi_d x\|_1 \neq 0$, for each $d \in [D]$. Thus, we can interpret Co-L1 as attempting to solve the optimization problem (1.9), which is a weighted version of the "$\ell_{p,q}$ mixed norm" problem from [31] for $p = 1$ and $q \to 0$. This establishes Part 2 of Theorem 1.

1.2.4 Bayesian MAP Interpretation of Co-L1

The MAP estimate [38] of x from y is

$$x_{\mathsf{MAP}} \triangleq \arg\max_x p(x|y) = \arg\min_x \big\{ -\log p(x|y) \big\} \tag{1.27}$$

$$= \arg\min_x \big\{ -\log p(x) - \log p(y|x) \big\}, \tag{1.28}$$

where (1.27) used the monotonicity of log and (1.28) used Bayes rule. In the case of a noiseless likelihood (e.g., AWGN with variance $\sigma^2 \to 0$), the second term in (1.28) is $+\infty$ unless $y = \Phi x$, and so

$$x_{\mathsf{MAP}} = \arg\min_x \big\{ -\log p(x) \big\} \text{ s.t. } y = \Phi x. \tag{1.29}$$

Recall that, with shape parameter κ and scale parameter θ, the Gamma pdf 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 see that $\Gamma(\lambda_d; 0, \infty) \propto 1_{\lambda_d > 0}/\lambda_d$, which is Jeffrey's non-informative hyperprior [7, 24, 37] for the Laplace scale parameter λ_d. Then, according to (1.10)-(1.11), the prior equals

$$p(x) = \int_{\mathbb{R}^D} p(x|\lambda)p(\lambda)\mathrm{d}\lambda \tag{1.30}$$

$$\propto \prod_{d=1}^{D} \int_0^\infty \left(\frac{\lambda_d}{2}\right)^{L_d} \exp(-\lambda_d \|\Psi_d x\|_1) \frac{\exp(-\lambda_d \delta)}{\lambda_d} \mathrm{d}\lambda_d \tag{1.31}$$

$$= \prod_{d=1}^{D} \frac{(L_d - 1)!}{\big(2(\|\Psi_d x\|_1 + \delta)\big)^{L_d}}, \tag{1.32}$$

which implies that

$$-\log p(x) = \text{const} + \sum_{d=1}^{D} L_d \log\big(\|\Psi_d x\|_1 + \delta\big). \tag{1.33}$$

Thus (1.29), (1.33), and (1.8) imply

$$x_{\mathsf{MAP}} = \arg\min_x R_{\mathsf{ls}}^D(x;0) \;\; \text{s.t.} \;\; y = \Phi x. \tag{1.34}$$

Finally, applying the MM algorithm to this optimization problem (as detailed in Sec. 1.2.1), we arrive at the $\varepsilon = 0$ version of Algorithm 1. This establishes Part 3 of Theorem 1.

1.2.5 Variational EM Interpretation of Co-L1

The variational expectation-maximization (VEM) algorithm [36, 8] is an iterative approach to maximum-likelihood (ML) estimation that generalizes the EM algorithm from [20]. We now provide a brief review of the VEM algorithm and describe how it can be applied to estimate λ in (1.12).

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

$$\log p(y;\lambda) = \int q(x) \log p(y;\lambda)\mathrm{d}x \tag{1.35}$$

$$= \int q(x) \log\left[\frac{p(x,y;\lambda)}{q(x)} \frac{q(x)}{p(x|y;\lambda)}\right] \mathrm{d}x \tag{1.36}$$

$$= \underbrace{\int q(x) \log \frac{p(x,y;\lambda)}{q(x)} \mathrm{d}x}_{\triangleq F(q(x);\lambda)} + \underbrace{\int q(x) \log \frac{q(x)}{p(x|y;\lambda)} \mathrm{d}x}_{\triangleq D_{\mathsf{KL}}(q(x)\|p(x|y;\lambda))}, \tag{1.37}$$

for an arbitrary pdf $q(x)$, 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 $F(q(x);\lambda)$ is a lower bound on $\log p(y;\lambda)$. The EM algorithm performs ML estimation by iterating

$$q^{(t)}(x) = \arg\min_q D_{\mathsf{KL}}\big(q(x)\big\|p(x|y;\lambda^{(t)})\big) \tag{1.38}$$

$$\lambda^{(t+1)} = \arg\max_\lambda F(q^{(t)}(x);\lambda), \tag{1.39}$$

where the "E" step (1.38) tightens the lower bound and the "M" step (1.39) maximizes the lower bound.

The EM algorithm places no constraints on $q(x)$, in which case the solution to (1.38) is simply $q^{(t)}(x) = p(x|y;\lambda^{(t)})$, i.e., the posterior pdf of x under $\lambda = \lambda^{(t)}$. In many applications, however, this posterior is too difficult to compute and/or use in (1.39). To circumvent this problem, the VEM algorithm constrains $q(x)$ to some family of distributions Q that makes (1.38)-(1.39) tractable.

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

$$q(x) \propto \lim_{\tau\to 0} \exp\Big(\tfrac{1}{\tau}\log p(x|y;\lambda)\Big), \tag{1.40}$$

which has the effect of concentrating the mass in $q(x)$ at its mode. Plugging this $q(x)$ and $p(x,y;\lambda) = p(y|x)p(x;\lambda)$ into (1.37), we see that the M step (1.39) reduces to

$$\lambda^{(t+1)} = \arg\max_\lambda \log p(x;\lambda)\big|_{x=x_{\mathsf{MAP}}^{(t)}} \tag{1.41}$$

$$\text{for } x_{\mathsf{MAP}}^{(t)} \triangleq \arg\max_x p(x|y;\lambda^{(t)}), \tag{1.42}$$

where (1.42) can be interpreted as the E step. For the particular $p(x;\lambda)$ in (1.12), we have that

$$\log p(x;\lambda) = \text{const} + \sum_{d=1}^{D}\big[L_d\log(\lambda_d) - \lambda_d(\|\Psi_d x\|_1 + \delta)\big], \tag{1.43}$$

and by zeroing the gradient w.r.t. λ, we find that (1.41) becomes

$$\lambda_d^{(t+1)} = \frac{L_d}{\big\|\Psi_d x_{\mathsf{MAP}}^{(t)}\big\|_1 + \delta}, \;\; d \in [D]. \tag{1.44}$$

Meanwhile, from the noiseless MAP expression (1.29) and (1.43), we find that (1.42) becomes

$$x_{\mathsf{MAP}}^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)}\|\Psi_d x\|_1 \;\text{ s.t. }\; y = \Phi x. \tag{1.45}$$

In conclusion, our VEM algorithm iterates the steps (1.44)-(1.45), which match the steps in Algorithm 1 for $\varepsilon = 0$. This establishes Part 4 of Theorem 1.

1.2.6 Co-L1 for Complex-Valued x

In Theorem 1 and Sections 1.2.1-1.2.5, real-valued x was assumed for ease of explanation. However, real-valuedness was employed only in defining the Laplacian pdfs (1.10) and (1.12). As we now show, the Co-L1 algorithm in Algorithm 1 can also be justified based on a complex-valued Laplacian pdf. 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 show that, for $\varepsilon = 0$, Algorithm 1 results from VEM inference under an noiseless likelihood and the signal prior

$$p(x;\lambda) \propto \prod_{d=1}^{D} \left(\frac{\lambda_d}{2\pi}\right)^{2L_d} \exp\left(-\lambda_d(\|\Psi_d x\|_1 + \delta)\right), \tag{1.46}$$

which, when $\delta = 0$, is i.i.d. Laplacian on $z_d = \Psi_d x \in \mathbb{C}^{L_d}$ with deterministic scale parameter $\lambda_d > 0$. To show this, we follow the steps in Sec. 1.2.5 up to the log-prior in (1.43), which now becomes

$$\log p(x;\lambda) = \text{const} + \sum_{d=1}^{D} \left[2L_d \log(\lambda_d) - \lambda_d(\|\Psi_d x\|_1 + \delta)\right]. \tag{1.47}$$

Zeroing the gradient w.r.t. λ, we find that the VEM update in (1.41) becomes

$$\lambda_d^{(t+1)} = \frac{2L_d}{\left\|\Psi_d x_{\mathsf{MAP}}^{(t)}\right\|_1 + \delta}, \;\; d \in [D], \tag{1.48}$$

which differs from its real-valued counterpart (1.44) in a constant scaling of 2. However, this scaling does not affect $x_{\mathsf{MAP}}^{(t+1)}$ in (1.45) and thus does not affect the output $x^{(t)}$ of Algorithm 1, and thus can be ignored.

1.2.7 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 from [13]. For clarity, and for later use in Sec. 1.3, we summarize this latter algorithm in Algorithm 2, and note that the synthesis-CS formulation follows from the special case that $\Psi = I$.

Comparing Algorithm 2 to Algorithm 1, we see that IRW-L1 coincides with Co-L1 in the case that every sub-dictionary Ψ_d has dimension one, i.e., $L_d = 1 \; \forall d$ 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.

Algorithm 2 The IRW-L1 Algorithm

```
input: $\Psi = [\psi_1, \dots, \psi_L]^T$, $\Phi$, $y$, $\varepsilon \geq 0$, $\delta \geq 0$
initialization: $W^{(1)} = I$
for $t = 1, 2, 3, \dots$
    $x^{(t)} \leftarrow \arg\min_x \|W^{(t)} \Psi x\|_1$ s.t. $\|y - \Phi x\|_2 \leq \varepsilon$
    $W^{(t+1)} \leftarrow \operatorname{diag}\left\{ \frac{1}{\delta + |\psi_1^T x^{(t)}|}, \dots, \frac{1}{\delta + |\psi_L^T x^{(t)}|} \right\}$
end
output: $x^{(t)}$
```

Corollary 1 (IRW-L1). *The IRW-L1 algorithm from Algorithm 2 has the following interpretations:*

1. *MM applied to (1.2) under the log-sum penalty*

$$R_{\mathsf{ls}}^L(x; \delta) = \sum_{l=1}^{L} \log(\delta + |\psi_l^T x|), \tag{1.49}$$

recalling the definition of R_{ls}^L from (1.8),

2. *as $\delta \to 0$, an approximate solution to the ℓ_0 problem*

$$\arg\min_x \sum_{l=1}^{L} 1_{|\psi_l^T x| > 0} \;\; \text{s.t.} \;\; \|y - \Phi x\|_2 \leq \varepsilon, \tag{1.50}$$

3. *for $\varepsilon = 0$, MM applied to Bayesian MAP estimation under a noiseless likelihood and the hierarchical prior*

$$p(x|\lambda) = \prod_{l=1}^{L} \frac{\lambda_l}{2} \exp\left(-\lambda_l |\psi_l^T x|\right) \tag{1.51}$$

$$\lambda \sim \text{i.i.d. } \Gamma(0, \delta^{-1}), \tag{1.52}$$

where $z_l = \psi_l^T x$ is Laplacian given λ_l, and λ_l is Gamma distributed with scale parameter δ^{-1} and shape parameter zero, which becomes Jeffrey's non-informative hyperprior $p(\lambda_l) \propto 1_{\lambda_l > 0}/\lambda_l$ when $\delta = 0$.

4. *for $\varepsilon = 0$, variational EM under a noiseless likelihood and the prior*

$$p(x; \lambda) \propto \prod_{l=1}^{L} \frac{\lambda_l}{2} \exp\left(-\lambda_l (|\psi_l^T x| + \delta)\right), \tag{1.53}$$

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

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

1.3 The Co-IRW-L1 algorithm

We now propose the Co-IRW-L1-δ algorithm, which is summarized in Algorithm 3. Co-IRW-L1-δ 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-δ algorithm uses sub-dictionary dependent weights λ_d that are updated at each iteration t using a sparsity metric on $\Psi_d x^{(t)}$. But, like with IRW-L1, the Co-IRW-L1-δ algorithm also uses diagonal weight matrices $W_d^{(t)}$ that are updated at each iteration. As with both Co-L1 and IRW-L1, the computational burden of Co-IRW-L1-δ is dominated by the constrained ℓ_1 minimization problem in line 4 of Algorithm 3, which is readily solved by existing techniques like Douglas-Rachford splitting.

Algorithm 3 The Real-Valued Co-IRW-L1-δ Algorithm

```
1: input: $\{\Psi_d\}_{d=1}^D$, $\Phi$, $y$, $\varepsilon \geq 0$, $\delta_d > 0\ \forall d$, $\rho \geq 0$,
2: initialization: $\lambda_d^{(1)} = 1, W_d^{(1)} = I,\ \forall d \in [D]$
3: for $t = 1,2,3,\dots$
4:     $x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1$ s.t. $\|y - \Phi x\|_2 \leq \varepsilon$
5:     $\lambda_d^{(t+1)} \leftarrow \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log\left(1+\rho+\frac{|\psi_{d,l}^T x^{(t)}|}{\delta_d}\right)\right]^{-1} + 1,\ \forall d \in [D]$
6:     $W_d^{(t+1)} \leftarrow \operatorname{diag}\left\{\frac{1}{\delta_d(1+\rho)+|\psi_{d,1}^T x^{(t)}|}, \cdots, \frac{1}{\delta_d(1+\rho)+|\psi_{d,L_d}^T x^{(t)}|}\right\},\ \forall d \in [D]$
7: end
8: output: $x^{(t)}$
```

THE Co-IRW-L1-δ algorithm can be interpreted in various ways, as we detail below. For clarity, we first consider fixed regularization parameters δ and later, in Sec. 1.3.6, 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 case where x is real-valued and later, in Sec. 1.3.7, discuss the complex-valued case.

Theorem 2 (Co-IRW-L1-δ). *The real-valued Co-IRW-L1-δ algorithm in Algorithm 3 has the following interpretations:*

1. *MM applied to (1.2) under the log-sum-log penalty*

$$R_{\mathsf{lsl}}(x;\delta,\rho) \triangleq \sum_{d=1}^{D}\sum_{l=1}^{L_d}\log\left[\left(\delta_d(1+\rho)+|\psi_{d,l}^T x|\right)\sum_{i=1}^{L_d}\log\left(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d}\right)\right], \tag{1.54}$$

2. *as $\rho \to 0$ and $\delta_d \to 0\ \forall d$, an approximate solution to the $\ell_0 + \ell_{0,0}$ problem*

$$\arg\min_x \|\Psi x\|_0 + \sum_{d=1}^{D} L_d\, 1_{\|\Psi_d x\|_0>0} \ \textit{ s.t. } \ \|y-\Phi x\|_2 \le \varepsilon, \tag{1.55}$$

3. *for $\varepsilon = 0$, MM applied to Bayesian MAP estimation under a noiseless likelihood and the hierarchical prior*

$$p(x|\lambda;\delta) = \prod_{d=1}^{D}\prod_{l=1}^{L_d}\frac{\lambda_d}{2\delta_d}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-(\lambda_d+1)} \tag{1.56}$$

$$p(\lambda) = \prod_{d=1}^{D} p(\lambda_d), \ \ p(\lambda_d) \propto \begin{cases}\frac{1}{\lambda_d} & \lambda_d > 0\\ 0 & \textit{else}\end{cases}, \tag{1.57}$$

where, when $\rho = 0$, the variables $z_d = \Psi_d x \in \mathbb{R}^{L_d}$ are i.i.d. generalized-Pareto [15] given λ_d, and $p(\lambda_d)$ is Jeffrey's non-informative hyperprior [7, 24, 37] for the random shape parameter λ_d.

4. *for $\varepsilon = 0$, variational EM under a noiseless likelihood and the prior*

$$p(x;\lambda,\delta) = \prod_{d=1}^{D}\prod_{l=1}^{L_d}\frac{\lambda_d-1}{2\delta_d}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-\lambda_d}, \tag{1.58}$$

where, when $\rho = 0$, the variables $z_d = \Psi_d x \in \mathbb{R}^{L_d}$ are i.i.d. generalized-Pareto with deterministic shape parameter $\lambda_d > 1$ and scale parameter $\delta_d > 0$.

Proof. See Sections 1.3.1 to 1.3.5 below.

As with Co-L1, the MM interpretation implies convergence (in the sense of an asymptotic stationary point condition) when $\rho > 0$, as detailed in Sec. 1.3.2.

1.3.1 Log-Sum-Log MM Interpretation of Co-IRW-L1-δ

Consider the optimization problem

$$\arg\min_x R_{\mathsf{lsl}}(x;\delta,\rho) \ \text{ s.t. } \ \|y-\Phi x\|_2 \le \varepsilon, \tag{1.59}$$

with R_{lsl} defined in (1.54). We attack this optimization problem using the MM approach detailed in Sec. 1.2.1. The difference is that now the function g is defined as

$$g(v) = \sum_{d=1}^{D} \sum_{k \in K_d} \log \left[\left(\delta_d (1+\rho) + v_k \right) \sum_{i \in K_d} \log \left(1 + \rho + \frac{v_i}{\delta_d} \right) \right] \tag{1.60}$$

$$= \sum_{d=1}^{D} \left[L_d \log \sum_{i \in K_d} \log \left(1 + \rho + \frac{v_i}{\delta_d} \right) + \sum_{k \in K_d} \log \left(\delta_d (1+\rho) + v_k \right) \right], \tag{1.61}$$

which has a gradient of

$$[\nabla g(v^{(t)})]_k = \left(\frac{L_{d(k)}}{\sum_{i \in K_{d(k)}} \log \left(1 + \rho + \frac{v_i^{(t)}}{\delta_{d(k)}} \right)} + 1 \right) \frac{1}{\delta_{d(k)} (1+\rho) + v_k^{(t)}} \tag{1.62}$$

when $d(k) \neq 0$ and otherwise $[\nabla g(v^{(t)})]_k = 0$. Thus, recalling (1.18), MM prescribes

$$v^{(t+1)} = \arg\min_{v \in C} \sum_{d=1}^{D} \sum_{k \in K_d} \left(\frac{L_d}{\sum_{i \in K_d} \log \left(1 + \rho + \frac{v_i^{(t)}}{\delta_d} \right)} + 1 \right) \left(\frac{v_k}{\delta_d (1+\rho) + v_k^{(t)}} \right) \tag{1.63}$$

or equivalently

$$x^{(t+1)} = \arg\min_{x} \sum_{d=1}^{D} \sum_{l=1}^{L_d} \lambda_d^{(t+1)} \left(\frac{|\psi_{d,l}^T x|}{\delta_d (1+\rho) + |\psi_{d,l}^T x^{(t)}|} \right) \quad \text{s.t.} \ \|y - \Phi x\|_2 \leq \varepsilon \tag{1.64}$$

for

$$\lambda_d^{(t+1)} = \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log \left(1 + \rho + \frac{|\psi_{d,l}^T x^{(t)}|}{\delta_d} \right) \right]^{-1} + 1, \tag{1.65}$$

which coincides with Algorithm 3. This establishes Part 1 of Theorem 2.

1.3.2 Convergence of Co-IRW-L1-δ

The convergence of Co-IRW-L1-δ (in the sense of an asymptotic stationary point condition) for $\rho > 0$ can be shown using the same procedure as in Sec. 1.2.2. To do this, we only need to verify that the gradient ∇g in (1.62) is Lipschitz continuous when $\rho > 0$, which was done in [2, App. C].

1.3.3 Approximate $\ell_0 + \ell_{0,0}$ Interpretation of Co-IRW-L1-δ

Recalling the discussion in Sec. 1.2.3, we now consider the behavior of the $R_{\mathsf{lsl}}(x;\delta,\rho)$ regularizer in (1.54) as $\rho \to 0$ and $\delta_d \to 0\ \forall d$. For this, it helps to decouple (1.54) into two terms:

$$\begin{aligned} &R_{\mathsf{lsl}}(x;\delta,\rho) \qquad (1.66)\\ &= \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left(\delta_d(1+\rho)+|\psi_{d,l}^T x|\right) + \sum_{d=1}^{D}\sum_{l=1}^{L_d} \log\left[\sum_{i=1}^{L_d} \log\left(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d}\right)\right]. \end{aligned}$$

As $\delta_d \to 0\ \forall d$, the first term in (1.66) contributes an infinite valued “reward” for each pair (d,l) such that $|\psi_{d,l}^T x| = 0$, or a finite valued cost otherwise. As for the second term, we see that $\lim_{\rho\to 0,\delta_d\to 0} \sum_{i=1}^{L_d} \log\left(1+|\psi_{d,i}^T x|/\delta_d+\rho\right) = 0$ if and only if $|\psi_{d,i}^T x| = 0\ \forall i \in [L_d]$, i.e., if and only if $\|\Psi_d x\|_0 = 0$. And when $\|\Psi_d x\|_0 = 0$, the second term in (1.66) contributes L_d infinite valued rewards. In summary, as $\rho \to 0$ and $\delta_d \to 0\ \forall d$, the first term in (1.66) behaves like $\|\Psi x\|_0$ and the second term like the weighted $\ell_{0,0}$ quasi-norm $\sum_{d=1}^{D} L_d 1_{\|\Psi_d x\|_0>0}$, as stated in (1.55). This establishes Part 2 of Theorem 2.

1.3.4 Bayesian MAP Interpretation of Co-IRW-L1-δ

To show that Co-IRW-L1-δ can be interpreted as Bayesian MAP estimation under the hierarchical prior (1.56)-(1.57), we first compute the prior $p(x)$. To start,

$$\begin{aligned} p(x) &= \int_{\mathbb{R}^D} p(\lambda)p(x|\lambda)\mathrm{d}\lambda \qquad (1.67)\\ &\propto \prod_{d=1}^{D} \int_0^\infty \frac{1}{\lambda_d} \prod_{l=1}^{L_d} \frac{\lambda_d}{2\delta_d}\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-(\lambda_d+1)} \mathrm{d}\lambda_d. \qquad (1.68) \end{aligned}$$

Writing $(1+\rho+|\psi_{d,l}^T x|/\delta_d)^{-(\lambda_d+1)} = \exp(-(\lambda_d+1)Q_{d,l})$ for $Q_{d,l} \triangleq \log(1+\rho+|\psi_{d,l}^T x|/\delta_d)$, we get

$$p(x) \propto \prod_{d=1}^{D} \frac{1}{(2\delta_d)^{L_d}} \int_0^\infty \lambda_d^{L_d-1} e^{-(\lambda_d+1)\sum_{l=1}^{L_d} Q_{d,l}} \mathrm{d}\lambda_d. \qquad (1.69)$$

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

$$p(x) \propto \prod_{d=1}^{D} \frac{e^{-Q_d}}{(2\delta_d Q_d)^{L_d}} \underbrace{\int_0^\infty \tau_d^{L_d-1} e^{-\tau_d} \mathrm{d}\tau_d}_{(L_d-1)!} \tag{1.70}$$

$$\propto \prod_{d=1}^{D} \left[\frac{1}{\delta_d \sum_{i=1}^{L_d} \log(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d})} \right]^{L_d} \prod_{l=1}^{L_d} \frac{1}{1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}} \tag{1.71}$$

$$= \prod_{d=1}^{D} \prod_{l=1}^{L_d} \left[\Big(\delta_d(1+\rho)+|\psi_{d,l}^T x|\Big) \sum_{i=1}^{L_d} \log\left(1+\rho+\frac{|\psi_{d,i}^T x|}{\delta_d}\right) \right]^{-1}, \tag{1.72}$$

which implies that

$$-\log p(x) = \text{const} + R_{\mathsf{lsl}}(x;\delta,\rho) \tag{1.73}$$

for $R_{\mathsf{lsl}}(x;\delta,\rho)$ defined in (1.54).

Plugging (1.73) into noiseless MAP expression (1.29), we have

$$x_{\mathsf{MAP}} = \arg\min_x R_{\mathsf{lsl}}(x;\delta,\rho) \;\text{ s.t. }\; y = \Phi x, \tag{1.74}$$

which is equivalent to the optimization problem in (1.59) when $\varepsilon = 0$. We showed in Sec. 1.3.1 that, by applying the MM algorithm to (1.59), we arrive at Algorithm 3. This establishes Part 3 of Theorem 2.

1.3.5 Variational EM Interpretation of Co-IRW-L1-δ

To justify the variational EM (VEM) interpretation of Co-IRW-L1-δ, we closely follow the approach used for Co-L1 in Sec. 1.2.5. The main difference is that now the prior takes the form of $p(x;\lambda,\delta)$ from (1.58). Thus, (1.43) becomes

$$\log p(x;\lambda,\delta) = \sum_{d=1}^{D} \sum_{l=1}^{L_d} \left[\log\left(\frac{\lambda_d - 1}{\delta_d}\right) - \lambda_d \log\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right) \right] + \text{const} \tag{1.75}$$

and by zeroing the gradient w.r.t. λ we see that the M step (1.44) becomes

$$\frac{1}{\lambda_d^{(t+1)} - 1} = \frac{1}{L_d} \log\left(1+\rho+\frac{|\psi_{d,l}^T x_{\mathsf{MAP}}^{(t)}|}{\delta_d}\right), \;\; d \in [D], \tag{1.76}$$

where again $x_{\mathsf{MAP}}^{(t)}$ denotes the MAP estimate of x under $\lambda = \lambda^{(t)}$. From (1.29) and (1.58), we see that

$$x_{\mathsf{MAP}}^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \sum_{l=1}^{L_d} \log\big(|\psi_{d,l}^T x| + \delta_d(1+\rho)\big) \text{ s.t. } y = \Phi x, \tag{1.77}$$

which (for $\rho = 0$) is a $\lambda^{(t)}$-weighted version of the IRW-L1 log-sum optimization problem (recall Part 1 of Corollary 1). To solve (1.77), we apply MM with inner iteration i. With a small modification of the MM derivation from Sec. 1.2.1, we obtain the 2-step iteration

$$x_{\mathsf{MAP}}^{(i)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(i)} \Psi_d x\|_1 \text{ s.t. } y = \Phi x \tag{1.78}$$

$$W_d^{(i+1)} = \operatorname{diag}\left\{ \frac{1}{\delta_d(1+\rho) + |\psi_{d,1}^T x^{(i)}|}, \cdots, \frac{1}{\delta_d(1+\rho) + |\psi_{d,L_d}^T x^{(i)}|} \right\}, \tag{1.79}$$

with $\lambda_d^{(t)}$ fixed at the value appearing in (1.77). Next, by using only a single MM iteration per VEM iteration, the MM index "i" can be equated with the VEM index "t," in which case the VEM algorithm becomes

$$x^{(t)} = \arg\min_x \sum_{d=1}^{D} \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1 \text{ s.t. } y = \Phi x \tag{1.80}$$

$$W_d^{(t+1)} = \operatorname{diag}\left\{ \frac{1}{\delta_d(1+\rho) + |\psi_{d,1}^T x^{(t)}|}, \cdots, \frac{1}{\delta_d(1+\rho) + |\psi_{d,L_d}^T x^{(t)}|} \right\}, \forall d \tag{1.81}$$

$$\lambda_d^{(t+1)} = \left[\frac{1}{L_d} \log\left(1 + \rho + \frac{|\psi_{d,l}^T x^{(t)}|}{\delta_d} \right) \right]^{-1} + 1, \ \forall d, \tag{1.82}$$

which matches the steps in Algorithm 3 under $\varepsilon = 0$. This establishes Part 4 of Theorem 2.

1.3.6 Co-IRW-L1

Until now, we have considered the Co-IRW-L1-δ parameters δ 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-δ that adapts the δ vector at every iteration. The resulting procedure, which we will refer to as Co-IRW-L1, is summarized in Algorithm 4.

In the case of real-valued x, the expression for $\log p(x;\lambda,\delta)$ in line 6 of Algorithm 4 is given in (1.75) for $\lambda_d > 1$ and $\delta_d > 0$. Although there does not appear to be a closed-form solution to the joint maximization problem in line 6, it is over two real parameters and thus can be solved numerically without a significant computational burden.

Algorithm 4 The Co-IRW-L1 Algorithm

```
1: input: $\{\Psi_d\}_{d=1}^D$, $\Phi$, $y$, $\varepsilon \geq 0, \rho \geq 0$
2: if $x \in \mathbb{R}^n$, use $\Lambda = (1,\infty)$ and $\log p(x;\lambda,\delta)$ from (1.75);
   if $x \in \mathbb{C}^n$, use $\Lambda = (2,\infty)$ and $\log p(x;\lambda,\delta)$ from (1.84).
3: initialization: $\lambda_d^{(1)} = 1, W_d^{(1)} = I, \; \forall d \in [D]$
4: for $t = 1,2,3,\ldots$
5:     $x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^D \lambda_d^{(t)} \|W_d^{(t)} \Psi_d x\|_1 \;\text{ s.t. }\; \|y - \Phi x\|_2 \leq \varepsilon$
6:     $(\lambda_d^{(t+1)}, \delta_d^{(t+1)}) \leftarrow \arg\max_{\lambda_d \in \Lambda, \delta_d > 0} \log p(x^{(t)};\lambda,\delta), \; d \in [D]$
7:     $W_d^{(t+1)} \leftarrow \text{diag}\left\{ \frac{1}{\delta_d^{(t+1)}(1+\rho) + |\psi_{d,1}^T x^{(t)}|}, \ldots, \frac{1}{\delta_d^{(t+1)}(1+\rho) + |\psi_{d,L_d}^T x^{(t)}|} \right\}, \; d \in [D]$
8: end
9: output: $x^{(t)}$
```

Algorithm 4 can be interpreted as a generalization of the VEM approach to Co-IRW-L1-δ that is summarized in Part 4 of Theorem 2 and detailed in Sec. 1.3.5. Whereas Co-IRW-L1-δ used VEM to estimate the λ parameters in the prior (1.58) for a fixed value of δ, Co-IRW-L1 uses VEM to *jointly* estimate (λ, δ) in (1.58). Thus, Co-IRW-L1 can be derived by repeating the steps in Sec. 1.3.5, except that now the maximization of $\log p(x;\lambda,\delta)$ in (1.75) is performed jointly over (λ,δ), as reflected by line 6 of Algorithm 4.

1.3.7 Co-IRW-L1 for Complex-Valued x

In Sections 1.3.1-1.3.6, the signal x was assumed to be real-valued. We now extend the previous results to the case of complex-valued x. For this, we focus on the Co-IRW-L1 algorithm, since Co-IRW-L1-δ follows as the special case where δ is fixed at a user-supplied value.

Recalling that Co-IRW-L1 was constructed by generalizing the VEM interpretation of Co-IRW-L1-δ, we reconsider this VEM interpretation for the case of complex-valued x. In particular, we assume an AWGN likelihood and the following complex-valued extension of the prior (1.58):

$$p(x;\lambda,\delta) \propto \prod_{d=1}^{D} \prod_{l=1}^{L_d} \frac{(\lambda_d - 1)(\lambda_d - 2)}{2\pi\delta_d^2} \left(1 + \rho + \frac{|\psi_{d,l}^T x|}{\delta_d}\right)^{-\lambda_d}, \tag{1.83}$$

which is now i.i.d. generalized-Pareto on $z_d = \Psi_d x \in \mathbb{C}^{L_d}$ with deterministic shape parameter $\lambda_d > 2$ and deterministic scale parameter $\delta_d > 0$. In this case, the log-prior (1.75) changes to

$$\log p(x;\lambda,\delta) \\ = \text{const} + \sum_{d=1}^{D}\sum_{l=1}^{L_d}\left[\log\left(\frac{(\lambda_d-1)(\lambda_d-2)}{\delta_d^2}\right) - \lambda_d \log\left(1+\rho+\frac{|\psi_{d,l}^T x|}{\delta_d}\right)\right], \quad (1.84)$$

which is then maximized over (λ,δ) in line 6 of Algorithm 4.

1.4 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) x from m-sample noisy compressed measurements $y = \Phi x + e$, 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. 1.1.4, Co-L1 can be considered as the composite extension of the standard L1-regularized L2-constrained approach to analysis CS, i.e., (1.2) under the non-composite L1 regularizer $R(x) = \|\Psi x\|_1$. Similarly, Co-IRW-L1 can be considered as the composite extension of the standard IRW approach to the same L1 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.

1.4.1 Experimental Setup

For the dynamic MRI experiment, we constructed Φ using randomly sub-sampled Fourier measurements at each time instant with a varying sampling pattern across time. More details are given in Sec. 1.4.4. For the other experiments, we used a “spread spectrum” operator [39] of the form $\Phi = DFC$, where $C \in \mathbb{R}^{n\times n}$ is diagonal matrix with i.i.d equiprobable ± 1 entries, $F \in \mathbb{C}^{n\times n}$ is the discrete Fourier transform (DFT), and $D \in \mathbb{R}^{m\times n}$ is a row-selection operator that selects m rows of $FC \in \mathbb{C}^{n\times n}$ uniformly at random.

In all cases, the noise e 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 $\|y\|_2^2/(m\sigma^2)$ and the recovery SNR of signal estimate $\widehat{x}$ as $\|x\|_2^2/\|x-\widehat{x}\|_2^2$.

Note that, when x is real-valued, the measurements y will be complex-valued due to the construction of Φ. Thus, to allow the use of real-valued L1 solvers, we split each complex-valued element of y (and the corresponding rows of Φ and e)

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 FCx, we ensured that D selected at most one sample from each complex-conjugate pair.

To implement the existing non-composite L1 and IRW-L1 methods, we used the Matlab codes linked[2] to the paper [14], which are based on Douglas-Rachford splitting [18]. All default settings were retained except that the maximum number of reweighting iterations was increased from 10 to 25, which resulted in improved recovery SNR. Then, to implement the weighted-ℓ_1 minimization step in Co-L1 and Co-IRW-L1, we used a similar Douglas-Rachford splitting technique. The maximum number of reweighting iterations for Co-L1 and Co-IRW-L1 was set at 25. For Co-L1, IRW-L1, and Co-IRW-L1, the t-indexed iterations in Algorithm 1, Algorithm 2, and Algorithm 4, respectively, were terminated when $\|x^{(t)} - x^{(t-1)}\|_2 / \|x^{(t)}\|_2 < 1 \times 10^{-8}$. In all experiments we used $\varepsilon = 0.8\sqrt{\sigma^2 m}$ and $\delta = 0 = \rho$.

1.4.2 *Synthetic 2D Finite-Difference Signals*

Our first experiment aims to answer the following question. If we know that the sparsity of $\Psi_1 x$ differs from the sparsity of $\Psi_2 x$, 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 (1.4).

To investigate this question, we constructed 2D signals with finite-difference structure in both the vertical and horizontal domains. In particular, we constructed $X = x_1 1^T + 1 x_2^T$, where both $x_1 \in \mathbb{R}^{48}$ and $x_2 \in \mathbb{R}^{48}$ are finite-difference signals and $1 \in \mathbb{R}^{48}$ contains only ones. The locations of the transitions in x_1 and 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 x_1 and x_2 was fixed at 28, but the ratio of the number of transitions in x_1 to the number in x_2, denoted by α, was varied from 1 to 27. The case $\alpha = 1$ corresponds to X having 14 vertical transitions and 14 horizontal transitions, while the case $\alpha = 27$ corresponds to X having 27 vertical transitions and a single horizontal transition. (See Fig. 1.1 for examples.) Finally, the signal $x \in \mathbb{R}^n$ appearing in our model (1.1) was created by vectorizing X, yielding a total of $n = 48^2 = 2304$ pixels.

Given x, noisy observations $y = \Phi x + e$ were generated using the random "spread spectrum" measurement operator Φ described earlier at a sampling ratio of $m/n = 0.3$, with additive white Gaussian noise (AWGN) e scaled to achieve a measurement SNR of 40 dB. All recovery algorithms used vertical and horizontal finite-difference operators Ψ_1 and Ψ_2, respectively, with $\Psi = [\Psi_1^T, \Psi_2^T]^T$ in the non-composite case.

Figure 1.2 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

[2] Matlab codes for [14] are provided at `https://github.com/basp-group/sopt`

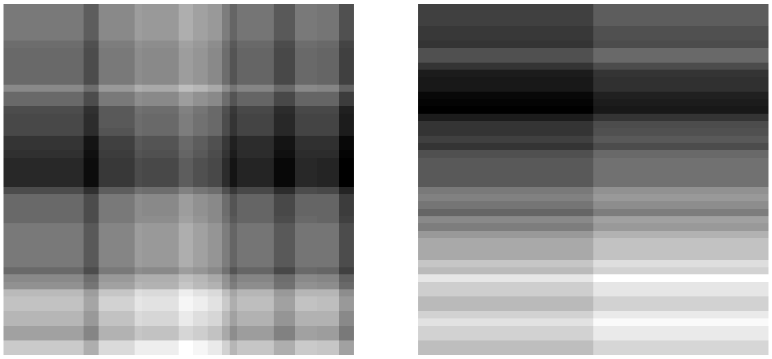


Fig. 1.1: Examples of the 2D finite-difference signal X used in the first experiment. On the left is a realization generated under a transition ratio of $\alpha = 14/14 = 1$, and on the right is a realization generated under $\alpha = 27/1 = 27$.

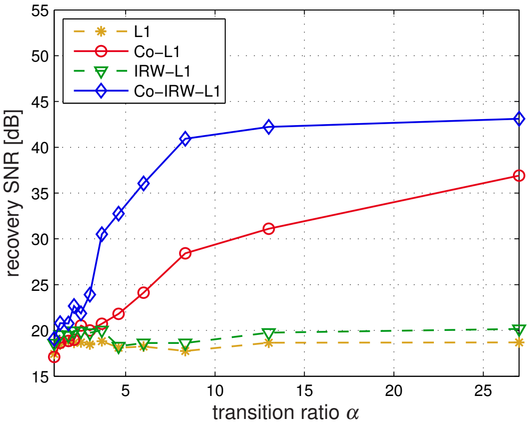


Fig. 1.2: Recovery SNR versus transition ratio α for the first experiment, which used 2D finite-difference signals, spread-spectrum measurements at $m/n = 0.3$, AWGN at 40 dB, and finite-difference operators for Ψ_d. Each recovery SNR represents the median value from 45 independent trials.

figure represents the median value from 45 trials, each using an independent realization of the triple (Φ, x, e). The figure shows that the recovery SNR of both L1 and IRW-L1 is roughly invariant to the transition ratio α, which makes sense because the overall sparsity of Ψx 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.

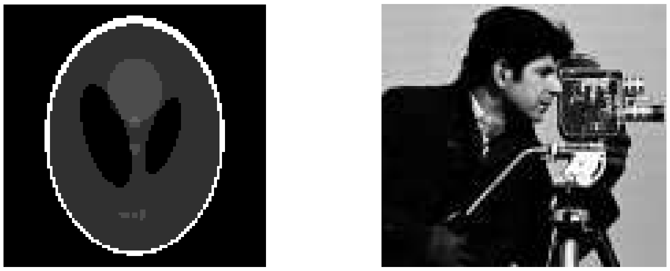


Fig. 1.3: Left: the Shepp-Logan phantom of size $n = 96 \times 96$. Right: the cropped Cameraman image of size $n = 96 \times 104$.

1.4.3 Shepp-Logan and Cameraman 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 $n = 96 \times 96$ Shepp-Logan phantom and the $n = 96 \times 104$ cropped Cameraman image shown in Fig. 1.3, and we constructed compressed noisy measurements y using spread-spectrum Φ and AWGN e at a measurement SNR of 30 dB in the Shepp-Logan case and 40 dB in the Cameraman case.

All algorithms used analysis operator $\Psi \in \mathbb{R}^{7n\times n}$ constructed from the undecimated Daubechies-1 2D wavelet transform (UWT-db1) with two levels of decomposition. However, the Co-L1 and Co-IRW-L1 algorithms treated each of the seven subbands of UWT-db1 as a separate sub-dictionary $\Psi_d \in \mathbb{R}^{n\times n}$ in their composite regularizers.

Fig. 1.4 shows recovery SNR versus sampling ratio m/n for the Shepp-Logan phantom, while Fig. 1.5 shows the same for the Cameraman image. Each recovery SNR represents the median value from 7 independent realizations of (Φ, e). 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 $m/n \geq 0.4$. Although not shown, similar results were observed with a level-three decomposition of UWT-db1, and at higher (50 dB) and lower (25 dB) measurement SNRs.

1.4.4 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 $\{y_t\}_{t=1}^{T}$ constructed as

$$y_t = \Phi_t x_t + e_t, \tag{1.85}$$

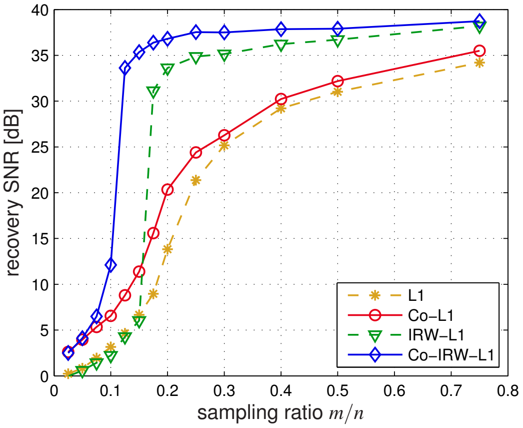


Fig. 1.4: Recovery SNR versus sampling ratio m/n for the Shepp-Logan phantom. Measurements were constructed using a spread-spectrum operator and AWGN at 30 dB SNR, and recovery used the UWT-db1 2D wavelet transform at two levels of decomposition. Each recovery SNR represents the median value from 7 independent trials.

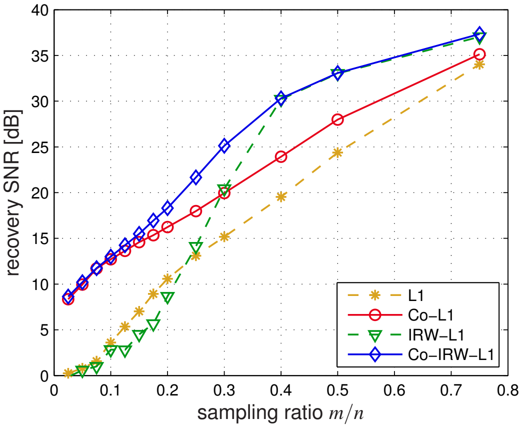


Fig. 1.5: Recovery SNR versus sampling ratio m/n for the cropped Cameraman image. Measurements were constructed using a spread-spectrum operator and AWGN at 40 dB SNR, and recovery used the UWT-db1 2D wavelet transform at two levels of decomposition. Each SNR value represents the median value from 7 independent trials.

where $x_t \in \mathbb{R}^{n_1 n_2}$ 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 $e_t \in \mathbb{R}^m_1$ 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

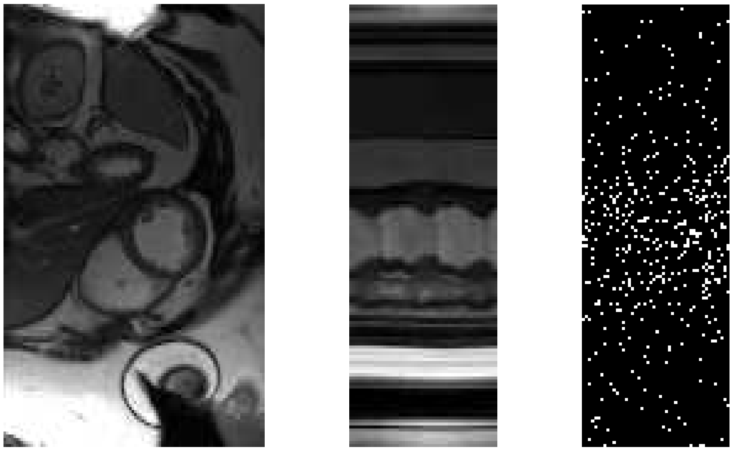


Fig. 1.6: Left: A 144×85 spatial slice from the $144 \times 85 \times 48$ dMRI dataset. Middle: The 144×48 spatio-temporal slice used for the dMRI experiment. Right: a realization of the variable-density k-space sampling pattern, versus time, at $m/n = 0.15$.

and to sample more densely at low frequencies, where most of the signal energy lies (e.g., [3]). Putting (1.85) into the form of our measurement model (1.1), we get

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} \Phi_1 & & \\ & \ddots & \\ & & \Phi_T \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} e_1 \\ \vdots \\ e_T \end{bmatrix}}_{e}, \tag{1.86}$$

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 x of dimensions $n_1 = 144$, $n_2 = 85$, and $T = 48$. The left pane in Fig. 1.6 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. 1.6, as opposed to the full 3D cine, from subsampled k-t-domain measurements. For this purpose, we constructed measurements $\{y\}_{t=1}^{T}$ 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. 1.6 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 $\Psi \in \mathbb{R}^{3n \times n}$ from a vertical concatenation of the db1-db3 Daubechies or-

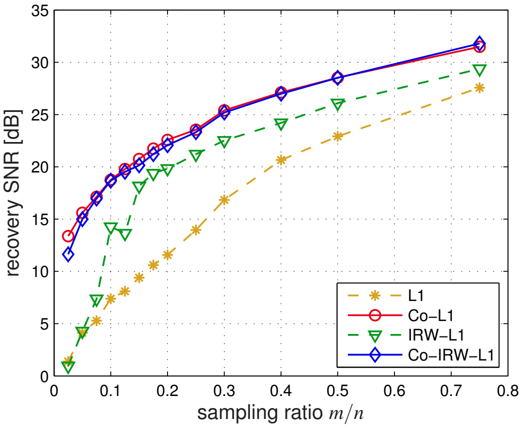
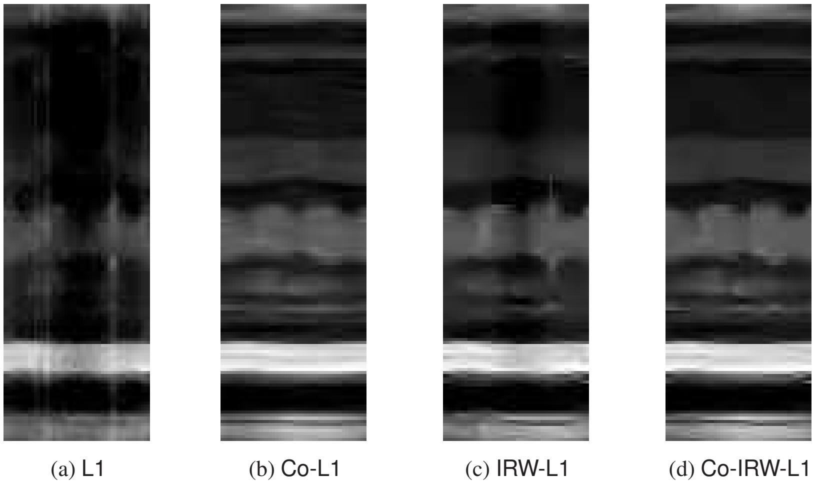


Fig. 1.7: Recovery SNR versus sampling ratio m/n for the dMRI experiment. Each SNR value represents the median value from 7 independent trials. Measurements were constructed using variable-density sub-sampled Fourier operator and AWGN at 30 dB measurement SNR, and recovery used a concatenation of db1-db3 orthogonal 2D wavelet bases at three levels of decomposition.

thogonal 2D discrete wavelet bases, each with three levels of decomposition. For the Co-L1 and Co-IRW-L1 algorithms, we assigned each of the 30 sub-bands in Ψ to a separate sub-dictionary $\Psi_d \in \mathbb{R}^{L_d \times n}$. 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. 1.7 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 (Φ, e). The figure shows that Co-L1 and Co-IRW-L1 outperform their non-composite counterparts by ≥ 2 dB at all tested values of m/n, with larger gains at small m/n. Interestingly, Co-L1 and Co-IRW-L1 gave nearly identical recovery SNR in this experiment, which suggests that—for each d—the analysis coefficients *within* $\Psi_d x$ were of a similar magnitude. Although not shown here, we obtained similar results with other cine datasets and with an UWT-db1-based analysis operator.

For qualitative comparison, Fig. 1.8 shows the spatio-temporal profile recovered by each of the four algorithms under test at $m/n = 0.15$ for a typical realization of (Φ, e). Compared to the ground-truth profile shown in the middle pane of Fig. 1.6, 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.



(a) L1 (b) Co-L1 (c) IRW-L1 (d) Co-IRW-L1

Fig. 1.8: Recovered dMRI spatio-temporal profiles at $m/n = 0.15$

	Shepp-Logan	Cameraman	MRI
L1	20.8	23.1	29.3
Co-L1	32.7	34.2	86.4
IRW-L1	45.9	48.4	54.1
Co-IRW-L1	72.1	96.4	131

Table 1.1: Computation times (in seconds) for the presented experimental studies. The times are averaged over trial runs and different sampling ratios.

1.4.5 Algorithm Runtime

Table 1.1 reports the average runtimes of the L1, Co-L1, IRW-L1, and Co-IRW-L1 algorithms for the experiments in Sections 1.4.3 and 1.4.4. There we see that the runtime of Co-L1 ranged between 1.5× to 3× that of L1, and the runtime of Co-IRW-L1 ranged between 1.5× to 3× the runtime of IRW-L1.

1.5 Conclusions

Motivated by the observation that a given signal x admits sparse representations in multiple dictionaries Ψ_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 [44, 17, 45] from the L1 penalty $\|\Psi x\|_1$ to composite L1 penalties of the form (1.4) while self-adjusting the regularization weights λ_d. Our second algorithm, Co-IRW-L1, extends the well-known IRW-L1 algorithm [13, 14] to the same family of composite penalties while self-adjusting the regularization weights λ_d and the regularization parameters δ_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 [34]. 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., 1.5×-3×) increase in runtime.

Acknowledgements This work has been supported in part by NSF grants CCF-1218754 and CCF-1018368 and DARPA grant N66001-11-1-4090. An early version of this work was presented at the 2015 ISMRM Annual Meeting and Exhibition.

References

1. Afonso, M.V., Bioucas-Dias, J.M., Figueiredo, M.A.T.: Fast image recovery using variable splitting and constrained optimization. IEEE Trans. Image Process. **19**(9), 2345–2356 (2010)
2. Ahmad, R., Schniter, P.: Iteratively reweighted ℓ_1 approaches to sparse composite regularization. IEEE Trans. Comp. Imag. **10**(2), 220–235 (2015)
3. Ahmad, R., Xue, H., Giri, S., Ding, Y., Craft, J., Simonetti, O.P.: Variable density incoherent spatiotemporal acquisition (VISTA) for highly accelerated cardiac MRI. Magnetic Resonance in Medicine pp. n/a–n/a (2014). DOI 10.1002/mrm.25507. URL http://dx.doi.org/10.1002/mrm.25507
4. Babacan, S.D., Nakajima, S., Do, M.N.: Bayesian group-sparse modeling and variational inference. IEEE Trans. Signal Process. **62**(11), 2906–2921 (2014)
5. Becker, S., Bobin, J., Candès, E.J.: NESTA: A fast and accurate first-order method for sparse recovery. SIAM J. Imag. Sci. **4**(1), 1–39 (2011)
6. Belge, M., Kilmer, M.E., Miller, E.L.: Efficient determination of multiple regularization parameters in a generalized L-curve framework. Inverse Problems **18**(4), 1161–1183 (2002)
7. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York (1985)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2007)
9. Borwein, J.M., Lewis, A.S.: Convex analysis and nonlinear optimization. Springer, New York (2006)
10. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2010)
11. Brezinski, C., Redivo-Zaglia, M., Rodriguez, G., Seatzu, S.: Multi-parameter regularization techniques for ill-conditioned linear systems. Numerische Mathematik **94**(2), 203–228 (2003)
12. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
13. Candès, E.J., Wakin, M.B., Boyd, S.: Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Anal. App. **14**(5), 877–905 (2008)

14. Carrillo, R.E., McEwen, J.D., Van De Ville, D., Thiran, J.P., Wiaux, Y.: Sparsity averaging for compressive imaging. IEEE Signal Process. Lett. **20**(6), 591–594 (2013)
15. Cevher, V.: Learning with compressible priors. In: Proc. Neural Inform. Process. Syst. Conf., pp. 261–269. Vancouver, B.C. (2009)
16. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressive sensing. In: Proc. IEEE Int. Conf. Acoust. Speech & Signal Process., pp. 3869–3872. Las Vegas, NV (2008)
17. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Scientific Comput. **20**(1), 33–61 (1998)
18. Combettes, P.L., Pesquet, J.C.: A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Sel. Topics Signal Process. **1**(4), 6564–574 (2007)
19. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery,. Commun. Pure & Appl. Math. **63**(1), 1–38 (2010)
20. Dempster, A., Laird, N.M., Rubin, D.B.: Maximum-likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. **39**, 1–17 (1977)
21. Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus synthesis in signal priors. Inverse Problems **23**, 947–968 (2007)
22. Eldar, Y.C., Kutyniok, G.: Compressed Sensing: Theory and Applications. Cambridge Univ. Press, New York (2012)
23. Figueiredo, M.A.: Adaptive sparseness for supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1150–1159 (2003)
24. Figueiredo, M.A.T., Nowak, R.D.: Wavelet-based image estimation: An empirical Bayes approach using Jeffreys' noninformative prior. IEEE Trans. Image Process. **10**(9), 1322–1331 (2001)
25. Figueiredo, M.A.T., Nowak, R.D.: Majorization-minimization algorithms for wavelet-based image restoration. IEEE Trans. Image Process. **16**(12), 2980–2991 (2007)
26. Fornasier, M., Naumova, V., Pereverzyev, S.V.: Multi-parameter regularization techniques for ill-conditioned linear systems. SIAM J. Numer. Anal. **52**(4), 1770–1794 (2014)
27. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser, New York (2013)
28. Gazzola, S., Novati, P.: Multi-parameter Arnoldi-Tikhonov methods. Electron. Trans. Numer. Anal. **40**, 452–475 (2013)
29. Hunter, D.R., Lange, K.: A tutorial on MM algorithms. The American Statistician **58**(1), 30–37 (2004)
30. Khajehnejad, M.A., Amin, M., Xu, W., Avestimehr, A.S., Hassibi, B.: Improved sparse recovery thresholds with two-step reweighted ℓ_1 minimization. In: Proc. IEEE Int. Symp. Inform. Thy., pp. 1603–1607 (2010)
31. Kowalski, M.: Sparse regression using mixed norms. Appl. Computational Harmonic Anal. **27**(2), 303–324 (2009)
32. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM J. Imag. Sci. **6**(2), 938–983 (2013)
33. Lu, S., Pereverzev, S.V.: Regularization Theory for Ill-posed Problems. Walter de Gruyter, Berlin (2013)
34. Mairal, J.: Optimization with first-order surrogate functions. In: Proc. Int. Conf. Mach. Learning, vol. 28, pp. 783–791 (2013)
35. Mairal, J., Bach, F., Ponce, J.: Sparse modeling for image and vision processing. Found. Trends Comput. Vision **8**(2-3), 85–283 (2014)
36. Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: M.I. Jordan (ed.) Learning in Graphical Models, pp. 355–368. MIT Press (1998)
37. Oliveira, J.P., Bioucas-Dias, J.M., Figueiredo, M.A.T.: Adaptive total variation image deblurring: A majorization-minimization approach. Signal Process. **89**(9), 1683–1693 (2009)
38. Poor, H.V.: An Introduction to Signal Detection and Estimation, 2nd edn. Springer, New York (1994)
39. Puy, G., Vandergheynst, P., Gribonval, R., Wiaux, Y.: Universal and efficient compressed sensing by spread spectrum and application to realistic Fourier imaging techniques. EURASIP J. Appl. Signal Process. **2012:6**, 1–13 (2012)

40. Rakotomamonjy, A.: Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms. Signal Process. **91**, 1505–1526 (2011)
41. Rao, B.D., Kreutz-Delgado, K.: An affine scaling methodology for best basis selection. IEEE Trans. Signal Process. **47**, 187–200 (1999)
42. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
43. Tan, Z., Eldar, Y., Beck, A., Nehorai, A.: Smoothing and decomposition for analysis sparse recovery. IEEE Trans. Signal Process. **62**(7), 1762–1774 (2014)
44. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. B **58**(1), 267–288 (1996)
45. Tibshirani, R.J.: Solution path of the generalized lasso. Ann. Statist. **39**(3), 1335–1371 (2011)
46. Wipf, D., Nagarajan, S.: Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. IEEE J. Sel. Topics Signal Process. **4**(2), 317–329 (2010)
47. Xu, P., Fukuda, Y., Liu, Y.: Multiple parameter regularization: numerical solutions and applications to the determination of geopotential from precise satellite orbits. J. Geodesy **80**(1), 17–27 (2006)