Received: 14 May 2019 | Accepted: 31 October 2020

DOI: 10.1111/rssb.12407

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Iterative Alpha Expansion for estimating gradient-sparse signals from linear measurements

Sheng Xu | **Zhou Fan**

Department of Statistics and Data Science, Yale University, New Haven, USA

Correspondence
Sheng Xu, Department of Statistics and Data Science, Yale University, New Haven, USA.
Email: sheng.xu@yale.edu

Abstract

We consider estimating a piecewise-constant image, or a gradient-sparse signal on a general graph, from noisy linear measurements. We propose and study an iterative algorithm to minimize a penalized least-squares objective, with a penalty given by the "ℓ_0-norm" of the signal's discrete graph gradient. The method uses a non-convex variant of proximal gradient descent, applying the alpha-expansion procedure to approximate the proximal mapping in each iteration, and using a geometric decay of the penalty parameter across iterations to ensure convergence. Under a cut-restricted isometry property for the measurement design, we prove global recovery guarantees for the estimated signal. For standard Gaussian designs, the required number of measurements is independent of the graph structure, and improves upon worst-case guarantees for total-variation (TV) compressed sensing on the 1-D line and 2-D lattice graphs by polynomial and logarithmic factors respectively. The method empirically yields lower mean-squared recovery error compared with TV regularization in regimes of moderate undersampling and moderate to high signal-to-noise, for several examples of changepoint signals and gradient-sparse phantom images.

1 | INTRODUCTION

Consider an unknown signal $\mathbf{x}_* \in \mathbb{R}^p$ observed via n noisy linear measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x}_* + \mathbf{e} \in \mathbb{R}^n.$$

We study the problem of estimating $\mathbf{x}_*$, under the assumption that its coordinates correspond to the p vertices of a given graph $G = (V, E)$, and $\mathbf{x}_*$ is gradient sparse. By this, we mean that

$$\|\nabla \mathbf{x}_*\|_0 \equiv \sum_{(i,j)\in E} \mathbf{1}\{x_{*,i} \neq x_{*,j}\} \tag{1}$$

is much smaller than the total number of edges $|E|$. Special cases of interest include the 1-D line graph, where variables have a sequential order and $\mathbf{x}_*$ has a changepoint structure, and the 2-D lattice graph, where coordinates of $\mathbf{x}_*$ represent pixels of a piecewise-constant image.

This problem has been studied since early pioneering works in compressed sensing (Candès et al., 2006a, b; Donoho, 2006). Among widely used approaches for estimating $\mathbf{x}_*$ are those based on constraining or penalizing the total-variation (TV) semi-norm (Rudin et al., 1992), which may be defined (anisotropically) for a general graph as

$$\|\nabla \mathbf{x}\|_1 \equiv \sum_{(i,j)\in E} |x_i - x_j|.$$

These are examples of ℓ_1-analysis methods (Candès et al., 2011; Elad et al., 2007; Nam et al., 2013), which regularize the ℓ_1-norm of a general linear transform of $\mathbf{x}$ rather than of its coefficients in an orthonormal basis. Related fused-lasso methods have been studied for different applications of regression and prediction in Tibshirani et al. (2005), Rinaldo (2009), Tibshirani (2011) and Padilla et al. (2017). Other graph-based regularization methods were studied in Krishnamuthy et al. (2013), Li et al. (2018) and Kim and Gao (2019), and generalizations to trend-filtering methods that regularize higher-order discrete derivatives of $\mathbf{x}$ were studied in Kim et al. (2009) and Wang et al. (2016).

The reconstruction error of TV regularization depends on the structure of the graph (Cai & Xu, 2015; Needell & Ward, 2013a, b). More generally, the error of ℓ_1-analysis methods with sparsifying transform ∇ depends on sparse conditioning properties of the pseudo-inverse $\nabla^\dagger$ (Candès et al., 2011). For direct measurements $\mathbf{A}=\mathbf{I}$, these and related issues were discussed in Hütter and Rigollet (2016), Dalalyan et al. (2017) and Fan and Guan (2018), which showed in particular that TV regularization may not achieve the same recovery guarantees as analogous ℓ_0-regularization methods on certain graphs including the 1-D line. In this setting of $\mathbf{A} = \mathbf{I}$, different computational approaches also exist to approximately minimize an ℓ_0-regularized objective on general graphs (Boykov et al., 1999; Kleinberg & Tardos, 2002; Xu et al., 2011).

Motivated by this line of work, our current paper studies an alternative to TV regularization in the more difficult setting of indirect linear measurements, where $\mathbf{A} \neq \mathbf{I}$. Our procedure is based similarly on the idea of minimizing a possibly non-convex objective

$$F(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_{(i,j)\in E} c(x_i, x_j) \tag{2}$$

for an edge-associated cost function c. We will focus attention in this work on the specific choice of an ℓ_0-regularizer

$$c(x_i, x_j) = \mathbf{1}\{x_i \neq x_j\}, \tag{3}$$

which matches Equation (1), although the algorithm may be applied with more general choices of metric edge cost. For the above ℓ_0 edge cost, the resulting objective takes the form

$$F(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\nabla \mathbf{x}\|_0.$$

For $\mathbf{A} = \mathbf{I}$, Fan and Guan (2018) analysed the alpha-expansion algorithm of Boykov et al. (1999) for minimizing this objective, and showed that it can achieve statistically rate-optimal estimation guarantees. We review this method in Section 2. Its algorithmic idea is specific to $\mathbf{A} = \mathbf{I}$, where the objective (2) decomposes as a sum of terms involving only individual variables x_i and pairs (x_i, x_j), and this idea does not easily extend to indirect linear measurements. In this work, we instead study an approach of applying this method to minimize $F(\mathbf{x})$ using a non-convex and non-smooth variant of proximal gradient descent: For parameters $\gamma \in (0, 1)$ and $\eta > 0$, we iteratively compute $\mathbf{x}_{k+1}$ from $\mathbf{x}_k$ via

$$\begin{aligned}
&\mathbf{a}_{k+1} \leftarrow \mathbf{x}_k - \eta \mathbf{A}^\top (\mathbf{A}\mathbf{x}_k - \mathbf{y}) \\
&\mathbf{x}_{k+1} \text{“}\leftarrow\text{”} \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{a}_{k+1}\|_2^2 + \lambda_k \sum_{(i,j)\in E} c(x_i, x_j) \\
&\lambda_{k+1} \leftarrow \lambda_k \cdot \gamma
\end{aligned}$$

The update for $\mathbf{x}_{k+1}$ is carried out approximately, using the alpha-expansion idea. We call this algorithm ITALE, for ITerative ALpha Expansion.

There are two important differences between ITALE and standard proximal gradient methods for convex problems (Beck & Teboulle, 2009; Parikh & Boyd, 2014). First, since the edge cost $c(x_i, x_j)$ is non-convex, the minimization problem for updating $\mathbf{x}_{k+1}$ is also non-convex. That such an algorithm should converge is not as evident as for proximal gradient methods applied with convex penalties. Second, to ensure that the algorithm indeed converges, we must start with a large initialization for the penalty $\lambda_{\max}$ and geometrically decay this penalty across iterations. This is the case even if we were only interested in one final tuning parameter λ in the objective (2). This type of penalty decay was studied previously in a convex setting by Xiao and Zhang (2013), but the purpose there was to improve the convergence rate rather than to ensure convergence.

In practice, for γ sufficiently close to 1, we directly interpret the sequence of ITALE iterates $\mathbf{x}_k$ as approximate minimizers of the objective function (2) for penalty parameters $\lambda = \lambda_k / \eta$ along a regularization path. We comment more on this approach in Section 2. We select the iterate k using cross-validation on the prediction error for $\mathbf{y}$, and we use the final estimate $\widehat{\mathbf{x}}^{\text{ITALE}} = \mathbf{x}_k$.

Despite $F(\mathbf{x})$ being non-convex and non-smooth, we provide global recovery guarantees for ITALE. For example, under exact gradient sparsity $\|\nabla \mathbf{x}_*\|_0 = s_*$, if $\mathbf{A}$ consists of

$$n \gtrsim s_* \log(1 + |E| / s_*) \tag{4}$$

linear measurements with i.i.d. $\mathcal{N}(0, 1/n)$ entries, then the ITALE iterate $\mathbf{x}_k$ for the ℓ_0-regularizer (3) and a penalty value $\lambda_k \asymp \|\mathbf{e}\|_2^2 / s_*$ satisfies, with high probability,

$$\|\mathbf{x}_k - \mathbf{x}_*\|_2 \lesssim \|\mathbf{e}\|_2. \tag{5}$$

More generally, we provide recovery guarantees when $\mathbf{A}$ satisfies a certain cut-restricted isometry property, described in Definition 1 below. Note that Equation (5) is the optimal worst-case error guarantee for deterministic measurement errors $\mathbf{e}$, which is the typical setting studied in the compressed sensing literature (Blumensath & Davies, 2009; Candès et al., 2006a, b; Needell & Tropp, 2009) and also the setting that we study in this work.

Even for i.i.d. Gaussian design, we are not aware of previous polynomial-time algorithms which provably achieve this guarantee for either the 1-D line or the 2-D lattice. In particular, connecting with the previous discussion, similar existing results for TV regularization in noisy or noiseless settings require $n \gtrsim s_* (\log |E|)^3$ Gaussian measurements for the 2-D lattice and $n \gtrsim \sqrt{|E| s_*} \log |E|$

measurements for the 1-D line (Cai & Xu, 2015; Needell & Ward, 2013b). Applying thresholding or ℓ_1-regularization instead to a representation of $\mathbf{x}_*$ in a spanning tree wavelet basis, as proposed and studied in Padilla et al. (2017), would reduce this requirement for n to be optimal up to a logarithmic factor. The requirement for n in ITALE is instead optimal up to a constant factor, for any bounded-degree graph.

Figure 1 compares in simulation $\widehat{\mathbf{x}}^{\mathrm{ITALE}}$ using the ℓ_0-regularizer (3) with $\widehat{\mathbf{x}}^{\mathrm{TV}}$ (globally) minimizing the TV-regularized objective

$$F^{\mathrm{TV}}(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\nabla\mathbf{x}\|_1. \tag{6}$$

The example depicts a synthetic image of a human chest slice, previously generated by Gong et al. (2017) using the XCAT digital phantom (Segars et al., 2010). The design $\mathbf{A}$ is an undersampled and reweighted Fourier matrix, using a sampling scheme described in Section 3 and similar to that proposed in Krahmer and Ward (2014) for TV-regularized compressed sensing. In a low-noise setting, a detailed comparison of the recovered images reveals that $\widehat{\mathbf{x}}^{\mathrm{ITALE}}$ provides a sharper reconstruction than $\widehat{\mathbf{x}}^{\mathrm{TV}}$. As noise increases, $\widehat{\mathbf{x}}^{\mathrm{TV}}$ becomes blotchy, while $\widehat{\mathbf{x}}^{\mathrm{ITALE}}$ begins to lose finer image details. Quantitative comparisons of recovery error are provided in Section 4.2 and are favourable towards ITALE in lower noise regimes.

ITALE is similar to some methods oriented towards ℓ_0-regularized sparse regression and signal recovery (Bertsimas et al., 2016; Tropp & Gilbert, 2007; Zhang, 2011), including notably the Iterative Hard Thresholding (IHT) Blumensath and Davies (2009) and CoSaMP Needell and Tropp (2009) methods in compressed sensing. We highlight here several differences:

1. For sparsity in an orthonormal basis, forward stepwise selection and orthogonal matching pursuit provide greedy "ℓ_0" approaches to variable selection, also with provable guarantees

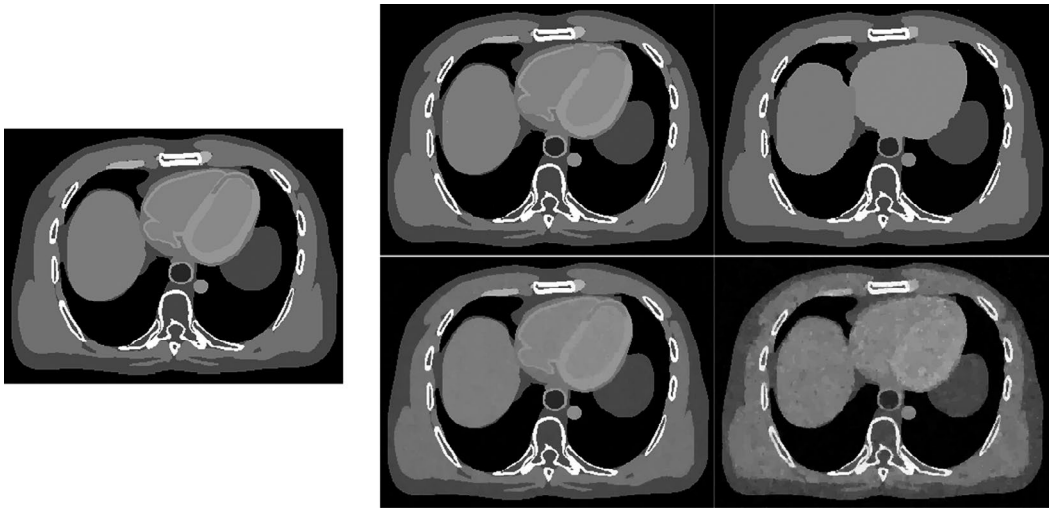

FIGURE 1 Left: Original image slice from the XCAT digital phantom. Top row: $\widehat{\mathbf{x}}^{\mathrm{ITALE}}$ from 20% undersampled and reweighted Fourier measurements, in low noise ($\sigma = 4$, left) and medium noise ($\sigma = 16$, right) settings. Bottom row: $\widehat{\mathbf{x}}^{\mathrm{TV}}$ for the same measurements. The iterate k in ITALE and tuning parameter λ for TV were both selected using fivefold cross-validation on the squared prediction error for $\mathbf{y}$

(Elenberg et al., 2018; Tropp & Gilbert, 2007; Zhang, 2011). However, such methods do not have direct analogues for gradient sparsity in graphs, as one cannot select a single edge difference $x_i - x_j$ to be non-zero without changing other edge differences.

2. IHT and CoSaMP enforce sparsity of $\mathbf{x}_{k+1}$ in each iteration by projecting to the s largest coordinates of $\mathbf{a}_{k+1}$, for user-specified s. In contrast, ITALE uses a Lagrangian form that penalizes (rather than constrains) $\|\nabla \mathbf{x}_{k+1}\|_0$. This is partly for computational reasons, as we are not aware of fast algorithms that can directly perform such a projection step onto the (non-convex) set $\{\mathbf{x}\colon \|\nabla \mathbf{x}\|_0 \leq s\}$ for general graphs. This Lagrangian form complicates the theoretical convergence analysis, as it requires establishing simultaneous control of the gradient sparsity $\|\nabla \mathbf{x}_k\|_0$ and the error $\|\mathbf{x}_k - \mathbf{x}_*\|_2$ in each iteration.
3. In contrast to general-purpose mixed-integer optimization procedures studied in Bertsimas et al. (2016), each iterate of ITALE (and hence also the full algorithm, for a polynomial number of iterations) is provably polynomial-time in the input size $(n, p, |E|)$ (Fan & Guan, 2018). On our personal computer, for the $p = 360 \times 270 = 97200$ image of Figure 1, computing the 60 iterates constituting a full ITALE solution path required about 20 min, using the optimized alpha-expansion code of Boykov and Kolmogorov (2004).

While our theoretical focus is on ℓ_0-regularization, we expect that for certain regimes of undersampling and signal-to-noise, improved empirical recovery may be possible with edge costs $c(x_i, x_j)$ interpolating between the ℓ_0 and ℓ_1 penalties. These are applicable in the ITALE algorithm and would be interesting to investigate in future work.

2 | MODEL AND ALGORITHM

Let $G = (V, E)$ be a given connected graph on the vertices $V = \{1, \ldots, p\}$, with undirected edge set E. We assume throughout that $p \geq 3$. For a signal vector $\mathbf{x}_* \in \mathbb{R}^p$, measurement matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$, and measurement errors $\mathbf{e} \in \mathbb{R}^n$, we observe

$$\mathbf{y} = \mathbf{A}\mathbf{x}_* + \mathbf{e} \in \mathbb{R}^n. \tag{7}$$

Denote by $\nabla \in \{-1, 0, 1\}^{|E| \times p}$ the discrete gradient matrix on the graph G, defined by

$$\nabla \mathbf{x} = (x_i - x_j\colon\ (i,j) \in E) \in \mathbb{R}^{|E|}.$$

Here, we may fix an arbitrary ordering of the vertex pair (i, j) for each edge. We study estimation of $\mathbf{x}_*$, assuming that $\mathbf{x}_*$ has (or is well approximated by a signal having) small exact gradient sparsity $\|\nabla \mathbf{x}_*\|_0$.

Our proposed algorithm is an iterative approach called ITALE, presented as Algorithm 1. It is based around the idea of minimizing the objective (2). In this objective, the cost function $c\colon \mathbb{R}^2 \to \mathbb{R}$ must satisfy the metric properties

$$c(x,y) = c(y,x) \geq 0, \quad c(x,x) = 0 \Leftrightarrow x = 0, \quad c(x,z) \leq c(x,y) + c(y,z), \tag{8}$$

but is otherwise general. Importantly, c may be non-smooth and non-convex. The algorithm alternates between constructing a surrogate signal $\mathbf{a}_{k+1}$ in line 3, denoising this surrogate signal in line 4, and geometrically decaying the penalty parameter λ_k used for the denoiser in line 5. We discuss these steps in more detail below.

The surrogate signal $\mathbf{a}_{k+1}$ that is computed in line 3 may be written as

$$\begin{aligned}\mathbf{a}_{k+1} &= \mathbf{x}_k - \eta \mathbf{A}^\top (\mathbf{A}\mathbf{x}_k - \mathbf{y}) \\ &= \mathbf{x}_* + (\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A})(\mathbf{x}_k - \mathbf{x}_*) + \eta \mathbf{A}^\top \mathbf{e}.\end{aligned}$$

This is a noisy version of the true signal $\mathbf{x}_*$, with two sources of noise $(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A})(\mathbf{x}_k - \mathbf{x}_*)$ and $\eta \mathbf{A}^\top \mathbf{e}$. Line 4 denoises this signal by applying the alpha-expansion graph cut procedure from Boykov et al. (1999) to approximately solve the minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^p} \frac{1}{2} \| \mathbf{x} - \mathbf{a}_{k+1} \|_2^2 + \lambda_k \sum_{(i,j) \in E} c(x_i, x_j).$$

This sub-routine is denoted as $\text{AlphaExpansion}(\mathbf{a}_{k+1}, \lambda_k, \delta)$, and is described in Algorithm 2 for completeness. At a high level, the alpha-expansion method encodes the above objective function in the structure of an edge-weighted augmented graph, and iterates over global moves that swap the signal value on a subset of vertices for a given new value, by finding a minimum graph cut. The original alpha-expansion algorithm of Boykov et al. (1999) is in the setting of a discrete Potts model. To apply this to a continuous signal domain, we restrict coordinate values of $\mathbf{x}$ to a discrete grid

$$\delta \mathbb{Z} = \{ k\delta : k \in \mathbb{Z} \}$$

for a small user-specified parameter $\delta > 0$.

The geometric decay of λ_k in line 5 may be understood by examining the two sources of error $(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A})(\mathbf{x}_k - \mathbf{x}_*)$ and $\eta \mathbf{A}^\top \mathbf{e}$ in $\mathbf{a}_{k+1}$. Assuming that $\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A}$ has a small operator norm when restricted to gradient-sparse vectors, the first error term decays geometrically across iterations, whereas the second error term is fixed in every iteration. When $\mathbf{e} \neq 0$, this suggests choosing λ_k to also decay geometrically up to a final positive constant $\lambda_* > 0$, after which we may fix $\lambda_k = \lambda_*$ and run the iterations to convergence. In this approach, the best choice for λ_* would depend on the size of $\eta \mathbf{A}^\top \mathbf{e}$, and this may be set in practice using cross-validation.

We do not directly use this approach, because this requires a separate run for each different value of λ_* to perform the cross-validation. Instead, Algorithm 1 performs only a single proximal gradient step for each λ_k, starting from a value $\lambda_{\max} > \lambda_*$ that oversmooths the surrogate signal and ending at a value $\lambda_{\min} < \lambda_*$ that undersmooths the surrogate signal (when $\mathbf{e} \neq 0$). For γ sufficiently close to 1, we directly interpret each iterate $\mathbf{x}_k$ as an approximate minimizer of the objective (2) for a different penalty $\lambda \equiv \lambda_k / \eta$. We apply cross-validation to select the iterate $\mathbf{x}_k$ that represents the final estimate $\widehat{\mathbf{x}}^{\text{ITALE}}$, and this corresponds to selecting a penalty λ in Equation (2). Thus, Algorithm 1 computes an estimate for each tuning parameter along a regularization path, in a single pass of the proximal gradient descent. We find that this works well in practice and yields substantial savings in computational cost, and our theoretical analysis will also be for the algorithm in this form.

Algorithm 1 Iterative Alpha Expansion

Input: $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and parameters $\gamma \in (0, 1)$, $\lambda_{\max} > \lambda_{\min} > 0$, and $\eta, \delta > 0$.

1: Initialize $\mathbf{x}_0 \leftarrow \mathbf{0}$, $\lambda_0 \leftarrow \lambda_{\max}$

2: **for** $k = 0, 1, 2, \ldots, K$ until $\lambda_K < \lambda_{\min}$ **do**

3: $\quad \mathbf{a}_{k+1} \leftarrow \mathbf{x}_k - \eta \mathbf{A}^\top (\mathbf{A}\mathbf{x}_k - \mathbf{y})$

4: $\mathbf{x}_{k+1} \leftarrow \text{AlphaExpansion}(\mathbf{a}_{k+1}, \lambda_k, \delta)$

5: $\lambda_{k+1} \leftarrow \lambda_k \cdot \gamma$

6: **end for**

Output: $\mathbf{x}_1, \ldots, \mathbf{x}_K$

Algorithm 2 AlphaExpansion($\mathbf{a}$, λ, δ) subroutine

Input: $\mathbf{a} \in \mathbb{R}^p$, cost function c: $\mathbb{R}^2 \to \mathbb{R}$, parameters $\lambda, \delta > 0$.

1: Let $a_{\min}, a_{\max}$ be the minimum and maximum values of $\mathbf{a}$. Initialize $\mathbf{x} \in \mathbb{R}^p$ arbitrarily.

2: **loop**

3: **for** each $z \in \delta\mathbb{Z} \cap [a_{\min}, a_{\max}]$ **do**

4: Construct the following edge-weighted augmentation $G_{z,\mathbf{x}}$ of the graph G:

5: Introduce a source vertex s and a sink vertex t, connect s to each $i \in \{1, \ldots, p\}$ with weight $\frac{1}{2}(a_i - z)^2$, and connect t to each $i \in \{1, \ldots, p\}$ with weight $\frac{1}{2}(a_i - x_i)^2$ if $x_i \neq z$, or weight ∞ if $x_i = z$.

6: **for** each edge $\{i, j\} \in E$ **do**

7: **if** $x_i = x_j$ **then**

8: Assign weight $\lambda c(x_i, z)$ to $\{i, j\}$.

9: **else**

10: Introduce a new vertex $v_{i,j}$, and replace edge $\{i, j\}$ by the three edges $\{i, v_{i,j}\}, \{j, v_{i,j}\}$, and $\{t, v_{i,j}\}$, with weights $\lambda c(x_i, z)$, $\lambda c(x_j, z)$, and $\lambda c(x_i, x_j)$ respectively.

11: **end if**

12: **end for**

13: Find the minimum s-t cut (S, T) of $G_{z,\mathbf{x}}$ such that $s \in S$ and $t \in T$.

14: For each $i \in \{1, \ldots, p\}$, update $x_i \leftarrow z$ if $i \in T$, and keep x_i unchanged if $i \in S$.

15: **end for**

16: If $\mathbf{x}$ was unchanged for each z above, then return $\mathbf{x}$.

17: **end loop**

Output: $\mathbf{x}$

We make a few additional remarks regarding parameter tuning in practice:

1. Using conservative choices for $\lambda_{\max}$ (large), γ (close to 1), and δ (small) increases the total runtime of the procedure, but does not degrade the quality of recovery. In our experiments, we fix $\gamma = 0.9$ and set δ in each iteration to yield 300 grid values for $\delta\mathbb{Z} \cap [a_{\min}, a_{max}]$ in Algorithm 2.
2. We do not specify $\lambda_{\min}$. Instead, we monitor the gradient sparsity $\|\nabla \mathbf{x}_k\|_0$ across iterations, and terminate the algorithm when $\|\nabla \mathbf{x}_K\|_0$ exceeds a certain fraction (e.g. 50%) of the total number of edges $|E|$.
3. The parameter η should be matched to the scaling and restricted isometry properties of the design matrix $\mathbf{A}$. For sub-Gaussian and Fourier designs scaled by $1/\sqrt{n}$ as in Propositions 1 and 2 below, we set $\eta = 1$.
4. The most important tuning parameter is the iterate k for which we take the final estimate $\hat{\mathbf{x}}^{\text{ITALE}} = \mathbf{x}_k$. In our examples, we apply fivefold cross-validation on the mean-squared prediction error for $\mathbf{y}$ to select k. Note that η should be rescaled by the number of training samples in each fold, for example,

for fivefold cross-validation with training sample size $0.8n$, we set $\eta = 1/0.8$ instead of $\eta = 1$ in the cross-validation runs.

3 | RECOVERY GUARANTEES

We provide in this section theoretical guarantees on the recovery error $\| \widehat{\mathbf{x}}^{\mathrm{ITALE}} - \mathbf{x}_* \|_2$, where $\widehat{\mathbf{x}}^{\mathrm{ITALE}} \equiv \mathbf{x}_k$ for a deterministic (non-adaptive) choice of iterate k. Throughout this section, ITALE is assumed to be applied with the ℓ_0 edge cost $c(x_i, x_j) = \mathbf{1}\{x_i \neq x_j\}$.

3.1 | cRIP condition

Our primary assumption on the measurement design $\mathbf{A}$ will be the following version of a restricted isometry property.

Definition 1 Let $\kappa > 0$, and let ρ: $[0, \infty) \to [0, \infty)$ be any function satisfying $\rho'(s) \geq 0$ and $\rho''(s) \leq 0$ for all $s > 0$. A matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ satisfies the (κ, ρ)-**cut-restricted isometry property** (cRIP) if, for every $\mathbf{x} \in \mathbb{R}^p$ with $\|\nabla \mathbf{x}\|_0 \geq 1$, we have

$$\left(1 - \kappa - \sqrt{\rho(\|\nabla \mathbf{x}\|_0)}\right) \|\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}\|_2 \leq \left(1 + \kappa + \sqrt{\rho(\|\nabla \mathbf{x}\|_0)}\right) \|\mathbf{x}\|_2.$$

This definition depends implicitly on the structure of the underlying graph G, via its discrete gradient matrix ∇. Examples of the function ρ are given in the two propositions below.

This condition is stronger than the usual RIP condition in compressed sensing (Candès et al., 2006a, b); in two ways: First, Definition 1 requires quantitative control of $\|\mathbf{A}\mathbf{x}\|_2$ for *all* vectors $\mathbf{x} \in \mathbb{R}^p$, rather than only those with sparsity $\|\nabla \mathbf{x}\|_0 \leq s$ for some specified s. We use this in our analysis to handle regularization of $\|\nabla \mathbf{x}\|_0$ in Lagrangian (rather than constrained) form. Second, approximate isometry is required for signals with small gradient sparsity $\|\nabla \mathbf{x}\|_0$, rather than small sparsity $\|\mathbf{x}\|_0$. This requirement is similar to the D-RIP condition of Candès et al. (2011) for general sparse analysis models, and is also related to the condition of Needell and Ward (2013b) that $\mathbf{A}\mathcal{H}^{-1}$ satisfies the usual RIP condition, where $\mathcal{H}^{-1}$ is the inverse Haar-wavelet transform on the 2-D lattice.

Despite this strengthening of the required RIP condition, the following shows that Definition 1 still holds for sub-Gaussian designs $\mathbf{A}$. For a random vector $\mathbf{a}$, we denote its sub-Gaussian norm as $\|\mathbf{a}\|_{\psi_2} = \sup_{\mathbf{u}:\|\mathbf{u}\|_2=1} \sup_{k \geq 1} k^{-1/2} \mathbb{E}[|\mathbf{u}^{\mathsf{T}}\mathbf{a}|^k]^{1/k}$, and say that $\mathbf{a}$ is sub-Gaussian if $\|\mathbf{a}\|_{\psi_2} \leq K$ for a constant $K > 0$.

Proposition 1 *Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ have i.i.d. rows $\mathbf{a}_i / \sqrt{n}$, where $\mathrm{Cov}[\mathbf{a}_i] = \mathbf{\Sigma}$ and $\|\mathbf{a}_i\|_{\psi_2} \leq K$. Suppose that the largest and smallest eigenvalues of $\mathbf{\Sigma}$ satisfy $\sigma_{\max}(\mathbf{\Sigma}) \leq (1 + \kappa)^2$ and $\sigma_{\min}(\mathbf{\Sigma}) \geq (1 - \kappa)^2$ for a constant $\kappa \in (0, 1)$. Then for any $\mathrm{k} > 0$ and some constant $\mathrm{C} > 0$ depending only on K, κ,k, with probability at least $1 - |E|^{-k}$, the matrix $\mathbf{A}$ satisfies (κ, ρ)-cRIP for the function*

$$\rho(s) = \frac{Cs \log(1 + |E|/s)}{n}.$$

Here, κ depends on the condition number of the design covariance, and $\rho(s)$ does not depend on the structure of the graph other than its total number of edges. The proof is a standard union bound argument, which we defer to Appendix B of the online supplementary material.

For large 2-D images, using Fourier measurements with matrix multiplication implemented by an FFT can significantly reduce the runtime of Algorithm 1. As previously discussed in Lustig et al. (2007), Needell and Ward (2013b) and Krahmer and Ward (2014), uniform random sampling of Fourier coefficients may not be appropriate for reconstructing piecewise-constant images, as these typically have larger coefficients in the lower Fourier frequencies. We instead study a non-uniform sampling and reweighting scheme similar to that proposed in Krahmer and Ward (2014) for TV compressed sensing, and show that Definition 1 also holds for this reweighted Fourier matrix.

For $p = N_1 N_2$ and N_1, N_2 both powers of 2, let $\mathcal{F} \in \mathbb{C}^{p \times p}$ be the 2-D discrete Fourier matrix on the lattice graph G of size $N_1 \times N_2$, normalized such that $\mathcal{F}\mathcal{F}^* = \mathbf{I}$. We define this as the Kronecker product $\mathcal{F} = \mathcal{F}^1 \otimes \mathcal{F}^2$, where $\mathcal{F}^1 \in \mathbb{C}^{N_1 \times N_1}$ is the 1-D discrete Fourier matrix with entries

$$\mathcal{F}^1_{jk} = \frac{1}{\sqrt{N_1}} \cdot e^{2\pi \mathbf{i} \cdot \frac{(j-1)(k-1)}{N_1}},$$

and $\mathcal{F}^2 \in \mathbb{C}^{N_2 \times N_2}$ is defined analogously. (Thus rows closer to $N_1/2 + 1$ in $\mathcal{F}^1$ correspond to higher frequency components.) Let $\mathcal{F}^*_{(i,j)}$ denote row (i, j) of $\mathcal{F}$, where we index by pairs $(i, j) \in \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$ corresponding to the Kronecker structure. We define a sampled Fourier matrix as follows: Let ν_1 be the probability mass function on $\{1, \ldots, N_1\}$ given by

$$\nu_1(i) \propto \frac{1}{C_0 + \min(i - 1, N_1 - i + 1)}, \quad C_0 \geq 1. \tag{9}$$

Define similarly ν_2 on $\{1, \ldots, N_2\}$, and let $\nu = \nu_1 \times \nu_2$. For a given number of measurements n, draw $(i_1, j_1), \ldots, (i_n, j_n) \overset{\text{iid}}{\sim} \nu$, and set

$$\tilde{\mathbf{A}} = \frac{1}{\sqrt{n}} \begin{pmatrix} \mathcal{F}^*_{(i_1,j_1)} / \sqrt{\nu(i_1, j_1)} \\ \vdots \\ \mathcal{F}^*_{(i_n,j_n)} / \sqrt{\nu(i_n, j_n)} \end{pmatrix} \in \mathbb{C}^{n \times p}. \tag{10}$$

Proposition 2 *Let* G *be the 2-D lattice graph of size $N_1 \times N_2$, where N_1, N_2 are powers of 2 and $1/K < N_1/N_2 < K$ for a constant $\mathrm{K} > 0$. Set $p = N_1 N_2$ and let $\tilde{\mathbf{A}}$ be the matrix defined in (10). Then for some constants C, $t_0 > 0$ depending only on* K*, and for any $t > t_0$, with probability at least $1 - e^{-(\log n)(\log p)^3} - p^{-t}$, $\tilde{\mathbf{A}}$ satisfies the (κ, ρ)-cRIP with $\kappa = 0$ and*

$$\rho(s) = Cts \frac{(\log p)^8 \log n}{n}.$$

The proof follows closely the ideas of Rudelson and Vershynin (2008, Theorem 3.3), and we defer this to Appendix B of the online supplementary material.

This proposition pertains to the complex analogue of Definition 1, where $\tilde{\mathbf{A}}$, $\mathbf{x}$ are allowed to be complex valued, and $\|\cdot\|_2$ denotes the complex ℓ_2-norm. For a real-valued signal $\mathbf{x}_* \in \mathbb{R}^p$, Algorithm 1 may be applied to $\tilde{\mathbf{y}} = \tilde{\mathbf{A}}\mathbf{x}_* + \mathbf{e} \in \mathbb{C}^n$ by separating real and imaginary parts of $\tilde{\mathbf{y}}$ into a real vector

$\mathbf{y} \in \mathbb{R}^{2n}$. The corresponding $\mathbf{A} \in \mathbb{R}^{2n\times p}$ satisfies $\|\mathbf{A}\mathbf{x}\|_2^2 = \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2$, so the same cRIP condition holds (in the real sense) for $\mathbf{A}$.

3.2 | Recovery error bounds

To illustrate the idea of analysis, we first establish a result showing that ITALE can yield exact recovery in a setting of no measurement noise. We require $\mathbf{x}_*$ to be gradient sparse with coordinates belonging exactly to $\delta\mathbb{Z}$, as the ITALE output has this latter property. Discretization error will be addressed in our subsequent result.

Theorem 1 *Suppose* $\mathbf{e}=\mathbf{0}$ *and* $\mathbf{x}_* \in (\delta\mathbb{Z})^p$, *and denote* $s_* = \max(\|\nabla\mathbf{x}_*\|_0, 1)$. *Suppose* $\sqrt{\eta}\cdot\mathbf{A}$ *satisfies* (κ, ρ)*-cRIP, where* $\kappa \in [0, \sqrt{3/2}-1)$. *Set* $t(\kappa) = 1-4\kappa-2\kappa^2 \in (0, 1]$, *and choose tuning parameters*

$$(1-t(\kappa)/4)^2 < \gamma < 1, \quad \lambda_{\max} > \|\mathbf{x}_*\|_2^2.$$

For some constants $C, c > 0$ *depending only on* κ, *if* $\rho(s_*) \le c$, *then each iterate* $\mathbf{x}_k$ *of Algorithm 1 satisfies*

$$\|\mathbf{x}_k - \mathbf{x}_*\|_2 \le C\sqrt{\lambda_{\max} s_*}\cdot\gamma^{k/2}. \tag{11}$$

In particular, $\mathbf{x}_k = \mathbf{x}_*$ *for all sufficiently large* k.

Thus, in this noiseless setting, the iterates exhibit linear convergence to the true signal $\mathbf{x}_*$. The required condition $\rho(s_*) \le c$ translates into a requirement of

$$n \gtrsim s_*\log(1+|E|/s_*)$$

measurements for $\mathbf{A}$ having i.i.d. $\mathcal{N}(0, 1/n)$ entries, by Proposition 1, or

$$n \gtrsim s_*(\log p)^8\log\log p$$

weighted Fourier measurements for the 2-D lattice graph, as defined in Proposition 2. For these designs, (κ, ρ)-cRIP holds for $\sqrt{\eta}\cdot\mathbf{A}$ where $\kappa = 0$ and $\eta = 1$.

Proof (Proof of Theorem 1) Denote

$$s_k = \|\nabla\mathbf{x}_k\|_0, \quad \mathbf{r}_k = \mathbf{x}_k - \mathbf{x}_*.$$

As shown in Fan and Guan (2018, Lemma S2.1) (see also Boykov et al., 1999, Theorem 6.1), the output $\mathbf{x}_{k+1}$ of the sub-routine $\mathrm{AlphaExpansion}(\mathbf{a}_{k+1}, \lambda_k, \delta)$ has the deterministic guarantee

$$\frac{1}{2}\|\mathbf{x}_{k+1}-\mathbf{a}_{k+1}\|_2^2 + \lambda_k\|\nabla\mathbf{x}_{k+1}\|_0 \le \min_{\mathbf{x}\in(\delta\mathbb{Z})^p}\left(\frac{1}{2}\|\mathbf{x}-\mathbf{a}_{k+1}\|_2^2 + 2\lambda_k\|\nabla\mathbf{x}\|_0\right). \tag{12}$$

Applying this optimality condition (12) to compare $\mathbf{x}_{k+1}$ with $\mathbf{x}_* = \mathbf{x}_k - \mathbf{r}_k$, we obtain

$$\|\mathbf{x}_{k+1}-\mathbf{a}_{k+1}\|_2^2 + 2\lambda_k s_{k+1} \le \|\mathbf{x}_k - \mathbf{r}_k - \mathbf{a}_{k+1}\|_2^2 + 4\lambda_k s_*. \tag{13}$$

Let $\mathcal{S}_k$ be the partition of $\{1, \ldots, p\}$ induced by the piecewise-constant structure of $\mathbf{x}_k$: Each element of $\mathcal{S}_k$ corresponds to a connected subgraph of G on which $\mathbf{x}_k$ takes a constant value. Let $\mathcal{S}_{k+1}$, $\mathcal{S}_*$ similarly be the partitions induced by $\mathbf{x}_{k+1}$, $\mathbf{x}_*$, and denote by $\mathcal{S}$ the common refinement of $\mathcal{S}_k$, $\mathcal{S}_{k+1}$, $\mathcal{S}_*$. Defining the boundary

$$\partial \mathcal{S} = \{\, (i,j) \in E\colon\ i,j \text{ belong to different elements of } \mathcal{S} \,\},$$

observe that each edge $(i, j) \in \partial S$ must be such that at least one of $\mathbf{x}_k$, $\mathbf{x}_{k+1}$, or $\mathbf{x}_*$ takes different values at its two endpoints. Then

$$|\,\partial \mathcal{S}\,| \le s_* + s_k + s_{k+1}. \tag{14}$$

Let $\mathbf{P}\colon \mathbb{R}^p \to \mathbb{R}^p$ be the orthogonal projection onto the subspace of signals taking a constant value over each element of $\mathcal{S}$, and let $\mathbf{P}^\perp = \mathbf{I} - \mathbf{P}$. Then $\mathbf{x}_{k+1}$, $\mathbf{x}_k, \mathbf{r}_k$ all belong to the range of $\mathbf{P}$, so an orthogonal decomposition yields

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{a}_{k+1}\|_2^2 &= \|\mathbf{x}_{k+1} - \mathbf{P}\mathbf{a}_{k+1}\|_2^2 + \|\mathbf{P}^\perp \mathbf{a}_{k+1}\|_2^2,\\ \|\mathbf{x}_k - \mathbf{r}_k - \mathbf{a}_{k+1}\|_2^2 &= \|\mathbf{x}_k - \mathbf{r}_k - \mathbf{P}\mathbf{a}_{k+1}\|_2^2 + \|\mathbf{P}^\perp \mathbf{a}_{k+1}\|_2^2.\end{aligned}$$

Applying this, the definition (in the noiseless setting $\mathbf{e}=\mathbf{0}$)

$$\mathbf{a}_{k+1} = \mathbf{x}_k - \eta \mathbf{A}^\top(\mathbf{A}\mathbf{x}_k - \mathbf{y}) = \mathbf{x}_k - \eta \mathbf{A}^\top \mathbf{A}\mathbf{r}_k,$$

and the condition $\mathbf{P}\mathbf{x}_k = \mathbf{x}_k$ to Equation (13), we obtain

$$\|\,\mathbf{x}_{k+1} - \mathbf{x}_k + \eta \mathbf{P}\mathbf{A}^\top \mathbf{A}\mathbf{r}_k\,\|_2^2 \le \|\,\eta \mathbf{P}\mathbf{A}^\top \mathbf{A}\mathbf{r}_k - \mathbf{r}_k\,\|_2^2 + \lambda_k(4s_* - 2s_{k+1}).$$

Applying the triangle inequality and $\mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{r}_{k+1} - \mathbf{r}_k$,

$$\left(\|\mathbf{r}_{k+1}\|_2 - \|\mathbf{r}_k - \eta \mathbf{P}\mathbf{A}^\top \mathbf{A}\mathbf{r}_k\|_2\right)_+^2 \le \|\,\mathbf{r}_k - \eta \mathbf{P}\mathbf{A}^\top \mathbf{A}\mathbf{r}_k\,\|_2^2 + \lambda_k(4s_* - 2s_{k+1}). \tag{15}$$

We derive from these two consequences: First, lower-bounding the left side by 0 and rearranging,

$$\lambda_k s_{k+1} \le \frac{1}{2}\|\,\mathbf{r}_k - \eta \mathbf{P}\mathbf{A}^\top \mathbf{A}\mathbf{r}_k\,\|_2^2 + 2\lambda_k s_* \le \|\,\mathbf{r}_k\,\|_2^2 + \|\,\sqrt{\eta}\mathbf{A}\mathbf{P}\,\|_{\text{op}}^2 \cdot \|\,\sqrt{\eta}\mathbf{A}\mathbf{r}_k\,\|_2^2 + 2\lambda_k s_*. \tag{16}$$

The condition (14) and definition of $\mathbf{P}$ imply, for any $\mathbf{u} \in \mathbb{R}^p$, that $\|\nabla(\mathbf{P}\mathbf{u})\|_0 \le s_* + s_k + s_{k+1}$. The definition of $\mathbf{r}_k$ implies $\|\nabla \mathbf{r}_k\|_0 \le s_* + s_k$. Setting

$$\tau_k = \kappa + \sqrt{\rho(s_* + s_k + s_{k+1})}, \quad \zeta_k = \kappa + \sqrt{\rho(s_* + s_k)}$$

we deduce from the (κ, ρ)-cRIP condition for $\sqrt{\eta} \cdot \mathbf{A}$ that

$$\|\,\sqrt{\eta}\mathbf{A}\mathbf{P}\,\|_{\text{op}}^2 = \sup_{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|_2 = 1} \|\,\sqrt{\eta}\mathbf{A}\mathbf{P}\mathbf{u}\,\|_2^2 \le (1 + \tau_k)^2, \quad \|\,\sqrt{\eta}\mathbf{A}\mathbf{r}_k\,\|_2^2 \le (1 + \zeta_k)^2 \|\,\mathbf{r}_k\,\|_2^2. \tag{17}$$

Note that since $\rho(s)$ and $\sqrt{\rho(s)}$ are both non-negative and concave by Definition 1, we have

$$\rho'(s) \le (\rho(s) - \rho(0))/s \le \rho(s)/s, \frac{d}{ds}[\sqrt{\rho(s)}] \le (\sqrt{\rho(s)} - \sqrt{\rho(0)})/s \le \sqrt{\rho(s)}/s.$$

The function

$$f_k(s) = (1 + \kappa + \sqrt{\rho(s_* + s_k + s)})^2$$

is also increasing and concave, and by the above, its derivative at $s = 0$ satisfies

$$f_k'(0) \le d_k/(s_* + s_k), \quad d_k \equiv 2(1+\kappa)\sqrt{\rho(s_* + s_k)} + \rho(s_* + s_k).$$

Thus

$$(1+\tau_k)^2 = f_k(s_{k+1}) \le f_k(0) + f_k'(0) \cdot s_{k+1} \le (1+\zeta_k)^2 + d_k s_{k+1}/s_*. \tag{18}$$

Applying this and Equation (17) to Equation (16), we get

$$\begin{aligned}\lambda_k s_{k+1} &\le \left(1 + (1+\tau_k)^2(1+\zeta_k)^2\right)\|\mathbf{r}_k\|_2^2 + 2\lambda_k s_* \\ &\le \left(1 + (1+\zeta_k)^4 + (1+\zeta_k)^2 d_k s_{k+1}/s_*\right)\|\mathbf{r}_k\|_2^2 + 2\lambda_k s_*.\end{aligned}$$

Rearranging gives

$$(\lambda_k - (1+\zeta_k)^2 d_k \|\mathbf{r}_k\|_2^2/s_*) \cdot s_{k+1} \le (1 + (1+\zeta_k)^4) \cdot \|\mathbf{r}_k\|_2^2 + 2\lambda_k s_*. \tag{19}$$

Second, applying the (κ, ρ)-cRIP condition for $\sqrt{\eta} \cdot \mathbf{A}$ again, we have for every $\mathbf{u} \in \mathbb{R}^p$

$$\begin{aligned}|\mathbf{u}^\top(\eta\mathbf{P}\mathbf{A}^\top\mathbf{A}\mathbf{P} - \mathbf{P})\mathbf{u}| &= |\|\sqrt{\eta}\mathbf{A}\mathbf{P}\mathbf{u}\|_2^2 - \|\mathbf{P}\mathbf{u}\|_2^2| \\ &\le \max(|1-(1-\tau_k)^2|, |1-(1+\tau_k)^2|)\|\mathbf{P}\mathbf{u}\|_2^2 = (2\tau_k + \tau_k^2)\|\mathbf{P}\mathbf{u}\|_2^2,\end{aligned}$$

So $\|\eta\mathbf{P}\mathbf{A}^\top\mathbf{A}\mathbf{P} - \mathbf{P}\|_{\text{op}} \le 2\tau_k + \tau_k^2$. Then, as $\mathbf{r}_k = \mathbf{P}\mathbf{r}_k$, we get from Equation (15) that

$$(\|\mathbf{r}_{k+1}\|_2 - (2\tau_k + \tau_k^2)\|\mathbf{r}_k\|_2)_+^2 \le (2\tau_k + \tau_k^2)^2\|\mathbf{r}_k\|_2^2 + \lambda_k(4s_* - 2s_{k+1}).$$

Taking the square root and applying $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$,

$$\|\mathbf{r}_{k+1}\|_2 \le (4\tau_k + 2\tau_k^2)\|\mathbf{r}_k\|_2 + \sqrt{\lambda_k(4s_* - 2s_{k+1})_+}$$

Applying the definitions of τ_k and $t(\kappa)$,

$$4\tau_k + 2\tau_k^2 \le 1 - t(\kappa) + 4(1+\kappa)\sqrt{\rho(s_* + s_k + s_{k+1})} + 2\rho(s_* + s_k + s_{k+1}).$$

Thus

$$\|\mathbf{r}_{k+1}\|_2 \le [1 - t(\kappa) + 4(1+\kappa)\sqrt{\rho(s_* + s_k + s_{k+1})} + 2\rho(s_* + s_k + s_{k+1})] \cdot \|\mathbf{r}_k\|_2 + \sqrt{4\lambda_k s_*}. \tag{20}$$

We now claim by induction on k that, if $\rho(s_*) \le c_0$ for a sufficiently small constant $c_0 > 0$, then

$$s_k \le \frac{90}{t(\kappa)^2}s_*, \quad \|\mathbf{r}_k\|_2 \le \frac{4\sqrt{\lambda_k s_*}}{t(\kappa)} \tag{21}$$

for every k. For $k = 0$, these are satisfied as $s_0 = 0$ and $\lambda_0 = \lambda_{\max} \geq \|\mathbf{r}_0\|_2^2 = \|\mathbf{x}_*\|_2^2$. Assume inductively that these hold for k. Note that for any $t \geq 1$, non-negativity and concavity yield $\rho(ts_*) \leq t\rho(s_*)$. In particular, assuming Equation (21) and applying $\kappa < \sqrt{3/2} - 1$ and $\rho(s_*) \leq c_0$, we get for small enough c_0 that $(1 + \zeta_k)^2 < 2$. Then applying Equation (21) to Equation (19), we get for a constant $C \equiv C(\kappa) > 0$ not depending on c_0 that

$$\left(1 - C\sqrt{c_0}\right) \lambda_k s_{k+1} \leq \left(\frac{80}{t(\kappa)^2} + 2\right) \lambda_k s_*.$$

Then for small enough c_0,

$$s_{k+1} \leq \left(1 - C\sqrt{c_0}\right)^{-1} \frac{82}{t(\kappa)^2} s_* < \frac{90}{t(\kappa)^2} s_*.$$

Applying Equation (21) and this bound to Equation (20), for sufficiently small c_0, we have

$$\|\mathbf{r}_{k+1}\|_2 \leq \left(1 - \frac{3}{4}t(\kappa)\right) \|\mathbf{r}_k\|_2 + \sqrt{4\lambda_k s_*} \leq \left(\frac{4}{t(\kappa)} - 1\right) \sqrt{\lambda_k s_*}.$$

Applying $\sqrt{\lambda_k} = \sqrt{\lambda_{k+1}/\gamma} \leq \sqrt{\lambda_{k+1}}(1 - t(\kappa)/4)^{-1}$, we obtain from this

$$\|\mathbf{r}_{k+1}\|_2 \leq 4\sqrt{\lambda_{k+1} s_*}/t(\kappa).$$

This completes the induction and establishes Equation (21) for every k.

The bound (11) follows from Equation (21), the definition of $\mathbf{r}_k$, and $\lambda_k = \lambda_{\max}\gamma^k$. Since $\mathbf{x}_k, \mathbf{x}_* \in (\delta\mathbb{Z})^p$, for k large enough such that the right side of Equation (11) is less than δ^2, we must have $\mathbf{x}_k = \mathbf{x}_*$.

We now extend this result to a robust recovery guarantee in the presence of measurement and discretization error. In this setting, ITALE is not guaranteed to converge to a global minimizer of the non-convex objective (2). Instead, we provide a direct bound on the estimation error of a suitably chosen ITALE iterate. The proof is an extension of the above argument, which we defer to Appendix A of the online supplementary material.

Theorem 2 *Suppose $\sqrt{\eta} \cdot \mathbf{A}$ satisfies (κ, ρ)-cRIP, where $\kappa \in [0, \sqrt{3/2} - 1)$. Choose tuning parameters γ, $\lambda_{\max}$ as in Theorem 1. Then for some constants C, C', $c > 0$ depending only on κ, the following holds: Let $\mathbf{x} \in (\delta\mathbb{Z})^p$ be any vector satisfying $\rho(s) \leq c$ where $s \equiv \max(\|\nabla \mathbf{x}\|_0, 1)$. Let D be the maximum vertex degree of G, and define*

$$E(\mathbf{x}) = \left(1 + \sqrt{D\rho(s)}\right) \cdot \left(\|\mathbf{x} - \mathbf{x}_*\|_2 + \frac{\|\mathbf{x} - \mathbf{x}_*\|_1}{\sqrt{s}}\right) + \sqrt{\eta} \cdot \|\mathbf{e}\|_2.$$

Suppose $\lambda_{\max} \geq CE(\mathbf{x})^2/s \geq \lambda_{\min}$, and let k_* be the last iterate of Algorithm 1 where $\lambda_{k_*} \geq CE(\mathbf{x})^2/s$. Then $\widehat{\mathbf{x}} \equiv \mathbf{x}_{k_*}$ satisfies

$$\|\widehat{\mathbf{x}} - \mathbf{x}_*\|_2 \leq C' E(\mathbf{x}).$$

Here, $\mathbf{x} \in (\delta\mathbb{Z})^p$ is any deterministic vector that approximates $\mathbf{x}_*$ and satisfies $\|\nabla\mathbf{x}\|_0 \leq s$, and the theorem should be interpreted for $\mathbf{x}$ being the best such approximation to $\mathbf{x}_*$. The quantity $E(\mathbf{x})$ above is the combined measurement error and approximation error of $\mathbf{x}_*$ by $\mathbf{x}$. For any $\mathbf{A}$ scaled such that it satisfies (κ, ρ)-cRIP with $\eta = 1$, and for G with maximum degree $D \lesssim 1$, we get

$$\|\widehat{\mathbf{x}} - \mathbf{x}_*\|_2 \lesssim \|\mathbf{x}_* - \mathbf{x}\|_2 + \frac{\|\mathbf{x}_* - \mathbf{x}\|_1}{\sqrt{s}} + \|\mathbf{e}\|_2. \tag{22}$$

This guarantee is similar to those for compressed sensing of sparse signals in Candès et al. (2006b), Needell and Tropp (2009) and Blumensath and Davies (2009). Note that, as in these works, we are assuming a setting of deterministic and possibly adversarial measurement error $\mathbf{e}$.

If $\mathbf{x}_*$ has exact gradient sparsity $\|\nabla\mathbf{x}_*\|_0 \leq s$, then also $\mathbf{x} \in (\delta\mathbb{Z})^p$ obtained by entrywise rounding to $\delta\mathbb{Z}$ satisfies $\|\nabla\mathbf{x}\|_0 \leq s$. Hence, applying Equation (22) with this $\mathbf{x}$ and choosing $\delta \ll \|\mathbf{e}\|_2 / p$ further ensures

$$\|\widehat{\mathbf{x}} - \mathbf{x}_*\|_2 \lesssim \|\mathbf{e}\|_2$$

that is, the discretization error is negligible in the above bound. It is clear that this is the rate-optimal error bound for worst-case error $\mathbf{e}$, as may be seen by taking $\mathbf{e} = \mathbf{A1}$ where $\mathbf{1}$ is the all-1's vector. The required number of measurements is the same as in Theorem 1 for the noiseless setting, which is $n \gtrsim s_* \log(1 + |E| / s_*)$ for i.i.d. Gaussian designs. This is the claim (5) stated in Section 1.

When $\mathbf{x}_*$ is not exactly gradient sparse, the error (22) depends also on the errors $\|\mathbf{x}_* - \mathbf{x}\|_2$ and $\|\mathbf{x}_* - \mathbf{x}\|_1$ of the approximation by a gradient-sparse vector $\mathbf{x}$. This dependence is similar to the guarantees of Needell and Ward (2013b) for TV regularization, although we note that Needell and Ward (2013b) provided bounds in terms of $\nabla\mathbf{x}_* - \nabla\mathbf{x}$ rather than $\mathbf{x}_* - \mathbf{x}$.

4 | SIMULATIONS

We compare $\widehat{\mathbf{x}}^{\text{ITALE}}$ using the ℓ_0 edge cost (3) to $\widehat{\mathbf{x}}^{\text{TV}}$ minimizing the TV-regularized objective (6), for several signals on the 1-D and 2-D lattice graphs. We used software developed by Boykov and Kolmogorov (2004), to implement the alpha-expansion sub-routine of Algorithm 2. For convenience, we further made an R package `ITALE` to realize Algorithm 1. To minimize the TV-regularized objective (6), we used the generalized lasso path algorithm from Tibshirani (2011) in the 1-D examples and the FISTA algorithm from Beck and Teboulle (2009) in the 2-D examples. All parameters were set as described in Section 2 for ITALE.

4.1 | 1-D changepoint signals

We tested ITALE on two simulated signals for the linear chain graph, with different changepoint structures: the 'spike' signal depicted in Figures 2 and 3, and the 'wave' signal depicted in Figures 4 and 5. The two signals both have $p = 1000$ vertices with $s_* = 9$ break points. The spike signal consists of short segments of length 10 with elevated mean, while the breaks of the wave signal are equally spaced.

We sampled i.i.d. random Gaussian measurements $A_{ij} \sim \mathcal{N}(0, 1)$. The measurement error $\mathbf{e}$ was generated as i.i.d. Gaussian noise $e_k \sim \mathcal{N}(0, \sigma^2)$. To provide an intuitive understanding of the tested

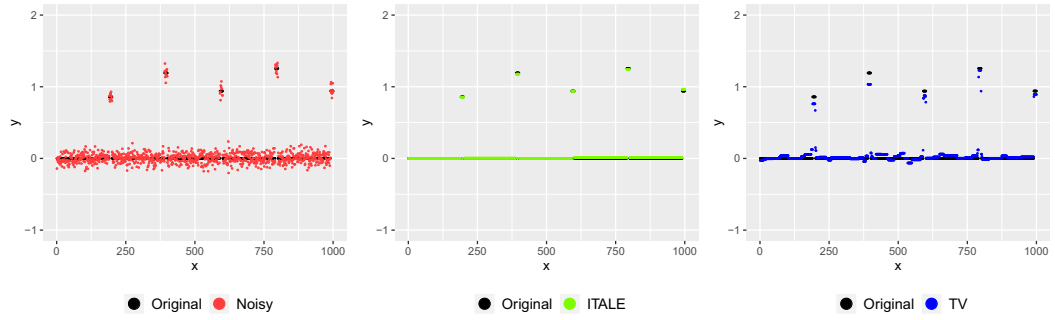

FIGURE 2 Left: True spike signal $\mathbf{x}_*$ (black) and a depiction of $\mathbf{x}_* + \mathbf{A}^\top \mathbf{e}/n$ (red) under low noise $\sigma = 1$ for i.i.d. measurements $A_{ij} \sim \mathcal{N}(0, 1)$ with 15% undersampling. Middle and right: True signal (black), $\hat{\mathbf{x}}^{\text{ITALE}}$ (green), and $\hat{\mathbf{x}}^{\text{TV}}$ (blue) for one simulation

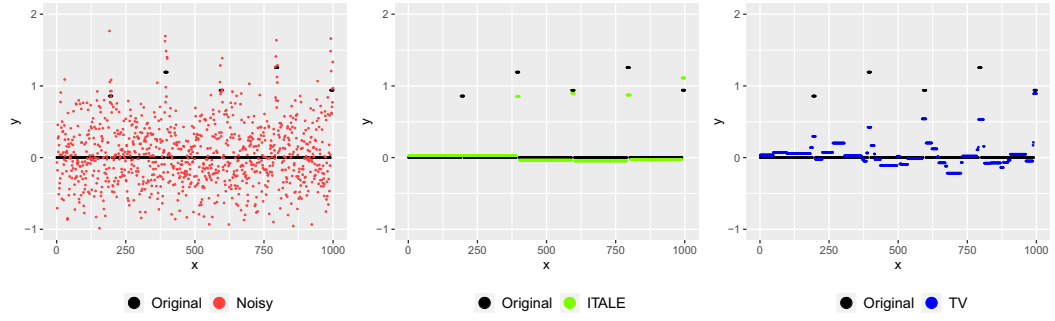

FIGURE 3 Same setting as Figure 2, for noise level $\sigma = 6$

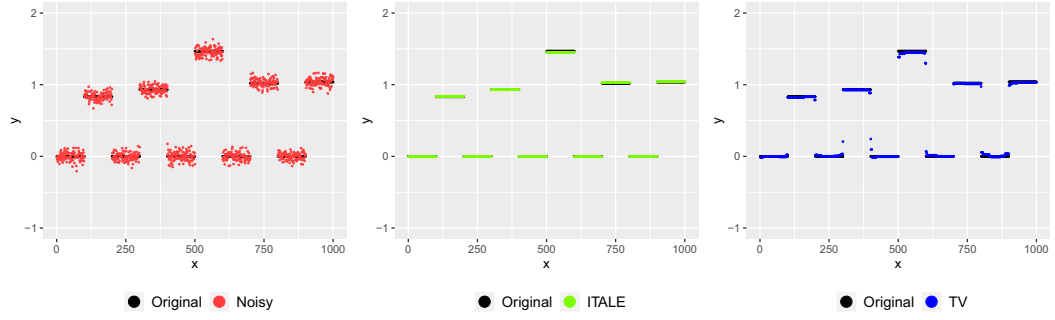

FIGURE 4 Left: True wave signal $\mathbf{x}_*$ (black) and a depiction of $\mathbf{x}_* + \mathbf{A}^\top \mathbf{e}/n$ (red) under low noise σ=1 for i.i.d. measurements $A_{ij} \sim \mathcal{N}(0, 1)$ with 15% undersampling. Middle and right: True signal (black), $\hat{\mathbf{x}}^{\text{ITALE}}$ (green), and $\hat{\mathbf{x}}^{\text{TV}}$ (blue) for one simulation

signal-to-noise, we plot $\mathbf{x}_* + \mathbf{A}^\top \mathbf{e}/n$ in red in Figures 2–5, corresponding to two different tested noise levels. Recall that ITALE denoises $\mathbf{a}_{k+1} = \mathbf{x}_* + (\mathbf{I} - \mathbf{A}^\top \mathbf{A}/n)(\mathbf{x}_k - \mathbf{x}_*) + \mathbf{A}^\top \mathbf{e}/n$ in each iteration (corresponding to $\eta = 1/n$ for this normalization of $\mathbf{A}$), so that $\mathbf{x}_* + \mathbf{A}^\top \mathbf{e}/n$ represents the noisy signal in an ideal setting if $\mathbf{x}_k \equiv \mathbf{x}_*$ is a perfect estimate from the preceding iteration.

Tables 1 and 2 display the root-mean-squared estimation errors $\text{RMSE} = \sqrt{\|\hat{\mathbf{x}} - \mathbf{x}_*\|_2^2/p}$, for undersampling ratio n/p from 10% to 50%, and a range of noise levels σ that yielded RMSE values

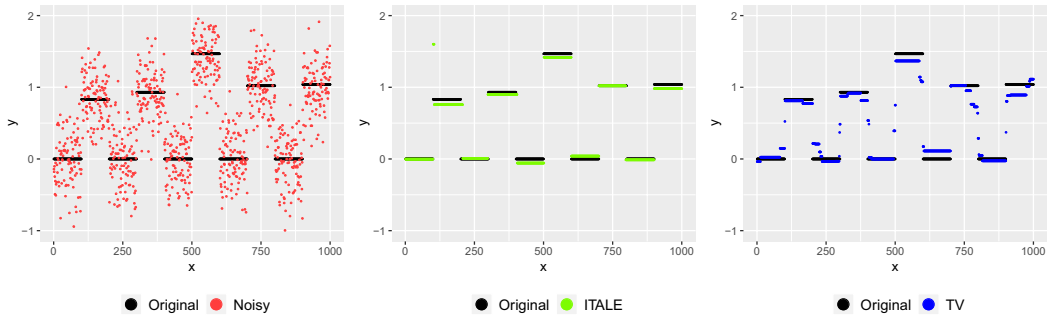

FIGURE 5 Same setting as Figure 4, for noise level $\sigma = 6$

TABLE 1 RMSE for the 1-D spike signal, averaged over 20 simulations

n/p		$\sigma = 0$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$	$\sigma = 6$	$\sigma = 7$
10%	ITALE	**0.000**	**0.014**	**0.060**	**0.090**	**0.144**	**0.173**	**0.199**	**0.216**
	TV	**0.000**	0.047	0.092	0.129	0.160	0.189	0.213	0.228
15%	ITALE	**0.000**	**0.009**	**0.023**	**0.049**	**0.076**	**0.104**	**0.133**	**0.153**
	TV	**0.000**	0.030	0.060	0.088	0.114	0.136	0.158	0.175
20%	ITALE	**0.000**	**0.007**	**0.015**	**0.032**	**0.056**	**0.076**	**0.099**	**0.123**
	TV	**0.000**	0.022	0.045	0.067	0.089	0.109	0.128	0.146
30%	ITALE	**0.000**	**0.006**	**0.012**	**0.021**	**0.031**	**0.049**	**0.065**	**0.079**
	TV	**0.000**	0.017	0.035	0.052	0.070	0.087	0.104	0.120
40%	ITALE	**0.000**	**0.005**	**0.010**	**0.015**	**0.025**	**0.041**	**0.051**	**0.063**
	TV	**0.000**	0.014	0.028	0.043	0.057	0.071	0.085	0.098
50%	ITALE	**0.000**	**0.005**	**0.010**	**0.015**	**0.023**	**0.033**	**0.040**	**0.051**
	TV	**0.000**	0.013	0.026	0.038	0.051	0.064	0.075	0.088

TABLE 2 RMSE for the 1-D wave signal, averaged over 20 simulations

n/p		$\sigma = 0$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$	$\sigma = 6$	$\sigma = 7$
10%	ITALE	0.036	0.040	0.118	0.150	0.198	0.236	0.262	0.315
	TV	**0.000**	**0.032**	**0.064**	**0.093**	**0.120**	**0.143**	**0.168**	**0.189**
15%	ITALE	**0.000**	**0.009**	**0.025**	**0.059**	0.090	0.111	0.143	0.176
	TV	**0.000**	0.023	0.046	0.068	**0.089**	**0.109**	**0.127**	**0.144**
20%	ITALE	**0.000**	**0.007**	**0.017**	**0.039**	**0.061**	**0.079**	**0.103**	**0.121**
	TV	**0.000**	0.019	0.037	0.056	0.074	0.092	0.108	0.124
30%	ITALE	**0.000**	**0.006**	**0.012**	**0.019**	**0.035**	**0.051**	**0.065**	**0.085**
	TV	**0.000**	0.014	0.028	0.042	0.056	0.070	0.084	0.097
40%	ITALE	**0.000**	**0.005**	**0.011**	**0.018**	**0.027**	**0.037**	**0.052**	**0.064**
	TV	**0.000**	0.012	0.024	0.037	0.049	0.061	0.073	0.085
50%	ITALE	**0.000**	**0.005**	**0.010**	**0.016**	**0.024**	**0.033**	**0.044**	**0.055**
	TV	**0.000**	0.011	0.022	0.033	0.043	0.054	0.065	0.075

between 0 and roughly 0.2. Each reported error value is an average across 20 independent simulations. In these results, the iterate k in ITALE and penalty parameter λ in TV were both selected using fivefold cross-validation. Best-achieved errors over all k and λ are reported in Appendix C of the online supplementary material, and suggest the same qualitative conclusions. Standard deviations of the best-achieved errors are also reported in Appendix C; those for cross-validation are similar and omitted for brevity.

In the spike example, ITALE yielded lower RMSE in all of the above settings of undersampling and signal-to-noise. Figures 2 and 3 display one instance each of the resulting estimates $\widehat{\mathbf{x}}^{\text{ITALE}}$ and $\widehat{\mathbf{x}}^{\text{TV}}$ at 15% undersampling, illustrating some of their differences and typical features. Under optimal tuning, $\widehat{\mathbf{x}}^{\text{TV}}$ returns an undersmoothed estimate even in a low-noise setting where ITALE can often correctly estimate the changepoint locations. With higher noise, ITALE begins to miss changepoints and oversmooth.

In the wave example, with undersampling ranging between 15% and 50%, ITALE yielded lower RMSE at most tested noise levels. Figures 4 and 5 depict two instances of the recovered signals at 15% undersampling. For 10% undersampling, the component $(\mathbf{I} - \mathbf{A}^{\mathsf{T}}\mathbf{A}/n)(\mathbf{x}_k - \mathbf{x}_*)$ of the effective noise was sufficiently high such that ITALE often did not estimate the true changepoint structure, and TV usually outperformed ITALE in this case. The standard deviations of RMSE reported in Appendix C indicate that the ITALE estimates are a bit more variable than the TV estimates in all tested settings, but particularly so in this 10% undersampling regime.

4.2 | 2-D phantom images

Next, we tested ITALE on three 2-D image examples, corresponding to piecewise-constant digital phantom images of varying complexity: the Shepp–Logan digital phantom depicted in Figure 6, a digital brain phantom from Fessler and Hero (1994) depicted in Figure 7, and the XCAT chest slice from Gong et al. (2017) as previously depicted in Figure 1.

Each image $\mathbf{x}_*$ was normalized to have pixel value in [0, 1]. We sampled a random Fourier design matrix as specified in Equation (10), fixing the constant $C_0 = 10$ in the weight distribution (9) for this design. This value of C_0 yielded the best recovery across several tested values for both ITALE and TV. The measurement error $\mathbf{e}$ was generated as i.i.d. Gaussian noise $e_k \sim \mathcal{N}(0, \sigma^2)$, applied to the measurements $\mathcal{F}^*_{(i,j)}\mathbf{x}_*/\sqrt{\nu(i,j)}$ before the $1/\sqrt{n}$ normalization. Tables 3, 4 and 5 display the average RMSE of the estimates $\widehat{\mathbf{x}}^{\text{ITALE}}$ and $\widehat{\mathbf{x}}^{\text{TV}}$ across 20 independent simulations of $\mathbf{e}$, with tuning parameters selected by fivefold cross-validation. Best-achieved errors and standard deviations are reported in Appendix C.

For the simpler Logan–Shepp and brain phantom images, which exhibit stronger gradient sparsity, ITALE yielded lower RMSE in nearly all tested undersampling and signal-to-noise regimes. For the XCAT chest phantom, with undersampling ranging between 15% and 50%, ITALE yielded lower RMSE at a range of tested noise levels, and in particular for those settings of higher signal-to-noise. With 10% undersampling for the XCAT phantom, ITALE was not able to recover some details of the XCAT image even with no measurement noise, and RMSE was higher than TV at all tested noise levels. Results of Appendix C indicate that this is partially due to sub-optimal selection of the tuning parameter using fivefold cross-validation, caused by the further reduction of undersampling from 10% to 8% in the size of the training data in each fold.

Examples of recovered signals $\widehat{\mathbf{x}}^{\text{ITALE}}$ and $\widehat{\mathbf{x}}^{\text{TV}}$ are depicted for the Shepp–Logan and brain phantoms in Figures 5 and 7, at 15% and 20% undersampling for two low-noise and medium-noise settings. The qualitative comparisons are similar to those in the 1-D simulations, and to those previously

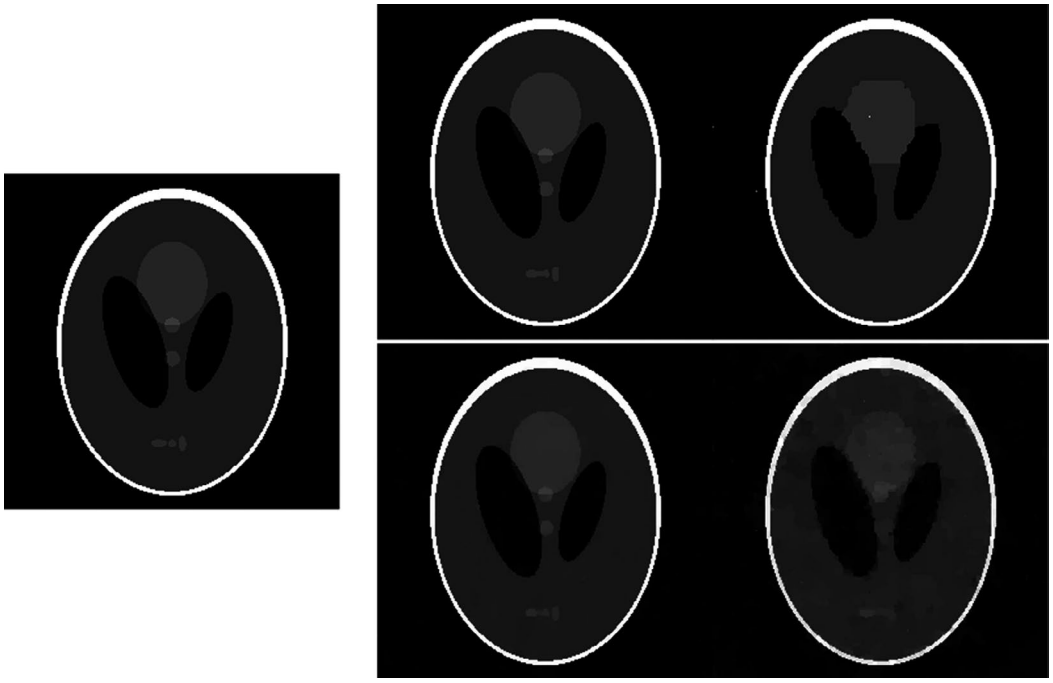

FIGURE 6 Left: Original Shepp–Logan phantom. Top row: $\hat{\mathbf{x}}^{\mathrm{ITALE}}$ from 15% undersampled and reweighted Fourier measurements, in low noise (σ=4, left) and medium noise ($\sigma = 16$, right) settings. Bottom row: $\hat{\mathbf{x}}^{\mathrm{TV}}$ for the same measurements

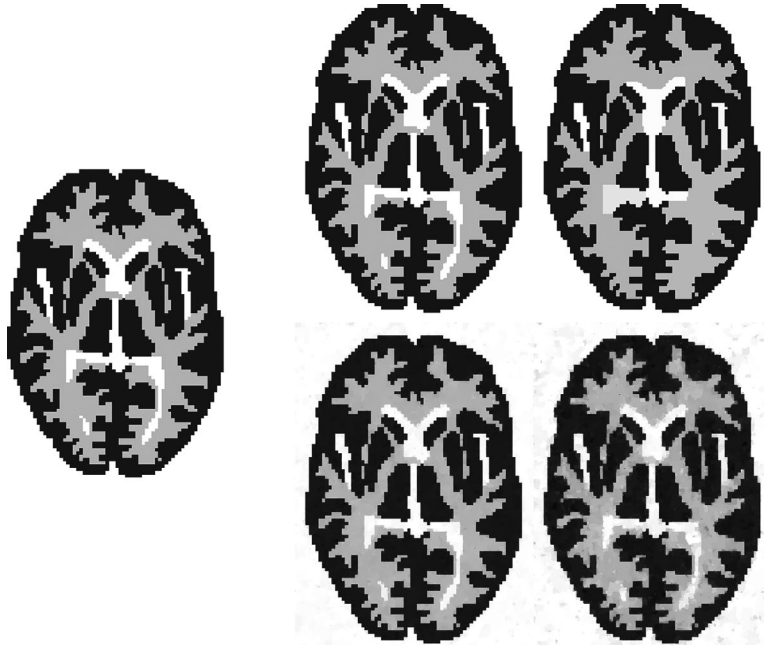

FIGURE 7 Left: Original brain phantom. Top row: $\hat{\mathbf{x}}^{\mathrm{ITALE}}$ from 20% undersampled reweighted Fourier measurements, in low noise ($\sigma = 16$, left) and medium noise ($\sigma = 40$, right) settings. Bottom row: $\hat{\mathbf{x}}^{\mathrm{TV}}$ for the same measurements

TABLE 3 RMSE for the Shepp–Logan phantom, averaged over 20 simulations

n/p		$\sigma = 0$	$\sigma = 4$	$\sigma = 8$	$\sigma = 12$	$\sigma = 16$	$\sigma = 20$	$\sigma = 24$	$\sigma = 28$
10%	ITALE	**0.001**	**0.006**	**0.012**	**0.018**	**0.028**	**0.036**	**0.051**	0.071
	TV	0.005	0.011	0.021	0.031	0.040	0.049	0.057	**0.064**
15%	ITALE	**0.000**	**0.003**	**0.011**	**0.013**	**0.018**	**0.028**	**0.034**	**0.042**
	TV	0.001	0.009	0.016	0.024	0.031	0.038	0.046	0.053
20%	ITALE	**0.000**	**0.002**	**0.009**	**0.012**	**0.014**	**0.024**	**0.028**	**0.034**
	TV	**0.000**	0.007	0.014	0.020	0.027	0.033	0.039	0.045
30%	ITALE	**0.000**	**0.002**	**0.006**	**0.011**	**0.013**	**0.015**	**0.021**	**0.028**
	TV	**0.000**	0.006	0.012	0.017	0.022	0.027	0.032	0.036
40%	ITALE	**0.000**	**0.001**	**0.005**	**0.010**	**0.012**	**0.013**	**0.015**	**0.021**
	TV	**0.000**	0.005	0.010	0.015	0.019	0.023	0.028	0.032
50%	ITALE	**0.000**	**0.001**	**0.004**	**0.008**	**0.011**	**0.013**	**0.014**	**0.017**
	TV	**0.000**	0.005	0.009	0.013	0.018	0.022	0.025	0.028

TABLE 4 RMSE for the brain phantom, averaged over 20 simulations

n/p		$\sigma = 0$	$\sigma = 8$	$\sigma = 16$	$\sigma = 24$	$\sigma = 32$	$\sigma = 40$	$\sigma = 48$	$\sigma = 56$
10%	ITALE	0.003	**0.002**	**0.011**	**0.027**	**0.044**	**0.062**	0.081	0.097
	TV	**0.002**	0.014	0.028	0.041	0.054	0.066	**0.078**	**0.088**
15%	ITALE	**0.000**	**0.001**	**0.007**	**0.018**	**0.030**	**0.044**	**0.059**	**0.073**
	TV	0.001	0.011	0.022	0.032	0.043	0.053	0.062	**0.073**
20%	ITALE	**0.000**	**0.001**	**0.005**	**0.011**	**0.025**	**0.035**	**0.047**	**0.060**
	TV	**0.000**	0.010	0.019	0.028	0.038	0.047	0.055	0.062
30%	ITALE	**0.000**	**0.001**	**0.003**	**0.008**	**0.015**	**0.026**	**0.033**	**0.043**
	TV	**0.000**	0.008	0.015	0.023	0.030	0.037	0.046	0.052
40%	ITALE	**0.000**	**0.001**	**0.002**	**0.006**	**0.010**	**0.020**	**0.026**	**0.034**
	TV	**0.000**	0.007	0.013	0.020	0.026	0.032	0.038	0.044
50%	ITALE	**0.000**	**0.000**	**0.002**	**0.004**	**0.008**	**0.014**	**0.022**	**0.028**
	TV	**0.000**	0.006	0.012	0.018	0.023	0.029	0.035	0.040

depicted for the XCAT chest slice in Figure 1: As measurement noise increases, ITALE begins to lose finer details, while TV begins to yield an undersmoothed and blotchy image. These observations are also similar to previous comparisons that have been made for algorithms oriented towards ℓ_0 versus TV regularization for direct measurements $\mathbf{A} = \mathbf{I}$, in Xu et al. (2011), Fan and Guan (2018) and Kim and Gao (2019).

5 | CONCLUSION

We have studied recovery of piecewise-constant signals over arbitrary graphs from noisy linear measurements. We have proposed an iterative algorithm, ITALE, to minimize an ℓ_0-edge-penalized least-squares objective. Under a cut-restricted isometry property for the measurement design, we have established global recovery guarantees for the estimated signal, in noisy and noiseless settings.

TABLE 5 RMSE for the XCAT chest slice phantom, averaged over 20 simulations

n/p		$\sigma = 0$	$\sigma = 4$	$\sigma = 8$	$\sigma = 12$	$\sigma = 16$	$\sigma = 20$	$\sigma = 24$	$\sigma = 28$
10%	ITALE	0.063	0.065	0.070	0.075	0.082	0.091	0.099	0.108
	TV	**0.009**	**0.019**	**0.032**	**0.043**	**0.053**	**0.061**	**0.068**	**0.073**
15%	ITALE	**0.002**	**0.007**	**0.024**	0.036	0.055	0.070	0.079	0.088
	TV	0.005	0.014	**0.024**	**0.034**	**0.042**	**0.050**	**0.057**	**0.063**
20%	ITALE	**0.002**	**0.005**	**0.014**	**0.023**	**0.032**	0.045	0.062	0.076
	TV	**0.002**	0.011	0.020	0.028	0.036	**0.043**	**0.050**	**0.055**
30%	ITALE	**0.002**	**0.004**	**0.011**	**0.018**	**0.025**	**0.031**	**0.041**	0.050
	TV	**0.002**	0.008	0.016	0.023	0.030	0.036	0.042	**0.047**
40%	ITALE	0.002	**0.003**	**0.009**	**0.015**	**0.020**	**0.027**	**0.033**	**0.040**
	TV	**0.001**	0.007	0.014	0.020	0.026	0.031	0.036	0.042
50%	ITALE	0.002	**0.003**	**0.008**	**0.013**	**0.018**	**0.023**	**0.028**	**0.033**
	TV	**0.001**	0.006	0.012	0.018	0.023	0.028	0.033	0.037

In the field of compressed sensing, for signals exhibiting sparsity in an orthonormal basis, ℓ_1-regularization (Candès et al., 2006a, b; Donoho, 2006) and discrete iterative algorithms (Blumensath & Davies, 2009; Needell & Tropp, 2009; Tropp & Gilbert, 2007) constitute two major approaches for signal recovery. It has been observed that for recovering piecewise-constant signals, regularizing the signal gradient in a sparse analysis framework can yield better empirical recovery than regularizing signal coefficients in such a basis. Whereas ℓ_1-regularization extends naturally to the sparse analysis setting, iterative algorithms have received less attention. By applying the alpha-expansion idea for MAP estimation in discrete Markov random fields, ITALE provides a computationally tractable approach for 'iterative thresholding' recovery of gradient-sparse signals, with provable recovery guarantees.

In contrast to sparse signal recovery over an orthonormal basis, the comparison of ℓ_1 versus ℓ_0 regularization for gradient-based sparsity is graph dependent. Using an ℓ_0-based approach, we establish signal recovery guarantees on the 1-D and 2-D lattice graphs with numbers of measurements optimal up to a constant factor, which were not previously available for TV regularization. This difference is closely connected to slow and fast rates of convergence for lasso and best-subset regression for correlated regression designs (Bühlmann et al., 2013; Dalalyan et al., 2017, Zhang et al., 2014). ITALE provides a polynomial-time approach for ℓ_0-regularization in a special graph-based setting, and we believe it is an interesting question whether similar algorithmic ideas may be applicable to other classes of sparse regression problems.

ACKNOWLEDGEMENTS

We thank Brian Caffo for pointing us to the XCAT digital phantom, Feng Ruan for helpful discussions, and the anonymous reviewers for suggestions that helped improve our paper. This research was supported in part by NSF grant DMS-1916198.

REFERENCES

Beck, A. & Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2, 183–202.

Bertsimas, D., King, A. & Mazumder, R. (2016) Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44, 813–852.

Blumensath, T. & Davies, M.E. (2009) Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27, 265–274.

Boykov, Y. & Kolmogorov, V. (2004) An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 9, 1124–1137.

Boykov, Y., Veksler, O. & Zabih, R. (1999) Fast approximate energy minimization via graph cuts. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1, IEEE. pp. 377–384.

Bühlmann, P., Rütimann, P., van de Geer, S. & Zhang, C.-H. (2013) Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143, 1835–1858.

Cai, J.-F. & Xu, W. (2015) Guarantees of total variation minimization for signal recovery. *Information and Inference: A Journal of the IMA*, 4, 328–353.

Candès, E.J., Romberg, J. & Tao, T. (2006a) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52, 489.

Candès, E.J., Romberg, J.K. & Tao, T. (2006b) Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59, 1207–1223.

Candès, E.J., Eldar, Y. C., Needell, D. & Randall, P. (2011) Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31, 59–73.

Dalalyan, A.S., Hebiri, M. & Lederer, J. (2017) On the prediction performance of the Lasso. *Bernoulli*, 23, 552–581.

Donoho, D.L. (2006) Compressed sensing. *IEEE Transactions on Information Theory*, 52, 1289–1306.

Elad, M., Milanfar, P. & Rubinstein, R. (2007) Analysis versus synthesis in signal priors. *Inverse Problems*, 23, 947.

Elenberg, E.R., Khanna, R., Dimakis, A.G. & Negahban, S. (2018) Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46, 3539–3568.

Fan, Z. & Guan, L. (2018) Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46, 3217–3245.

Fessler, J.A. & Hero, A.O. (1994) Space-alternating generalized EM algorithms for penalized maximum-likelihood image reconstruction. *Technical Report*, Technical Report 286, Communications and Signal Processing Laboratory, Department of EECS, the University of Michigan.

Gong, C., Han, C., Gan, G., Deng, Z., Zhou, Y., Yi, J. et al. (2017) Low-dose dynamic myocardial perfusion CT image reconstruction using pre-contrast normal-dose CT scan induced structure tensor total variation regularization. *Physics in Medicine & Biology*, 62, 2612.

Hütter, J.-C. & Rigollet, P. (2016) Optimal rates for total variation denoising. In: *Conference on Learning Theory*, pp. 1115–1146.

Kim, Y. & Gao, C. (2019) Bayesian model selection with graph structured sparsity. *arXiv preprint arXiv:1902.03316*.

Kim, S.-J., Koh, K., Boyd, S. & Gorinevsky, D. (2009) ℓ_1 trend filtering. *SIAM Review*, 51, 339–360.

Kleinberg, J. & Tardos, E. (2002) Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM (JACM)*, 49, 616–639.

Krahmer, F. & Ward, R. (2014) Stable and robust sampling strategies for compressive imaging. *IEEE Transactions on Image Processing*, 23, 612–622.

Krishnamuthy, A., Sharpnack, J. & Singh, A. (2013) Recovering graph-structured activations using adaptive compressive measurements. In: *2013 Asilomar Conference on Signals, Systems and Computers*, IEEE. pp. 765–769.

Li, Y., Mark, B., Raskutti, G. & Willett, R. (2018) Graph-based regularization for regression problems with highly-correlated designs. In: *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE. pp. 740–742.

Lustig, M., Donoho, D. & Pauly, J.M. (2007) Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58, 1182–1195.

Nam, S., Davies, M.E., Elad, M. & Gribonval, R. (2013) The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34, 30–56.

Needell, D. & Tropp, J.A. (2009) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26, 301–321.

Needell, D. & Ward, R. (2013a) Near-optimal compressed sensing guarantees for total variation minimization. *IEEE Transactions on Image Processing*, 22, 3941–3949.

Needell, D. & Ward, R. (2013b) Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6, 1035–1058.

Padilla, O.H.M., Sharpnack, J. & Scott, J.G. (2017) The dfs fused lasso: Linear-time denoising over general graphs. *The Journal of Machine Learning Research*, 18, 6410–6445.

Parikh, N. & Boyd, S. (2014) Proximal algorithms. *Foundations and Trends® in Optimization*, 1, 127–239.

Rinaldo, A. (2009) Properties and refinements of the fused lasso. *The Annals of Statistics*, 37, 2922–2952.

Rudelson, M. & Vershynin, R. (2008) On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61, 1025–1045.

Rudin, L.I., Osher, S. & Fatemi, E. (1992) Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60, 259–268.

Segars, W., Sturgeon, G., Mendonca, S., Grimes, J. & Tsui, B.M. (2010) 4D XCAT phantom for multimodality imaging research. *Medical Physics*, 37, 4902–4915.

Tibshirani, R.J. (2011) The solution path of the generalized lasso. PhD thesis, Stanford University.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Tropp, J.A. & Gilbert, A.C. (2007) Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53, 4655–4666.

Wang, Y.-X., Sharpnack, J., Smola, A.J. & Tibshirani, R.J. (2016) Trend filtering on graphs. *The Journal of Machine Learning Research*, 17, 3651–3691.

Xiao, L. & Zhang, T. (2013) A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23, 1062–1091.

Xu, L., Lu, C., Xu, Y. & Jia, J. (2011) Image smoothing via l_0 gradient minimization. *ACM Transactions on Graphics (TOG)*, 30, 174.

Zhang, T. (2011) Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57, 6215–6221.

Zhang, Y., Wainwright, M.J. & Jordan, M.I. (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In: *Conference on Learning Theory*, pp. 921–948.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Xu S, Fan Z. Iterative Alpha Expansion for estimating gradient-sparse signals from linear measurements. *J R Stat Soc Series B*. 2021;83:271–292. https://doi.org/10.1111/rssb.12407