Appl. Comput. Harmon. Anal. ••• (••••) •••–•••

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Theoretical guarantees for graph sparse coding

Yael Yankelevsky *, Michael Elad

Computer Science Department, Technion - Israel Institute of Technology, Haifa 3200003, Israel

ARTICLE INFO

Article history:
Received 29 August 2018
Received in revised form 13 March 2019
Accepted 20 March 2019
Available online xxxx
Communicated by Charles K. Chui

Keywords:
Sparse representations
Graph sparse coding
Graph regularization
Manifold learning
Signal recovery
Orthogonal matching pursuit
Basis pursuit

ABSTRACT

Over the last decade, the sparse representation model has led to remarkable results in numerous signal and image processing applications. To incorporate the inherent structure of the data and account for the fact that not all support patterns are equally likely, this model was enriched by enforcing various structural sparsity patterns. One plausible such extension of classic sparse coding, instigated by the emergence of graph signal processing, is graph regularized sparse coding. This model explicitly considers the intrinsic geometrical structure of the data domain, and has been successfully employed in various applications. However, emphasis was given to developing algorithmic solutions, and to date, the theoretical foundations to this problem have been lagging behind. In this work, we fill this gap and present a novel theoretical analysis of the graph regularized sparse coding problem, providing worst-case guarantees for the stability of the obtained solution, as well as for the success of several pursuit techniques. Furthermore, we formulate the conditions for which the superiority of the graph regularized sparse coding solution over the structure-agnostic sparse coding counterpart is established.

1. Introduction

Sparse coding (SC) has become a popular paradigm for data representation and has been proven effective in practical image processing and computer vision tasks such as denoising [1–3], inpainting [4], deblurring [5], super-resolution [6], object recognition or tracking [7–10] and classification [11–13]. In this framework, one assumes a signal $\mathbf{y} \in \mathbb{R}^N$ to be a sparse combination of a few columns (or atoms) from a collection $\mathbf{D} \in \mathbb{R}^{N \times K}$, termed the dictionary. Put differently, $\mathbf{y} = \mathbf{D}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^K$ is a sparse vector. Finding such a vector can be formulated as the following optimization problem:

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{1}$$

This is known as the (P_0) problem for obtaining exact recovery.

* Corresponding author.
E-mail addresses: yaelyan@cs.technion.ac.il (Y. Yankelevsky), elad@cs.technion.ac.il (M. Elad).

https://doi.org/10.1016/j.acha.2019.03.003
1063-5203/

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

When dealing with natural signals, the (P_0) problem is often relaxed to consider model deviations as well as measurement noise, leading to the (P_0^ϵ) problem:

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 \leq \epsilon^2. \tag{2}$$

In this setup one assumes $\mathbf{y}=\mathbf{D}\mathbf{x}+\mathbf{v}$ where $\mathbf{v}$ is a nuisance vector of finite energy $\|\mathbf{v}\|_2 \leq \epsilon$.

Given multiple signals to be sparse coded over the same dictionary, these problems can be solved for each signal independently using standard SC techniques. However, such methods fail to consider the geometrical structure of the data domain and disregard the locality and similarity among the signals to be coded. Furthermore, due to the typical redundancy of the representative dictionary, small variations in the data may result in very distinct representations, compromising the coding robustness.

To mitigate these limitations, several SC methods have been proposed that model the dependencies between dictionary elements and enforce structural sparsity patterns, for example by adding spatial consistency constraints [14], a hierarchical tree structure [15–18], or block-sparsity [19,20]. Other works introduce joint- or group-sparsity for simultaneous coding of multiple signals, encouraging their representations to use identical or correlated subsets of atoms [21–27].

Another approach, motivated by the recent progress in spectral graph theory and manifold learning, is graph regularized sparse coding (GRSC), which explicitly exploits the local geometric structure of the data to alleviate the representation instability. The underlying assumption is that in many real applications, the data is likely to reside on or near a low-dimensional manifold embedded in the high-dimensional ambient space [28–30]. Moreover, if two data points are close in the intrinsic data manifold, then their representations in any other domain are assumed to be close as well. Encoding the manifold structure by a graph, its Laplacian matrix can thus be incorporated into the sparse coding framework as a regularizer that preserves these similarities in the data domain. In recent years, such regularization has become prevalent in image processing for describing pairwise relationships between image pixels or patches [31–38].

GRSC and its extensions have been successfully employed for tasks of denoising [37,39,40], action recognition [41,42], classification and clustering [43,33,44–49]. However, while various algorithms have been proposed for obtaining the GRSC solution, no theoretical guarantees are currently known for the success of these methods.

In this work, we address the theoretical aspects of the GRSC problem. To this end, we generalize mathematical quantities such as the ℓ_0 norm, ERC and RIP to their counterparts in the graph constrained setting, capturing both local properties for each signal and global measures for the ensemble. Doing so, we offer the first meaningful analysis of the stability of the GRSC solution and terms of success of pursuit algorithms, as well as formulate the conditions for which its superiority is established over the classic, structure-agnostic sparse coding.

The paper is organized as follows. Section 2 revisits the graph sparse representation model and introduces its reformulation that will serve our analysis. In Section 3 we present the main stability result for this model, followed by a thorough analysis in Section 4, highlighting the conditions for which it is most beneficial. Section 5 presents stability results for common pursuit algorithms, that are then numerically validated in Section 6. We finally conclude in Section 7 and discuss further research directions.

2. Graph regularized sparse coding

Consider a set of signals $\{\mathbf{y}_1,...,\mathbf{y}_M\} \in \mathbb{R}^N$, constituting the columns of the data matrix $\mathbf{Y} \in \mathbb{R}^{N\times M}$. Let us construct a weighted graph $\mathcal{M}$ with M nodes (vertices), each node representing a signal or data point. The weight w_{ij} assigned to the edge connecting the i-th and j-th nodes is designed to be inversely proportional to the distance between them. A common choice is applying a Gaussian kernel function,

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

$$w_{ij} = \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|_2^2}{\epsilon_{\mathcal{M}}}\right), \tag{3}$$

where $\epsilon_{\mathcal{M}}$ is a properly chosen kernel scale parameter. The graph Laplacian matrix $\mathbf{L} \in \mathbb{R}^{M\times M}$ is then defined as $\mathbf{L} = \mathbf{D}^{\mathcal{M}} - \mathbf{W}$, where the graph adjacency matrix $\mathbf{W}$ consists of the edge weights w_{ij}, and the degree matrix $\mathbf{D}^{\mathcal{M}}$ is a diagonal matrix whose entries are $\mathbf{D}^{\mathcal{M}}_{ii} = \sum_j w_{ij}$.

The graph regularized sparse coding problem is formulated as:

$$\arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \beta Tr(\mathbf{X}\mathbf{L}\mathbf{X}^T) \quad \text{s.t.} \quad \|\mathbf{x}_i\|_0 \le T \quad \forall i, \tag{4}$$

where $\mathbf{D} \in \mathbb{R}^{N\times K}$ is the dictionary, and $\mathbf{X} \in \mathbb{R}^{K\times M}$ is the sparse representations matrix corresponding to the data in $\mathbf{Y}$, having the individual signal representations $\mathbf{x}_i$ as its columns. A slightly different formulation, representing the equivalent of the (P_0^ϵ) problem for the GRSC setting, becomes

$$\arg\min_{\mathbf{X}} \|\mathbf{X}\|_{0,\infty} \quad \text{s.t.} \quad \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \beta Tr(\mathbf{X}\mathbf{L}\mathbf{X}^T) \le \epsilon^2. \tag{5}$$

Sparsity is here measured via the $\ell_{0,\infty}$ mixed norm, counting the maximal number of non-zeros in the columns of $\mathbf{X}$. That is, formally, $\|\mathbf{X}\|_{0,\infty} = \max_i \|\mathbf{x}_i\|_0$.

Observe that $Tr(\mathbf{X}\mathbf{L}\mathbf{X}^T) = \frac{1}{2}\sum_{i,j} w_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. Minimizing this term therefore encourages similar signals, having a large proximity measure w_{ij}, to have similar sparse representations, thus satisfying the commonly known manifold assumption [29]. In other words, the added regularization limits the degree of freedom in the sparse coding task and favors solutions preserving the local intrinsic geometry, i.e. varying smoothly along the geodesics of the underlying data manifold.

From a different perspective, the GRSC can be thought of as a hybrid model, combining two different approaches for non-linear dimensionality reduction methods: manifold learning and sparse representations. Yet unlike other manifold embeddings (e.g. [29,30]), the geometry preserving sparse representations have the merit of being reversible. That is, by multiplying with the dictionary one can shift back from the embedded domain to the original signal space, for serving signal recovery tasks.

Note that graph regularized sparse coding (GRSC) concerns an ensemble of signals, given through the columns of a matrix $\mathbf{Y}$, rather than a single signal $\mathbf{y}$. Furthermore, due to the imposed graph constraint, the signals are jointly coded, i.e. the problems are no longer independent but are instead coupled through the manifold Laplacian $\mathbf{L}$. This calls for the development of new pursuit algorithms suited for the graph regularized setting.

Indeed, several works have recently studied this problem. Zheng et al. [33] proposed to solve the ℓ_1 counterpart of Equation (4),

$$\arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \beta Tr(\mathbf{X}\mathbf{L}\mathbf{X}^T) + \gamma \sum_i \|\mathbf{x}_i\|_1, \tag{6}$$

using a coordinate descent approach and subgradient methods. Other previously proposed methods are based on the feature sign search algorithm [43] or a modified sequential quadratic programming [34]. A similar approach was taken in [38] by applying the Laplacian regularization on the reconstructed data $\mathbf{D}\mathbf{X}$ rather than on the sparse representation $\mathbf{X}$. In [39] we proposed a different solution based on the Alternating Direction Method of Multipliers (ADMM) [50], which enables simultaneous update of all columns of $\mathbf{X}$.

Nevertheless, despite the growing interest in this problem and the various algorithms proposed for obtaining its solution, its theoretical foundation is yet to be established, which is the aim of this work.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

2.1. GRSC reformulated

In order to analyze the GRSC problem given in Equation (5), we modify its formulation to use an effective dictionary that integrates both the original dictionary $\mathbf{D}$ and the manifold Laplacian $\mathbf{L}$. For that purpose, let us denote the vectorized versions of $\mathbf{X}$ and $\mathbf{Y}$ by $\mathbf{Z} \triangleq vec(\mathbf{X}) = \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_M^T\right]^T$ and $\tilde{\mathbf{Y}} \triangleq vec(\mathbf{Y}) = \left[\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_M^T\right]^T$, respectively. Additionally, define the global dictionary $\tilde{\mathbf{D}} = \mathbf{I}_M \otimes \mathbf{D}$ and the global manifold Laplacian $\tilde{\mathbf{L}} = \mathbf{L} \otimes \mathbf{I}_K$, where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix and $\otimes$ is the Kronecker product.

With slight abuse of notation, we define the block-sparsity $\ell_{0,\infty}$ measure as $\|\mathbf{Z}\|_{0,\infty} = \max_i \|\mathbf{Z}_i\|_0$, counting the maximal number of non-zeros in non-overlapping segments of length K from the vector $\mathbf{Z}$. This vector norm is equivalent to the standard $\|\mathbf{X}\|_{0,\infty}$ norm applied to the matrix form of the representations in Equation (5).

Symmetrically, we define the $\ell_{\infty,0}$ sparsity measure $\|\mathbf{Z}\|_{\infty,0} = \|\mathbf{X}^T\|_{0,\infty}$, counting the maximal number of non-zeros in each *row* of the corresponding representation matrix $\mathbf{X}$.

Note that since both $\mathbf{L}$ and $\mathbf{I}_K$ are symmetric and positive semi-definite, so is $\tilde{\mathbf{L}}$. Therefore, utilizing its eigendecomposition $\tilde{\mathbf{L}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, we can extract the square root $\tilde{\mathbf{Q}} = \sqrt{\tilde{\mathbf{L}}} = \sqrt{\mathbf{\Lambda}}\mathbf{V}^T$ such that $\tilde{\mathbf{L}} = \tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}$. Using this notion and the above relations, the GRSC problem (i.e. (5)) becomes

$$\arg\min_{\mathbf{Z}} \|\mathbf{Z}\|_{0,\infty} \quad \text{s.t.} \quad \left\|\hat{\mathbf{Y}} - \mathbf{A}\mathbf{Z}\right\|_2^2 \leq \epsilon^2, \tag{7}$$

where we have denoted $\mathbf{A} = \begin{bmatrix} \tilde{\mathbf{D}} \\ \sqrt{\beta}\tilde{\mathbf{Q}} \end{bmatrix}$ and $\hat{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{Y}} \\ 0 \end{bmatrix}$.

We will hereafter refer to $\tilde{\mathbf{D}}$ as the global dictionary and to $\mathbf{A}$ as the generalized effective dictionary. Equation (7) will be referred to as the $(P^{\epsilon}_{0,\infty})$ problem in this block-sparsity context.

Before proceeding, we define the following terms that will be used throughout the analysis: The support of a sparse vector $\mathbf{Z}$, representing the set of indices corresponding to its non-zero entries, will be denoted by Ω. The minimal and maximal absolute values of the vector $\mathbf{Z}$ within its support will be denoted $Z_{\min} = \min_{i\in\Omega} |\mathbf{Z}_i|$ and $Z_{\max} = \max_{i\in\Omega} |\mathbf{Z}_i|$. The smallest and largest manifold weights will be denoted by $L_{\min} = \min_{i\neq j} |\mathbf{L}_{ij}|$ and $L_{\max} = \max_{i\neq j} |\mathbf{L}_{ij}|$. The minimal and maximal node degrees in the manifold graph will be denoted by $\Delta_{\min} = \min_i \mathbf{L}_{ii}$ and $\Delta_{\max} = \max_i \mathbf{L}_{ii}$. Finally, we denote $\Theta_L = \sqrt{\frac{1+\beta\Delta_{\max}}{1+\beta\Delta_{\min}}}$.

Having those definitions at hand, we can turn to analyze the reformulated $(P^{\epsilon}_{0,\infty})$ problem.

3. Stability of the $(P^{\epsilon}_{0,\infty})$ solution

Assume a block-sparse vector $\mathbf{Z}$ with $\|\mathbf{Z}\|_{0,\infty} \leq s$ and $\|\mathbf{Z}\|_{\infty,0} \leq \eta$ satisfies $\|\hat{\mathbf{Y}} - \mathbf{A}\mathbf{Z}\|_2^2 \leq \epsilon^2$. Suppose we solve the above $(P^{\epsilon}_{0,\infty})$ problem and obtain a solution $\hat{\mathbf{Z}}$. How close is it to the original $\mathbf{Z}$?

Much like the (P^{ϵ}_0) problem defined in Equation (2), one cannot claim the uniqueness of a solution to the $(P^{\epsilon}_{0,\infty})$ problem (7), but instead can guarantee that it will be close enough to the true underlying block-sparse vector $\mathbf{Z}$ that generated the data.

This kind of stability results have traditionally been derived by leveraging the Restricted Isometry Property (RIP) [51,52]. Similar stability claims can be formulated in terms of the mutual coherence [53], by exploiting its relationship with the RIP property [54].

Recall that the mutual coherence, which quantifies the similarity of the atoms in the dictionary $\mathbf{D}$, was defined in [55] as:

$$\mu(\mathbf{D}) = \max_{i\neq j} \frac{\left|\mathbf{d}_i^T\mathbf{d}_j\right|}{\|\mathbf{d}_i\|_2\|\mathbf{d}_j\|_2}. \tag{8}$$

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

Assuming hereafter that the atoms are normalized such that $\|\mathbf{d}_i\|_2 = 1\ \forall i$, it is equivalently given by $\mu(\mathbf{D}) = \max_{i\neq j} \left|\mathbf{d}_i^T \mathbf{d}_j\right|$.

Evidently, if each original signal is coded with T or fewer non-zeros, i.e. $\|\mathbf{Z}\|_{0,\infty} = T$, then the sparsity of the global vector $\mathbf{Z}$ in the ℓ_0 norm sense is as high as $\|\mathbf{Z}\|_0 = TM$, making the traditional guarantees infeasible. Instead we shall derive alternative guarantees for the $\ell_{0,\infty}$ norm.

For better readability, the theorem proofs are deferred to Appendix A.

We first generalize the RIP definition [51] to the $\ell_{0,\infty}$ case, which we name Block-RIP.

Definition 1. A matrix $\mathbf{A}$ is said to have an asymmetric[1] k-BRIP (Block-RIP) with constants δ_k^L, δ_k^H if these are the smallest quantities such that

$$(1-\delta_k^L)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k^H)\|\mathbf{x}\|_2^2 \tag{9}$$

for every vector $\mathbf{x}$ satisfying $\|\mathbf{x}\|_{0,\infty} = k$.

Similarly to the RIP, computing the BRIP is hard or practically impossible, and so we shall bound it using the mutual coherence.

Theorem 2 *(Upper bounding the BRIP via the mutual coherence). For a support Ω with $\ell_{0,\infty}$ norm equal to k and with $\ell_{\infty,0}$ norm equal to η, the k-BRIP of the generalized effective dictionary $\mathbf{A}$ can be upper bounded by*

$$\begin{aligned} \delta_k^H &\leq (k-1)\mu(\mathbf{D}) + \beta[2\Delta_{\max} - (M-\eta)L_{\min}], \\ \delta_k^L &\leq (k-1)\mu(\mathbf{D}) - \beta(M-\eta)L_{\min}. \end{aligned} \tag{10}$$

Exploiting the result of Theorem 2, we devise a stability theorem for the graph $(P^{\epsilon}_{0,\infty})$ problem.

Theorem 3 *(Stability of the solution to the $(P^{\epsilon}_{0,\infty})$ problem). Consider a sparse vector $\mathbf{Z}$ such that $\|\mathbf{Z}\|_{\infty,0} = \eta$ and $\|\mathbf{Z}\|_{0,\infty} = s < \frac{1}{2}\left(1 + \frac{1+\beta(M-\eta)L_{\min}}{\mu(\mathbf{D})}\right)$, and a generalized effective dictionary $\mathbf{A}$ satisfying the BRIP property for $\ell_{0,\infty} = 2s$ with coefficients $\delta_{2s}^L, \delta_{2s}^H$. Then, the distance between the true sparse vector $\mathbf{Z}$ and the solution to the $(P^{\epsilon}_{0,\infty})$ problem, $\hat{\mathbf{Z}}$, is bounded by*

$$\|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_{2s}^L} \leq \frac{4\epsilon^2}{1-(2s-1)\mu(\mathbf{D}) + \beta(M-\eta)L_{\min}}. \tag{11}$$

Notice that since the manifold regularization constrains the rows of the original representation matrix $\mathbf{X}$, the devised stability term and corresponding bound are no longer expressed solely as a function of the dictionary, but instead describe a relation between the column sparsity s and the row sparsity η, integrating additional properties of the dictionary and the manifold graph Laplacian. While the column sparsity has a local (per-signal) flavor, the row sparsity represents a global measure for the signal ensemble.

4. Better together? A deeper analysis of guaranteeable performance

The major question arising in light of the newly formed bound is whether it offers any advantage compared with the traditional one. Put differently, given a collection of signals sampled from some manifold and the graph Laplacian modeling this manifold, is GRSC guaranteed to yield better results than individual sparse coding of these signals, which is oblivious to the underlying structure? And if so, under what conditions?

[1] By defining an asymmetric form of the RIP one can obtain tighter bounds than using the common symmetric RIP, as shown in [56].

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

4.1. Comparing the stability terms

Comparing the obtained bounds in an attempt to answer the aforementioned questions, observe that adding the manifold constraint with some $\beta > 0$ improves the stability term, since by definition $\eta \leq M$ and $L_{\min} \geq 0$.

Also note that for $\beta = 0$ we obtain the standard worst-case stability bound as devised for the (P_0^ϵ) problem. That is, our stability requirement in the $\ell_{0,\infty}$ sense aligns with the ℓ_0 sparsity requirement for each individual signal, and the error bound coincides with the sum of individual errors.

Another case for which we revert to the classic stability condition is if there exists some dictionary atom that is chosen by all ensemble signals, hence $\eta = M$. Given the richness and redundancy of the typical dictionary, it is reasonable to assume such case is rare. Moreover, observe that the sparser the rows of $\mathbf{X}$, the less sparse each column ought to be to guarantee better stability. A geometric interpretation suggests that a wider distribution of the chosen atoms is an indication of the diversity of the signal ensemble, in which case there is a higher potential gain from capitalizing on the manifold assumption.

Finally, we should note that the above devised stability condition could be further improved by replacing the term $(M-\eta)L_{\min}$ with $S_\eta(\mathbf{L})$, the sum of the $M-\eta$ smallest weights in $\mathbf{L}$. Clearly, $S_\eta(\mathbf{L}) \geq (M-\eta)L_{\min}$. This modification better accommodates sparse graphs, and leads to an improved stability condition as long as the minimal number of non-zeros in each row of $\mathbf{L}$ is larger than η.

To conclude, except for the above mentioned cases where the terms are identical, in all other cases the GRSC stability terms necessarily improve compared with non-regularized sparse coding. Moreover, the stronger the enforced regularity (i.e. larger β), the larger the graph weights, or the more diverse the ensemble (i.e. smaller η) – the looser and better the sparsity requirement.

4.2. Comparing the guaranteed error bounds

So far we have shown that GRSC requires a more relaxed sparsity condition to yield a stable global solution compared with the classical SC. For cardinalities satisfying the stability terms for both GRSC and standard SC, we may compare the resulting stability bounds and study the conditions for which incorporating graph constraints leads to lower guaranteed approximation error.

Allegedly, the larger β, the better the obtained stability bound. While this may seem true, recall that β is the relative weight of manifold smoothness, which implicitly effects the noise level as well.

Revisiting Equation (4), let us denote the data error by $\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 = \epsilon_D^2$ and the manifold smoothness error by $Tr(\mathbf{X}\mathbf{L}\mathbf{X}^T) = \epsilon_G^2$. Since for GRSC $\epsilon^2 = \epsilon_D^2 + \beta\epsilon_G^2$, it is evident that increasing β comes along with increasing ϵ^2. Therefore, though improving the stability terms, the stability bound may overall not improve. To obtain a tighter bound, we seek conditions for which

$$\frac{4\epsilon_D^2 + 4\beta\epsilon_G^2}{1-(2s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}} < \frac{4\epsilon_D^2}{1-(2s-1)\mu(\mathbf{D})}. \tag{12}$$

This is satisfied whenever

$$\frac{\epsilon_G^2}{(M-\eta)L_{\min}} < \frac{\epsilon_D^2}{1-(2s-1)\mu(\mathbf{D})}, \tag{13}$$

which can be translated to a lower bound on the block-sparsity level s, namely

$$s > \frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D})}\right) - \frac{\epsilon_D^2(M-\eta)L_{\min}}{2\mu(\mathbf{D})\epsilon_G^2}. \tag{14}$$

Combined with the known stability terms, we deduce that the GRSC stability bound is superior to that of classic SC whenever

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

$$\frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D})}\right)-\frac{\epsilon_D^2(M-\eta)L_{\min}}{2\mu(\mathbf{D})\epsilon_G^2}<s<\frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D})}\right). \tag{15}$$

Put differently, if the manifold Laplacian is scaled such that

$$\frac{(M-\eta)L_{\min}}{\epsilon_G^2}>\frac{1-\mu(\mathbf{D})}{\epsilon_D^2} \tag{16}$$

with respect to the dictionary related properties, GRSC will lead to better stability bounds for all feasible cardinalities. Surprisingly, this relation does not depend on the choice of β. Yet the smaller η or higher the degree of smoothness (i.e. smaller ϵ_G), the more pronounced the theoretical advantage of GRSC.

To summarize the comparison, we have established that GRSC guarantees a stable solution for a wider range of cardinalities. Moreover, given cardinalities for which standard SC has guaranteed stability as well, the stability bound provided by GRSC is tighter if the cardinality is lower bounded, which is a reasonable assumption in non-extreme cases.

4.3. Numerical evaluation

Following the above analysis, we now provide a numerical experiment demonstrating the obtained bounds.

To minimize the mutual coherence, the dictionary is built as a Grassmannian matrix of size 64×128 [54], yielding a mutual coherence as low as $\mu(\mathbf{D})\approx0.1$. As a realistic estimation for the other parameter values, we use $M=50$, $\epsilon_D=0.1\sqrt{M}$, $\epsilon_G=0.1$, $L_{\min}=0.02$ and $\beta=1$.

For these settings, stability of the standard (P_0^ϵ) solution is guaranteed for cardinalities up to 5, while for the graph constrained $(P_{0,\infty}^\epsilon)$ this maximal cardinality can go up to 10, depending on η.

Fig. 1a illustrates the comparison between the theoretical guarantees provided for classic SC ($\beta=0$) and GRSC ($\beta=1$). All different combinations of s and η are divided into 4 regions: In the blue and red regions stability is guaranteed for both cases, where in the red region the GRSC stability bounds are stronger than those obtained for SC, whereas for the blue region the classic SC bounds are stronger (based on the analysis in Section 4.2). The green region represents cases for which stability is no longer guaranteed for classic SC yet still guaranteed for GRSC (according to Theorem 3). In the black region, stability is no longer guaranteed for either case. For $\eta=M=50$, the two bounds unite. As these results indicate, it is theoretically beneficial to incorporate the manifold regularization and jointly code the ensemble of signals.

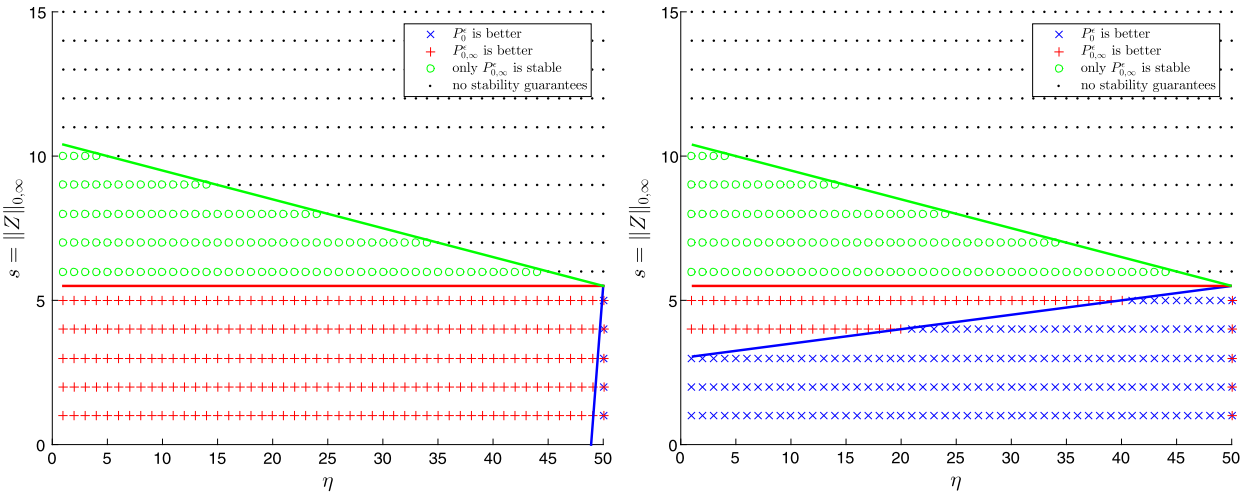
Repeating the experiment for an increased ϵ_G reveals a different picture. The stability terms are independent of ϵ_G and thus will not change, however the superiority of the GRSC stability bounds over those of classic SC is compromised. Setting $\epsilon_G=1$, observe that there are now cases where adding the manifold constraint does not help improve the theoretical stability, as demonstrated in Fig. 1b. The reasoning is that large values of η indicate that some atoms are chosen abundantly. Combined with the higher value of ϵ_G, this implies that for this setup, the manifold assumption is weaker and thus its potential contribution is expectedly lower.

To further evaluate the influence of the choice of β, we set $\eta=10$ and $\epsilon_G=1$ and plot the error bounds as a function of β for 3 different values of s, representing the different regimes illustrated in Fig. 1b. As can be observed in Fig. 2, when GRSC yields better bounds compared with classic SC, the bounds improve more as β grows. On the contrary, when GRSC degrades the bounds, they degrade more as β grows.

4.4. Analyzing the noiseless case

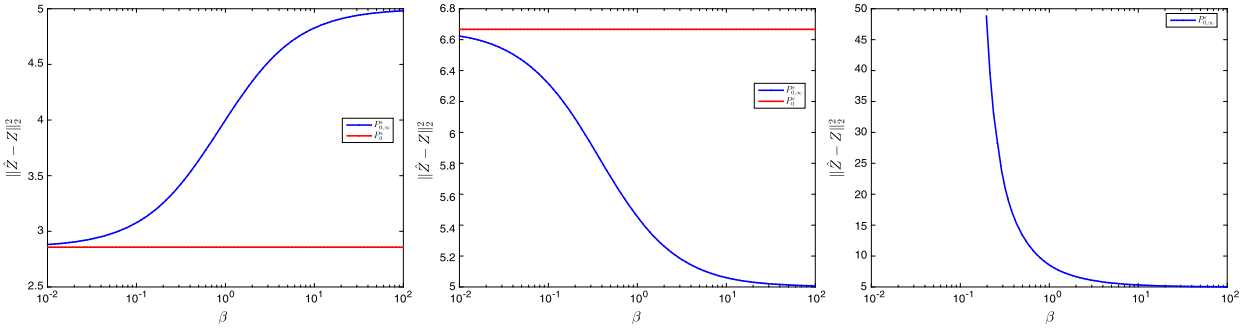
A final remark concerns the analysis of the noiseless setup. While the $(P_{0,\infty}^\epsilon)$ is meaningless for $\epsilon=0$, a more realistic scenario is the semi-noiseless case, where $\mathbf{Y}=\mathbf{DX}$ yet $Tr(\mathbf{XLX}^T)>0$. Recall that if the true cardinality satisfies the traditional (P_0) condition, exact recovery is guaranteed without adding the graph

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003



(a) $\epsilon_G = 0.1$ (b) $\epsilon_G = 1$

Fig. 1. Comparison of the theoretical stability guarantees for P_0^ϵ (classic SC) and $P_{0,\infty}^\epsilon$ (GRSC), for different levels of manifold regularity ϵ_G.



(a) $s = 2$ (b) $s = 4$ (c) $s = 6$

Fig. 2. Comparison of the theoretical stability bounds for P_0^ϵ (classic SC) and $P_{0,\infty}^\epsilon$ (GRSC) as a function of β, for $\epsilon_G = 1$, $\eta = 10$ and different values of s.

constraints. However, for larger cardinalities that do not satisfy this requirement but do satisfy the more relaxed $(P_{0,\infty}^\epsilon)$ condition (for $\epsilon^2 = \beta\epsilon_G^2$), one could still guarantee stable recovery. Thus by using GRSC we extend the previous recovery range.

In the numerical setup presented above, for example, classic SC guarantees unique recovery for cardinalities up to (and including) $s = 5$. As for GRSC, stable recovery can be guaranteed up to a higher maximal cardinality of $s = 10$, depending on η. Specifically, for $\eta = 10$, GRSC accommodates cardinalities of $s \leq 9$. While classic SC provides no guarantee beyond $s = 5$, GRSC can guarantee recovery with a bounded error, which is at worst $\|\hat{\mathbf{Z}} - \mathbf{Z}\|_2^2 \leq 0.4$ for the maximal cardinality $s = 9$.

5. Stability guarantees for pursuit algorithms

Up until now we have shown that the solution to the $(P_{0,\infty}^\epsilon)$ problem will be close to the true sparse vector $\mathbf{Z}$. However, we have not yet guaranteed the feasibility of obtaining such a solution. It is therefore important to know whether this solution can be approximated by pursuit algorithms.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

In this section, we address such a question for the Orthogonal Matching Pursuit (OMP) [57], Basis Pursuit (BP) [58], and Thresholding Algorithms.[2]

Classic known bounds for these algorithms typically involve both the sparsity of $\mathbf{Z}$ and the signal-to-noise ratio, which relates to the term $\frac{\epsilon}{|Z_{\min}|}$. In the context of GRSC, such results provide weak and nearly meaningless bounds, because they rely on the global ℓ_0 sparsity measure rather than on the block-sparsity $\ell_{0,\infty}$, and because they are based on the global error energy ϵ which could be quite high.

So, largely following [59], we harness the inherent locality of the global dictionary atoms in order to replace the global error bound with a local block-error, as well as replace the ℓ_0 norm with the block-sparsity $\ell_{0,\infty}$. For this purpose, and due to the special structure of our problem, we need further assumptions about the noise distribution.

Throughout the following analysis, we slightly change notation and suppose a clean signal $\mathbf{X}$, having a representation $\mathbf{AZ}$ over a generalized effective dictionary $\mathbf{A} = \begin{bmatrix} \tilde{\mathbf{D}} \\ \sqrt{\beta}\tilde{\mathbf{Q}} \end{bmatrix}$, is contaminated with noise $\mathbf{E} = \begin{bmatrix} \mathbf{E_D} \\ \mathbf{E_L} \end{bmatrix}$ to create the measurement $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ such that $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \epsilon$. Let us denote the top part of $\mathbf{Y}$, corresponding to the dictionary, by $\mathbf{Y_D}$, and its bottom part corresponding to the Laplacian by $\mathbf{Y_L}$.

To prove the recovery guarantees for all considered algorithms, we assume that $\mathbf{E_L}$ has energy $\sqrt{\beta}\epsilon_G$, or $\|\mathbf{E_L}\|_2^2 = \beta\epsilon_G^2$, and that $\mathbf{E_D}$ has energy $\|\mathbf{E_D}\|_2^2 = \epsilon_D^2$. We further denote by ϵ_l the highest energy of all N-dimensional non-overlapping blocks extracted from the upper part of $\mathbf{E}$, i.e. $\mathbf{E_D}$ has block-energy ϵ_l. Finally, define the effective noise $\epsilon_{eff} = \epsilon_l + \beta\epsilon_G\sqrt{\Delta_{\max}}$.

The proofs for all theorems in this section are provided in Appendix B.

5.1. Stability of the Thresholding algorithm

We commence by developing a stability theorem for the thresholding algorithm, which is not only far simpler than OMP and BP, but is also closer in spirit to the ADMM based graph sparse coding algorithm we proposed in [39].

Recall that the thresholding algorithm is a simplification of OMP that relies on a single projection, choosing the support as the largest inner products of $|\mathbf{A}^T\hat{\mathbf{Y}}|$ [54].

Theorem 4 *(Thresholding algorithm performance in the presence of noise). Suppose a clean signal* $\mathbf{X}$ *has a representation* $\mathbf{AZ}$ *over a generalized effective dictionary* $\mathbf{A}$*, and that it is contaminated with noise* $\mathbf{E}$ *to create the measurement* $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ *such that* $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \epsilon$.

If $\mathbf{Z}$ *satisfies*

$$s = \|\mathbf{Z}\|_{0,\infty} < \frac{1}{1+\Theta_L}\left(1 + \frac{1}{\mu(\mathbf{D})}\frac{|Z_{\min}|}{|Z_{\max}|}\right) - \frac{\epsilon_l}{\mu(\mathbf{D})|Z_{\max}|} \tag{17}$$

where $|Z_{\min}|, |Z_{\max}|$ *are the minimum and maximum absolute values of the vector* $\mathbf{Z}$ *within its support, then we are guaranteed that the thresholding algorithm finds the correct support, and its solution, denoted* $\mathbf{Z_{THR}}$*, satisfies*

$$\|\mathbf{Z_{THR}} - \mathbf{Z}\|_2^2 \leq \frac{\epsilon^2}{1 - (s-1)\mu(\mathbf{D}) + \beta(M-\eta)L_{\min}}. \tag{18}$$

Recall that $\Theta_L = \sqrt{\frac{1+\beta\Delta_{\max}}{1+\beta\Delta_{\min}}}$ where $\Delta_{\min}, \Delta_{\max}$ are the minimal and maximal node degrees in the manifold graph. Therefore, note that for $\beta = 0$, or for a homogeneous manifold graph (i.e. when all node

[2] These algorithms are not the common approach for solving the GRSC problem, albeit in the reformulated problem they could provide an alternative with provable guaranteed performance.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

degrees are the same), we have $\Theta_L = 1$ and so the obtained stability condition almost reduces to the classic known one. A minor difference is that the values $|Z_{\min}|, |Z_{\max}|$ are computed globally over the ensemble. This results in a tighter and more strict sparsity requirement, dictated by the signal with the most extreme entries. Further, observe that while jointly coding the ensemble of signals, the stability condition restricts the sparsity in the $\ell_{0,\infty}$ sense, and considers the local (per-signal) noise level ϵ_l rather than the global one.

5.2. Stability guarantee of OMP

Next, we provide guarantees for stable recovery of OMP, generalizing similar claims from [60] to the GRSC setting.

Theorem 5 *(Stable recovery of OMP in the presence of noise). Suppose a clean signal* $\mathbf{X}$ *has a representation* $\mathbf{AZ}$ *over a generalized effective dictionary* $\mathbf{A}$*, and that it is contaminated with noise* $\mathbf{E}$ *to create the measurement* $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ *such that* $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \epsilon$.

If $\mathbf{Z}$ *satisfies*

$$s = \|\mathbf{Z}\|_{0,\infty} < \frac{1}{1+\Theta_L}\left(1 + \frac{1 - \Theta_L \beta \eta L_{\max}}{\mu(\mathbf{D})}\right) - \frac{1}{\mu(\mathbf{D})} \cdot \frac{\epsilon_{eff}}{|Z_{\min}|} \tag{19}$$

where $|Z_{\min}|$ *is the minimum absolute value of the vector* $\mathbf{Z}$ *within its support, then running OMP for* $\|\mathbf{Z}\|_0$ *iterations, we are guaranteed that OMP finds the correct support, and its solution, denoted* $\mathbf{Z_{OMP}}$*, satisfies*

$$\|\mathbf{Z_{OMP}} - \mathbf{Z}\|_2^2 \leq \frac{\epsilon^2}{1 - (s-1)\mu(\mathbf{D}) + \beta(M - \eta)L_{\min}}. \tag{20}$$

Note again that when $\beta = 0$ we collapse to the known bounds. Yet, for $\beta > 0$, the sparsity requirement for OMP to succeed is more strict compared with both the corresponding requirement for the stability of $(P^{\epsilon}_{0,\infty})$ and the requirement for the non-regularized case ($\beta = 0$). This could be explained by the fact that the problem is indeed more challenging: ensemble coding implies satisfying the requirements for each individual signal, while also forcing additional constraints on the relations between them. This also reflects the fact that a greedy strategy for the GRSC problem is more sensitive to errors due to the multitude of signals and their cross dependences. However, given that the requirement is fulfilled, not only is the support correctly recovered but, under mild assumptions, the coefficient values are guaranteed to be closer to those of the true solution.

5.3. Stability guarantee of basis pursuit

Similarly to the classic (P^{ϵ}_0) problem, the solution of the $(P^{\epsilon}_{0,\infty})$ problem can be approximated using the Basis Pursuit Denoising (BPDN) algorithm, by relaxing the $\ell_{0,\infty}$ norm with the convex ℓ_1. The BPDN in its Lagrangian form is defined as

$$\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{Y} - \mathbf{AZ}\|_2^2 + \lambda \|\mathbf{Z}\|_1. \tag{21}$$

The stability of BPDN was proven in [61] given that the Exact Recovery Condition (ERC) is satisfied. Recall that as proposed in [62], the ERC is met for a support Ω whenever

$$\theta = ERC(\Omega) = 1 - \max_{i \notin \Omega} \|\mathbf{D}_{\Omega}^{\dagger}\mathbf{d}_i\|_1 > 0, \tag{22}$$

where we have denoted by $\mathbf{D}_{\Omega}^{\dagger}$ the Moore-Penrose pseudoinverse of the dictionary restricted to support Ω. Equivalently, the ERC was shown to hold whenever the total number of non-zeros in Ω is upper bounded.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

Exploiting the unique structure of the generalized effective dictionary, we can now establish similar claims for the GRSC model in terms of the $\ell_{0,\infty}$ norm.

Theorem 6 *(ERC for the graph sparse coding problem). Given a generalized effective dictionary* $\mathbf{A}$*, the ERC condition* (22) *is met for every support* Ω *that satisfies*[3]

$$s = \|\Omega\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1 + \beta(M-\eta)L_{\min} - \beta\eta L_{\max}}{\mu(\mathbf{D})}\right) \tag{23}$$

where $\|\Omega\|_{\infty,0} = \eta$ *and* $L_{\min}, L_{\max}$ *are the smallest and largest manifold weights, respectively.*

To understand the above result, note that in order for the sparsity requirement to be meaningful, and specifically improve over the structure agnostic SC, we must have η small enough to obey $\eta < \frac{M}{1+L_{\max}/L_{\min}}$. Generally, the more homogeneous the manifold weights, the better the ERC guarantees we could obtain. In the optimal case, if $L_{\max}/L_{\min} = 1$, then $\eta < M/2$ guarantees the ERC holds for sparse matrices X with a higher cardinality $\|\mathbf{X}\|_{0,\infty}$. Otherwise, the more varied the weights, the smaller should η be to improve the ERC condition over standard SC. In other words, varied manifold weights necessitate sampling a more diverse ensemble or using a richer dictionary, either of which will result in more distributed atom selection.

Having extended the ERC for the GRSC model, we follow the analysis presented in [61,59] and propose a stability claim for the Lagrangian BPDN algorithm (21), proving that it manages to approximate the solution of the $(P_{0,\infty})$ problem.

Theorem 7 *(Stable recovery of BP in the presence of noise). Suppose a clean signal* $\mathbf{X}$ *has a representation* $\mathbf{AZ}$ *over a generalized effective dictionary* $\mathbf{A}$*, and that it is contaminated with noise* $\mathbf{E}$ *to create the measurement* $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ *such that* $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \epsilon$*.*

If $\mathbf{Z}$ *satisfies*

$$s = \|\mathbf{Z}\|_{0,\infty} < \frac{1}{3}\left(1 + \frac{1 + \beta(M-\eta)L_{\min} - 2\beta\eta L_{\max}}{\mu(\mathbf{D})}\right), \tag{24}$$

we are guaranteed that $\mathbf{Z_{BP}}$*, the solution to the Lagrangian BP formulation with parameter* $\lambda = 4\epsilon_{eff}$*, satisfies the following:*

1. *The support of* $\mathbf{Z_{BP}}$ *is contained in* $\Omega = supp(\mathbf{Z})$*.*
2. $\|\mathbf{Z_{BP}} - \mathbf{Z}\|_\infty < \frac{15}{2}\epsilon_{eff}$*.*
3. *The support of* $\mathbf{Z_{BP}}$ *contains every index* $i \in \Omega$ *for which* $|\mathbf{Z}_i| > \frac{15}{2}\epsilon_{eff}$*.*
4. *The minimizer of the problem,* $\mathbf{Z_{BP}}$*, is unique.*

A corollary of the third point is that the complete support must be recovered whenever $\frac{\epsilon_{eff}}{|Z_{\min}|} < \frac{2}{15}$.

Observe that compared with the ERC, the stability condition here implies a stricter sparsity requirement. Explicitly, this condition is only meaningful when $\eta < \frac{M}{1+2L_{\max}/L_{\min}}$, implying that in the optimal case of homogeneous manifold weights, no more than one third of the signal ensemble may choose to use the same atom.

Comparing this result with the stability claims for the Thresholding and OMP, the stability terms here are no longer sensitive to $|Z_{\min}|, |Z_{\max}|$. Further, the difference between the entries in $\mathbf{Z_{BP}}$ and $\mathbf{Z}$ is bounded in terms of the effective noise level ϵ_{eff}. As a consequence, all atoms with coefficients above this measure are guaranteed to be recovered.

[3] This is a slight abuse of notation. Formally, the $\ell_{0,\infty}$ norm should apply to a sparse vector rather than the support Ω.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

5.4. An iterative thresholding alternative

The theoretical guarantees provided for classic pursuit algorithms are formed in terms of the $\ell_{0,\infty}$ norm, albeit the algorithms themselves are applied globally, considering the ℓ_0 or ℓ_1 norm of the vector $\mathbf{Z}$, and ignoring the problem structure. Furthermore, while the OMP and BP algorithms can be practically used to solve the reformulated GRSC problem (7) for low dimensional data, the tensor products leading to this formulation dictate a very high dimensional regime, for which using these or similar methods may be infeasible. To accommodate high dimensional setups, several dedicated algorithms have been proposed that directly tackle the GRSC objective in its original form. In what follows we aim at bridging the gap between these algorithms and the classic ones, so as to combine the theoretical benefits with the practical aspects.

One approach for doing so is based on migrating to the ℓ_1 norm and solving Equation (6). Revisiting the ADMM-GRSC algorithm proposed in [39], this algorithm can be extended to solve Equation (6) by replacing the hard-thresholding operator with a soft one. By equivalence to the reformulated problem (21) and due to the convex nature of the objective, the modified ADMM algorithm is guaranteed to converge to the global BP solution, thus maintaining the validity of the theoretical guarantees derived in the previous section.

Another related method for minimizing the objective in Equation (21) uses the Iterative Soft-Thresholding Algorithm (ISTA) [63], consisting of iterative updates of the form

$$\mathbf{Z}^{(k)} = \mathcal{S}_{\lambda/c}\left(\mathbf{Z}^{(k-1)} + \frac{1}{c}\mathbf{A}^T(\hat{\mathbf{Y}} - \mathbf{A}\mathbf{Z}^{(k-1)})\right), \tag{25}$$

where the operator $\mathcal{S}_{\lambda/c}$ applies entry-wise soft-thresholding with threshold λ/c. For an appropriate choice of the parameter c, satisfying $\frac{1}{c} < \frac{2}{\|\mathbf{A}\|_2}$,[4] this algorithm will again converge to the minimizer of the global BP problem [63]. Yet, this substitute approach is far more practical in higher dimensions and does not require explicit construction of the generalized effective dictionary $\mathbf{A}$. In fact, plugging in the definitions of $\hat{\mathbf{Y}}$ and $\mathbf{A}$ for the GRSC setting, Equation (25) can be equivalently expressed in matrix form as

$$\mathbf{X}^{(k)} = \mathcal{S}_{\lambda/c}\left(\mathbf{X}^{(k-1)} + \frac{1}{c}\mathbf{D}^T(\mathbf{Y} - \mathbf{D}\mathbf{X}^{(k-1)}) - \frac{\beta}{c}\mathbf{X}^{(k-1)}\mathbf{L}\right). \tag{26}$$

This implies that the global dictionary and Laplacian need not be computed in practice and the projection step relies directly on the given matrices $\mathbf{Y}, \mathbf{D}$, and $\mathbf{L}$. The difference with respect to the standard ISTA (obtained for $\beta = 0$) boils down to the additional required multiplication $\mathbf{XL}$, which is governed by M, the size of the data ensemble. The ISTA thus offers an efficient way for solving Equation (21) in practical scenarios, while still equipped with theoretical guarantees. It also enjoys the potential benefit of accelerated convergence by using the FISTA [64].

As an alternative to the above, one may revert to the $\ell_{0,\infty}$ sparsity measure and develop a similar method, replacing ISTA with an iterative hard thresholding algorithm. However, no convergence guarantees could be claimed under such formulation.

6. Experiments

In this section we provide numerical results that validate the above presented theoretical bounds and demonstrate the performance of the OMP and BP algorithms in practice.

[4] Note that $\|\mathbf{A}\|_2$, the maximal singular value of the generalized effective dictionary $\mathbf{A}$, could be easily bounded based on our previous results (see for example the proof of Theorem 2).

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

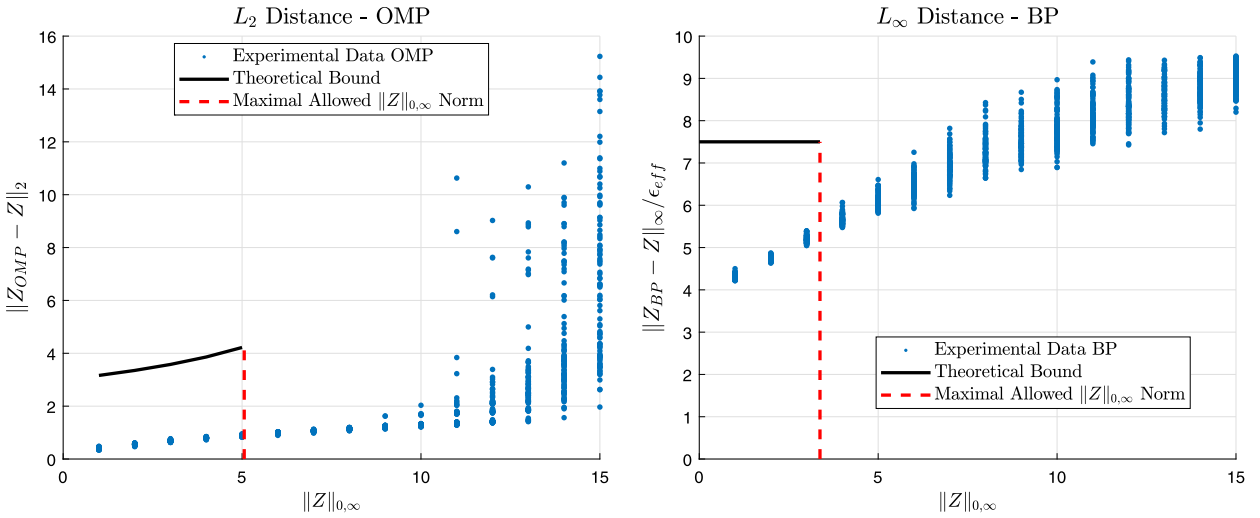


Fig. 3. Numerical evaluation of the coefficient stability: (a) The empirical distance $\|\mathbf{Z_{OMP}} - \mathbf{Z}\|_2$ as a function of the $\ell_{0,\infty}$ norm; (b) The distance $\|\mathbf{Z_{BP}} - \mathbf{Z}\|_\infty / \epsilon_{eff}$ as a function of the $\ell_{0,\infty}$ norm.

The experimental procedure is as follows. First, we generate the graph simulating a manifold of $M = 100$ signals. For this purpose, we draw a random sparse matrix with 20 non-zeros per signal having absolute values between 1 and 3. The graph weights are constructed by applying a Gaussian kernel to the pairwise distances between the sparse vectors, resulting in $\Theta_L \approx 1$.

Note that this graph construction is slightly different than the procedure described in Section 2, which was based on pairwise distances between signals in $\mathbf{Y}$. This is due to the fact that the bounds we wish to validate are given in terms of certain graph properties, such as $L_{\max}$ and Θ_L. In order to guarantee that these properties remain fixed throughout the experiment and that all realizations of $\mathbf{Y}$ share the same underlying manifold graph, we here construct it using an auxiliary set of vectors. Consequently, the signal collections will be generated so as to fit this given manifold structure.

As described in Section 4.3, the dictionary is built as a Grassmannian matrix of size 64×128, having a mutual coherence of $\mu(\mathbf{D}) \approx 0.1$. Using this dictionary and the manifold Laplacian, the generalized effective dictionary $\mathbf{A}$ is then assembled for $\beta = 0.01$.

For each evaluated sparsity level s between 1 and $s_{\max} = 20$, we draw a sparse matrix $\mathbf{X}$ by randomly choosing the support to satisfy $\|\mathbf{X}\|_{0,\infty} = s$, and then draw the non-zero entries as random uniform variables in the range $[-3, -1] \cup [1, 3]$.

Once $\mathbf{X}$ is generated, we compute $\mathbf{Y_0} = \mathbf{DX}$. Next, we contaminate each signal with a zero-mean additive white Gaussian noise, creating the measurements $\mathbf{Y} = \mathbf{Y_0} + \mathbf{V_n}$, where $\|\mathbf{V_n}\|_2 = \epsilon_D = \sqrt{10 - \beta Tr(\mathbf{XLX}^T)}$.

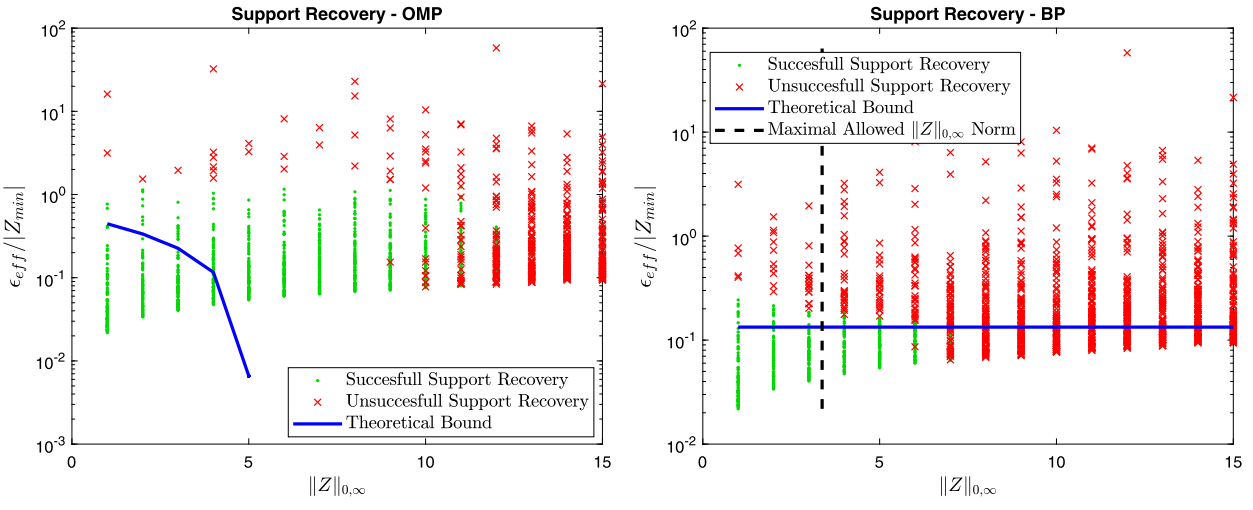
Given these noisy measurements, we construct the effective measurement vector $\hat{\mathbf{Y}}$ and attempt to recover the underlying sparse matrix $\mathbf{X}$ (or rather, its vectorized version $\mathbf{Z}$) using both OMP and BP,[5] and comparing with classic OMP applied to each signal independently.

For each realization we compute the minimal absolute entry $Z_{\min}$, the $\ell_{\infty,0}$ norm η, and the effective noise ϵ_{eff}. We perform 100 such experiments per each cardinality.

First, we corroborate the stability bound of OMP as posed in Theorem 5. Fig. 3a presents the empirical distance between the true sparse representations and the estimated ones as a function of the $\ell_{0,\infty}$ norm of the original vector $\mathbf{Z}$, validating that this distance is indeed below the theoretical bound. Since both

[5] For BP we use the Lagrangian formulation of the LARS algorithm as implemented in [65], with the parameter $\lambda = 4\epsilon_{eff}$.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003



(a) (b)

Fig. 4. Support recovery: the ratio $\frac{\epsilon_{eff}}{|Z_{\min}|}$ as a function of the $\ell_{0,\infty}$ norm, and the theoretical bound for successful support recovery using (a) OMP; and (b) BP.

η and the ratio $\frac{\epsilon_{eff}}{|Z_{\min}|}$ are realization dependent, instead of delimiting the cardinality range satisfying the requirement in (19), we show the more optimistic condition

$$\|\mathbf{Z}\|_{0,\infty} < \frac{1}{1+\Theta_L}\left(1+\frac{1-\Theta_L\beta\eta_{\min}L_{\max}}{\mu(\mathbf{D})}\right) \tag{27}$$

where $\eta_{\min}$ is the minimal empirical value of η over the experiments. Yet, the empirical results remain stable, even for much higher cardinalities than predicted by theory.

Next, we verify the stability guarantees for BP as posed in Theorem 7. Fig. 3b depicts the ratio $\frac{\|\mathbf{Z_{BP}}-\mathbf{Z}\|_\infty}{\epsilon_{eff}}$ for each realization as a function of the $\ell_{0,\infty}$ norm of $\mathbf{Z}$, validating that it is indeed below $\frac{15}{2}$ as long as the $\ell_{0,\infty}$ norm satisfies the requirement in Equation (24). It can be observed once more that the empirical results are stable for cardinalities far beyond the theoretically guaranteed range.

In the sequel, we would like to corroborate the assertions regarding the recovery of the true support. To allow a wider range of coefficient scales and noise levels, we repeat the above described set of experiments, where instead of fixing the total noise energy, we randomly pick the degree of manifold smoothness ϵ_G as well as the data noise ϵ_D, so that their levels differ per realization. The coefficient magnitudes are adjusted accordingly.

Fig. 4a shows the ratio $\frac{\epsilon_{eff}}{|Z_{\min}|}$ for each realization as a function of the $\ell_{0,\infty}$ norm. Each point is marked according to the success or failure of OMP in recovering the complete support. Additionally, we plot the theoretical condition for the success of OMP from Equation (19), reformulated as a bound on the aforementioned ratio, namely

$$\frac{\epsilon_{eff}}{|Z_{\min}|} < \frac{1}{1+\Theta_L}\left(1+\mu(\mathbf{D})-\Theta_L\beta\eta L_{\max}\right)-s\mu(\mathbf{D}). \tag{28}$$

As the results indicate, the empirical results agree with the theoretical claims.

Similarly assessing the support recovery using BP, we again plot the ratio $\frac{\epsilon_{eff}}{|Z_{\min}|}$ as a function of the $\ell_{0,\infty}$ norm for each realization, marking it according to the success or failure of BP in recovering the complete support. In accordance with our theorem, the complete support is indeed recovered when $\frac{\epsilon_{eff}}{|Z_{\min}|} < \frac{2}{15}$ (and beyond).

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

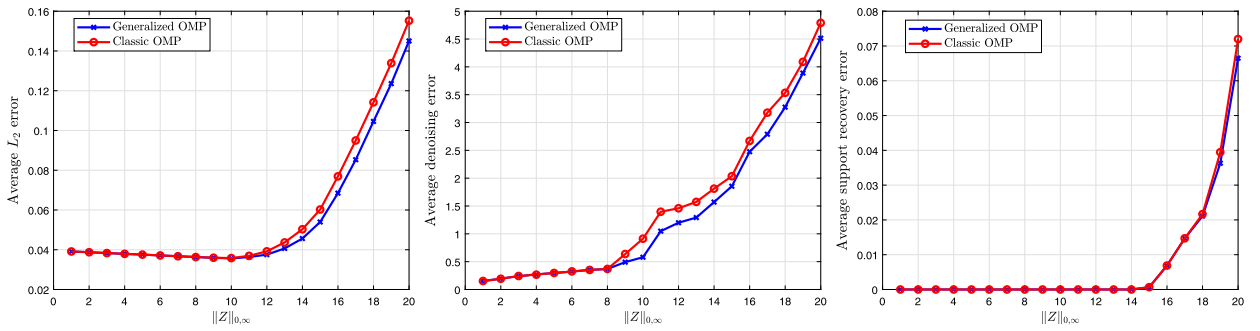


Fig. 5. Comparing the generalized OMP with the structure-agnostic classic OMP: (a) Average approximation error $\|\hat{\mathbf{Y}} - \mathbf{A}\hat{\mathbf{Z}}\|_2/\sqrt{NM}$; (b) Average denoising error $\|\hat{\mathbf{Y}}_{\mathbf{0}} - \mathbf{A}\hat{\mathbf{Z}}\|_2/\sqrt{NM}$; (c) Average support recovery error. To indicate the accuracy of support recovery, we measure the distance between the estimated support $\hat{\Omega}$ and the true one Ω as $dist(\hat{\Omega}, \Omega) = \frac{\max(|\hat{\Omega}|,|\Omega|)-|\hat{\Omega}\cap\Omega|}{\max(|\hat{\Omega}|,|\Omega|)}$.

As can be seen from the above presented results, the theoretical bounds are far from being tight. Nonetheless, recall that the worst-case analysis under an adversarial noise assumption provides rather pessimistic theoretical bounds, and a similar loose flavor was observed in the traditional sparse representation model as well [66]. A probabilistic study could likely result in tighter bounds that better reflect the expected performance in practice, yet such an analysis is beyond the scope of this work.

Last, we compare the average errors obtained using OMP (for the generalized effective dictionary) with those obtained using classic OMP (without the manifold constraint). The plots presented in Fig. 5 show that while the proven theoretical guarantees pose a more strict sparsity requirement for the generalized OMP, in practice it performs the same or better than classic (per-signal) OMP.

7. Conclusions

In this work we have presented a formal analysis of the graph sparse coding problem. To this end, we have generalized concepts such as the RIP and ERC to the GRSC setting, and reformulated the GRSC objective using the $\ell_{0,\infty}$ norm and a generalized effective dictionary, introducing the corresponding $(P^{\epsilon}_{0,\infty})$ problem. By doing so, we were able to provide the first known results for recovery of the geometry preserving sparse representations under adversarial noise assumptions, as well as meaningful stability guarantees for corresponding pursuit algorithms. In addition, we formulated the conditions for which these devised theoretical bounds are superior to the classic ones, shedding light on the desired properties of the data manifold that impact the guaranteed stability.

This work paves the way for several future directions. First, as previously mentioned, it only considers worst-case analysis. Better bounds could likely be obtained by extending this study to an average-performance analysis. This naturally requires further assumptions on the model and noise distribution. Nevertheless, such results would close the gap between the current bounds and the empirical results, which were shown to obey far relaxed sparsity conditions.

By reformulating the GRSC problem we have demonstrated how, under the devised conditions, popular pursuit algorithms such as OMP and BP succeed in finding its solution. However, as the construction of the generalized effective dictionary entails an inevitable dimension increase, these methods are only feasible for low dimensional data. In practice, several other algorithm have been proposed for solving the GRSC problem. Expanding on the direction presented in Section 5.4, it should be of interest to provide theoretical guarantees for these methods as well. We believe such claims could emerge from a similar analysis to the one taken throughout this work. These directions and more are left for future research.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

Acknowledgments

The research leading to these results has received funding from the European Research Council under European Union's Seventh Framework Program, ERC Grant agreement no. 320649, and from the Israel Science Foundation (ISF) grant number 1770/14.

Appendix A. Theorem proofs

In order to prove Theorem 2, we first state and prove the following properties.

Lemma 8. *Given a global dictionary* $\tilde{\mathbf{D}} = \mathbf{I}_M \otimes \mathbf{D}$*, its mutual coherence satisfies* $\mu(\tilde{\mathbf{D}}) = \mu(\mathbf{D})$.

Proof of Lemma 8. Since we assume the atoms of the original dictionary $\mathbf{D}$ to have unit norms, this property holds for the atoms of the global dictionary $\tilde{\mathbf{D}}$ as well, thus $\|\tilde{\mathbf{d}}_i\|_2 = \|\mathbf{d}_i\|_2 = 1 \quad \forall i$.

Let Ω_j denote the subset of atoms belonging to the same block in $\tilde{\mathbf{D}}$ as the j-th atom, i.e. $\Omega_j = \left\{i : \lceil \frac{i}{K} \rceil = \lceil \frac{j}{K} \rceil\right\}$. Observe that atoms from different blocks have non-overlapping supports, thus

$$\tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_j = \begin{cases} 0 & i \notin \Omega_j \\ \mathbf{d}_{i^*}^T \mathbf{d}_{j^*} & otherwise \end{cases} \tag{A.1}$$

where i^* denotes the corresponding index of the atom in the original dictionary $\mathbf{D}$. Therefore, by definition,

$$\mu(\tilde{\mathbf{D}}) = \max_{i \neq j} \frac{\left|\tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_j\right|}{\|\tilde{\mathbf{d}}_i\|_2 \|\tilde{\mathbf{d}}_j\|_2} = \max_{i \neq j} \left|\tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_j\right| = \max_{i \neq j} \left|\mathbf{d}_i^T \mathbf{d}_j\right| = \mu(\mathbf{D}). \quad \square \tag{A.2}$$

Lemma 9. *Consider a global dictionary* $\tilde{\mathbf{D}} = \mathbf{I}_M \otimes \mathbf{D}$ *and a support* Ω *with* $\ell_{0,\infty}$ *norm equal to* k*. Let* $\mathbf{G}_\Omega = \tilde{\mathbf{D}}_\Omega^T \tilde{\mathbf{D}}_\Omega$*, where* $\tilde{\mathbf{D}}_\Omega$ *is the matrix* $\tilde{\mathbf{D}}$ *restricted to the columns indicated by the support* Ω*. Then, the eigenvalues of this Gram matrix, given by* $\lambda_i(\mathbf{G}_\Omega)$*, are bounded by*

$$1 - (k-1)\mu(\mathbf{D}) \leq \lambda_i(\mathbf{G}_\Omega) \leq 1 + (k-1)\mu(\mathbf{D}). \tag{A.3}$$

Proof of Lemma 9. From Gerschgorin's disk theorem, the eigenvalues of the gram matrix $\mathbf{G}_\Omega$ reside in the union of its Gerschgorin disks, where the disk corresponding to the j-th row of $\mathbf{G}_\Omega$ is defined as

$$\left|\lambda(\mathbf{G}_\Omega) - \mathbf{G}_\Omega(j,j)\right| \leq \sum_{t \neq j} \left|\mathbf{G}_\Omega(j,t)\right|. \tag{A.4}$$

Since the atoms are normalized, $\mathbf{G}_\Omega(j,j) = 1 \; \forall j$, implying that all Gerschgorin disks are co-centered at 1, thus all eigenvalues reside inside the circle with the largest radius. The radius of each circle equals the sum of absolute values of the off-diagonal entries in the corresponding row of $\mathbf{G}_\Omega$, which implies

$$\left|\lambda_i(\mathbf{G}_\Omega) - 1\right| \leq \max_j \sum_{t \neq j} \left|\mathbf{G}_\Omega(j,t)\right| = \max_j \sum_{t \neq j;\; t,j \in \Omega} \left|\tilde{\mathbf{d}}_j^T \tilde{\mathbf{d}}_t\right|. \tag{A.5}$$

By definition of the mutual coherence and using Lemma 8, $\left|\tilde{\mathbf{d}}_j^T \tilde{\mathbf{d}}_t\right| \leq \mu(\tilde{\mathbf{D}}) = \mu(\mathbf{D})$ for atoms j, t included in the same block, while $\left|\tilde{\mathbf{d}}_j^T \tilde{\mathbf{d}}_t\right| = 0$ if these atoms are farther apart. Therefore,

$$\left|\lambda_i(\mathbf{G}_\Omega) - 1\right| \leq \max_j \sum_{t \neq j;\; t,j \in \Omega} \left|\tilde{\mathbf{d}}_j^T \tilde{\mathbf{d}}_t\right| \leq (k-1)\mu(\mathbf{D}) \tag{A.6}$$

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

where $k-1$ is the maximal number of non-zero elements in a block after omitting the diagonal entry. From Equation (A.6) we obtain the desired claim:

$$1-(k-1)\mu(\mathbf{D}) \leq \lambda_i(\mathbf{G}_\Omega) \leq 1+(k-1)\mu(\mathbf{D}). \quad \square \tag{A.7}$$

Lemma 10. *Consider the extended manifold Laplacian $\tilde{\mathbf{L}} = \mathbf{L} \otimes \mathbf{I}_K$ and a support Ω with $\ell_{0,\infty}$ norm equal to k and with $\ell_{\infty,0}$ norm equal to η. Let $\tilde{\mathbf{L}}_\Omega$ denote the matrix $\tilde{\mathbf{L}}$ symmetrically restricted to the rows and columns indicated by the support Ω. Then, the eigenvalues of $\tilde{\mathbf{L}}_\Omega$ are bounded by*

$$(M-\eta)L_{\min} \leq \lambda_i(\tilde{\mathbf{L}}_\Omega) \leq 2\Delta_{\max} - (M-\eta)L_{\min}. \tag{A.8}$$

Proof of Lemma 10. First, note that $\tilde{\mathbf{L}}$ is symmetric and positive semi-definite, and its eigenvalues are identical to those of $\mathbf{L}$, up to multiplicity. Consequently, these eigenvalues are bounded by $0 \leq \lambda_i(\tilde{\mathbf{L}}) \leq \lambda_{\max}(\mathbf{L})$. However, better bounds may be obtained by relying on the properties of the support Ω.

Using once again Gerschgorin's disk theorem, the eigenvalues are bounded by the union of disks formed by the rows of $\tilde{\mathbf{L}}_\Omega$. Consider the circle formed by the row of $\tilde{\mathbf{L}}_\Omega$ corresponding to some $i \in \Omega$. Due to the symmetric restriction of $\tilde{\mathbf{L}}_\Omega$, the circle center is $\tilde{\mathbf{L}}_{ii} = \mathbf{L}_{c_i,c_i} > 0$ where $c_i = \lceil \frac{i}{K} \rceil$. The Gerschgorin circle defined by this row is thus

$$\left|\lambda - \tilde{\mathbf{L}}_{ii}\right| \leq \sum_{j\in\Omega; j\neq i} \left|\tilde{\mathbf{L}}_{ij}\right| \tag{A.9}$$

providing the lower bound

$$\lambda \geq \tilde{\mathbf{L}}_{ii} - \sum_{j\in\Omega; j\neq i} \left|\tilde{\mathbf{L}}_{ij}\right| = \sum_{j\neq i} \left|\tilde{\mathbf{L}}_{ij}\right| - \sum_{j\in\Omega; j\neq i} \left|\tilde{\mathbf{L}}_{ij}\right| = \sum_{j\notin\Omega} \left|\tilde{\mathbf{L}}_{ij}\right|, \tag{A.10}$$

where we have used the fact that like the original $\mathbf{L}$, $\tilde{\mathbf{L}}$ is also a valid Laplacian matrix, and so $\tilde{\mathbf{L}}_{ii} > 0\ \forall i$, $\tilde{\mathbf{L}}_{ij} \leq 0\ \forall j \neq i$, and $\tilde{\mathbf{L}}_{ii} = -\sum_{j\neq i} \tilde{\mathbf{L}}_{ij} = \sum_{j\neq i} |\tilde{\mathbf{L}}_{ij}|$.

Finally, recall that there are only M non-zero entries in each row of $\tilde{\mathbf{L}}$, at most η of which are sampled in the support Ω. Summing only over $j \notin \Omega$, we thus encounter at least $M-\eta$ non-zero entries, implying

$$\lambda \geq (M-\eta) \min_{i,j} |\mathbf{L}_{ij}| = (M-\eta)L_{\min}. \tag{A.11}$$

Clearly, this same inequality holds for every row $i \in \Omega$, thus providing a lower bound for the eigenvalues of $\tilde{\mathbf{L}}_\Omega$. Similarly, for the upper bound,

$$\lambda \leq \tilde{\mathbf{L}}_{ii} + \sum_{j\in\Omega; j\neq i} \left|\tilde{\mathbf{L}}_{ij}\right| = 2\tilde{\mathbf{L}}_{ii} - \sum_{j\notin\Omega} \left|\tilde{\mathbf{L}}_{ij}\right|. \tag{A.12}$$

Using the relation obtained for the lower bound, and since the inequality holds for every $i \in \Omega$,

$$\lambda \leq 2\tilde{\mathbf{L}}_{ii} - \sum_{j\notin\Omega} \left|\tilde{\mathbf{L}}_{ij}\right| \leq 2\max_i \mathbf{L}_{ii} - (M-\eta)L_{\min} = 2\Delta_{\max} - (M-\eta)L_{\min}. \tag{A.13}$$

From Equations (A.11) and (A.13) we obtain the desired claim:

$$(M-\eta)L_{\min} \leq \lambda_i(\tilde{\mathbf{L}}_\Omega) \leq 2\Delta_{\max} - (M-\eta)L_{\min}. \quad \square \tag{A.14}$$

Based on these properties, we proceed to prove Theorem 2.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

Proof of Theorem 2. We shall prove the upper bounds on the BRIP via the mutual coherence. Observe that

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}\mathbf{x} + \beta\mathbf{x}^T\tilde{\mathbf{L}}\mathbf{x} = \|\tilde{\mathbf{D}}\mathbf{x}\|_2^2 + \beta\mathbf{x}^T\tilde{\mathbf{L}}\mathbf{x}. \tag{A.15}$$

Denote the support of $\mathbf{x}$ by Ω and let $\mathbf{x}_\Omega$ be the vector $\mathbf{x}$ restricted to the support Ω, thus containing only the non-zero entries.

For the first part, consider the sub-dictionary $\tilde{\mathbf{D}}_\Omega$, obtained by restricting the columns of $\tilde{\mathbf{D}}$ to the support Ω. Then, $\|\tilde{\mathbf{D}}\mathbf{x}\|_2 = \|\tilde{\mathbf{D}}_\Omega\mathbf{x}_\Omega\|_2$. Using the eigenvalue bound from Lemma 9, and since $\|\mathbf{x}\|_2 = \|\mathbf{x}_\Omega\|_2$, for every such $\mathbf{x}$ we have that

$$\begin{aligned}(1-(k-1)\mu(\mathbf{D}))\|\mathbf{x}\|_2^2 &\le \lambda_{\min}(\tilde{\mathbf{D}}_\Omega^T\tilde{\mathbf{D}}_\Omega)\|\mathbf{x}\|_2^2 \le \|\tilde{\mathbf{D}}_\Omega\mathbf{x}_\Omega\|_2^2 \\ &\le \lambda_{\max}(\tilde{\mathbf{D}}_\Omega^T\tilde{\mathbf{D}}_\Omega)\|\mathbf{x}\|_2^2 \le (1+(k-1)\mu(\mathbf{D}))\|\mathbf{x}\|_2^2\end{aligned} \tag{A.16}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximal and minimal eigenvalues, respectively.

As for the second term, $\mathbf{x}^T\tilde{\mathbf{L}}\mathbf{x} = \mathbf{x}_\Omega^T\tilde{\mathbf{L}}_\Omega\mathbf{x}_\Omega$, and from Lemma 10 we obtained

$$\begin{aligned}(M-\eta)L_{\min}\|\mathbf{x}\|_2^2 &\le \lambda_{\min}(\tilde{\mathbf{L}}_\Omega)\|\mathbf{x}\|_2^2 \le \mathbf{x}_\Omega^T\tilde{\mathbf{L}}_\Omega\mathbf{x}_\Omega \\ &\le \lambda_{\max}(\tilde{\mathbf{L}}_\Omega)\|\mathbf{x}\|_2^2 \le [2\Delta_{\max} - (M-\eta)L_{\min}]\|\mathbf{x}\|_2^2.\end{aligned} \tag{A.17}$$

Combining the above results, we have

$$\begin{aligned}(1-(k-1)\mu(\mathbf{D}) + \beta(M-\eta)L_{\min})\|\mathbf{x}\|_2^2 &\le \|\mathbf{A}\mathbf{x}\|_2^2 \\ &\le (1+(k-1)\mu(\mathbf{D}) + \beta[2\Delta_{\max} - (M-\eta)L_{\min}])\,\|\mathbf{x}\|_2^2.\end{aligned} \tag{A.18}$$

Consequently, since δ_k^L, δ_k^H are defined as the smallest quantities that satisfy the above inequality, we conclude that $\delta_k^L \le (k-1)\mu(\mathbf{D}) - \beta(M-\eta)L_{\min}$ and $\delta_k^H \le (k-1)\mu(\mathbf{D}) + \beta[2\Delta_{\max} - (M-\eta)L_{\min}]$. □

Finally, exploiting the above proven bound, we may prove the main stability theorem.

Proof of Theorem 3. The solution to the $(P_{0,\infty}^\epsilon)$ problem satisfies $\|\hat{\mathbf{Y}} - \mathbf{A}\hat{\mathbf{Z}}\|_2 \le \epsilon$, and it must also satisfy $\|\hat{\mathbf{Z}}\|_{0,\infty} \le \|\mathbf{Z}\|_{0,\infty} = s$ (since $\hat{\mathbf{Z}}$ is the solution with the minimal $\ell_{0,\infty}$ norm).

Define $\boldsymbol{\Delta} = \mathbf{Z} - \hat{\mathbf{Z}}$. Using the triangle inequality,

$$\|\mathbf{A}\boldsymbol{\Delta}\|_2 = \|\mathbf{A}\mathbf{Z} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \mathbf{A}\hat{\mathbf{Z}}\|_2 \le \|\hat{\mathbf{Y}} - \mathbf{A}\mathbf{Z}\|_2 + \|\hat{\mathbf{Y}} - \mathbf{A}\hat{\mathbf{Z}}\|_2 \le 2\epsilon, \tag{A.19}$$

therefore we have that $\|\mathbf{A}\boldsymbol{\Delta}\|_2^2 \le 4\epsilon^2$. Furthermore, since the $\ell_{0,\infty}$ norm satisfies the triangle inequality as well,

$$\|\boldsymbol{\Delta}\|_{0,\infty} = \|\mathbf{Z} - \hat{\mathbf{Z}}\|_{0,\infty} \le \|\mathbf{Z}\|_{0,\infty} + \|\hat{\mathbf{Z}}\|_{0,\infty} \le 2s. \tag{A.20}$$

Using the BRIP of $\mathbf{A}$, we have that

$$(1-\delta_{2s}^L)\|\boldsymbol{\Delta}\|_2^2 \le \|\mathbf{A}\boldsymbol{\Delta}\|_2^2 \le 4\epsilon^2, \tag{A.21}$$

where in the first inequality we have used the lower bound provided by the definition of the BRIP. Finally, we obtain the following stability claim:

$$\|\boldsymbol{\Delta}\|_2^2 = \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \le \frac{4\epsilon^2}{1-\delta_{2s}^L}. \tag{A.22}$$

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

Using the bound of the BRIP from Theorem 2, we obtain

$$\|\boldsymbol{\Delta}\|_2^2 = \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_{2s}^L} \leq \frac{4\epsilon^2}{1-(2s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}}. \tag{A.23}$$

For the last inequality to hold and assure that $\delta_{2s}^L < 1$, we have assumed $s = \|\mathbf{Z}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1+\beta(M-\eta)L_{\min}}{\mu(\mathbf{D})}\right)$. □

Appendix B. Stability proofs for pursuit algorithms

Suppose a clean signal $\mathbf{X}$, having a block-sparse representation (in the $\ell_{0,\infty}$ sense) $\mathbf{AZ}$ over a generalized effective dictionary $\mathbf{A}$, is contaminated with noise to create the measurement $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ such that $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \epsilon$. Denoting by Ω the support of $\mathbf{Z}$, we can write

$$\mathbf{Y} = \mathbf{AZ} + \mathbf{E} = \sum_{t\in\Omega} \mathbf{a}_t \mathbf{Z}_t + \mathbf{E}. \tag{B.1}$$

From Equation (B.1) we can express

$$\begin{aligned} \mathbf{Y_D} &= \sum_{t\in\Omega} \tilde{\mathbf{d}}_t \mathbf{Z}_t + \mathbf{E_D} \quad ; \quad \|\mathbf{E_D}\|_2 = \epsilon_D = \sqrt{M}\epsilon_l \\ \mathbf{Y_L} &= \sum_{t\in\Omega} \sqrt{\beta}\tilde{\mathbf{q}}_t \mathbf{Z}_t + \mathbf{E_L} \quad ; \quad \|\mathbf{E_L}\|_2 = \sqrt{\beta}\epsilon_G \end{aligned} \tag{B.2}$$

Utilizing these settings and properties, we prove stability guarantees for several pursuit algorithms.

B.1. Stability proof for the thresholding algorithm

Proof of Theorem 4 – thresholding stability guarantee. Success of the thresholding algorithm is guaranteed by the requirement

$$\min_{i\in\Omega} \frac{|\mathbf{a}_i^T \mathbf{Y}|}{\|\mathbf{a}_i\|_2} > \max_{j\notin\Omega} \frac{|\mathbf{a}_j^T \mathbf{Y}|}{\|\mathbf{a}_j\|_2}. \tag{B.3}$$

First, observe that due to the special dictionary structure, every effective atom satisfies

$$\|\mathbf{a}_i\|_2 = \sqrt{\tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_i + \beta \tilde{\mathbf{q}}_i^T \tilde{\mathbf{q}}_i} = \sqrt{1 + \beta \tilde{\mathbf{L}}_{ii}} \tag{B.4}$$

where we relied on the normalization assumption for the original dictionary atoms. Further, since $\mathbf{E_L}$ is such that the lower portion of $\mathbf{Y}$ (denoted $\mathbf{Y_L}$) is zero,

$$\mathbf{a}_i^T \mathbf{Y} = \tilde{\mathbf{d}}_i^T \mathbf{Y_D} = \tilde{\mathbf{d}}_i^T \left(\mathbf{E_D} + \sum_{t\in\Omega} \mathbf{Z}_t \tilde{\mathbf{d}}_t \right) = \tilde{\mathbf{d}}_i^T \mathbf{E_D} + \sum_{t\in\Omega} \mathbf{Z}_t \tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_t. \tag{B.5}$$

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

Therefore, addressing the left-hand-side term of Equation (B.3),

$$\begin{aligned}\min_{i\in\Omega}\left|\mathbf{a}_i^T\mathbf{Y}\right| &= \min_{i\in\Omega}\left|\tilde{\mathbf{d}}_i^T\mathbf{E_D}+\sum_{t\in\Omega}\mathbf{Z}_t\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t\right| = \min_{i\in\Omega}\left|\tilde{\mathbf{d}}_i^T\mathbf{E_D}+\mathbf{Z}_i+\sum_{t\in\Omega;t\neq i}\mathbf{Z}_t\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t\right| \\ &\geq \min_{i\in\Omega}|\mathbf{Z}_i| - \max_{i\in\Omega}\left|\mathbf{E_D}^T\tilde{\mathbf{d}}_i\right| - \max_{i\in\Omega}\left|\sum_{t\in\Omega;t\neq i}\mathbf{Z}_t\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t\right|\end{aligned} \tag{B.6}$$

where we have again exploited the fact that the columns of $\tilde{\mathbf{D}}$ are normalized, and used the reverse triangle inequality.

Based on the Cauchy-Schwarz inequality and the atom normalization, one could bound the inner product of the noise and the atom $\tilde{\mathbf{d}}_i$ by $|\mathbf{E_D}^T\tilde{\mathbf{d}}_i| \leq \|\mathbf{E_D}\|_2\|\tilde{\mathbf{d}}_i\|_2 = \|\mathbf{E_D}\|_2 = \sqrt{M}\epsilon_l$. However, such bound would disregard the local nature of the atoms. Due to their limited support we have that $\tilde{\mathbf{d}}_i = \mathbf{R}_i^T\mathbf{R}_i\tilde{\mathbf{d}}_i$ where $\mathbf{R}_i$ is an operator extracting an N-dimensional segment from an NM-dimensional vector.[6] Based on this observation,

$$\left|\mathbf{E_D}^T\tilde{\mathbf{d}}_i\right| = \left|\mathbf{E_D}^T\mathbf{R}_i^T\mathbf{R}_i\tilde{\mathbf{d}}_i\right| \leq \|\mathbf{R}_i\mathbf{E_D}\|_2\|\tilde{\mathbf{d}}_i\|_2 \leq \epsilon_l \tag{B.7}$$

where we have used the facts that $\|\mathbf{R}_i\mathbf{E_D}\|_2 \leq \epsilon_l \ \ \forall i$ and $\|\mathbf{R}_i\tilde{\mathbf{d}}_i\|_2 = \|\tilde{\mathbf{d}}_i\|_2 = 1$. By exploiting the locality of the atoms, together with the assumption regarding the maximal local[7] energy of the noise, we are able to derive a much tighter bound.

As for the last term,

$$\left|\sum_{t\in\Omega;t\neq i}\mathbf{Z}_t\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t\right| \leq \sum_{t\in\Omega;t\neq i}\left|\mathbf{Z}_t\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t\right| = \sum_{t\in\Omega;t\neq i}|\mathbf{Z}_t|\left|\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t\right| \leq (s-1)\mu(\mathbf{D})\left|Z_{\max}\right| \tag{B.8}$$

where we have used the fact that, by definition, the absolute inner product between atoms is upper bounded by $\mu(\mathbf{D})$. Also notice that $\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_t = 0$ for every atom too far from $\tilde{\mathbf{d}}_i$ as the atoms do not overlap. Summing over the support Ω, there are only $s = \|\mathbf{Z}\|_{0,\infty}$ atoms for which this inner product is non-zero.

Combining the above, we obtain

$$\min_{i\in\Omega}\left|\mathbf{a}_i^T\mathbf{Y}\right| \geq \left|Z_{\min}\right| - \epsilon_l - (s-1)\mu(\mathbf{D})\left|Z_{\max}\right|. \tag{B.9}$$

Finally, the left-hand-side term in Equation (B.3) can be bounded by

$$\min_{i\in\Omega}\frac{\left|\mathbf{a}_i^T\mathbf{Y}\right|}{\|\mathbf{a}_i\|_2} \geq \frac{\min_{i\in\Omega}\left|\mathbf{a}_i^T\mathbf{Y}\right|}{\max_{i\in\Omega}\|\mathbf{a}_i\|_2} \geq \frac{\left|Z_{\min}\right| - \epsilon_l - (s-1)\mu(\mathbf{D})\left|Z_{\max}\right|}{\sqrt{1+\beta\Delta_{\max}}} \tag{B.10}$$

where $\Delta_{\max}$ denotes the maximal manifold degree $\Delta_{\max} = \max_i \tilde{\mathbf{L}}_{ii}$.

Following similar steps, we can upper-bound the right-hand-side term by

$$\max_{j\notin\Omega}\frac{\left|\mathbf{a}_j^T\mathbf{Y}\right|}{\|\mathbf{a}_j\|_2} \leq \frac{\max_{j\notin\Omega}\left|\mathbf{a}_j^T\mathbf{Y}\right|}{\min_{j\notin\Omega}\|\mathbf{a}_j\|_2} \leq \frac{\epsilon_l + s\mu(\mathbf{D})\left|Z_{\max}\right|}{\sqrt{1+\beta\Delta_{\min}}} \tag{B.11}$$

[6] Denoting by $\mathbf{0}_{m\times n}$ an all-zero matrix of size $m \times n$, and by $\mathbf{I}_n$ an identity matrix of size $n \times n$, then $\mathbf{R}_i = \left[\mathbf{0}_{N\times(b_i-1)N}\ ;\ \mathbf{I}_N\ ;\ \mathbf{0}_{N\times(M-b_i)N}\right] \in \mathbb{R}^{N\times NM}$, where $b_i = \lceil\frac{i}{K}\rceil$ is the index of the block to which the atom $\tilde{\mathbf{d}}_i$ belongs.

[7] Locality in this context refers to the noise level per individual signal rather than the global noise level for the entire ensemble.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

where $\Delta_{\min}$ denotes the minimal manifold degree $\Delta_{\min} = \min_i \tilde{\mathbf{L}}_{ii}$.

Using these bounds, the requirement in Equation (B.3) is clearly satisfied if

$$\frac{|Z_{\min}| - \epsilon_l - (s-1)\mu(\mathbf{D})|Z_{\max}|}{\sqrt{1+\beta\Delta_{\max}}} > \frac{\epsilon_l + s\mu(\mathbf{D})|Z_{\max}|}{\sqrt{1+\beta\Delta_{\min}}} \tag{B.12}$$

or, equivalently, denoting $\Theta_L = \sqrt{\frac{1+\beta\Delta_{\max}}{1+\beta\Delta_{\min}}}$,

$$s = \|\mathbf{Z}\|_{0,\infty} < \frac{1}{1+\Theta_L}\left(1 + \frac{1}{\mu(\mathbf{D})}\frac{|Z_{\min}|}{|Z_{\max}|}\right) - \frac{\epsilon_l}{\mu(\mathbf{D})|Z_{\max}|}. \tag{B.13}$$

If this condition is fulfilled, the true support Ω is recovered, and thus the thresholding algorithm amounts to a simple least-squares solution:

$$\mathbf{Z}^{\mathbf{\Omega}}_{\mathbf{THR}} = \arg\min_{\mathbf{Z}} \|\mathbf{A}_\Omega \mathbf{Z} - \mathbf{Y}\|_2^2, \tag{B.14}$$

where $\mathbf{A}_\Omega$ is the generalized effective dictionary $\mathbf{A}$ restricted to the support Ω. Denoting by $\mathbf{Z}_\Omega$ the (dense) portion of the true sparse vector $\mathbf{Z}$ corresponding to the support Ω, the solution to the above problem is simply given by

$$\mathbf{Z}^{\mathbf{\Omega}}_{\mathbf{THR}} = \mathbf{A}_\Omega^\dagger \mathbf{Y} = \mathbf{A}_\Omega^\dagger(\mathbf{A}\mathbf{Z} + \mathbf{E}) = \mathbf{A}_\Omega^\dagger(\mathbf{A}_\Omega \mathbf{Z}_\Omega + \mathbf{E}) = \mathbf{Z}_\Omega + \mathbf{A}_\Omega^\dagger \mathbf{E} \tag{B.15}$$

where $\mathbf{A}_\Omega^\dagger$ denotes the Moore-Penrose pseudoinverse of the sub-dictionary $\mathbf{A}_\Omega$. Thus, relying on the induced norm properties,

$$\|\mathbf{Z}^{\mathbf{\Omega}}_{\mathbf{THR}} - \mathbf{Z}_\Omega\|_2^2 = \|\mathbf{A}_\Omega^\dagger \mathbf{E}\|_2^2 \le \|\mathbf{A}_\Omega^\dagger\|_2^2 \|\mathbf{E}\|_2^2 \le \frac{1}{\lambda_{\min}(\mathbf{A}_\Omega^T \mathbf{A}_\Omega)} \|\mathbf{E}\|_2^2. \tag{B.16}$$

As proven in Theorem 2 using Lemma 9 and Lemma 10,

$$\lambda_{\min}(\mathbf{A}_\Omega^T \mathbf{A}_\Omega) \ge \lambda_{\min}(\tilde{\mathbf{D}}_\Omega^T \tilde{\mathbf{D}}_\Omega) + \beta\lambda_{\min}(\tilde{\mathbf{L}}_\Omega) \ge 1 - (s-1)\mu(\mathbf{D}) + \beta(M-\eta)L_{\min}. \tag{B.17}$$

Using this bound,

$$\|\mathbf{Z}^{\mathbf{\Omega}}_{\mathbf{THR}} - \mathbf{Z}_\Omega\|_2^2 \le \frac{\|\mathbf{E}\|_2^2}{\lambda_{\min}(\mathbf{A}_\Omega^T \mathbf{A}_\Omega)} \le \frac{\epsilon^2}{1 - (s-1)\mu(\mathbf{D}) + \beta(M-\eta)L_{\min}}. \quad \square \tag{B.18}$$

B.2. Stability proof for OMP

Proof of Theorem 5 – OMP stability guarantee. We shall first prove that the first step of OMP succeeds in recovering an element from the correct support.

Suppose, without loss of generality, that $\mathbf{Z}$ has its largest coefficient in absolute value in $\mathbf{Z}_i$. The greedy algorithm operates by projecting $\mathbf{Y}$ onto each atom $\mathbf{a}_j$ in turn, selecting an atom index where the projection magnitude is highest. Therefore, for the first step of OMP to choose the atom $i \in \Omega$, we require

$$\frac{|\mathbf{a}_i^T \mathbf{Y}|}{\|\mathbf{a}_i\|_2} > \max_{j \notin \Omega} \frac{|\mathbf{a}_j^T \mathbf{Y}|}{\|\mathbf{a}_j\|_2} \tag{B.19}$$

where $\|\mathbf{a}_i\|_2 = \sqrt{\tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_i + \beta \tilde{\mathbf{q}}_i^T \tilde{\mathbf{q}}_i} = \sqrt{1 + \beta \tilde{\mathbf{L}}_{ii}}$. Recall that due to the special structure of our problem,

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

$$\left|\mathbf{a}_i^T\mathbf{Y}\right| = \left|\tilde{\mathbf{d}}_i^T\mathbf{Y_D} + \sqrt{\beta}\tilde{\mathbf{q}}_i^T\mathbf{Y_L}\right|. \tag{B.20}$$

In the first step the bottom part of $\mathbf{Y}$ is all zeros, i.e. $\mathbf{Y_L} = 0$. Therefore in practice, the requirement translates to[8]

$$\left|\tilde{\mathbf{d}}_i^T\mathbf{Y_D}\right| > \max_{j\notin\Omega}\left|\tilde{\mathbf{d}}_j^T\mathbf{Y_D}\right|. \tag{B.21}$$

However, in order to serve the proof for the next steps of OMP, we shall ignore this fact and consider the more general case.

For the left-hand-side term in Equation (B.19),

$$\begin{aligned}
\left|\mathbf{a}_i^T\mathbf{Y}\right| &= \left|\tilde{\mathbf{d}}_i^T\mathbf{Y_D} + \sqrt{\beta}\tilde{\mathbf{q}}_i^T\mathbf{Y_L}\right| \\
&= \left|\sum_{t\in\Omega}\mathbf{Z}_t\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i + \mathbf{E_D}^T\tilde{\mathbf{d}}_i + \sum_{t\in\Omega}\beta\mathbf{Z}_t\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_i + \sqrt{\beta}\mathbf{E_L}^T\tilde{\mathbf{q}}_i\right| \\
&= \left|\sum_{t\in\Omega}\mathbf{Z}_t\left(\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i + \beta\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_i\right) + \mathbf{E_D}^T\tilde{\mathbf{d}}_i + \sqrt{\beta}\mathbf{E_L}^T\tilde{\mathbf{q}}_i\right| \\
&\geq \left|\sum_{t\in\Omega}\mathbf{Z}_t\left(\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i + \beta\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_i\right)\right| - \left|\mathbf{E_D}^T\tilde{\mathbf{d}}_i\right| - \sqrt{\beta}\left|\mathbf{E_L}^T\tilde{\mathbf{q}}_i\right|
\end{aligned} \tag{B.22}$$

where the last step stems from the reverse triangle inequality.

Our next step is to bound each term individually. Based on the Cauchy-Schwarz inequality,

$$\left|\mathbf{E_L}^T\tilde{\mathbf{q}}_i\right| \leq \|\mathbf{E_L}\|_2\|\tilde{\mathbf{q}}_i\|_2 \leq \sqrt{\beta}\epsilon_G\sqrt{\Delta_{\max}} \tag{B.23}$$

where we have used the facts that $\|\tilde{\mathbf{q}}_i\|_2^2 = \tilde{\mathbf{q}}_i^T\tilde{\mathbf{q}}_i = \tilde{\mathbf{L}}_{ii}$ and $\Delta_{\max} = \max_i\tilde{\mathbf{L}}_{ii}$.

Similarly to the proof for the Thresholding algorithm (Equation (B.7)),

$$\left|\mathbf{E_D}^T\tilde{\mathbf{d}}_i\right| \leq \epsilon_l. \tag{B.24}$$

For the first term, using the reverse triangle inequality, the normalization of the atoms $\tilde{\mathbf{d}}_i$ and the fact that $|\mathbf{Z}_i| \geq |\mathbf{Z}_t|$ we obtain

$$\begin{aligned}
&\left|\sum_{t\in\Omega}\mathbf{Z}_t\left(\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i + \beta\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_i\right)\right| = \left|\mathbf{Z}_i\left(\|\tilde{\mathbf{d}}_i\|_2^2 + \beta\tilde{\mathbf{L}}_{ii}\right) + \sum_{t\in\Omega;t\neq i}\mathbf{Z}_t\left(\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i + \beta\tilde{\mathbf{L}}_{ti}\right)\right| \\
&\geq |\mathbf{Z}_i|(1+\beta\tilde{\mathbf{L}}_{ii}) - \sum_{t\in\Omega;t\neq i}|\mathbf{Z}_t|\left|\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i + \beta\tilde{\mathbf{L}}_{ti}\right| \\
&\geq |\mathbf{Z}_i|(1+\beta\tilde{\mathbf{L}}_{ii}) - |\mathbf{Z}_i|(s-1)\mu(\mathbf{D}) - \beta|\mathbf{Z}_i|\tilde{\mathbf{L}}_{ii} = |\mathbf{Z}_i|\left(1-(s-1)\mu(\mathbf{D})\right)
\end{aligned} \tag{B.25}$$

where for the last inequality we have used the fact that $\tilde{\mathbf{L}}$ is a valid graph Laplacian, satisfying $\tilde{\mathbf{L}}_{ii} = \sum_{t\neq i}\left|\tilde{\mathbf{L}}_{ti}\right| \geq \sum_{t\in\Omega;t\neq i}\left|\tilde{\mathbf{L}}_{ti}\right|$, as well as the atom localization, similarly to the derivation in Equation (B.8).

[8] Note that the manifold constraint has no influence on the first chosen atom. This is expected since the notion of internal smoothness is meaningless for a vector with a single non-zero entry.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

As a result of combining (B.23), (B.24) and (B.25), we obtain

$$\begin{aligned}\left|\mathbf{a}_i^T\mathbf{Y}\right| &\geq \left|\sum_{t\in\Omega}\mathbf{Z}_t\left(\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_i+\beta\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_i\right)\right| - \left|\mathbf{E_D}^T\tilde{\mathbf{d}}_i\right| - \sqrt{\beta}\left|\mathbf{E_L}^T\tilde{\mathbf{q}}_i\right| \\ &\geq |\mathbf{Z}_i|(1-(s-1)\mu(\mathbf{D})) - \epsilon_l - \beta\epsilon_G\sqrt{\Delta_{\max}}.\end{aligned} \tag{B.26}$$

An upper bound for the right hand side of Equation (B.19) can be constructed by following the same rationale, using again the triangle inequality:

$$\begin{aligned}j\notin\Omega \quad \left|\mathbf{a}_j^T\mathbf{Y}\right| &= \left|\sum_{t\in\Omega}\mathbf{Z}_t\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_j+\mathbf{E_D}^T\tilde{\mathbf{d}}_j+\sum_{t\in\Omega}\beta\mathbf{Z}_t\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_j+\sqrt{\beta}\mathbf{E_L}^T\tilde{\mathbf{q}}_j\right| \\ &\leq \left|\sum_{t\in\Omega}\mathbf{Z}_t\left(\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_j+\beta\tilde{\mathbf{q}}_t^T\tilde{\mathbf{q}}_j\right)\right| + \left|\mathbf{E_D}^T\tilde{\mathbf{d}}_j\right| + \sqrt{\beta}\left|\mathbf{E_L}^T\tilde{\mathbf{q}}_j\right| \\ &\leq \sum_{t\in\Omega}|\mathbf{Z}_t|\left|\tilde{\mathbf{d}}_t^T\tilde{\mathbf{d}}_j+\beta\tilde{\mathbf{L}}_{tj}\right| + \epsilon_l + \beta\epsilon_G\sqrt{\Delta_{\max}} \\ &\leq |\mathbf{Z}_i|s\mu(\mathbf{D}) + \beta|\mathbf{Z}_i|\eta L_{\max} + \epsilon_l + \beta\epsilon_G\sqrt{\Delta_{\max}}.\end{aligned} \tag{B.27}$$

For the last inequality we have used the fact that, as shown in Lemma 10, each row in $\tilde{\mathbf{L}}$ has only M non-zero entries, at most η of which are sampled in the support Ω, thus $\sum_{t\in\Omega}|\tilde{\mathbf{L}}_{jt}| \leq \eta L_{\max}$.

Requiring an inequality between the obtained bounds for both sides of Equation (B.19), and plugging in $\epsilon_{eff} = \epsilon_l + \beta\epsilon_G\sqrt{\Delta_{\max}}$, we conclude that OMP succeeds if

$$\frac{|\mathbf{Z}_i|\left(1-(s-1)\mu(\mathbf{D})\right)-\epsilon_{eff}}{\sqrt{1+\beta\tilde{\mathbf{L}}_{ii}}} > \max_{j\notin\Omega}\frac{|\mathbf{Z}_i|(s\mu(\mathbf{D})+\beta\eta L_{\max})+\epsilon_{eff}}{\sqrt{1+\beta\tilde{\mathbf{L}}_{jj}}} \tag{B.28}$$

or

$$\frac{|\mathbf{Z}_i|\left(1-(s-1)\mu(\mathbf{D})\right)-\epsilon_{eff}}{\sqrt{1+\beta\Delta_{\max}}} > \frac{|\mathbf{Z}_i|(s\mu(\mathbf{D})+\beta\eta L_{\max})+\epsilon_{eff}}{\sqrt{1+\beta\Delta_{\min}}}. \tag{B.29}$$

From this, it follows that

$$s = \|\mathbf{Z}\|_{0,\infty} < \frac{1}{1+\Theta_L}\left(1+\frac{1-\Theta_L\beta\eta L_{\max}}{\mu(\mathbf{D})}\right) - \frac{1}{\mu(\mathbf{D})}\cdot\frac{\epsilon_{eff}}{|\mathbf{Z}_i|}. \tag{B.30}$$

In order for this condition to hold for every i, the theorem assumption in Equation (19) is that the above holds for $|Z_{\min}|$ instead of $|\mathbf{Z}_i|$, as $|\mathbf{Z}_i| \geq |Z_{\min}|$. Therefore, the first step of OMP will succeed in finding an atom i from within the support Ω.

To address the success of subsequent iterations of OMP, define the sparse vector obtained after $k < \|\mathbf{Z}\|_0$ iterations as $\mathbf{\Gamma}^k$, and denote its support by Ω^k. Assuming that the algorithm has so far succeeded in identifying the correct atoms, $\Omega^k = supp\{\mathbf{\Gamma}^k\} \subset supp\{\mathbf{Z}\}$. The next step of the algorithm is updating the residual, which is performed by

$$\mathbf{Y}^k = \mathbf{Y} - \sum_{i\in\Omega^k}\mathbf{a}_i\mathbf{\Gamma}_i^k. \tag{B.31}$$

Since the clean signal $\mathbf{X} = \mathbf{AZ}$ is also a linear combination of atoms from the correct support, we could express

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

$$\mathbf{X}^k = \mathbf{X} - \sum_{i\in\Omega^k} \mathbf{a}_i\boldsymbol{\Gamma}_i^k = \mathbf{A}(\mathbf{Z}-\boldsymbol{\Gamma}^k) = \mathbf{A}\mathbf{Z}^k, \tag{B.32}$$

and so the objective is to recover the support of the sparse vector $\mathbf{Z}^k$, corresponding to $\mathbf{X}^k$. This sparse vector is defined as

$$\mathbf{Z}_i^k = \begin{cases} \mathbf{Z}_i - \boldsymbol{\Gamma}_i^k & \text{if } i \in \Omega^k \\ \mathbf{Z}_i & \text{if } i \notin \Omega^k \end{cases} \tag{B.33}$$

Note that $supp\{\mathbf{Z}^k\} \subseteq supp\{\mathbf{Z}\}$ and so

$$\|\mathbf{Z}^k\|_{0,\infty} \leq \|\mathbf{Z}\|_{0,\infty}. \tag{B.34}$$

That is, the $\ell_{0,\infty}$ norm of the underlying solution of $\mathbf{X}^k$ does not increase as the iterations proceed.

From the above definitions, we have that

$$\mathbf{Y}^k - \mathbf{X}^k = \mathbf{Y} - \sum_{i\in\Omega^k} \mathbf{a}_i\boldsymbol{\Gamma}_i^k - \mathbf{X} + \sum_{i\in\Omega^k} \mathbf{a}_i\boldsymbol{\Gamma}_i^k = \mathbf{Y} - \mathbf{X} = \mathbf{E}, \tag{B.35}$$

hence the noise level is preserved, both locally and globally, i.e. all ϵ, ϵ_l, ϵ_D and ϵ_G remain the same.

From Equation (B.33) and the fact that $|\Omega^k| = k$, it follows that $\mathbf{Z}^k$ differs from $\mathbf{Z}$ in at most k places. As such, $\|\mathbf{Z}^k\|_\infty$ is greater than the $(k+1)$-th largest element in absolute value in $\mathbf{Z}$, implying $\|\mathbf{Z}^k\|_\infty \geq |Z_{\min}|$. Combining this with Equation (B.34) and the theorem assumption, we obtain

$$\|\mathbf{Z}^k\|_{0,\infty} < \frac{1}{1+\Theta_L}\left(1 + \frac{1-\Theta_L\beta\eta L_{\max}}{\mu(\mathbf{D})}\right) - \frac{1}{\mu(\mathbf{D})}\cdot\frac{\epsilon_{eff}}{\|\mathbf{Z}^k\|_\infty}. \tag{B.36}$$

Similar to the first iteration, the above inequality together with the fact that the noise level is preserved, guarantees the success of the next iteration of OMP. Consequently, OMP is guaranteed to recover the true support after $\|\mathbf{Z}\|_0$ iterations.

Finally, having recovered the true support, the coefficients are estimated by solving

$$\mathbf{Z}_{\mathbf{OMP}}^{\boldsymbol{\Omega}} = \arg\min_{\mathbf{Z}} \|\mathbf{A}_\Omega\mathbf{Z} - \mathbf{Y}\|_2^2. \tag{B.37}$$

Thus, similarly to the proof for the Thresholding algorithm, the solution satisfies

$$\|\mathbf{Z}_{\mathbf{OMP}} - \mathbf{Z}\|_2^2 \leq \frac{\epsilon^2}{1-(s-1)\mu(\mathbf{D}) + \beta(M-\eta)L_{\min}}. \quad \square \tag{B.38}$$

B.3. Stability proof for BP

We commence by proving the $\ell_{0,\infty}$ condition for satisfying the ERC in the GRSC model.

Proof of Theorem 6. For the ERC to be satisfied, we must require that $\|\mathbf{A}_\Omega^\dagger\mathbf{a}_i\|_1 < 1 \quad \forall i \notin \Omega$. Using properties of induced norms,

$$\|\mathbf{A}_\Omega^\dagger\mathbf{a}_i\|_1 = \|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\mathbf{A}_\Omega^T\mathbf{a}_i\|_1 \leq \|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_1\|\mathbf{A}_\Omega^T\mathbf{a}_i\|_1. \tag{B.39}$$

Addressing the second term, observe that

$$\mathbf{A}_\Omega^T\mathbf{a}_i = \tilde{\mathbf{D}}_\Omega^T\tilde{\mathbf{d}}_i + \beta\tilde{\mathbf{Q}}_\Omega^T\tilde{\mathbf{q}}_i. \tag{B.40}$$

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

As derived in Equation (B.8), $\|\tilde{\mathbf{D}}_\Omega^T\tilde{\mathbf{d}}_i\|_1 \leq s\mu(\mathbf{D})$. As for $\tilde{\mathbf{Q}}_\Omega^T\tilde{\mathbf{q}}_i$, this is essentially a subset of the i-th row (or column) in $\tilde{\mathbf{L}}$, hence

$$\|\tilde{\mathbf{Q}}_\Omega^T\tilde{\mathbf{q}}_i\|_1 = \sum_{j\in\Omega}|\tilde{\mathbf{L}}_{ij}| \leq \eta \max_{i\neq j}|\tilde{\mathbf{L}}_{ij}| = \eta L_{\max}. \tag{B.41}$$

We have used the fact that, as shown in Lemma 10, each row in $\tilde{\mathbf{L}}$ has only M non-zero entries, at most η of which are sampled in the support Ω. Combining the two bounds and using the triangle inequality, we get

$$\|\mathbf{A}_\Omega^T\mathbf{a}_i\|_1 \leq \|\tilde{\mathbf{D}}_\Omega^T\tilde{\mathbf{d}}_i\|_1 + \beta\|\tilde{\mathbf{Q}}_\Omega^T\tilde{\mathbf{q}}_i\|_1 \leq s\mu(\mathbf{D}) + \beta\eta L_{\max}. \tag{B.42}$$

Next, we address the term $\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_1$, which, from symmetry of the matrix, equals $\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty$. Using the Ahlberg-Nilson-Varah bound [67] and similar steps to those presented in Lemma 9 and Lemma 10, we have that

$$\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty \leq \frac{1}{1-(s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}}. \tag{B.43}$$

This holds if the gram matrix $\mathbf{A}_\Omega^T\mathbf{A}_\Omega$ is strictly diagonally dominant, i.e. if $1-(s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min} > 0$, which is indeed the case given the theorem assumption.

Plugging the above into Equation (B.39), we obtain the condition

$$\|\mathbf{A}_\Omega^\dagger\mathbf{a}_i\|_1 \leq \frac{s\mu(\mathbf{D})+\beta\eta L_{\max}}{1-(s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}} < 1, \tag{B.44}$$

leading to the claimed relationship

$$s = \|\Omega\|_{0,\infty} < \frac{1}{2}\left(1+\frac{1+\beta(M-\eta)L_{\min}-\beta\eta L_{\max}}{\mu(\mathbf{D})}\right). \tag{B.45}$$

As this condition is the same for all columns $\mathbf{a}_i$, we conclude that if it is met then the ERC is satisfied. □

Next, we prove the main stability theorem for BP.

Proof of Theorem 7 – BP stability guarantee. Denote by $\mathbf{X_{LS}}$ the best ℓ_2 approximation of $\mathbf{Y}$ over the support Ω, i.e. $\mathbf{X_{LS}} = \mathbf{A}_\Omega\mathbf{A}_\Omega^\dagger\mathbf{Y}$, and denote by $\mathbf{Z_{LS}} = \mathbf{A}_\Omega^\dagger\mathbf{Y}$ the optimal coefficient vector, satisfying $\mathbf{X_{LS}} = \mathbf{A}_\Omega\mathbf{Z_{LS}}$.

In [61, Theorem 8], Tropp proved that if the ERC is met with constant $\theta = 1 - \max_{i\notin\Omega}\|\mathbf{A}_\Omega^\dagger\mathbf{a}_i\|_1$ for the support Ω, and $\|\mathbf{A}^T(\mathbf{Y}-\mathbf{X_{LS}})\|_\infty \leq \lambda\theta$, then:

1. The support of $\mathbf{Z_{BP}}$ is contained in Ω.
2. $\|\mathbf{Z_{BP}} - \mathbf{Z_{LS}}\|_\infty < \lambda\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty$.
3. The support of $\mathbf{Z_{BP}}$ contains every index i for which $|\mathbf{Z_{LS}}_i| > \lambda\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty$.
4. The minimizer of the problem, $\mathbf{Z_{BP}}$, is unique.

In Theorem 6 we have shown that the ERC is met if the $\ell_{0,\infty}$ norm of the support is less than $\frac{1}{2}\left(1+\frac{1+\beta(M-\eta)L_{\min}-\beta\eta L_{\max}}{\mu(\mathbf{D})}\right)$. The stricter assumption on s in the current theorem clearly satisfies this requirement, and so the ERC is met.

Using Equations (B.42) and (B.43), one can easily show that under the same condition from Equation (23), we have

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

$$\|\mathbf{A}^T(\mathbf{Y}-\mathbf{X_{LS}})\|_\infty \le 2\epsilon_{eff}. \tag{B.46}$$

Considering the theorem assumption on s and employing the inequality in Equation (B.44), we can lower bound the ERC constant by

$$\theta = 1-\max_{i\notin\Omega}\|\mathbf{A}_\Omega^\dagger\mathbf{a}_i\|_1 \ge 1-\frac{s\mu(\mathbf{D})+\beta\eta L_{\max}}{1-(s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}} > \frac{1}{2}. \tag{B.47}$$

Combining this with Equation (B.46), we have that for $\lambda = 4\epsilon_{eff}$,

$$\|\mathbf{A}^T(\mathbf{Y}-\mathbf{X_{LS}})\|_\infty \le 2\epsilon_{eff} < \theta\lambda. \tag{B.48}$$

Therefore we conclude that both conditions of [61, Theorem 8] are fulfilled in the GRSC setup, leading immediately to most of our theorem's results.

Concerning the second point, from Tropp's theorem we have $\|\mathbf{Z_{BP}}-\mathbf{Z_{LS}}\|_\infty < \lambda\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty$. Using Equation (B.43) we can upper bound

$$\begin{aligned}\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty &\le \frac{1}{1-(s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}}\\ &\le \frac{1}{1-(s-1)\mu(\mathbf{D})+\beta(M-\eta)L_{\min}-2\beta\eta L_{\max}}.\end{aligned} \tag{B.49}$$

Since we assumed

$$\begin{aligned}s=\|\mathbf{Z}\|_{0,\infty} &< \frac{1}{3}\left(1+\frac{1+\beta(M-\eta)L_{\min}-2\beta\eta L_{\max}}{\mu(\mathbf{D})}\right)\\ &\le \frac{1}{3}\left(3+\frac{1+3\beta(M-\eta)L_{\min}-6\beta\eta L_{\max}}{\mu(\mathbf{D})}\right),\end{aligned} \tag{B.50}$$

it stems that $(s-1)\mu(\mathbf{D})-\beta(M-\eta)L_{\min}+2\beta\eta L_{\max} > \frac{1}{3}$, thus from Equation (B.49)

$$\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty < \frac{3}{2}. \tag{B.51}$$

Consequently, using once more the assumption that $\lambda = 4\epsilon_{eff}$,

$$\|\mathbf{Z_{BP}}-\mathbf{Z_{LS}}\|_\infty < \lambda\|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty < 6\epsilon_{eff}. \tag{B.52}$$

Moreover, the distance from the real $\mathbf{Z}$ satisfies

$$\begin{aligned}\|\mathbf{Z_{LS}}-\mathbf{Z}\|_\infty &= \|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\mathbf{A}_\Omega^T(\mathbf{Y}-\mathbf{X})\|_\infty\\ &\le \|(\mathbf{A}_\Omega^T\mathbf{A}_\Omega)^{-1}\|_\infty\|\mathbf{A}_\Omega^T\mathbf{E}\|_\infty < \frac{3}{2}\epsilon_{eff}\end{aligned} \tag{B.53}$$

where we have used Equation (B.51) and the fact that

$$\begin{aligned}\|\mathbf{A}_\Omega^T\mathbf{E}\|_\infty &= \max_{i\in\Omega}\left|\mathbf{a}_i^T\mathbf{E}\right| = \max_{i\in\Omega}\left|\tilde{\mathbf{d}}_i^T\mathbf{E_D}+\sqrt{\beta}\tilde{\mathbf{q}}_i^T\mathbf{E_L}\right|\\ &\le \max_{i\in\Omega}\left|\tilde{\mathbf{d}}_i^T\mathbf{E_D}\right|+\max_{i\in\Omega}\left|\sqrt{\beta}\tilde{\mathbf{q}}_i^T\mathbf{E_L}\right| \le \epsilon_l+\beta\epsilon_G\sqrt{\Delta_{\max}} = \epsilon_{eff}.\end{aligned} \tag{B.54}$$

Finally, using the triangle inequality we obtain

$$\|\mathbf{Z_{BP}}-\mathbf{Z}\|_\infty \le \|\mathbf{Z_{BP}}-\mathbf{Z_{LS}}\|_\infty+\|\mathbf{Z_{LS}}-\mathbf{Z}\|_\infty < \frac{15}{2}\epsilon_{eff}. \quad \square \tag{B.55}$$

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

References

[1] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (12) (2006) 3736–3745.
[2] J. Mairal, G. Sapiro, M. Elad, Multiscale sparse image representation with learned dictionaries, in: 2007 IEEE International Conference on Image Processing, ICIP, vol. 3, 2007, pp. III–105–III–108.
[3] W. Dong, X. Li, L. Zhang, G. Shi, Sparsity-based image denoising via dictionary learning and structural clustering, in: CVPR, 2011, pp. 457–464.
[4] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, IEEE Trans. Image Process. 17 (1) (2008) 53–69.
[5] W. Dong, L. Zhang, G. Shi, X. Wu, Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization, IEEE Trans. Image Process. 20 (7) (2011) 1838–1857.
[6] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[7] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[8] Q. Zhang, B. Li, Discriminative K-SVD for dictionary learning in face recognition, in: CVPR, 2010, pp. 2691–2698.
[9] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: CVPR, 2011, pp. 625–632.
[10] B. Liu, J. Huang, C. Kulikowski, L. Yang, Robust visual tracking using local sparse appearance model and k-selection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2968–2981.
[11] K. Huang, S. Aviyente, Sparse representation for signal classification, in: NIPS, 2006, pp. 609–616.
[12] J. Mairal, F.R. Bach, J. Ponce, G. Sapiro, A. Zisserman, Supervised dictionary learning, in: NIPS, 2009, pp. 1033–1040.
[13] Z. Jiang, Z. Lin, L.S. Davis, Learning a discriminative dictionary for sparse coding via label consistent K-SVD, in: CVPR, 2011, pp. 1697–1704.
[14] K. Kavukcuoglu, M. Ranzato, R. Fergus, Y. Le-Cun, Learning invariant features through topographic filter maps, in: CVPR, 2009, pp. 1605–1612.
[15] R. Jenatton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for sparse hierarchical dictionary learning, in: ICML, 2010, pp. 487–494.
[16] S. Kim, E.P. Xing, Tree-guided group lasso for multi-task regression with structured sparsity, in: ICML, 2010, pp. 543–550.
[17] K. Yu, Y. Lin, J. Lafferty, Learning image representations from the pixel level via hierarchical sparse coding, in: CVPR, 2011, pp. 1713–1720.
[18] J. Huang, T. Zhang, D. Metaxas, Learning with structured sparsity, J. Mach. Learn. Res. 12 (Nov) (2011) 3371–3412.
[19] Y.C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, IEEE Trans. Inform. Theory 55 (11) (2009) 5302–5316.
[20] M.F. Duarte, Y.C. Eldar, Structured compressed sensing: from theory to applications, IEEE Trans. Signal Process. 59 (9) (2011) 4053–4085.
[21] J. Chen, X. Huo, Theoretical results on sparse representations of multiple-measurement vectors, IEEE Trans. Signal Process. 54 (12) (2006) 4634–4643.
[22] J.A. Tropp, A.C. Gilbert, M.J. Strauss, Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit, Signal Process. 86 (3) (2006) 572–588.
[23] J.A. Tropp, Algorithms for simultaneous sparse approximation. Part II: Convex relaxation, Signal Process. 86 (3) (2006) 589–602.
[24] R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst, Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms, J. Fourier Anal. Appl. 14 (5) (2008) 655–687.
[25] L. Jacob, G. Obozinski, J.-P. Vert, Group lasso with overlap and graph lasso, in: ICML, 2009, pp. 433–440.
[26] S. Bengio, F. Pereira, Y. Singer, D. Strelow, Group sparse coding, in: NIPS, 2009, pp. 82–89.
[27] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Non-local sparse models for image restoration, in: ICCV, 2009, pp. 2272–2279.
[28] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[29] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[30] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30.
[31] A. Elmoataz, O. Lezoray, S. Bougleux, Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing, IEEE Trans. Image Process. 17 (7) (2008) 1047–1060.
[32] S. Bougleux, A. Elmoataz, M. Melkemi, Local and nonlocal discrete regularization on weighted graphs for image and mesh processing, Int. J. Comput. Vis. 84 (2) (2009) 220–236.
[33] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2011) 1327–1336.
[34] K.N. Ramamurthy, J.J. Thiagarajan, P. Sattigeri, A. Spanias, Learning dictionaries with graph embedding constraints, in: Signals, Systems and Computers, ASILOMAR, 2012, pp. 1974–1978.
[35] P. Milanfar, A tour of modern image filtering: new insights and methods, both practical and theoretical, IEEE Signal Process. Mag. 30 (1) (2013) 106–128.
[36] A. Kheradmand, P. Milanfar, A general framework for regularized, similarity-based image restoration, IEEE Trans. Image Process. 23 (12) (2014) 5136–5151.
[37] X. Liu, D. Zhai, D. Zhao, G. Zhai, W. Gao, Progressive image denoising through hybrid graph Laplacian regularization: a unified framework, IEEE Trans. Image Process. 23 (4) (2014) 1491–1503.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003

[38] P.A. Forero, K. Rajawat, G.B. Giannakis, Prediction of partially observed dynamical processes over networks via dictionary learning, IEEE Trans. Signal Process. 62 (13) (2014) 3305–3320.
[39] Y. Yankelevsky, M. Elad, Dual graph regularized dictionary learning, IEEE Trans. Signal Inf. Process. Netw. 2 (4) (2016) 611–624.
[40] Y. Yankelevsky, M. Elad, Finding gems: multi-scale dictionaries for high-dimensional graph signals, IEEE Trans. Signal Process. 67 (7) (2019) 1889–1901.
[41] W. Liu, Z. Wang, D. Tao, J. Yu, Hessian regularized sparse coding for human action recognition, in: MultiMedia Modeling, Springer International Publishing, 2015, pp. 502–511.
[42] W. Liu, Z.J. Zha, Y. Wang, K. Lu, D. Tao, p-Laplacian regularized sparse coding for human activity recognition, IEEE Trans. Ind. Electron. 63 (8) (2016) 5120–5129.
[43] S. Gao, I.W.-H. Tsang, L.-T. Chia, P. Zhao, Local features are not Lonely–Laplacian sparse coding for image classification, in: CVPR, 2010, pp. 3555–3561.
[44] S. Gao, I.W.H. Tsang, L.T. Chia, Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 92–104.
[45] W. Liu, D. Tao, J. Cheng, Y. Tang, Multiview Hessian discriminative sparse coding for image annotation, Comput. Vis. Image Underst. 118 (2014) 50–60.
[46] M. Yin, J. Gao, Z. Lin, Laplacian regularized low-rank representation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 504–517.
[47] X. Zhu, X. Li, S. Zhang, C. Ju, X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE Trans. Neural Netw. Learn. Syst. 28 (6) (2017) 1263–1275.
[48] Y. Yankelevsky, M. Elad, Structure-aware classification using supervised dictionary learning, in: IEEE ICASSP, 2017, pp. 4421–4425.
[49] Y. Yankelevsky, M. Elad, Graph-constrained supervised dictionary learning for multi-label classification, in: IEEE ICSEE, 2016, pp. 1–5.
[50] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.
[51] E.J. Candes, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (12) (2005) 4203–4215.
[52] R. Giryes, M. Elad, RIP-based near-oracle performance guarantees for SP, CoSaMP, and IHT, IEEE Trans. Signal Process. 60 (3) (2012) 1465–1468.
[53] D.L. Donoho, M. Elad, V.N. Temlyakov, On Lebesgue-type inequalities for greedy approximation, J. Approx. Theory 147 (2) (2007) 185–195.
[54] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, 2010.
[55] D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization, Proc. Nat. Acad. Sci. 100 (5) (2003) 2197–2202.
[56] J.D. Blanchard, C. Cartis, J. Tanner, Compressed sensing: how sharp is the restricted isometry property?, SIAM Rev. 53 (1) (2011) 105–125.
[57] Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, vol. , 1993, pp. 40–44.
[58] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. 43 (1) (2001) 129–159.
[59] V. Papyan, J. Sulam, M. Elad, Working locally thinking globally: theoretical guarantees for convolutional sparse coding, IEEE Trans. Signal Process. 65 (21) (2017) 5687–5701.
[60] D.L. Donoho, M. Elad, V.N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inform. Theory 52 (1) (2006) 6–18.
[61] J.A. Tropp, Just relax: convex programming methods for identifying sparse signals in noise, IEEE Trans. Inform. Theory 52 (3) (2006) 1030–1051.
[62] J.A. Tropp, Greed is good: algorithmic results for sparse approximation, IEEE Trans. Inform. Theory 50 (10) (2004) 2231–2242.
[63] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math. 57 (11) (2004) 1413–1457.
[64] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202.
[65] SPAMS Package, http://spams-devel.gforge.inria.fr/.
[66] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (1) (2009) 34–81.
[67] J.H. Ahlberg, E.N. Nilson, Convergence properties of the spline fit, J. Soc. Indust. Appl. Math. 11 (1) (1963) 95–104.

Please cite this article in press as: Y. Yankelevsky, M. Elad, Theoretical guarantees for graph sparse coding, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.03.003