

DICTIONARY LEARNING FOR HIGH DIMENSIONAL GRAPH SIGNALS

Yael Yankelevsky and Michael Elad

Computer Science Department
Technion - Israel Institute of Technology
Haifa 32000, Israel

{yaelyan, elad}@cs.technion.ac.il

ABSTRACT

In recent years there is a growing interest in operating on graph signals. One systematic and productive such line of work is incorporating sparsity-inspired models to this data type, offering these signals a description as sparse linear combinations of atoms from a given dictionary. In this paper, we propose a dictionary learning algorithm for this task that is capable of handling high dimensional data. We incorporate the underlying graph topology by forcing the learned dictionary atoms to be sparse combinations of graph wavelet functions. The resulting atoms thus adhere to the underlying graph structure and possess a desired multi-scale property, yet they capture the prominent features of the data of interest. This results in both adaptive representations and an efficient implementation. Experimental results on different datasets, representing both synthetic and real network data, demonstrate the effectiveness of the proposed algorithm for graph signal processing.

Index Terms— Sparse representation, dictionary learning, graph signal processing, graph Laplacian, double-sparsity, graph wavelets.

1. INTRODUCTION

Numerous modern applications introduce signals having an underlying complicated geometric topology, which could be represented using a graph structure. Examples of such signals can be found in applications of transportation, energy, social or sensor networks [1]. In these applications, the vertices (or nodes) of the graph represent the discrete data domain, and the edge weights reflect the pairwise similarity between these vertices. A graph signal is a function that resides on the graph, assigning a real value to each vertex.

Given a weighted graph and a class of signals residing on it, we would like to construct an overcomplete dictionary such that the given graph signals can be well approximated as

linear combinations of only a few atoms from this dictionary. Naturally, when designing dictionaries for graph signals, an additional challenge is exploiting prior knowledge of the intrinsic underlying topology of the data.

In recent years, different directions have been taken in an attempt to design dictionaries for graph signals. Similarly to signals on the Euclidean domain, the choice of a dictionary often involves a tradeoff between analytic dictionaries, which are efficient to construct and apply, and dictionary learning methods, that are more adaptable, training the dictionary using a given set of signal realizations.

Various analytic dictionaries for graph signals have been proposed, generalizing transform-based dictionaries from the Euclidean domain to the graph settings. These include the graph Fourier transform [2], windowed graph Fourier transform [3], diffusion wavelets [4], and spectral graph wavelets [5], among others. Such methods offer efficient implementations while also accounting for the underlying topology, yet they are not adapted to the signals themselves.

The alternative approach of dictionary learning methods, e.g. the MOD [6] and K-SVD [7], do adapt to the data by training the dictionary using a given set of signal realizations. Assuming sufficient training signals are provided, these methods are able to detect the most important characteristics of the data. However, they do not explicitly incorporate the non-Euclidean underlying topology of the data domain.

Generalizing these methods to the graph setting, we have recently proposed a dual graph regularized dictionary learning algorithm (DGRDL) [8], enjoying both worlds. This algorithm incorporates the graph structure into the dictionary learning framework by introducing graph constraints in both the signal and manifold domains. Furthermore, the proposed scheme suggests the additional ability of learning the graph topology along with the dictionary in cases where this structure is not given.

Though exhibiting impressive performance, DGRDL is mostly suited for moderate size graphs. As data dimensions keep growing, this method becomes gradually cumbersome, both because of the induced size of the dictionary and the graph Laplacian matrix.

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.

In order to overcome this limitation and operate on much higher dimensional graph signals, this work proposes infusing structure into the learned dictionary by harnessing the double sparsity framework [9, 10]. By modeling the dictionary itself as the product of an efficient and topology-adapted analytic base dictionary and an adaptable sparse matrix, we lower the degrees of freedom of the problem, and with it, the computational cost of training and applying the dictionary, thus enabling treatment of higher dimensional graph signals.

More specifically, we propose to use a Haar wavelet basis that is adapted to the graph Laplacian, following [11]. As such, the proposed approach benefits from the multi-scale structure and the topology-awareness that this base dictionary brings, along with the ability to adapt to the signals. This approach can be viewed as a fusion of the analytic and the trainable paradigms.

Indeed, such a bridge between the two choices for getting the dictionary has already been proposed in earlier work imposing other parametric structures on the learned dictionary. Along this reasoning, Zhang et al. [12] suggest that the dictionary should be a collection of shift-invariant filters or sub-dictionaries of form $D_s = \chi \Lambda_s \chi^T$ where χ is the eigenbasis of the graph Laplacian L and $\Lambda_s \succeq 0$ are some diagonal matrices. Thanou et al. [13] further restrict the dictionary to a polynomial structure, $D_s = \sum_{k=0}^K \alpha_{s,k} L^k$, with additional constraints imposed in order to control the frequency behavior of the kernels. However, these techniques as well are limited in their ability to manage high-dimensional data, due to their direct use of the graph Laplacian. And so, while our proposed scheme can also be considered as belonging to this parametric regime, it offers a way to go beyond the dimensionality limitations that all current methods suffer from.

The rest of the paper is organized as follows: We present our sparse dictionary learning algorithm for graph signals in Section 2, including a detailed description of the base dictionary construction procedure. We then evaluate the performance of the proposed algorithm in Section 3, and conclude in Section 4.

2. DOUBLE SPARSE DICTIONARY LEARNING FOR GRAPH SIGNALS

The double sparsity approach [9] is based on a sparsity model of the dictionary atoms over a base dictionary, i.e. it defines the dictionary as a product $D = \Phi A$, where Φ is some fixed (perhaps analytic or structured) base dictionary, and A is a learned sparse matrix, having P non-zeros per column.

Formally, the dictionary learning problem in this case is given by

$$\begin{aligned} \arg \min_{A, X} \|Y - \Phi A X\|_F^2 \\ \text{s.t. } \|x_i\|_0 \leq T \quad \forall i, \\ \|a_j\|_0 \leq P \quad \forall j, \quad \|\Phi a_j\|_2 = 1 \quad \forall j, \end{aligned} \quad (1)$$

where $Y \in \mathbb{R}^{N \times M}$ is the data matrix containing the training examples in its columns, $X \in \mathbb{R}^{K \times M}$ contains the corresponding sparse representation vectors, $\Phi \in \mathbb{R}^{N \times N}$ is the base dictionary and $A \in \mathbb{R}^{N \times K}$ is a redundant ($K > N$) column-wise sparse matrix¹. A solution to this problem is obtained by an alternating optimization over A and X .

The above dictionary, ΦA , has a compact representation and provides efficient forward and adjoint operators, yet it can be effectively trained from given data, even when N and K are large (many thousands and beyond). Therefore, it naturally bridges the gap between analytic dictionaries, which have efficient implementations yet lack adaptability, and standard trained dictionaries, which are fully adaptable but non-efficient and costly to deploy.

We are about to bring this double-sparsity idea to the treatment of graph signals, and we will refer to the resulting algorithm as Sparse Dictionary Learning (SDL). Our approach is to use the very same setup as defined in Equation (1), and bring the graph topology into consideration by the choice of Φ .

The success of SDL heavily depends on a proper choice of the base dictionary Φ . Incorporating multi-scale properties in the learned dictionary is a vital property for representing large signals. In this work, we therefore choose to construct a Haar-like graph wavelet basis inspired by [11]. As an initial step, the underlying graph should be converted to a hierarchical tree by spectral partitioning.

Spectral graph partitioning methods are based on the Fiedler vector [14], which is the eigenvector corresponding to the smallest non-zero eigenvalue of the graph Laplacian matrix L . The Fiedler vector bisects the graph into two disjoint sets of nodes based on the sign of the corresponding vector entry. By applying the spectral bisection procedure recursively, full partitioning is obtained and the graph can be encoded as a hierarchical tree [15]. We note that the Fiedler vector itself may be efficiently computed using the Lanczos algorithm [16], without having to compute the full eigen-decomposition of L .

Equipped with a tree representation of the given data, we can now construct an orthonormal Haar-like wavelet basis in the spirit of the method proposed in [11], which will serve as our base dictionary Φ . Not only is Φ orthogonal by construction, the data geometry encapsulated by it is captured by a hierarchical tree of increasingly refined partitions. This achieves the desired multi-scale property of the basis functions, and consequently manifested as localization over the graph of their sparse linear combinations, which are the atoms of the dictionary $D = \Phi A$.

Furthermore, we emphasize that due to the sparsity of A , SDL requires a training of far less parameters than classic dictionary learning methods, hence learning in this case is feasi-

¹Note that the double-sparsity framework allows flexibility in the dimensions of Φ and A and it is not generally necessary for Φ to be square. Nonetheless, we here choose to use an orthogonal transform.

ble even given limited training data, or truly high dimensions.

3. EXPERIMENTS AND APPLICATIONS

In this section, we demonstrate the effectiveness of the proposed SDL method on synthetic examples and on real network data and show its potential use in data analysis applications of different dimensions.

3.1. Synthetic Experiment

We first describe an experiment on synthetic signals, similar to the one described in [8]. For this purpose, we generated a random graph consisting of N nodes randomly drawn from a uniform distribution. The edge weights between each pair of nodes were determined based on the Euclidean distances between them and using the Gaussian Radial Basis Function (RBF) $W_{ij} = \exp\left(\frac{-d^2(i,j)}{2\sigma^2}\right)$ with $\sigma = 0.5$.

A randomly drawn data matrix $Y_0 \in \mathbb{R}^{N \times 10N}$ was made smooth over the resulting graph Laplacian by solving

$$\arg \min_Y \|Y - Y_0\|_F^2 + \lambda \text{Tr}(Y^T L Y), \quad (2)$$

yielding $Y = (I + \lambda L)^{-1} Y_0$. Consequently, each signal was normalized to have unit norm. A subset of 40% of the generated signals were used for training and the rest for testing.

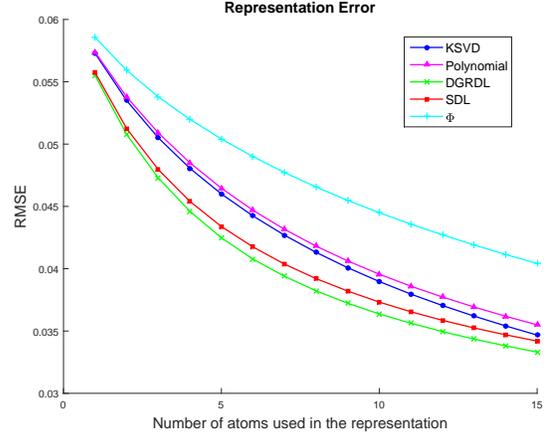
Using this data, several dictionaries were trained including the K-SVD [7], the graph Polynomial dictionary [13], DGRDL [8] and the proposed SDL with a graph Haar-like base dictionary. For a fair comparison, all these dictionaries are of the same size, $N \times 2N$. We also evaluated a direct use of the graph wavelet basis Φ (serving as a base dictionary for SDL), whose size is $N \times N$. The dictionaries were compared by their ability to obtain the best m -term approximation of the test data (for different sparsity levels), and performance was measured in terms of the normalized RMSE, $\frac{1}{\sqrt{NM}} \|Y - DX\|_F$.

The above described experiment was repeated for two different setups: the first with a moderate size graph of $N = 256$ nodes, and the second with a high-dimensional graph containing $N = 4096$ nodes. The dictionaries were trained with a fixed number of non-zeros in the sparse coding stage ($T = 12$ and $T = 25$, respectively). For SDL, the respective sparsity levels of the dictionary A were set to $P = 12$ and $P = 40$.

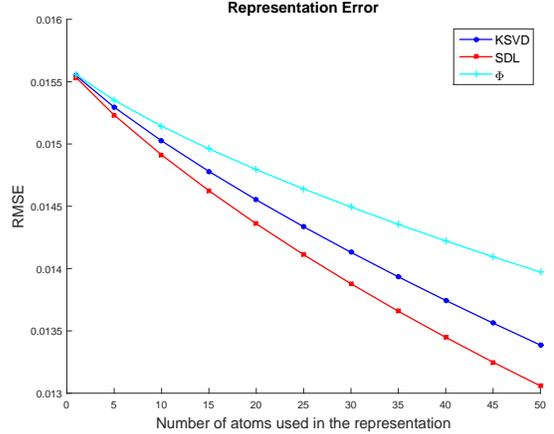
The representation errors presented in Figure 1a show that for a moderate size graph, the proposed SDL yields lower errors compared with K-SVD, the Polynomial method and the graph wavelets, and is almost as good as DGRDL while being significantly faster and more efficient. For this setup, the training process for SDL and K-SVD converged in just a few seconds, compared with several minutes for the Polynomial dictionary and DGRDL.

The representation errors obtained for a large graph setting are presented in Figure 1b, demonstrating the scalability

of the proposed SDL method to high dimensional data. For this data dimension, the Polynomial dictionary and DGRDL can no longer train in reasonable time, and were therefore omitted from the comparison. Nevertheless, SDL still outperforms K-SVD and the graph wavelet base dictionary Φ .



(a) $N = 256$



(b) $N = 4096$

Fig. 1: Comparison of the learned dictionaries in terms of normalized RMSE for representing synthetic data of different dimensions with various sparsity levels.

3.2. Traffic Network Data

Next, the proposed method was evaluated on real network data from the PeMS database [17]. The dataset consists of 2863 signals, representing the daily traffic loads measured at $N = 578$ predefined locations across the highways in Alameda County, California, between 2007 and 2014. All signals were normalized with respect to the one having the maximal energy. A random subset of 1400 signals constitutes the training set, and the rest are used for testing.

The initial graph Laplacian L was designed by connecting stations whose distance (in terms of the Euclidean distance between the GPS coordinates of the stations) is smaller than 13 kilometers and the edge weights are set to be inversely proportional to the distance.

The proposed SDL approach was again compared with K-SVD [7], the graph Polynomial dictionary [13] and DGRDL [8]. All evaluated dictionaries are of the same size of $N \times 2N$ and a sparsity threshold of $T = 6$ was used for training.

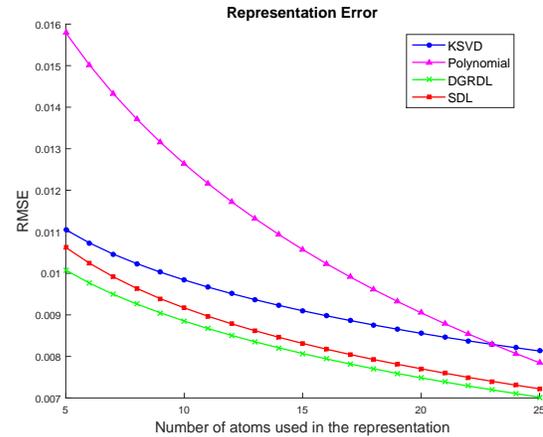
The different methods are evaluated on two tasks. We start by evaluating their ability to sparsely represent the test set data with different sparsity levels (number of used atoms). Next, we evaluate their performance for the common signal denoising problem, by adding Gaussian noise of different levels σ_n to the test signals and comparing recovery using each of the dictionaries in terms of the normalized Root Mean Squared Error (RMSE). Assuming each noisy test signal is modeled as $y_i = \Phi Ax_i + n_i$ where n_i denotes the added noise, the denoised signal $\hat{y}_i = \Phi A \hat{x}_i$ is obtained by seeking the sparse approximation of the noisy test signals with a known sparsity level of $T = 6$.

The representation errors for this experiment are presented in Figure 2a, and the corresponding denoising errors in Figure 2b. It can be observed that in both tasks, for all sparsity levels and all noise levels tested, the proposed sparse dictionary (SDL) yields significantly lower errors compared with K-SVD and the Polynomial graph dictionary, and is almost as good as DGRDL. These results coincide with those obtained for the synthetic experiment. Recall, however, that DGRDL is much more complex and its runtime is significantly longer, making its use impractical for larger dimensions.

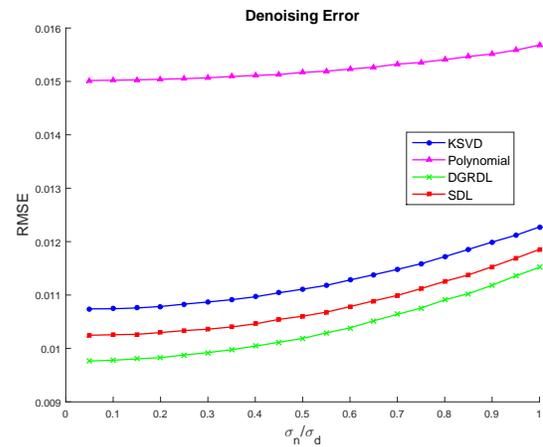
It should be emphasized again that the performance of SDL compared with K-SVD is expected to further improve as the training set becomes scarce. Furthermore, the results could be improved by re-training the dictionaries for every sparsity level. Instead, training was performed once for a fixed T and the generalization ability of the resulting dictionaries was challenged by evaluating them using different (both smaller and larger) sparsity levels. Nevertheless, as the experimental results demonstrate, the trained model fits the data well even in this setting.

4. CONCLUSIONS

This work presented an efficient way for handling high dimensional graph signals, going beyond the data sizes previously handled in sparsity based graph signal processing. At the core of the proposed dictionary learning algorithm stands a simple graph wavelet transform, possessing several desired properties: (i) a multi-scale nature that we deem vital for representing large signals, (ii) orthogonality and ease of construction, and (iii) adhering to the graph topology underlying the data.



(a)



(b)

Fig. 2: Comparison of the learned dictionaries in terms of normalized RMSE for different applications tested on the traffic dataset: (a) representation error for different sparsity levels, (b) denoising error for different noise levels σ_n (with respect to the data STD σ_d).

Employing this as the base dictionary within a double sparsity model adds flexibility and allows adapting the dictionary to the given data without compromising these properties.

As demonstrated through experiments on both synthetic and real data, the proposed SDL algorithm achieves superior performance to other tested methods. Moreover, it is able to take into account the underlying graph structure while retaining computational efficiency. This method is therefore suitable for much larger graphs for which DGRDL [8], the Polynomial dictionary [13] or even the structure agnostic K-SVD [7] collapse. As such, SDL opens the door to new challenges and problems that remained unattainable until now.

5. REFERENCES

- [1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.
- [2] A. Sandryhaila and J. M. F. Moura, "Discrete Signal Processing on Graphs: Graph Fourier Transform," in *ICASSP*, 2013, pp. 6167–6170.
- [3] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "A Windowed Graph Fourier Transform," in *IEEE Statistical Signal Processing Workshop (SSP)*, Aug. 2012, pp. 133–136.
- [4] R. R. Coifman and M. Maggioni, "Diffusion Wavelets," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 53–94, 2006.
- [5] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on Graphs via Spectral Graph Theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.
- [6] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of Optimal Directions for Frame Design," in *ICASSP*, 1999, vol. 5, pp. 2443–2446.
- [7] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. Signal Proc.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [8] Y. Yankelevsky and M. Elad, "Dual Graph Regularized Dictionary Learning," *IEEE Trans. Sig. Inf. Proc. over Net.*, vol. 2, no. 4, pp. 611–624, Dec. 2016.
- [9] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Proc.*, vol. 58, no. 3, pp. 1553–1564, March 2010.
- [10] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Trans. Signal Proc.*, vol. 64, no. 12, pp. 3180–3193, June 2016.
- [11] M. Gavish, B. Nadler, and R. R. Coifman, "Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 367–374.
- [12] X. Zhang, X. Dong, and P. Frossard, "Learning of Structured Graph Dictionaries," in *ICASSP*, 2012, pp. 3373–3376.
- [13] D. Thanou, D. I. Shuman, and P. Frossard, "Learning Parametric Dictionaries for Signals on Graphs," *IEEE Trans. Signal Proc.*, vol. 62, no. 15, pp. 3849–3862, Aug. 2014.
- [14] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," *Czechoslovak Mathematical Journal*, vol. 25, no. 4, pp. 619–633, 1975.
- [15] H. D. Simon, "Partitioning of unstructured problems for parallel processing," *Computing systems in engineering*, vol. 2, no. 2-3, pp. 135–148, 1991.
- [16] B. N. Parlett, H. Simon, and L. M. Stringer, "On estimating the largest eigenvalue with the lanczos algorithm," *Mathematics of computation*, vol. 38, no. 157, pp. 153–165, 1982.
- [17] T. Choe, A. Skabardonis, and P. Varaiya, "Freeway Performance Measurement System (PeMS): An Operational Analysis Tool," in *Proceedings of the 81st Transportation Research Board Annual Meeting, National Academies, Washington, D.C.*, Jan 2002.