Beyond Black-box Dictionary Learning for Waves

Harsha Vardhan Tetali* **K. Supreet Alguri** † **Joel B. Harley** ‡

Abstract

This work discusses an optimization framework to embed dictionary learning frameworks with the wave equation as a strategy for incorporating prior scientific knowledge into a machine learning algorithm. We modify dictionary learning to study ultrasonic guided wave based defect detection for non-destructive structural health monitoring systems. Specifically, this work involves altering the popular K-SVD algorithm for dictionary learning by enforcing prior knowledge about the ultrasound guided wave problem through a physics-based regularization derived from the wave equation. We confer it the name "wave-informed K-SVD." Training a dictionary on data simulated from a fixed string added with noise using both K-SVD and wave-informed K-SVD, we show an improved physical consistency of columns of dictionary matrix with the known modal behavior of different one-dimensional wave simulations is observed.

1 Introduction

Machine learning methods for analysis of wave data have recently gained prominence in guided wave structural health monitoring (SHM) systems. Though still in a nascent stage, the uncertainties involved in modeling waves propagation inspire usage of machine learning methods to model wave propagation. In this work, we want to step towards introducing theory guided data science into modeling wave propagation and detection. With present day research on waves gaining prominence (e.g. recent advancements in detection of gravitational waves), we believe the method discussed in this work can contribute towards widening the purview of theory guided machine learning on wave based data. The usual practice in guided wave structural health monitoring, is to compare measurements before damage (baseline data) and after damage (test data) to detect damages. From a practical viewpoint, baseline data is not available, thus data from surrogate structures (structures similar to the test structure) could replace baseline data, but even the slightest differences in material properties, such as thickness, temperature, and other effects, makes this data unreliable. The work presented in [2] overcomes this challenge and detects damage with surrogate information by using dictionary learning framework. Dictionary learning is a branch of machine learning that finds a dictionary matrix in which some training data is represented as a sparse linear combination of the columns of the dictionary. We start by justifying the usage of dictionary learning framework in ultrasound guided wave based defect detection in non-destructive structural health monitoring systems through the analytic solution of the wave equation. We then turn to the main goal of the paper, looking beyond black-box dictionary learning. We enforce structure into columns of dictionary matrix with the help of a physics based regularizer [6]. Since we work with wave data, our regularizer is derived by using the wave equation. We thus take a step towards developing a dictionary utilizing both theory and data. From an algorithmic point of view, we unwrap the popular dictionary learning algorithm, K-SVD [1], and integrate specific domain knowledge, in this case, the wave equation, to create a dictionary that is restricted to the particular domain. We do this for the simple case of a fixed string (i.e. one dimensional waves). We finally contrast the results obtained using K-SVD algorithm with

*Electrical & Computer Engineering, University of Florida, Gainesville, FL, vardhanh71@gmail.com
†Electrical & Computer Engineering, University of Utah, Salt Lake City, kssupreet@gmail.com
‡Electrical & Computer Engineering, University of Florida, Gainesville, FL, joel.harley@ufl.edu

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

the modified algorithm, which we confer "wave-informed K-SVD." We train dictionaries using both K-SVD and wave-informed K-SVD on wave data corrupted by white Gaussian noise. Wave-informed K-SVD learns columns of dictionary matrix that do not contain noise, while simple K-SVD learns noisy atoms. In the general paradigm of noise removal algorithms, we either have a pattern matching step (like matched filter) or a thresholding step (in a different domain, like in frequency based filtering, KL transform based denoising or dictionary learning based denoising [5]). Our algorithm falls mostly in the pattern matching category, where the we do not have a single pattern but an equation that defines a gamut of patterns. And the wave-informed KSVD picks those patterns as columns of the dictionary matrix which satisfy the wave equation as opposed to just picking a single pattern. This indicates an enforcement of structure by wave-informed KSVD on the atoms of the dictionary.

2 The Main Idea

The analytical solution of a linear homogeneous partial differential equation in $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$ with boundary conditions is often calculated by the method of separations of variables. We try to find a solution of the form: $f_i(\mathbf{x}, t) = u_{1,i}(x_1)u_{2,i}(x_2)\cdots u_{n,i}(x_n)v_{n,i}(t)$ where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. Since the partial differential equation is linear and homogeneous we form, $f(\mathbf{x}, t) = \sum_{i\in\mathbb{N}} a_i f_i(\mathbf{x}, t)$ for all $a_i \in \mathbb{R}$. So that $f(\mathbf{x}, t)$ is also a solution of the partial differential equation. Consider the one-dimensional wave equation:

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 u(x,t)}{\partial t^2} \tag{1}$$

As the general paradigm states, for the one-dimensional case, we assume a solution of the form:

$$u(x,t) = \sum_{n\in\mathbb{N}} a_n f_n(x,t) = \sum_{n\in\mathbb{N}} a_n u_n(x) v_n(t) \tag{2}$$

since $f_n(x,t) = u_n(x)v_n(t)$ (again its the one-dimensional case). Skipping the details of the solution method, we use this theoretical perspective to motivate the use of dictionary learning for decomposing a wave solution into its constituent modes. When discrete data of a wave traveling in one dimension is available, we can look at the function $f(x,t)$ in discrete terms, where x and t are both discrete, and treat it as a matrix, $\mathbf{Y}$. Similarly, a function of one variable in discrete terms can be seen as a vector. In discrete domain, the product of functions of two variables can analogously be seen as an outer product of two vectors, say $\mathbf{d}$ and $\mathbf{x}$ (in case of more than one spatial dimensions the usual practice is to vectorize spatial data and still have a matrix in space and time). Thus, the product $u_i(x)v_i(t)$ can analogously be seen as $\mathbf{d}_i\mathbf{x}_i^T$. Following the form in (2) any discrete data of waves uniformly sampled over space and time, represented by a matrix $\mathbf{Y}$, can be decomposed as: $\mathbf{Y} = \sum_{i=1}^{N} \mathbf{d}_i\mathbf{x}_i^T$, where $\mathbf{d}_i$ represents variation over space and $\mathbf{x}_i$ represents variation over time. Now stacking $\mathbf{d}_i$ as columns to form a matrix $\mathbf{D}$ and stacking $\mathbf{x}_i^T$ as rows to form a matrix $\mathbf{X}$, we have: $\mathbf{Y} = \mathbf{DX}$. Thus, factorizing the data matrix $\mathbf{Y}$ as the matrix product decouples the space and time characteristics of the propagating wave. It is usually convenient from a dictionary learning point of view if the matrix $\mathbf{X}$ is sparse in each column. We observe that this can be addressed when we have only a few frequencies present in data. We take the Fourier transform of each row of $\mathbf{Y}$ (i.e., Fourier transform over time). In this situation, we have each row $\mathbf{x}_i^T$ as the Fourier transform of the respective time characterization of each wave (note that each $\mathbf{d}_i\mathbf{x}_i^T$ is a wave with possibly a single mode). The sparsity level of each column of $\mathbf{X}$ is estimated by number of frequencies corresponding to each wavenumber present in the wave data. Therefore, we assume a fixed level of sparsity. This decoupling of space and time (or equivalently frequency) using dictionary learning has been used for an advantage in [2] to achieve baseline-free damage detection. Similar work is found in [4] where they use this decoupling to unveil anomalous regions in space. In the next section, we use this decoupling of space and frequency (or equivalently time) to derive a regularizer based on the wave equation to embed into K-SVD, the popular dictionary learning algorithm.

3 Wave Informed K-SVD

In this section, we assume that data is obtained from a one-dimensional medium, such as a string. Thus, the one-dimensional wave equation is the physics model enforced into the algorithm. We take

Fourier transforms and re-substitute $f_n(x,t)$ in equation (1), to get:

$$\frac{\partial^2 u_n(x)}{\partial x^2} v_n(t) = \frac{1}{v^2}\frac{\partial^2 v_n(t)}{\partial t^2} u_n(x) \Rightarrow \frac{\partial^2 u_n(x)}{\partial x^2} V_n(\omega) = \frac{-\omega^2}{v^2} V_n(\omega) u_n(x) \Rightarrow \frac{\partial^2 u_n(x)}{\partial x^2} = \frac{-\omega^2}{v^2} u_n(x) \quad (3)$$

when $V_n(\omega) \neq 0$ and v is the velocity of the wave. The above equation is an eigen-value problem. We then discretize everything above and write (3) as a discrete-space matrix form, where $\mathbf{L}$ is the negative of second difference matrix

$$\mathbf{L} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -2 \end{bmatrix} \quad (4)$$

Thus, we can concisely write, $\mathbf{Ld} = g\mathbf{d}$ for a suitable constant $g = \frac{\omega^2}{v^2}$ (observe that $\sqrt{g}$ is the wavenumber associated with the dictionary atom). To enforce structure into the atoms of the dictionary, we impose that each atom of the dictionary approximately satisfies the equation $\mathbf{Ld} = g\mathbf{d}$. Thus, we now have an objective function with an added regularization term. We choose regularization constants γ_i associated with each dictionary atom. Thus the objective function is, $\min_{\mathbf{X},\mathbf{D}}\{||\mathbf{Y} - \mathbf{DX}||_F^2 + \sum_{j=1}^{K} \gamma_j \|\mathbf{LD}_j - g_j \mathbf{D}_j\|_2^2\}$ subject to $||\mathbf{X}_i||_0 \leq s$. We solve this problem by iterating through the classical alternate minimization approach followed in [1]. In each iteration, the dictionary matrix and the coefficient matrix are updated. Observe that we embedded a physical constraint due to differential equations into the dictionary learning framework. We find $\mathbf{X}^{(0)}$ in the first step, by initializing the algorithm by a random dictionary matrix $\mathbf{D}^{(0)}$. Since, we assume a sparsity level s, we find orthogonal matching pursuit ([7]) as the best fit to solve for $\mathbf{X}$. This step of updating $\mathbf{X}$ for a fixed $\mathbf{D}$ is referred to as the sparse coding step. Sparse coding is performed in every iteration, once the dictionary is updated through the dictionary update step, described next. In the dictionary update step, we write the k-th dictionary atom update step in the t-th iteration as,

$$\left[\mathbf{D}_k^{(t+1)}, \widehat{\mathbf{X}}_k^{(t+1)}, g_k^{(t+1)}\right] = \arg\min_{\mathbf{d},\mathbf{u},g_k} \|\mathbf{E}_k^{(t)} - \mathbf{du}^T\|_F^2 + \gamma_k \|\mathbf{Ld} - g_k^{(t)}\mathbf{d}\|_2^2 \quad (5)$$

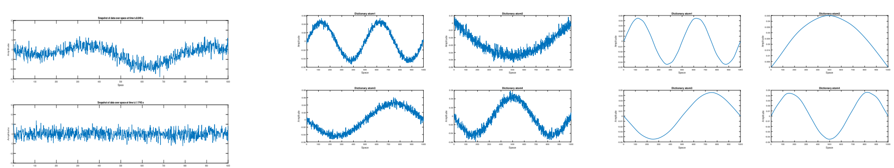
such that $\|\mathbf{D}_k^{(t+1)}\|_2 = 1$. We first derive an update rule for each $g_k^{(t)}$. We differentiate the function to be minimized in (5) with respect to the scalar $g_k^{(t)}$ at $\mathbf{d} = \mathbf{D}_k^{(t)}$ and $\mathbf{u} = \widehat{\mathbf{X}}_k^{(t)}$ and set it equal to zero to find the updated value $g_k^{(t+1)}$. For ease of differentiation, we choose λ to represent $g_k^{(t)}$, i.e., $\lambda = g_k^{(t)}$. Thus, we have, $\frac{\partial}{\partial\lambda}\left(\|\mathbf{E}_k^{(t)} - \mathbf{du}^T\|_F^2 + \gamma_k\|\mathbf{Ld} - \lambda\mathbf{d}\|_2^2\right)\Big|_{\lambda = g_k^{(t+1)}} = 0$ Differentiating at $g_k^{(t+1)}$, we obtain, $g_k^{(t+1)} = \mathbf{D}_k^{(t)T}\mathbf{LD}_k^{(t)}$ We next try to optimize for $\mathbf{u}$ and $\mathbf{d}$, differentiating the Lagrangian formed by the objective function defined in (5) with respect to $\mathbf{u}$ and setting it to zero gives $\mathbf{u} = \widetilde{\mathbf{E}}_k^{(t)T}\mathbf{d}$. Now re-substituting this in the objective and after dropping out terms that do not depend on our varying quantity, $\mathbf{D}_k^{(t)}$, and rearranging we get the objective to be $\arg\min_{\mathbf{D}_k, \|\mathbf{D}_k\|_2 = 1} \mathbf{D}_k^T\mathbf{BD}_k \equiv \arg\max_{\mathbf{D}_k, \|\mathbf{D}_k\|_2 = 1} -\mathbf{D}_k^T\mathbf{BD}_k$. And, $\mathbf{B} = \gamma_k\left(\mathbf{L} - g_k^{(t+1)}\mathbf{I}\right)\left(\mathbf{L} - g_k^{(t+1)}\mathbf{I}\right)^T - \widehat{\mathbf{E}}^{(t)}\widehat{\mathbf{E}}^{(t)T}$. The solution to this is to choose the top eigen-vector of $\mathbf{B}$. This is readily found in the optimization literature (for e.g. [3]). Comparing with the K-SVD algorithm, this algorithm gets modified in the SVD step where we now have to take the top eigenvector of the matrix $\widetilde{\mathbf{E}}_k^{(t)}\widetilde{\mathbf{E}}_k^{(t)T} - \gamma_k\left(\mathbf{L} - g_k\mathbf{I}\right)\left(\mathbf{L} - g_k\mathbf{I}\right)^T$ instead of $\widetilde{\mathbf{E}}_k^{(t)}\widetilde{\mathbf{E}}_k^{(t)T}$ (which is the same as the top left singular vector of $\widetilde{\mathbf{E}}_k^{(t)}$). We now summarize this in Algorithm 1.

4 Simulations

In this simulation, we have synthesized data of a string, fixed at both ends, oscillating in a combination of 4 different modes and a single velocity for a time period of 2s sampled at 2000Hz. In the spatial dimension 4000 points are taken each placed 0.01 space units apart. To testify the robustness of the algorithm we impose an exponential reduction of the wave amplitude with time. Additionally, we

Algorithm 1 wave-informed K-SVD, **Input**: $\mathbf{Y} \in \mathbb{R}^{m \times n}, K \in \mathbb{N}$

Intialize $\mathbf{D}^{(0)}$, $\mathbf{g}^{(0)} = (g_1^{(0)}, g_2^{(0)}, \cdots, g_K^{(0)})$ and $iter$ (no. of iterations)
Set $t = 0$
repeat
    Sparse Code Stage:
    i = 1,2,...,N ; $\min_{\mathbf{X}_i} \{||\mathbf{Y}_i - \mathbf{D}^{(t)}\mathbf{X}_i||_F^2\}$ subject to $||\mathbf{X}_i||_0 \leq s$
    Dictionary Update Stage:
    $g_k^{(t)} = \mathbf{D}_k^{(t)T}\mathbf{L}\mathbf{D}_k^{(t)}; k = 1, 2, ..., K$
    $\mathbf{E}_k^{(t)} = \mathbf{Y} - \sum_{j \neq k} \mathbf{D}_j^{(t)} \widehat{\mathbf{X}}^{(t)T}_j; k = 1, 2, ..., K$
    Let S contain indices of columns that are non-zero. Now $\widetilde{\mathbf{E}}_k^{(t)}$ is formed from $\mathbf{E}_k^{(t)}$ by selecting columns indicated by S.
    Eigen Value Decomposition of $\widetilde{\mathbf{E}}_k^{(t)}\widetilde{\mathbf{E}}_k^{(t)T} - \gamma_k (\mathbf{L} - g_k\mathbf{I})(\mathbf{L} - g_k\mathbf{I})^T = \mathbf{U}\mathbf{\Delta}\mathbf{U}^{-1}$
    Choose column $\mathbf{D}_k^{(t)}$ to be first column of U
    Update $\widetilde{\mathbf{X}}_k^{(t)} = \widetilde{\mathbf{E}}_k^{(t)T}\mathbf{D}_k^{(t)}$
    $\widehat{\mathbf{X}}_k^{(t)}$ is constructed from $\widetilde{\mathbf{X}}_k^{(t)}$ by placing the elements of the latter at the indices indicated by S, zeros otherwise.
    $t \leftarrow t + 1$
until $t == iter$



(a) Spatial data for two different time frames

(b) Columns of dictionary matrix from K-SVD

(c) Columns of dictionary matrix from wave-informed K-SVD

Figure 1: Experiment where the wave data is corrupted by Gaussian noise of SNR = -13 dB

corrupt the data with white Gaussian noise, $n(x, t)$. A continuous version of this data is represented by this equation: $y(x, t) = \sum_{k=1}^{4} \sin(kx) \sin(w_k t)e^{-4t} + n(x, t)$. Note that w_k is calculated using $w_k = vk$, where v is the velocity of the wave and k is the wave number. We demonstrate the effect of noise on the columns of dictionary matrix at SNR = -13dB. SNR is defined in the usual way, $10 \log(P_s/P_n)$, where P_s is the signal power and P_n is the noise power, we calculate the power over all space and time. A sampled version of $y(x, t)$ is the matrix $\widetilde{\mathbf{Y}}$. The columns of the data $\widetilde{\mathbf{Y}}$ represent the string along space whereas rows represent the string along time. We take the discrete Fourier transform on each row of $\widetilde{\mathbf{Y}}$ to form $\mathbf{Y}$. We now perform K-SVD and wave-informed K-SVD (with number of dictionary elements $K = 4$) and a sparsity of $s = 1$ (for the coefficient matrix $\mathbf{X}$) on $\mathbf{Y}$. Each dictionary atom has a different regularization, γ_k. We chose the $\gamma_k \propto 1/g_k^2$ with a proportionality constant, say γ_0, of around 10^5. This proportionality constant is observed to depend on the power of noise present. Comparing Figure 1b with Figure 1c, it is clear that K-SVD learns noisy atoms from data whereas the wave-informed K-SVD learns non-noisy versions which are closer to the actual sinusoids. We also observed a decrease in the noise levels of the dictionary atoms with increasing value of the regularization constant. This indicates an enforcement of wave physics into the algorithm.

5 Conclusion

The main goal of this work is to introduce an optimization framework to embed a physical constraint into a machine learning algorithm and step towards theory guided data science for data describing waves. We observe that the K-SVD algorithm has been naturally transformed into a modified version, with an alteration in the dictionary update step which reflects an enforcement of the physical constraint.

We can look at this algorithm as a filter that filters signals based on the physical domain they are described by. In future work, we want to analyse mathematically how the constraint introduced affects the dictionary update step in infusing structure in the dictionary atoms. We further want to extend this work to sparse autoencoders and delve into the question of model interpretability of the sparse autoencoder neural network for wave based training data.

6 Acknowledgments

This research is supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0126 and the National Science Foundation under award number 1839704.

References

[1] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311, 2006.

[2] K Supreet Alguri, Joseph Melville, and Joel B Harley. Baseline-free guided wave damage detection with surrogate data and dictionary learning. *The Journal of the Acoustical Society of America*, 143(6):3807–3818, 2018.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] Jeffrey M Druce, Jarvis D Haupt, and Stefano Gonella. Anomaly-sensitive dictionary learning for structural diagnostics from ultrasonic wavefields. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 62(7):1384–1396, 2015.

[5] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[6] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

[7] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE, 1993.