Compressible Dictionary Learning for Fast Sparse Approximation

Mehrdad Yaghoobi and Mike Davies

Institute for Digital Communications (IDCOM) & Joint Research Institute for Signal and Image Processing
School of Engineering, University of Edinburgh, Alexander Graham Bell Building, The King's Buildings, Mayfield Road, Edinburgh EH9 3JL, United Kingdom
Yaghoobi@ieee.org, Mike.Davies@ed.ac.uk

Abstract — *In order to find sparse approximations of a class of signals, an appropriate dictionary has to be known. When the dictionary is not available, dictionary learning methods can adapt a dictionary to the given class of signals. An important disadvantage of learnt dictionaries is a lack of structure that would allow fast implementation. Here, we propose a new dictionary model. The imposed model is not only flexible enough to allow the dictionary to adapt to the given class of signals, but also allows it to be implemented efficiently. Dictionary learning, with a compressibility assumption on the dictionary, is a well-defined non-convex optimization problem. We present a practical method to solve compressible dictionary learning approximately. This method can be shown to converge to a local minimum, or converge to a set of local minima. Some preliminary simulation results are shown in this poster.*

Compressible Dictionary Learning Problem

Compressible Dictionary

- *Compressible Matrix* : A matrix $\mathbf{\Psi}$ is compressible [1] if its entries obey a power law

$$|\mathbf{\Psi}|_{(k)} \le c_r k^{-r},$$

where $|\mathbf{\Psi}|_{(k)}$ is the k^{th} largest value of $\mathbf{\Psi}$, $r \ge 1$ and c_r is a constant.

- *Feature*: Let $\mathbf{\Psi}_K$ be the matrix with the K largest elements of $\mathbf{\Psi}$, and let the other elements be zero. $\mathbf{\Psi}_K$ is the best estimate for $\mathbf{\Psi}$ and

$$\|\mathbf{\Psi} - \mathbf{\Psi}_K\|_F \le c'_r K^{-r+1/2}.$$

- *Compressible Dictionary* : A dictionary $\mathbf{D}$ is called compressible if there exist a matrix $\mathbf{\Phi}$, called the mother dictionary, and a compressible matrix $\mathbf{\Psi}$ where

$$\mathbf{D} = \mathbf{\Phi}\mathbf{\Psi}.$$

- *Benefits:* Reducing the complexity of the approximation and a possible fast implementation.

Compressible Dictionary Learning Formulation

- A set of training samples $\{\mathbf{s}_l \in \mathbb{R}^d\}_{l\in\mathbb{L}}$, which build the matrix of training samples $\mathbf{S} \in \mathbf{R}^{d\times L=|\mathbb{L}|}$, and the mother dictionary $\mathbf{\Phi} \in \mathbb{R}^{d\times N}$ are given. The sparse approximation $\mathbf{\Theta}$ and the dictionary generator matrix $\mathbf{\Psi}$ are unknown.
- This problem can be formulated as a non-convex optimization problem as follows,

$$\min_{\mathbf{\Theta},\mathbf{\Psi}} \nu(\mathbf{\Theta},\mathbf{\Psi}) \;:\; \nu(\mathbf{\Theta},\mathbf{\Psi}) = \|\mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta} - \mathbf{S}\|_F^2 + \lambda\|\mathbf{\Theta}\|_{p,p}^p + \gamma\|\mathbf{\Psi}\|_{q,q}^q,$$

where $\|.\|_{p,p}^p$ is a sparsity measure in matrices form. Here we use $\|.\|_{1,1}^1$ which is an ℓ_1 norm (we simply use the notation $\|.\|_1$).

- The epigraph of the objective is a closed set. If we reduce ν in each step, the algorithm is stable (solution space is bounded).
- $\nu(\mathbf{\Theta},\mathbf{\Psi})$ is bi-convex using the ℓ_1 norm.
- A local minimum is found using a block-relaxed minimization algorithm using $\mathbf{\Psi}$ and $\mathbf{\Theta}$ as the blocks.
- Here, we use a majorization minimization method in each iterate of optimization based $\mathbf{\Psi}$ (keeping $\mathbf{\Theta}$ fixed) and $\mathbf{\Theta}$ (keeping $\mathbf{\Psi}$ fixed).

Compressible Dictionary Learning using Majorization Method

The Majorization Method

- A function μ majorizes ν when it satisfies the following conditions,

$$\nu(\omega) \le \mu(\omega,\xi),\ \forall\omega,\xi \in \Upsilon$$
$$\nu(\omega) = \mu(\omega,\omega),\ \forall\omega \in \Upsilon,$$

where Υ is the parameter space.

- The augmented function has an additional parameter ξ. We set this parameter to the current value of ω and find the optimal update for ω.

$$\omega_{new} = \arg\min_{\omega\in\Upsilon} \mu(\omega,\xi)$$

We then update ξ with ω_{new}.

- Augmented functions can be generated using Taylor series or Jensen's inequality.
- The Taylor series of a differentiable function $\nu(\omega)$ is,

$$\nu(\omega) = \nu(\xi) + d\nu(\xi)(\omega-\xi) + \frac{1}{2!}d^2\nu(\xi)(\omega-\xi)^2 + o(\omega^3).$$

If ν has bounded curvature ($d^2\nu < c_s$ for a finite constant c_s) this is majorized by,

$$\nu(\omega) \le \nu(\xi) + d\nu(\xi)(\omega-\xi) + \frac{c_s}{2}(\omega-\xi)^2, \forall\omega,\xi \in \Upsilon.$$

Therefore we can define $\mu(\omega,\xi)$ as follows,

$$\mu(\omega,\xi) = \nu(\xi) + d\nu(\xi)(\omega-\xi) + \frac{c_s}{2}(\omega-\xi)^2.$$

Majorizing Functions for CDL

- The majorizing function of $\nu_{\mathbf{\Psi}}(\mathbf{\Theta})$, assuming $\mathbf{\Psi}$ fixed, is found using the Taylor series. This can be reformulated by adding a strictly convex function $\pi_{\mathbf{\Psi}}(\mathbf{\Theta},\mathbf{\Theta}^{[n]})$ to $\nu_{\mathbf{\Psi}}(\mathbf{\Theta})$ which is found to be,

$$\pi_{\mathbf{\Psi}}(\mathbf{\Theta},\mathbf{\Theta}^{[n]}) = c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}\|\mathbf{\Theta} - \mathbf{\Theta}^{[n]}\|_F^2 - \|\mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta}^{[n]}\|_F^2,$$

where $c_{\mathbf{\Phi}} > \|\mathbf{\Phi}^T\mathbf{\Phi}\|$ and $c_{\mathbf{\Psi}} > \|\mathbf{\Psi}^T\mathbf{\Psi}\|$. The augmented objective $\mu_{\mathbf{\Psi}}(\mathbf{\Theta},\mathbf{\Theta}^{[n]})$ is then found by,

$$\mu_{\mathbf{\Psi}}(\mathbf{\Theta},\mathbf{\Theta}^{[n]}) = tr\{c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}\mathbf{\Theta}^T\mathbf{\Theta} - 2\mathbf{\Theta}^T(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{S} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta}^{[n]}) + c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}\mathbf{\Theta}^{[n]})\} + \lambda\|\Theta\|_1 + c_\Theta,$$

where c_Θ is a constant w.r.t. $\mathbf{\Theta}$.

- When $\mathbf{\Theta}$ is fixed, a similar technique can be used to generate the augmented objective for $\nu_{\mathbf{\Theta}}(\mathbf{\Psi})$. Here, $\pi_\Theta(\mathbf{\Psi},\mathbf{\Psi}^T)$ is calculated by,

$$\pi_{\mathbf{\Theta}}(\mathbf{\Psi},\mathbf{\Psi}^{[n]}) = c_{\mathbf{\Phi}}c_{\mathbf{\Theta}}\|\mathbf{\Psi} - \mathbf{\Psi}^{[n]}\|_F^2 - \|\mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta} - \mathbf{\Phi}\mathbf{\Psi}^{[n]}\mathbf{\Theta}\|_F^2.$$

where $c_{\mathbf{\Theta}} > \|\mathbf{\Theta}\mathbf{\Theta}^T\|$.The augmented objective $\mu_{\mathbf{\Psi}}(\mathbf{\Theta},\mathbf{\Theta}^{[n]})$ is now found to be,

$$\mu_\Theta(\mathbf{\Psi},\mathbf{\Psi}^{[n]}) = tr\{c_{\mathbf{\Phi}}c_{\mathbf{\Theta}}\mathbf{\Psi}^T\mathbf{\Psi} - 2\mathbf{\Psi}^T(\mathbf{\Phi}^T(\mathbf{S} - \mathbf{\Phi}\mathbf{\Psi}^{[n]}\mathbf{\Theta})\mathbf{\Theta}^T + c_{\mathbf{\Phi}}c_{\mathbf{\Theta}}\mathbf{\Psi}^{[n]})\} + \lambda\|\Psi\|_1 + c_\Psi,$$

where c_Ψ is a constant w.r.t. $\mathbf{\Psi}$.

Optimization of the Augmented Objectives

- The augmented objectives are convex and non-differentiable. Therefore, zero is in the sub-gradient of these objectives at the optimum.
- The update of $\mathbf{\Theta}$, given $\mathbf{\Theta}^{[n]}$, is found by,

$$0 \in \partial\mu_\Psi(\mathbf{\Theta},\mathbf{\Theta}^{[n]}),$$

where

$$\partial\mu_\Psi(\mathbf{\Theta},\mathbf{\Theta}^{[n]}) = 2c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}\mathbf{\Theta} - 2(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{S} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta}^{[n]}) + c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}\mathbf{\Theta}^{[n]}) + \lambda\partial\|\Theta\|_1.$$

This gives an update formula based on a soft-shrinkage operator as follows,

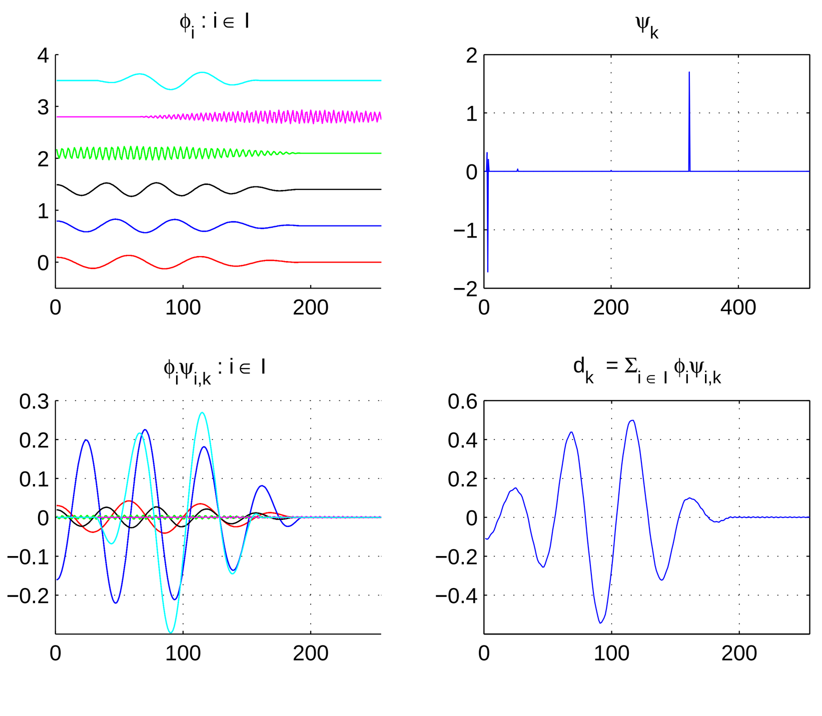
$$\mathbf{\Theta}^{[n+1]} = \mathcal{S}_{\lambda/2}\left[\frac{1}{c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}}(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{S} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{\Theta}^{[n]}) + c_{\mathbf{\Phi}}c_{\mathbf{\Psi}}\mathbf{\Theta}^{[n]})\right]$$

- A similar method can be used to find the update formula for $\mathbf{\Psi}$, given $\mathbf{\Psi}^{[n]}$. The $\mathbf{\Psi}^{[n+1]}$ is calculated by,

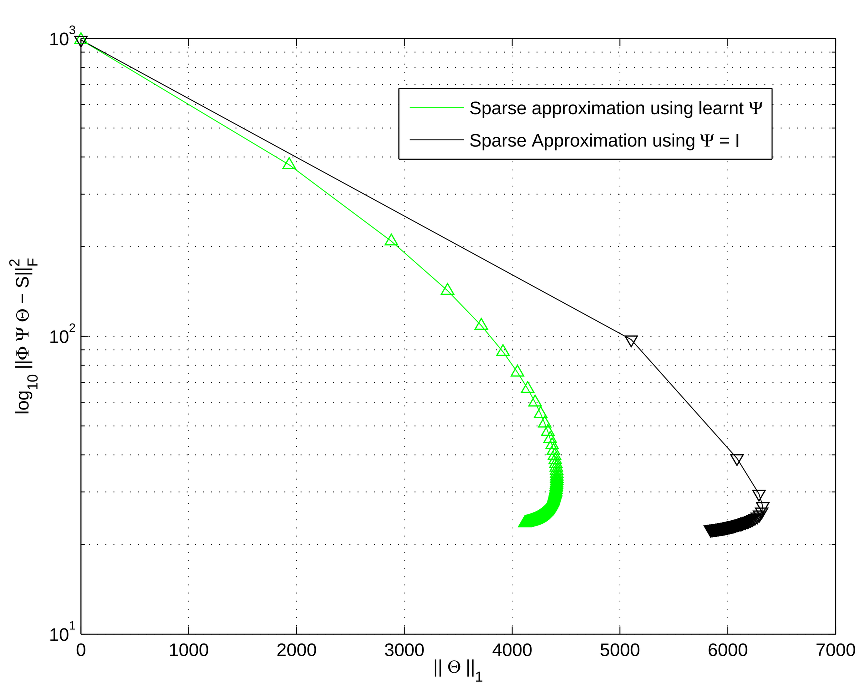
$$\mathbf{\Psi}^{[n+1]} = \mathcal{S}_{\gamma/2}\left[\frac{1}{c_{\mathbf{\Phi}}c_{\mathbf{\Theta}}}(\mathbf{\Phi}^T(\mathbf{S} - \mathbf{\Phi}\mathbf{\Psi}^{[n]}\mathbf{\Theta})\mathbf{\Theta}^T + c_{\mathbf{\Phi}}c_{\mathbf{\Theta}}\mathbf{\Psi}^{[n]})\right]$$

Simulation Results

Compressible Dictionary Learning for Audio Signals using a multi-scale MDCT mother dictionary.



A $\mathbf{\Psi} \in \mathbb{R}^{256\times512}$ is learnt by using a set of 8192 audio samples. The generation of an atom $\mathbf{d}_k$ is shown in this plot.



The phase plots (representation error v.s. ℓ_1) of 4096 evaluation signals.

Conclusion

- A novel model is introduced for the dictionary learning problem.
- The learnt compressible dictionary can be implemented efficiently using a suitable mother dictionary.
- The new model reduces the complexity of the approximation of the dictionary. Therefore fewer training samples are needed.
- The proposed algorithm is easy to implement and parallelizable.
- The convergence of the algorithm, to a connected set of local minima, is guaranteed.
- It is flexible and allows $\mathbf{\Theta}$ and $\mathbf{\Psi}$ to be updated in any cyclic sequence.

Acknowledgement

This work was fully supported by EPSRC grant EP/D000246. MY wants to acknowledge the support of Dr. Thomas Blumensath in this research.

[1] E. J. Candes, Justin K. Romberg and Terence Tao, *"Stable signal recovery from incomplete and inaccurate measurements"*, Communications on Pure and Applied Mathematics, Vol. 59, No. 8, p.p. 1207-1223, 2006.