Learning Dictionaries as Sums of Kronecker Products

Cássio Fraga Dantas, Rémi Gribonval
INRIA Rennes-Bretagne Atlantique, France

Renato R. Lopes, Michele N. da Costa
University of Campinas, Brazil

***Abstract*—The choice of an appropriate dictionary is a crucial step in the sparse representation of a given class of signals. Traditional dictionary learning techniques generally lead to unstructured dictionaries which are costly to deploy and do not scale well to higher dimensional signals. In order to overcome such limitation, we propose a learning algorithm that constrains the dictionary to be a sum of Kronecker products of smaller sub-dictionaries. A special case of the proposed structure is the widespread separable dictionary. This approach, named SuKro, is evaluated experimentally on an image denoising application.**

I. Introduction

Dictionary learning algorithms [1] generally lead to unstructured over-complete dictionaries which are very costly to operate with, limiting their applicability to relatively low-dimensional problems. In order to obtain computationally efficient dictionaries, some of the most recent works in the field employ parametric models in the training process, which produce structured dictionaries [2]–[8]. Among the countless possibilities, a promising one is learning separable dictionaries [2], which can be represented as the Kronecker product of two sub-dictionaries, i.e. $\mathbf{D} = \mathbf{B} \otimes \mathbf{C}$.

We propose a broader structure class of which the separable structure is a special case. It consists of a sum of separable terms, where the number of components serves as a fine tuner for the complexity-adaptability tradeoff:

$$\mathbf{D} = \sum_r \mathbf{B}^{(r)} \otimes \mathbf{C}^{(r)}. \tag{1}$$

II. Proposed technique

Consider a matrix $\mathbf{D} \in \mathbb{R}^{n_1 n_2 \times m_1 m_2}$ which is the Kronecker product of two sub-matrices $\mathbf{B} \in \mathbb{R}^{n_1 \times m_1}$ and $\mathbf{C} \in \mathbb{R}^{n_2 \times m_2}$. The *rearrangement operator*, defined in [9] and here denoted $\mathcal{R}(\cdot)$, reorganizes the elements of $\mathbf{D}$ in such a way that the rearranged matrix $\mathcal{R}(\mathbf{D}) \in \mathbb{R}^{m_1 n_1 \times m_2 n_2}$ has rank one and can be written as an outer product of the vectorized versions of $\mathbf{B}$ and $\mathbf{C}$.

$$\mathcal{R}(\mathbf{D}) = \operatorname{vec}(\mathbf{B}) \operatorname{vec}(\mathbf{C})^{\mathrm{T}}. \tag{2}$$

Now, let us consider a sum of α Kronecker products

$$\mathbf{D} = \sum_{r=1}^{\alpha} \mathbf{B}^{(r)} \otimes \mathbf{C}^{(r)} = \sum_{r=1}^{\alpha} \mathbf{D}^{(r)}. \tag{3}$$

After rearrangement, we obtain a rank-α matrix, since each term $\mathbf{D}^{(r)}$ leads to a rank-1 matrix as follows

$$\mathcal{R}(\mathbf{D}) = \sum_{r=1}^{\alpha} \mathcal{R}(\mathbf{D}^{(r)}) = \sum_{r=1}^{\alpha} \operatorname{vec}\Big(\mathbf{B}^{(r)}\Big) \operatorname{vec}\Big(\mathbf{C}^{(r)}\Big)^{\mathrm{T}}. \tag{4}$$

Therefore, by using (4), we can add a low-rank regularization term to the original dictionary learning optimization problem in order to learn a dictionary as a sum of few Kronecker products:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{X}} \ \tfrac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda \operatorname{rank}(\mathcal{R}(\mathbf{D})) \\ &\text{s.t.} \quad \forall i \ \|\mathbf{x}_i\|_0 \le t \ , \quad \forall j \ \|\mathbf{d}_j\|_2 = 1 \end{aligned} \tag{5}$$

where the parameter $\lambda \in \mathbb{R}^+$ controls the rank penalty.

III. Optimization Framework

As typically done in the literature [10]–[12], we solve the problem in (5) by alternately minimizing on the variables $\mathbf{D}$ and $\mathbf{X}$. On the sparse coding step (minimization with respect to $\mathbf{X}$), we use the existing Orthogonal Matching Pursuit (OMP) algorithm [13].

On the dictionary update step, we use the nuclear norm (denoted $\|\cdot\|_*$) as a convex relaxation of the rank operator [14], yielding

$$\text{Dictionary update:} \quad \min_{\mathbf{D}} \ \tfrac{1}{2}\|\mathbf{D}\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda \|\mathcal{R}(\mathbf{D})\|_* . \tag{6}$$

The Alternating Direction Method of Multipliers (ADMM) [15] can be employed to solve such problem. It introduces an auxiliary variable $\widetilde{\mathbf{D}} = \mathcal{R}(\mathbf{D})$ and a Lagrangian multiplier matrix $\mathbf{Z}$, as shown in Algorithm 1. The partial update with respect to $\widetilde{\mathbf{D}}$ (second step in Alg. 1) is the proximal operator associated to the nuclear norm. It consists in the singular value soft-thresholding operation [16].

Algorithm 1 Dictionary Update - ADMM

Initialize $\mathbf{D}_0$, $\widetilde{\mathbf{D}}_0$, $\mathbf{Z}_0$
while $\|\mathbf{Z}_{k+1} - \mathbf{Z}_k\|_F^2 < tol$ **do**
$\mathbf{D}_{k+1} = \mathbf{D}_k - \gamma \left[(\mathbf{D}_k\mathbf{X} - \mathbf{Y})\mathbf{X}^T + \mu\Big(\mathbf{D}_k - \mathcal{R}^{-1}(\widetilde{\mathbf{D}}_k - \mathbf{Z}_k)\Big)\right]$
$\widetilde{\mathbf{D}}_{k+1} = \operatorname{prox}_{\frac{\lambda}{\mu}\|\cdot\|_*}\left(\mathcal{R}(\mathbf{D}_{k+1}) + \mathbf{Z}_k\right)$
$\mathbf{Z}_{k+1} = \mathbf{Z}_k - \Big(\widetilde{\mathbf{D}}_{k+1} - \mathcal{R}(\mathbf{D}_{k+1})\Big)$
end while
Normalize columns of $\mathbf{D}$

IV. Experiments

We use a patch-based image denoising application to validate the proposed algorithm. The simulation set-up is the same as in [17] and is summarized in Table I. We compare our results to the unstructured K-SVD [11] dictionary, the SeDiL [2] separable dictionary, which uses a single Kronecker product and a different optimization strategy, and the ODCT analytic dictionary.

Figure 1 shows the denoised image PSNR as a function of the number of separable terms in the dictionary, which can be controlled by the parameter λ in (6). Note that even with very few separable terms, the results are close to the K-SVD. SuKro may even outperform K-SVD for high-noise scenarios, suggesting that imposing some structure on the dictionary may decrease the problem of overfitting. Also note that the one-term SuKro dictionary outperforms SeDiL. Naturally, as the number of separable terms increases, so does the denoising performance, since the structure becomes more flexible. The drawback is the increase in the dictionary application complexity.

Figure 2 illustrates the complexity-performance tradeoff. Besides providing very good denoising performance with quite reduced complexity costs, the proposed technique has the merit of providing a range of options on this tradeoff curve.

V. Conclusion

The proposed dictionary structure leads to fast operators while keeping a considerable degree of flexibility. Such tradeoff can be controlled through the number of terms in the summation. The image denoising simulations have shown very promising results, specially for higher noise scenarios. In such cases, the proposed structure manages to overcome an unstructured dictionary in terms of both computational complexity and denoising performance.

TABLE I
SIMULATION PARAMETERS

Sample dimension (n)	64
Number of atoms (m)	256
Training samples (N)	40000
Step-size (γ)	6×10^{-9}
Lagrangian penalty (μ)	10^7
Convergence tolerance (tol)	$\|\mathbf{D}\|_F \times 10^{-4}$
Alternate optimization iterations (N_{iter})	100
Dictionary initialization ($\mathbf{D}_0$)	ODCT
Size of sub-matrices $\mathbf{B}^{(r)}$ and $\mathbf{C}^{(r)}$	(8×16)
Size of $\mathcal{R}(\mathbf{D})$	(128×128)

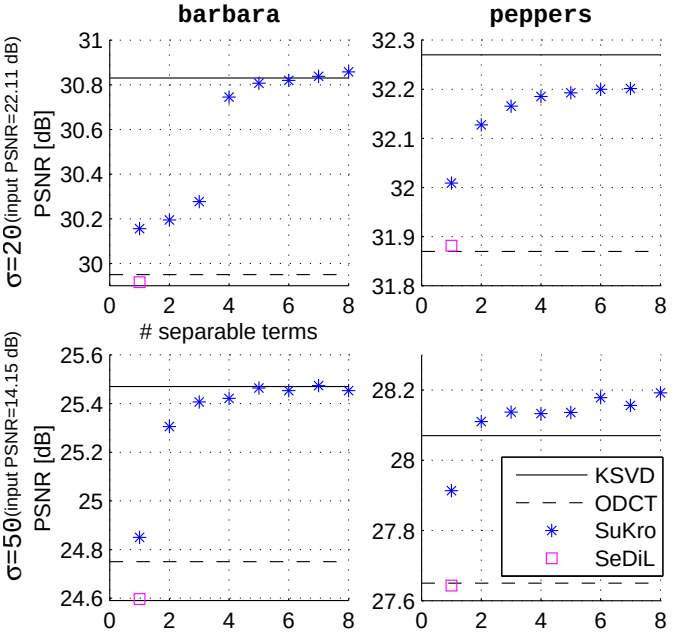

Fig. 1. PSNR vs. $\mathrm{rank}(\widetilde{\mathbf{D}})$ (i.e. the number of separable terms). σ is the noise standard deviation.

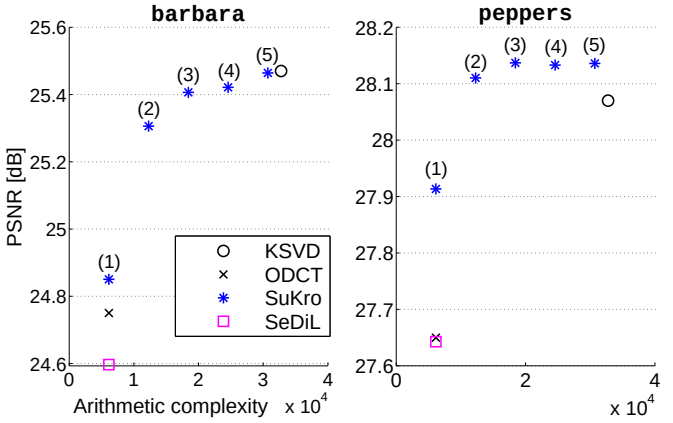

Fig. 2. Performance (PSNR) vs. Complexity, with noise standard deviation $\sigma = 50$. The number of separable terms is displayed between parentheses.

REFERENCES

[1] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[2] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 438–445.

[3] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM Journal on Imaging Sciences*, vol. 1, no. 3, pp. 228–247, 2008. [Online]. Available: http://dx.doi.org/10.1137/07070156X

[4] M. Yaghoobi and M. E. Davies, "Compressible dictionary learning for fast sparse approximations," in *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, Aug 2009, pp. 662–665.

[5] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *Signal Processing, IEEE Transactions on*, vol. 58, no. 3, pp. 1553–1564, 2010.

[6] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3180–3193, June 2016.

[7] L. L. Magoarou and R. Gribonval, "Chasing butterflies: In search of efficient dictionaries," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 3287–3291.

[8] O. Chabiron, F. Malgouyres, J.-Y. Tourneret, and N. Dobigeon, "Toward fast transform learning," *International Journal of Computer Vision*, vol. 114, no. 2-3, pp. 195–216, 2015.

[9] C. F. Van Loan and N. Pitsianis, "Approximation with Kronecker products," in *Linear algebra for large scale and real-time applications*. Springer, 1993, pp. 293–314.

[10] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5, 1999, pp. 2443–2446 vol.5.

[11] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[13] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, Nov 1993, pp. 40–44 vol.1.

[14] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010. [Online]. Available: http://dx.doi.org/10.1137/070697835

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan 2011. [Online]. Available: http://dx.doi.org/10.1561/2200000016

[16] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[17] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.