Learning non-structured, overcomplete and sparsifying transform

Dimche Kostadinov, Slava Voloshynovskiy and Sohrab Ferdowsi
Computer Science Department, University of Geneva, 7 Route de Drize, Carouge 1227 GE, Switzerland

I. TRANSFORM LEARNING

Transform learning has been introduced and studied in [1],[2], [3] and [4]. An optimal transform learning for structured and overcomplete matrix was proposed in [5]. However, several issues (optimality, convergence and computational complexity) related to learning an incoherent, well-conditioned, non-structured and overcomplete sparsifing transform still remain open.

Let $\mathbf{X} \in \Re^{N\times L}$ be a data matrix, having as columns data samples $\mathbf{x}_i \in \Re^N, i \in \mathcal{I} = \{1, 2, ..., L\}$. Assuming a sparsifing transform model [1], then the problem formulation for learning the overcomplete transform matrix $\mathbf{A} \in \Re^{M\times N}$, $(M > N)$ has the following form:

$$\underset{\mathbf{A},\mathbf{Y}}{min} g(\mathbf{A}, \mathbf{Y}) = \underset{\mathbf{A},\mathbf{Y}}{min} \|\mathbf{AX} - \mathbf{Y}\|_F^2 + \Omega_1(\mathbf{A}) + \lambda_1\|\mathbf{Y}\|_1, \qquad (1)$$

where $\|.\|_F$ and $\|.\|_1$ denote the Frobenius norm and the l_1 norm, respectively. The penalty $\Omega_1(\mathbf{A})$ is defined as: $\Omega_1(\mathbf{A}) = \frac{\lambda_2}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda_3}{2}\|\mathbf{AA}^T - \mathbf{I}\|_F^2 - \lambda_4 \log|\det \mathbf{A}^T\mathbf{A}|$ and λ_k are Lagrangian multipliers $\forall k \in \{1, 2, 3, 4\}$. The first term in (1) is known as *sparsification error* [1]. It represents the deviation of the transformed data from the exact sparse representation in the transform domain. The $\log|\det\ (\mathbf{A}^T\mathbf{A})|$ and $\|\mathbf{A}\|_F^2$ are functions of the singular values of $\mathbf{A}$ and together help regularize the conditioning of $\mathbf{A}$. Assuming that the expected coherence $\mu^2(\mathbf{A})$ between the rows $\mathbf{a}_m$ of $\mathbf{A}$ (*i.e.* $\mathbf{A}^T = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_M]$) is defined as: $\mu^2(\mathbf{A}) = \frac{2}{M(M-1)}\sum_{m_1\neq m_2}|\mathbf{a}_{m_1}\mathbf{a}_{m_2}^T|^2, \forall m_1, m_2 \in \{1, 2, .., M\}$ then the penalty $\|\mathbf{AA}^T - \mathbf{I}\|_F^2$ helps enforce a minimum expected coherence $\mu^2(\mathbf{A})$ and unit one row norm $\|\mathbf{a}_m\|_2 = 1$. The matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_L] \in \Re^{M\times L}$ has as columns the transformed data $\mathbf{y}_i, i \in \mathcal{I}$ that are constrained to have a small number of non-zeros by $\lambda_1\|\mathbf{y}_i\|_1$.

Similarity as in [4], [6] and [7] an alternating algorithm that has two steps (transform update step and sparse coding step) is proposed to solve (1) by iteratively updating $\mathbf{A}$ and $\mathbf{Y}$. It is important to highlight that the algorithm proposed in [4] solves the transform update step using projected conjugate gradient method. The main contribution in this work is the proposed closed form ϵ-close solution for the transform update step that results in a low computational complexity. Moreover, although an ϵ-close solution is used in the transform update step the proposed two step iterative algorithm is convergent [8].

A. Transform estimate

Given the current estimate of $\mathbf{Y}$, the estimate of the transform $\mathbf{A}$ is solution to the following problem $(P1) : \{\hat{\mathbf{A}}\} = argmin_{\mathbf{A}}\, \|\mathbf{AX} - \mathbf{Y}\|_F^2 + \Omega_1(\mathbf{A})$. The resulting ϵ-close solution w.r.t ϵ-close approximation $g_1(\mathbf{A}, \mathbf{Y}) \leq g(\mathbf{A}, \mathbf{Y})$ of the term $Tr\{\mathbf{AXY}^T\}$ in $(P1)$ is given by Proposition 1 (where $Tr\{\mathbf{AXY}^T\}$ appears from $Tr\{\mathbf{AXX}^T\mathbf{A}^T - 2\mathbf{AXY}^T + \mathbf{YY}^T\} = \|\mathbf{AX} - \mathbf{Y}\|_F^2$).

Proposition 1 *Given* $\mathbf{Y} \in \Re^{M\times L}$, $\forall\mathbf{X} \in \Re^{N\times L}$ *and* $M \geq N$, $\forall\lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0, \lambda_4 \geq 0$, *let the eigen value decomposition* $\mathbf{U}_X\mathbf{\Sigma}_X^2\mathbf{U}_X^T$ *of* $\mathbf{XX}^T + \lambda_1\mathbf{I}$ *and the singular value decomposition* $\mathbf{U}_{U_XXY}\mathbf{\Sigma}_{U_XXY}\mathbf{V}_{U_XXY}^T$ *of* $\mathbf{U}_X^T\mathbf{XY}^T$ *exists, then if and only if* $\sigma_X(n) > 0$, $\forall n \in \{1, 2, 3, ..., N\}$ *and* $\lambda_3 \geq 0$, $(P1)$ *has* ϵ*-close solution to the optimal solution of* $(P1)$ as:

$$\hat{\mathbf{A}} = \mathbf{V}_{U_XXY}\mathbf{U}_{U_XXY}^T\mathbf{\Sigma}_A\mathbf{\Sigma}_X^{-1}\mathbf{U}_X^T, \qquad (2)$$

where $\mathbf{\Sigma}_A$ *is diagonal matrix,* $\Sigma_A(n, n) = \sigma_A(n) \geq 0$, $\sigma_A(n), \forall i$ *are solutions to:*

$$\min_{\sigma_A(n)} \frac{\lambda_3}{\sigma_X{}^4}\sigma_A^4(n) + \left(\frac{\sigma_X^2(n) - 2\lambda_3}{\sigma_X^2(n)}\right)\sigma_A^2(n) - \frac{\sigma_{U_XXY}(n)}{\sigma_X(n)}\sigma_A(n) - 2\lambda_4 \log\frac{\sigma_A(n)}{\sigma_X(n)}, \qquad (3)$$

and the ϵ*-close solution is* $\epsilon = Tr\{(\mathbf{U}_{U_XXY}\ \mathbf{\Sigma}_{U_XXY}\ \mathbf{U}_{U_XXY}^T - \mathbf{\Sigma}_{U_XXY})\mathbf{\Sigma}_A\ \mathbf{\Sigma}_X^{-1}\} \leq 0$ *close to the optimal solution.*

B. Sparse coding

Given the current estimate of the transform $\mathbf{A}$, the sparse coding problem is formulated as $(P2) : \hat{\mathbf{Y}} = argmin_{\mathbf{Y}}\, \|\mathbf{AX} - \mathbf{Y}\|_F^2 + \lambda_1\|\mathbf{Y}\|_1$. Define $\mathbf{Q} = \mathbf{AX}$ then the closed form solution to $(P2)$ is $\hat{y}_i(m) = q_i(m)$, if $|q_i(m)| > \lambda_1$, and $\hat{y}_i(m) = 0$, otherwise, $\forall i \in \mathcal{I}$.

Lemma 1: *Given data* $\mathbf{X}$ *and a pair of initial transform and sparse data* $\{\mathbf{A}^0, \mathbf{Y}^0\}$, *let* $\{\mathbf{A}^k, \mathbf{Y}^k\}$ *denote the iterative sequence generated by the solutions* (2) *with* (3) *and the closed form solution of* $(P2)$. *Then, the sequence of the objective function values* $g\left(\mathbf{A}^k, \mathbf{Y}^k\right)$ *is monotone decreasing sequence, satisfying* $g\left(\mathbf{A}^{k+1}, \mathbf{Y}^{k+1}\right) \leq g_1(\mathbf{A}^{k+1}, \mathbf{Y}) \leq g\left(\mathbf{A}^{k+1}, \mathbf{Y}^k\right) \leq g\left(\mathbf{A}^k, \mathbf{Y}^k\right)$ *and converges to a finite value denoted as* $g^* = g^*\left(\mathbf{A}^0, \mathbf{Y}^0\right)$.

Proof: The proof is given in [8] (a submitted paper for review).

II. EXPERIMENTAL RESULTS

This section presents the preliminary results for image denoising application where the denoising problem formulation is similar to the one proposed in [4].

Peppers, Cameramen and Barbara at image resolution 256×256, 256×256 and 512×512, respectively, are used to evaluate the potential advantages of the proposed algorithm named (ϵTOL). We compare it with the transform learning based on: square matrix (TL-S) [6], non-structured overcomplete matrix (TL-O) [4], and K-SVD [7]. The results are shown on Tables 1, 2 and 3.

The properties of the proposed algorithm and the transform $\mathbf{A}$ using the Cameramen image as example are shown on Figure 1. The learned transforms for all the images have good conditioning numbers and low expected coherence (Table 1). The average execution time and the training data requirements are lower than those of the comparing algorithms. Moreover the run time is $\mathbf{3}\times$, $\mathbf{2}\times$ and $\mathbf{9}\times$ faster then TL-S, TL-O and K-SVD, respectively, as shown in Table 2. Further, the number of required parameters is small. Surprisingly, the number of noisy image patches required for learning is drastically lower than those of the reference algorithms. At the end as shown on Table 3 the recovery results are comparable or provide small improvement with respect to the reference algorithms.

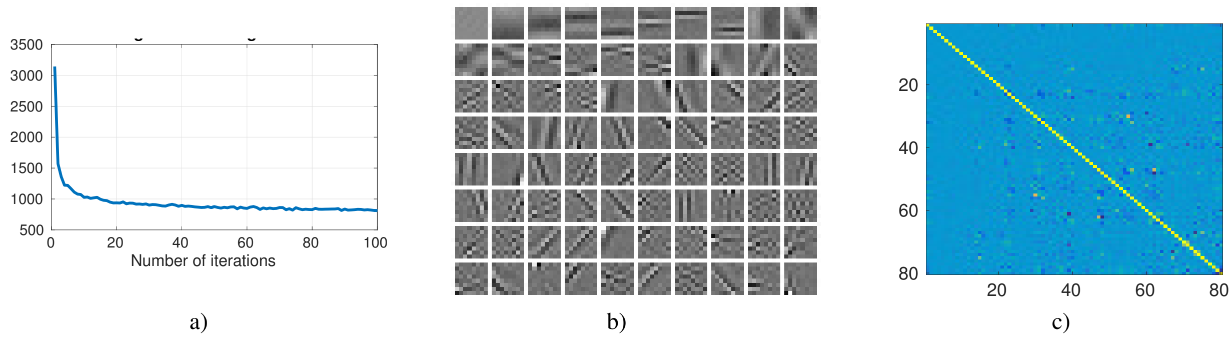

Fig. 1. The properties of the proposed algorithm and the transform $\mathbf{A}$ using the Cameraman image: a) The evolution of the sparsification error of the proposed algorithm; b) The actual basis vectors $\mathbf{a}_i$ (rows of $\mathbf{A}$) for the learned transform; c) The matrix $\mathbf{A}\mathbf{A}^T$ where the expected value of the values of the off diagonal elements represents the mutual coherence $\mu(\mathbf{A})$ (the transform was learned on overlapping 8×8 noisy image blocks, equivalently $N = 64$, and M was set to 80).

	Peppers	Cameramen	Barbara
C_n	1.31	1.22	1.28
$\mu^2(\mathbf{A})$	0.0023	0.0033	0.0031

TABLE I
THE CONDITIONING NUMBER C_n AND THE EXPECTED COHERENCE $\mu^2(\mathbf{A})$ OF $\mathbf{A}$.

	TL-S	TL-O	K-SVD	ϵTOL
$t_e[min]$	4.6	2.9	9.8	$\mathbf{1.15}$
$l_{data}[\%]$	$25 - 100$	$25 - 100$	100	$\mathbf{3 - 15}$

TABLE II
RUNNING TIME OF THE ALGORITHMS USING MATLAB IMPLEMENTATION DENOTED AS t_e AND THE PERCENTAGE OF THE TOTAL NUMBER OF TRAINING DATA USED IN THE LEARNING DENOTED AS l_{data}.

	σ	TL-S	TL-O	K-SVD	ϵTOL
Peppers	10	34.45	$\mathbf{34.49}$	34.2	34.44
	20	29.98	30.60	29.82	$\mathbf{30.63}$
Cameramen	10	$\mathbf{33.93}$	33.83	33.72	$\mathbf{33.93}$
	20	29.93	29.95	29.82	$\mathbf{30.12}$
Barbara	10	34.45	34.55	34.42	$\mathbf{34.60}$
	20	30.53	30.90	30.82	$\mathbf{30.91}$

TABLE III
DENOSING PERFORMANCE IN PSNR, WHERE σ IS THE NOISE STANDARD DEVIATION.

REFERENCES

[1] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms for image processing," in *19th IEEE International Conference on Image Processing, ICIP 2012, Lake Buena Vista, Orlando, FL, USA, September 30 - October 3, 2012*, 2012, pp. 681–684.

[2] S. Ravishankar and Y. Bresler, "Closed-form solutions within sparsifying transform learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 5378–5382.

[3] S. Ravishankar and Y. Bresler, "Doubly sparse transform learning with convergence guarantees," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. 2014, pp. 5262–5266, IEEE.

[4] S. Ravishankar and Y. Bresler, "Learning overcomplete sparsifying transforms for signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 3088–3092.

[5] B. Wen, S. Ravishankar, and Yoram Bresler, "Structured overcomplete sparsifying transform learning with convergence guarantees and applications," *International Journal of Computer Vision*, vol. 114, no. 2-3, pp. 137–167, 2015.

[6] S. Ravishankar and Y. Bresler, "ℓ_0 sparsifying transform learning with efficient optimal updates and convergence guarantees," *CoRR*, vol. abs/1501.02859, 2015.

[7] M. Aharon, M. Elad, and A. Bruckstein, "Svdd: An algorithm for designing overcomplete dictionaries for sparse representation," *Trans. Sig. Proc.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[8] D. Kostadinov and S. Voloshinovsky, "Learning non-structured, overcomplete and sparsifying transform," in *2017 IEEE International Conference on Image Processing, ICIP 2017, Beljing, China, September 27-30, 2017, submmited*.