Learning Overcomplete Signal Sparsifying Transforms

Saiprasad Ravishankar and Yoram Bresler
Department of Electrical and Computer Engineering and the Coordinated Science Laboratory,
University of Illinois, Urbana-Champaign, IL 61801, USA

I. Transform Learning

The formulations for learning synthesis [1] and analysis [2], [3] sparsifying dictionaries are typically non-convex and NP-hard, and the approximate algorithms are still computationally expensive. As an alternative, we recently introduced an approach for learning square sparsifying transforms $W \in \mathbb{R}^{m \times n}$, $m = n$ [4], which are competitive with overcomplete synthesis or analysis dictionaries in image denoising, at a fraction of the computational cost. In this work, we extend the learning to the overcomplete case, i.e., $m > n$.

The classical transform model for signal y is $Wy = x + e$, where $W \in \mathbb{R}^{m \times n}$ is a sparsifying transform, $x \in \mathbb{R}^m$ is sparse ($\|x\|_0 \ll m$), and e is a small residual in the transform domain [4]. Given a matrix $Y \in \mathbb{R}^{n \times N}$ whose columns are training signals, our formulation [4] for learning a square transform $W \in \mathbb{R}^{n \times n}$ is

$$\text{(P1)} \quad \min_{W, X; \|X_i\|_0 \leq s \ \forall \ i} \|WY - X\|_F^2 - \lambda \log \det W + \mu \|W\|_F^2$$

Here, X is a matrix, whose columns X_i are the sparse codes of the signals in Y. The term $\|WY - X\|_F^2$ is called *sparsification error*. The $-\log \det W$ (with $\det W > 0$) and $\|W\|_F^2$ penalties eliminate trivial solutions, and allow full control over the conditioning of W [4]. For the overcomplete transform case, we extend (P1) as follows.

$$\min_{W, X} \|WY - X\|_F^2 - \lambda \log \det \left(W^T W\right) + \eta \sum_{j \neq k} |\langle w_j, w_k \rangle|^p$$
$$s.t. \ \ \|X_i\|_0 \leq s \ \forall \ i, \ \|w_k\|_2 = 1 \ \forall \ k. \qquad \text{(P2)}$$

where we have replaced $\log \det W$ in (P1) with $\log \det \left(W^T W\right)$ in (P2), which now enforces full column rank of W. However, it cannot preclude repeated rows in W. Hence, we include an additional penalty on the inner products between the rows w_j and w_k (of unit norm) of the transform, to control their coherence. When $p = 2$, and W consists of orthonormal blocks, the incoherence penalty is constant (inactive) irrespective of the choice of the blocks, and thus fails to detect coherence. On the other hand, $p \leq 1$ enforces sparsity of the gram matrix WW^T (at most n rows of W can be mutually orthogonal). However, it cannot prevent repetition of a set of ($\leq n$) orthogonal rows, since the magnitudes of the non-zero inner products need not be small. Larger values of p (> 2) emphasize the peak coherence, and were found to work well in our experiments. Problem (P2) is however, non-convex.

Our algorithm for solving (P2) alternates between updating X (*sparse coding step*) and W (*transform update step*). The solution $\hat{X}$ of the sparse coding step is computed exactly by thresholding WY, and retaining the s largest coefficients (in magnitude) in each column. In the transform update step, we could solve for W using algorithms such as projected conjugate gradient (CG). However, we observed that the alternative strategy of employing the standard CG algorithm, followed by post-normalization of the rows of W led to better empirical performance in applications. Hence, we choose this alternative strategy, and also retain the $\|W\|_F^2$ penalty in the cost for CG, to prevent the scaling ambiguity [4].

This work was supported in part by the National Science Foundation (NSF) under grant CCF 10-18660.

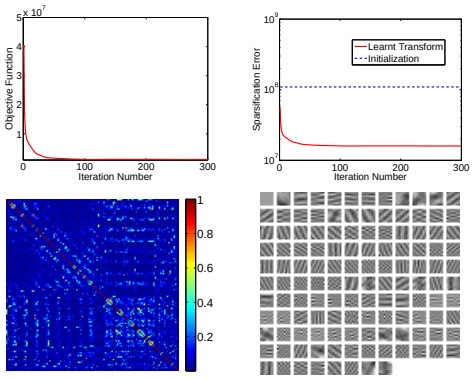

Fig. 1. Top: Evolution of the objective function, and sparsification error, along with the sparsification error of the initial transform. Bottom: Magnitude of WW^T (left), Rows of the learnt transform shown as patches (right).

The computational cost per iteration (of sparse coding and transform update) of the proposed algorithm scales as $O(mnN)$ for learning an $m \times n$ transform from N training vectors. This cost is typically much lower than the per-iteration cost of learning an $n \times K$ synthesis dictionary D using K-SVD [1], which scales as $O(Kn^2N)$ [5] (assuming that synthesis sparsity $s \propto n$).

II. Experimental Results

We learn a 128×64 transform from the 8×8 (zero mean) non-overlapping patches of the Barbara image [4] with $s = 11$, $p = 20$, $\lambda = \eta = 4 \times 10^5$. The algorithm is initialized with the (vertical) concatenation of the 2D DCT [4] and identity matrices. Fig. 1 shows the objective and sparsification error converging quickly for our algorithm. Moreover, the sparsification error improves significantly (by > 8 dB) over the iterations compared to that of the initial transform. The learnt transform in Fig. 1 exhibits geometric and frequency like structures, and is well-conditioned (condition number of 2.4). The magnitude of WW^T (Fig. 1) indicates mostly small values for the off-diagonals, with the mutual coherence (maximum off-diagonal magnitude in WW^T) being 0.88.

When applied to image denoising [6], adaptive overcomplete transforms provide promising PSNR improvements and speedups compared to overcomplete K-SVD [5]. They also denoise better than the adaptive square transform [4] learnt using Problem (P1), indicating the usefulness of overcompleteness.

References

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.
[2] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, "Noise aware analysis operator learning for approximately cosparse signals," in *Proc. ICASSP*, 2012, pp. 5409–5412.
[3] R. Rubinstein, T. Faktor, and M. Elad, "K-SVD dictionary-learning for the analysis sparse model," in *Proc. ICASSP*, 2012, pp. 5405–5408.
[4] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1072–1086, 2013.
[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.
[6] S. Ravishankar and Y. Bresler, "Learning overcomplete sparsifying transforms for signal processing," in *Proc. ICASSP*, 2013, to appear.