Closed-Form Optimal Updates in Transform Learning

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

While the idea of learning a synthesis [1] or analysis [2], [3] dictionary for sparse signal representation has received recent attention, these formulations are typically non-convex and NP-hard, and the approximate algorithms are still computationally expensive. In this work, we focus instead on the learning of square sparsifying transforms $W \in \mathbb{R}^{n \times n}$, and develop efficient algorithms.

The classical transform model for signal y is $Wy = x + e$, where $W \in \mathbb{R}^{m \times n}$ is a sparsifying transform, x is sparse, and e is a small transform domain residual [4]. The transform model is more general than both the analysis and noisy signal analysis models [4]. Moreover, transform sparse coding is easy and exact, involving thresholding Wy [4]. Importantly, the transform model holds well for natural images.

Given a matrix $Y \in \mathbb{R}^{n \times N}$, whose columns represent training signals, the following formulation learns an orthonormal transform.

$$\text{(P1)} \quad \min_{W,X} \|WY - X\|_F^2 \ s.t. \ W^T W = I_n, \ \|X_i\|_0 \leq s \ \forall \ i$$

Here, I_n is the $n \times n$ identity, and $X \in \mathbb{R}^{n \times N}$ is a matrix, whose columns X_i are the sparse codes of the training signals in Y. The term $\|WY - X\|_F^2$ is called *sparsification error*. However, orthonormality in (P1) is typically restrictive. Hence, (P1) is generalized as follows.

$$\text{(P2)} \quad \min_{W,X} \|WY - X\|_F^2 - \lambda \log |\det W| + \mu \|W\|_F^2 \quad s.t. \ \ \|X_i\|_0 \leq s \ \forall \ i$$

The $\log |\det W|$ and $\|W\|_F^2$ penalties help avoid trivial solutions in learning, and also allow full control over the condition number of W. We have recently proposed [4] an essentially identical formulation, and a computationally cheap alternating algorithm for solving it, which alternates between solving for X (*sparse coding step*) and W (*transform update step*). While the sparse coding step has an exact solution by thresholding, the transform update step was solved using an iterative method such as conjugate gradients (CG) [4]. In this work too, we propose an alternating algorithm (for both (P2) and (P1)), but present exact solutions (global minimizers) for each step.

For both (P1) and (P2), the sparse coding step is identical, and the solution $\hat{X}$ is computed exactly by thresholding WY, and retaining the s coefficients of largest magnitude in each column. In the transform update step of (P1), the optimal solution $\hat{W} = VU^T$, where $U\Sigma V^T$ denotes the full SVD of YX^T. The transform update step of (P2) also has an exact solution. Consider the (e.g., Cholesky) factorization $YY^T + \mu I_n = LL^T$ and let $L^{-1}YX^T$ have a full SVD of $Q\Sigma R^T$. Then, the closed-form solution for the transform update step of (P2) can be shown to be $\hat{W} = \frac{R}{2}\left(\Sigma + \left(\Sigma^2 + 2\lambda I_n\right)^{\frac{1}{2}}\right) Q^T L^{-1}$, where the square root is the positive-definite square root.

Although CG typically works well for the non-convex transform update step of (P2), its convergence to the global minimum of that step is not guaranteed. The closed-form solution for the transform update step not only overcomes this problem, but it also speeds up the computation by a factor of about J over CG, where J is the number of CG steps. The objective functions converge for our algorithms for (P1) and (P2). Empirical evidence suggests that the iterates too converge, and that transform learning is insensitive to initialization.

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

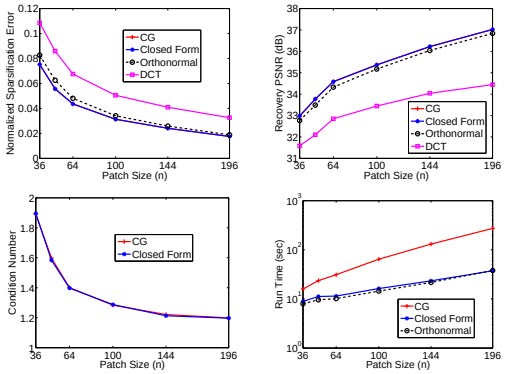

Fig. 1. Comparison of **DCT** with transforms learnt via (P2) by **CG** [4] and by **Closed Form** update, and **Orthonormal** transforms learnt via (P1). The **CG** and **Closed Form** curves overlap in all cases, except for run time.

II. Experimental Results

We learn sparsifying transforms from the $\sqrt{n} \times \sqrt{n}$ (zero mean) non-overlapping patches of the image Barbara [4] at various patch sizes n with $\lambda = \mu = 4 \times 10^5$, and $s = 0.17 \times n$ (rounded). Fig. 1 plots various metrics (cf. [4]) for the transforms learnt using various algorithms, and for the patch-based 2D DCT [4], versus patch size n. The learnt transforms provide better sparsification and recovery (compression) than the analytical DCT. The performance of the CG-based algorithm [4] is identical to the proposed one for (P2). However, the latter is much faster. Moreover, the adapted well-conditioned transforms (learnt via (P2)) provide better performance (upto 0.3 dB better recovery) compared to the adapted orthonormal ones (learnt via (P1)).

In the application to image denoising [5], the transforms learnt via the proposed efficient algorithms provide promising PSNR improvements and speedups (12x) compared to overcomplete K-SVD [6]. Note that the overcomplete K-SVD itself denoises better than the square K-SVD [4]. Qualitative results similar to those observed for image representation (Fig. 1) continue to hold for denoising.

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] ——, "Closed-form solutions within sparsifying transform learning," in *Proc. ICASSP*, 2013, to appear.

[6] 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.