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, USA

OVERVIEW

Problem Statement

- Learning overcomplete sparsifying transforms & applications.

Contributions

- Novel formulation for learning overcomplete transforms.
- Alternating algorithm involves a cheap sparse coding step and an iterative transform update step.
- Computational cost of overcomplete transform learning is lower than that of dictionary learning.
- Adaptive overcomplete transforms can denoise better than overcomplete synthesis K-SVD, while being faster.
- Overcomplete transforms also denoise better than square transforms.

SPARSE MODELS:

❑ **Synthesis Model (SM)**: Given signal y and dictionary $D \in \mathbb{R}^{n \times K}$ we have $y = Dx + e$, with $\|x\|_0 \ll K$, and e is a deviation term.

- Synthesis sparse coding:

$$\hat{x} = argmin_x \ \|y - Dx\|_2^2 \ s.t. \|x\|_0 \leq s$$

- This is NP-hard.
- Greedy and relaxation algorithms are computationally expensive.

❑ **Analysis Model (AM):** Given signal y and analysis dictionary $\Omega \in \mathbb{R}^{m \times n}$, $\|\Omega y\|_0 \ll m$.

❑ **Noisy Analysis Model (NSAM):** $y = q + e$, with Ωq sparse.

- Analysis sparse coding:

$$\hat{q} = argmin_q \ \|y - q\|_2^2 \ s.t. \|\Omega q\|_0 \leq s$$

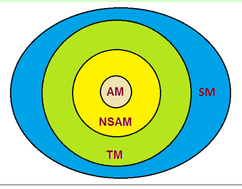
- This is NP-hard too.
- Algorithms - computationally expensive.

❑ **Transform Model (TM):** Given signal y and transform $W \in \mathbb{R}^{m \times n}$, $Wy = x + \eta$, with $\|x\|_0 \ll m$, and η an error term.

- Transform sparse coding:

$$\hat{x} = argmin_x \ \|Wy - x\|_2^2 \ s.t. \|x\|_0 \leq s$$

- $\hat{x}$ computed exactly, cheaply by thresholding Wy to the s largest magnitude elements.



LEARNING SPARSE MODELS:

- Synthesis and analysis learning formulations are typically non-convex and NP-hard, and the algorithms such as K-SVD [1] are computationally expensive.
- Square transform learning [2]

$$(P1) \quad \min_{W,X} \ \overbrace{\|WY - X\|_F^2}^{Sparsification\ Error} + \lambda \overbrace{(\|W\|_F^2 - \log|\det W|)}^{Regularizer}$$
$$s.t. \quad \|X_i\|_0 \leq s \ \forall \ i$$

  - $Y = [\, y_1 | \, y_2 | \, \ldots | \, y_N] \in \mathbb{R}^{n \times N}$: Matrix of training data.
  - X : Matrix of sparse codes.
- Sparsification error measures deviation of data in transform domain from perfect sparsity.
- $\log|\det W|$ enforces full rank, and prevents repeated, or zero rows.
- $\|W\|_F^2$ keeps objective bounded from below.
- Regularizer encourages reduction of condition number, κ.

OVERCOMPLETE TRANSFORM LEARNING:

- Difficulties in extending (P1) to overcomplete case -

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

  - $\log\det(W^T W)$ enforces full column rank of W.
  - However, it cannot prevent repeated rows, or zero rows.
- We propose incoherence penalty to prevent repeated rows.

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

  - w_k denotes a unit norm row of W.
  - $|\langle w_j, w_k \rangle|$ measures coherence (angle) between rows.
  - Problem (P2) is non-convex.
- Choice of p :
  - For $p = 2$, the incoherence penalty is constant when W is a unit norm tight frame, irrespective of the specific choice of W.
  - $p < 2$ encourages sparsity of gram matrix WW^T. However, magnitudes of non-zero inner products need not be small.
  - **Thus, $p \leq 2$ can fail to detect coherence of rows of W.**
  - Larger values of p (> 2) emphasize mutual coherence, and work well in applications.
- Properties of our formulation :
  - As $\lambda \to \infty$ with fixed η in (P2), the optimal $\widehat{W}^T$ approaches a unit norm tight frame, with $\widehat{W}^T \widehat{W} = \frac{m}{n} I_n$.
  - As $\eta, p \to \infty$ with fixed λ in (P2), the optimal $\widehat{W}^T$ approaches a Grassmanian frame.
- Equivalent solutions : given a minimizer $(\widehat{W}, \widehat{X})$, we can form equivalent minimizers by
  - Simultaneously permuting the rows of $\widehat{W}$ and $\widehat{X}$.
  - Pre-multiplying $\widehat{W}$ and $\widehat{X}$ with a diagonal matrix Γ with ± 1 entries.

ALGORITHM:

- Our algorithm alternates between updating X and W.
- Sparse coding step

$$\min_X \ \|WY - X\|_F^2 \ s.t. \quad \|X_i\|_0 \leq s \ \forall \ i$$

  - **Easy Problem:** Exact solution by setting to zero all but the s largest magnitude coefficients in each column of WY.
  - Alternative penalized form of sparse coding :

$$\min_X \ \|WY - X\|_F^2 + \gamma^2 \sum_{i=1}^{N} \|X_i\|_0$$

  - The solution $\hat{X}$ is again computed exactly by thresholding WY with a hard threshold of γ.
- Transform update step

$$\min_W \ \|WY - X\|_F^2 - \lambda \log\det(W^T W) + \eta \sum_{j \neq k} |\langle w_j, w_k \rangle|^p$$
$$s.t. \quad \|w_k\|_2 = 1 \ \forall \ k$$

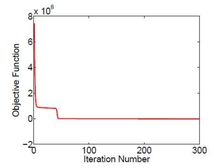
  - This is a non-convex problem with no analytical solution.
  - We could use the iterative projected gradient method, or the projected conjugate gradient (CG) method.
  - However, the alternative strategy of employing the standard CG followed by post-normalization of the rows led to better solutions.

COMPUTATIONAL ADVANTAGES:

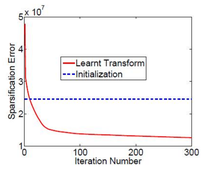
- Computational cost per iteration of proposed algorithm is $O(mnN)$ for N training signals, and $W \in \mathbb{R}^{m \times n}$.
- Cost per iteration of synthesis, or analysis K-SVD is $O(mn^2 N)$ for $D \in \mathbb{R}^{n \times m}$. Cost dominated by sparse coding.
- For images, with $p \times p$ patches, our algorithm provides a reduction of computations in the order by p^2.
- For 3D data, with $p \times p \times p$ patches, the reduction in order is p^3.
- Thus, transform learning may be better suited for big data applications compared to other learning models.

EXAMPLE 1: LEARNING ($s = 11$, $p = 20$)

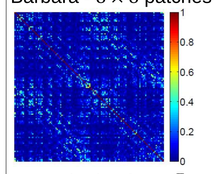
Barbara - 8×8 patches



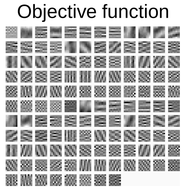
Objective function



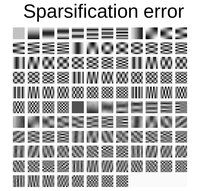
Sparsification error



Magnitude of WW^T
Peak coherence = 0.58



Learnt 128×64 W



Initial W

EXAMPLE 1: LEARNING

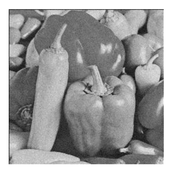
- Initialization: DCT + learnt square transform from (P1).
- Normalized sparsification error (NSE) = $\frac{\|WY - X\|_F^2}{\|WY\|_F^2}$.
  - measures the fraction of energy lost in sparse fitting.
  - NSE for learnt W is 0.07, while that for initialization is 0.13.
- Image recovered from sparse code is 2.5 dB better for learnt W.
- Learnt transform is well-conditioned ($\kappa = 2.6$).
- Atoms of learnt W exhibit geometric and frequency like structures.

APPLICATION: IMAGE DENOISING

$$\min_{W, \{x_j\}, \{\alpha_j\}} \ \overbrace{\sum_{j=1}^{M} \|Wx_j - \alpha_j\|_2^2}^{Sparsification\ Error} + \lambda \overbrace{Q(W)}^{Regularizer} + \tau \overbrace{\sum_{j=1}^{M} \|R_j y - x_j\|_2^2}^{Data\ Fidelity}$$
$$s.t. \quad \|\alpha_j\|_0 \leq s_j \ \forall \ j \qquad (P3)$$

- Goal : estimate image x from noisy measurement $y = x + h$.
- R_j extracts image patch. $R_j y \approx$ noiseless x_j, $Wx_j \approx \alpha_j$.
- Denoised image got by averaging x_j's at their 2D locations.
- (P3) is solved by an alternating scheme, and s_j are found adaptively.

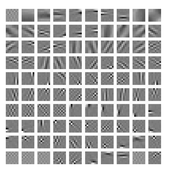
EXAMPLE 2: DENOISING



Noisy Image
PSNR = 28.10 dB

64×256 Synthesis D
PSNR = 34.21 dB

100×64 W ($\kappa = 2.1$)
PSNR = 34.49 dB



Atoms of learnt W

- Overcomplete transform-based denoising is better and 6x faster than overcomplete synthesis denoising.
- Square transform denoises slightly worse (34.36 dB) than overcomplete transform, but is 14x faster than K-SVD.

ACKNOWLEDGEMENT:
Research supported in part by the National Science Foundation under grant CCF 10-18660.

REFERENCES:

[1] Aharon M et al. IEEE Trans Sig Proc 2006; 54: 4311-22.
[2] Ravishankar S et al. IEEE Trans Sig Proc 2013; 61: 1072-86.