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, USA

OVERVIEW

Problem Statement

- Learning square sparsifying transforms & applications

Contributions

- We propose alternating algorithms for learning square transforms: (i) orthonormal, and (ii) unconstrained.
- Proposed algorithms
  - have efficient closed-form solutions for subproblems
  - achieve global minimum of non-convex steps
  - have low computational cost
  - encourage well-conditioning
- Adapted transforms provide better representations than analytical ones.
- In denoising and CS, adaptive square transforms perform comparably or better than learnt overcomplete synthesis dictionaries, but are significantly faster.

SPARSE MODELS:

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

❑ **Noisy Analysis Model:** $y = q + e$, 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:** 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]

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

  - $Y = [\,y_1|\,y_2|\,\ldots|\,y_N] \in \mathbb{R}^{n \times N}$: Matrix of training data.
  - X : Matrix of sparse codes.
- The objective in (P1) is lower bounded.
- Minimizing the regularizer encourages reduction of condition number (κ).
- The solution to (P1) has $\kappa = 1$ as $\lambda \to \infty$.
- The norms of rows of W controlled by controlling κ.

TRANSFORM LEARNING ALGORITHMS:

- Algorithm [2] alternates between updating X and W.
- Sparse coding step

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

  - Solution computed exactly by zeroing out all but the s largest magnitude coefficients in each column of WY.
- Transform update step

$$\min_W \ \|WY - X\|_F^2 + \lambda \ (\|W\|_F^2 - \log|\det W|)$$

  - Previously solved using Conjugate Gradients (CG) [2].
- Drawbacks of CG-based algorithm
  - CG not guaranteed to converge to global minimum of the non-convex transform update problem.
  - CG iterations can be time consuming.

❑ We derive closed-form solutions for transform update

1. Unconstrained W

$$\widehat{W} = \frac{U}{2}\left(\Sigma + (\Sigma^2 + 2\lambda I_n)^{\frac{1}{2}}\right) Q^H L^{-1}$$

  - $YY^H + \lambda I_n = LL^H$, and $L^{-1}YX^H = Q\Sigma U^H$.
  - Solution unique if and only if $L^{-1}YX^H$ has distinct, non-zero singular values.

2. Orthonormal Transform Learning: $W^H W = I_n$ in (P1)

$$\text{(P2)} \quad \min_{W,\,X} \ \|WY - X\|_F^2 \ s.t. \ W^H W = I_n \ , \ \|X_i\|_0 \leq s \ \forall \ i$$

  - Sparse coding step by thresholding.
  - Transform update step

$$\max_W \ Re\{trace(WYX^H)\} \ s.t. \ W^H W = I_n$$

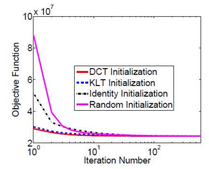
  - $\widehat{W} = VU^H$, where $YX^H = U\Sigma V^H$.

PROPERTIES OF ALGORITHMS:

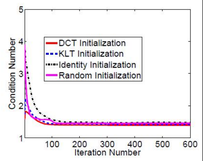
- Closed-form updates achieve global minimum of non-convex alternating optimization steps.
- The objective is monotone decreasing in our exact alternating algorithms. Moreover, since it is lower bounded, it converges.
- Computational cost per iteration of proposed algorithms: $O(Nn^2)$
  - Much lower than cost of K-SVD : $O(Nn^3)$ [2].
- Closed-form solution for transform update provides speedup of about L over CG involving L iterations.

CONVERGENCE WITH VARIOUS INITIALIZATIONS

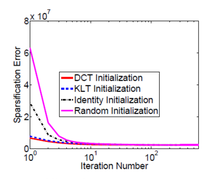

Barbara - 8 × 8 patches



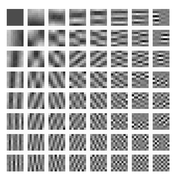
Objective Function



$\kappa(W)$

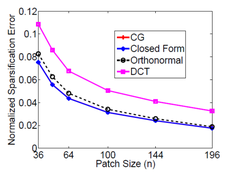


Sparsification Error ($s = 11$)

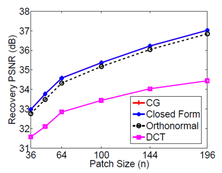

Learnt W – DCT Init

- Normalized sparsification error (NSE) = $\frac{\|WY - X\|_F^2}{\|WY\|_F^2}$.
  - measures the fraction of energy lost in sparse fitting.
  - NSE(W) ≈ 4.4%, NSE(DCT) = 6.8%.
- Recovery PSNR (rPSNR) = $\frac{255\sqrt{\#\ Pixels}}{\|Y - W^{-1}X\|_F}$.
  - Measures error in compressing image using X.
  - rPSNRs for learnt W about 1.7 dB better than DCT.
  - λ allows trade-off between NSE and $\kappa(W)$. rPSNR best at intermediate κ.

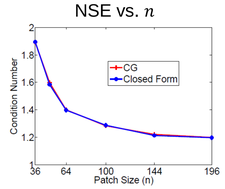
COMPARISON OF LEARNING ALGORITHMS



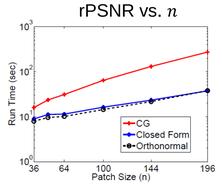
NSE vs. n



rPSNR vs. n



$\kappa(W)$ vs. n



Run time vs. n

APPLICATION: IMAGE DENOISING

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

- 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 efficient alternating scheme that uses closed-form updates, and s_j are found adaptively.

DENOISING EXAMPLE:


Noisy Image
PSNR = 24.60 dB


64×64 W ($\kappa = 1.3$)
PSNR = 31.66 dB


64×256 Synthesis D
PSNR = 31.50 dB

- Closed-form updates-based denoising is better and 17x faster than overcomplete K-SVD denoising.
- Square K-SVD (PSNR = 31.14 dB) denoises worse, slower.
- Our denoising PSNR typically increases with patch size n, while still providing speedups over K-SVD of lower n.
- Denoising gap between non-unitary and unitary adapted transforms typically increases with noise level.

APPLICATION: COMPRESSED SENSING MRI

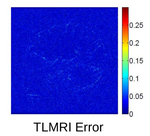
$$\min_{W,\,x,\,\{\alpha_j\}} \ \overbrace{\sum_{j=1}^{M} \|WR_j x - \alpha_j\|_2^2}^{Sparsification\ Error} + \eta^2 \overbrace{\sum_{j=1}^{M} \|\alpha_j\|_0}^{Sparsity} + \nu \overbrace{\|F_u x - y\|_2^2}^{Data\ Fidelity}$$
$$+ \lambda\, Q(W) \qquad \text{(P4)}$$

- Estimate image x from compressive measurements y.
- F_u : undersampled Fourier encoding matrix.
- (P4) solved efficiently by alternating scheme (TLMRI) that uses the proposed closed-form updates.

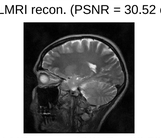
TLMRI EXAMPLE:

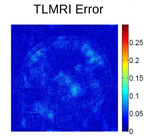
- 2D random sampling with 5x undersampling of k-space.
- TLMRI is better and 12x faster than overcomplete synthesis dictionary-based DLMRI [3].


TLMRI recon. (PSNR = 30.52 dB)



TLMRI Error


DLMRI recon. (PSNR = 28.70 dB)



DLMRI Error

ACKNOWLEDGEMENT:
Research supported in part by the National Science Foundation under grant CCF 10-18660. We thank Prof. M. Lustig, UC Berkeley, for providing the MR data used in our experiments.

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.
[3] Ravishankar S et al. IEEE Trans Med Imag 2011;30: 1028-41.