DOUBLY SPARSE TRANSFORM LEARNING WITH CONVERGENCE GUARANTEES

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

- Data-driven learning of sparsifying transforms, with strong convergence properties.

Contributions

- We propose a novel convex formulation for square sparsifying transform learning. We also introduce a non-convex variant of the convex problem.
- The transform is doubly sparse, which makes its learning, storage, and implementation efficient.
- Proposed algorithms
  - involve efficient closed-form updates
  - encourage well-conditioning
  - guarantee $O(1/k^2)$ convergence to a global optimum in the convex case
  - almost global convergence in the non-convex case
- Adapted transforms provide better image representations than analytical ones.

SPARSE SIGNAL MODELS

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

❑ **Noisy Analysis Model** [1]: $y = q + e$, with Ωq sparse and e a small error term in the *signal domain*.

- Analysis sparse coding:

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

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

❑ **Unstructured Transform Model** [2]: Given signal y and transform $W \in \mathbb{R}^{m \times n}$, $Wy = x + \eta$, with $\|x\|_0 \ll m$, and η a small error term in the *transform domain*.

- Natural signals are approximately sparse in Wavelets or DCT domains.
- Transform model is more general and cheaper than the analysis models.
- 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.
- Least squares signal estimate: $\hat{y} = W^{\dagger}\hat{x}$.

❑ **Doubly Sparse Transform Model** [3]: $W = B\Phi$, with $B \in \mathbb{R}^{n \times n}$ sparse, and Φ a fast analytical sparsifying transform.

- Doubly sparse transforms can be learnt, stored, and implemented efficiently.

PRIOR NON-CONVEX TRANSFORM LEARNING

- Square doubly sparse transform learning [3]

$$(P0) \quad \min_{B,X} \overbrace{\|B\Phi Y - X\|_F^2}^{Sparsification\ Error} + \lambda \overbrace{(\|B\|_F^2 - \log|\det B|)}^{Regularizer}$$
$$s.t. \ \ \|B\|_0 \leq r, \quad \|X_i\|_0 \leq s \ \forall \ i$$

  - $Y = [\, y_1 | \, y_2 | \, \dots | \, y_N] \in \mathbb{R}^{n \times N}$: Matrix of training data.
  - X : Matrix of sparse codes.
- The objective in (P0) is lower bounded.
- Regularizer prevents trivial solutions and controls scaling of B.
- Minimizing the regularizer encourages reduction of condition number of B.
- The solution to (P0) has $\kappa = 1$ as $\lambda \to \infty$.

CONVEX TRANSFORM LEARNING

- We model the sparse $B \in \mathbb{R}^{n \times n}$ as $B = I + A$, with I the identity, and $A = -A^T$ a skew-symmetric matrix.
  - Then, $W = B\Phi = \Phi + A\Phi$ is a perturbed version of (orthonormal) Φ.
  - $B^TB = I - A^2 \approx I$ when A is small $\Rightarrow W = B\Phi$ is well-conditioned.
  - Skew-symmetric structure observed for (P0) [3].

❑ Convex Doubly Sparse Formulation

$$(P1) \quad \min_{A,X} \overbrace{\|(I+A)Z - X\|_F^2}^{Sparsification\ Error} + \frac{\eta}{4} \overbrace{\|A + A^T\|_F^2}^{Skew-symm.} + \overbrace{\mu\|A\|_1 + \xi\|X\|_1}^{Sparsity}$$
$$s.t. \ \ A_{ii} = 0 \ \forall \ i \qquad (Z \triangleq \Phi Y)$$

  - $\|A\|_1$ penalty encourages sparsity and keeps A small.
- (P1) is a linear least squares problem in X and A, with additional ℓ_1 regularizer.
  - First convex transform learning formulation.
  - Quadratic part of cost not strictly convex, since it has linear variety of minimizers $\hat{X} = (I + \hat{A})Z$ with $\hat{A} = -\hat{A}^T$.
  - ℓ_1 regularizer encourages sparsity (as in Lasso).

❑ Alternative Non-Convex Variant of (P1)

$$(P2) \quad \min_{A,X} \|(I+A)Z - X\|_F^2 + \frac{\eta}{4}\|A + A^T\|_F^2 + \mu\|A\|_1$$
$$s.t. \ \ A_{ii} = 0 \ \forall \ i, \ \ \|X_i\|_0 \leq s \ \forall \ i$$

  - (P2) has fewer aspects of non-convexity than (P0).
- (P1) and (P2) generally distinct. Indeed, for fixed A :
  - $X = S_{\xi/2}(Z + AZ) = sign(Z + AZ) \odot (|Z + AZ| - \xi/2)_+$ for (P1).
  - $X = H_s(Z + AZ)$ for (P2), where $H_s(\cdot)$ zeros out all but the s largest magnitude elements in each column of a matrix.

DOSLIST ALGORITHM FOR (P1)

- It is a version of FISTA [4] that uses a block Lipschitz property.
- Let $f(A,X) = \|(I+A)Z - X\|_F^2 + (\eta/4)\|A + A^T\|_F^2$, where A has a zero diagonal. Then, for any A, A', X, X', we have

$$f(A',X') \leq f(A,X) + \langle \nabla_A f(A,X), A' - A \rangle + \frac{L_A}{2}\|A' - A\|_F^2 +$$
$$\langle \nabla_X f(A,X), X' - X \rangle + \frac{L_X}{2}\|X' - X\|_F^2 \qquad (1)$$

  - $\nabla_A f(A,X) = (\mathbf{1} - I) \odot (2(I+A)ZZ^T - 2XZ^T + \eta A + \eta A^T)$, $\nabla_X f(A,X) = 2(X - Z - AZ)$.

DOubly Sparse Learning by Iterative Soft Thresholding (DOSLIST) Algorithm

Input: $Z = \Phi Y$ – Data, L_A, L_X - constants satisfying condition (1), J – number of iterations.
Initialization: $Q^1 = A^0$, $R^1 = S_{\xi/2}(Z + A^0Z) = X^0$, $t_1 = 1$.
For k = 1 to J, Repeat

$$A^k = S_{\frac{\mu}{L_A}}\left(Q^k - \frac{1}{L_A}\nabla_A f(Q^k, R^k)\right)$$
$$X^k = S_{\frac{\xi}{L_X}}\left(R^k - \frac{1}{L_X}\nabla_X f(Q^k, R^k)\right)$$
$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$Q^{k+1} = A^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(A^k - A^{k-1})$$
$$R^{k+1} = X^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(X^k - X^{k-1})$$

End

DOSLAM ALGORITHM FOR (P2)

- The DOubly Sparse Learning by Alternating Minimization (DOSLAM) Algorithm alternates between updating X and A.
- Sparse coding step

$$\min_X \ \|(I+A)Z - X\|_F^2 \ \ s.t. \quad \|X_i\|_0 \leq s \ \forall \ i$$

  - Solution computed exactly as $X = H_s(Z + AZ)$.
- Transform update step

$$\min_{A;\, A_{ii}=0 \ \forall \ i} \ \|(I+A)Z - X\|_F^2 + \frac{\eta}{4}\|A + A^T\|_F^2 + \mu\|A\|_1$$

  - We use ISTA [4] here to obtain strong convergence.

CONVERGENCE OF DOSLIST

- **Global $O(1/k^2)$ Convergence:** *Let $\{A^k, X^k\}$ denote the iterate sequences generated by the DOSLIST algorithm with data Z and initial A^0. Let (A^*, X^*) be any minimizer of (P1). Then, for any $k \geq 1$, we have*

$$F(A^k, X^k) - F(A^*, X^*) \leq \frac{2C}{(k+1)^2} \qquad (2)$$

  - $C \triangleq L_A\|A^0 - A^*\|_F^2 + L_X\|X^0 - X^*\|_F^2$; $F(A,X)$ - the objective.
- **Scale-Invariance:** Set $\eta = \eta_0\sigma_1^2$, $\mu = \mu_0\sigma_1^2$, $\xi = \xi_0\sigma_1$, where σ_1 is the largest singular value of Z. Then
  - The optimal $\hat{A}$ in (P1) for αZ and Z are identical.
  - $L_A = C_1\sigma_1^2$, $L_X = C_2$, where C_1, C_2 are constants.
  - $\{A^k\}$ in DOSLIST is invariant to scaling of Z. Bound (2) is also homogenous to scaling of Z.
  - Note: $\{A^k\}$ in Standard FISTA depends on scale of Z.

CONVERGENCE OF DOSLAM

• By adding the constraint $\|X_i\|_0 \leq s \ \forall \ i$ as a penalty in the cost of (P2) using a barrier function, (P2) becomes unconstrained. We denote the unconstrained objective by $g(A,X)$.

• Let $\beta_j(u)$ denote the magnitude of the j^{th} largest element (magnitude-wise) of vector u.

• **Almost Global Convergence:** *Let $\{A^k, X^k\}$ denote the iterate sequence generated by DOSLAM with data Z and initial (A^0, X^0). Then, $\{g(A^k, X^k)\}$ is monotone decreasing, and converges to a finite value, say g^*. Moreover, every accumulation point (A,X) of the iterate sequence is a fixed point of the algorithm satisfying the following optimality condition.*

$$g(A + \Delta A, X + \Delta X) \geq g(A,X) = g^* \qquad (3)$$

CONVERGENCE OF DOSLAM

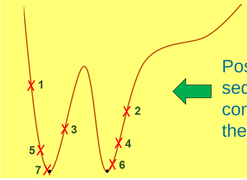
• Condition (3) holds for all $\Delta A \in \mathbb{R}^{n \times n}$ with zero diagonal and all $\Delta X \in \mathbb{R}^{n \times N}$ in the union of the half-space $\langle (I + A)Z - X, \Delta X \rangle \leq 0$ and local region $\|\Delta X\|_\infty \triangleq max_{i,j}|\Delta X_{ji}| < \min_i\{\beta_s(U_i): \|U_i\|_0 > s\}$, where $U = (I + A)Z$. If $\|U_i\|_0 \leq s \ \forall$ i, then ΔX can be arbitrary.

  - *Very large basin of attraction $\{\Delta A, \Delta X\}$.*

• All accumulation points are equivalent – have same cost.

• Thus, objective converges to local minimum and iterates converge to equivalence class of local minimizers.

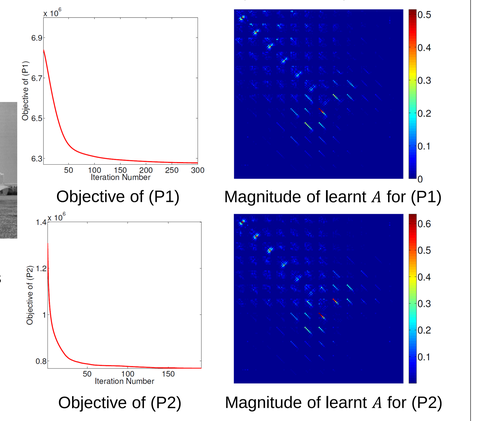
• Convergence irrespective of initialization or # ISTA iterations.



Possible behavior of DOSLAM: The iterate sequence here has two subsequences that converge to two equivalent local minima of the objective.

CONVERGENCE : Φ = 2D DCT, $n = 144$, $s = 24$

Cameraman
12 × 12 patches



Magnitude of learnt A for (P1)

Magnitude of learnt A for (P2)

COMPARISON OF LEARNING ALGORITHMS

- Normalized sparsification error (NSE) = $\|WY - X\|_F^2 / \|WY\|_F^2$.
  - measures the fraction of energy lost in sparse fitting.
- Recovery PSNR (rPSNR) = $255\sqrt{\#\ Pixels}\ /\ \|Y - W^{-1}X'\|_F$.
  - Measures error in recovering image from sparse X'.
  - While $X' = H_s(WY)$ works well, rPSNR improves by setting $supp(X') = supp(H_s(WY))$, and updating the non-zeros of X' to minimize $\|Y - W^{-1}X'\|_F$.
- Our algorithms perform similar to the prior *non-guaranteed* one.

Images	NSE (%)				rPSNR (dB)			
	P1	P2	P0	DCT	P1	P2	P0	DCT
Barbara	3.6	3.0	2.8	4.4	34.99	35.32	35.59	33.76
Brain [5]	0.5	0.4	0.4	0.6	45.32	45.72	45.77	44.12
Cameraman	0.6	0.5	0.4	1.1	43.34	44.19	44.39	40.27
Lena	3.0	2.6	2.4	3.1	37.74	37.97	38.13	37.13

ACKNOWLEDGEMENT

Research supported in part by the National Science Foundation (NSF) under grants CCF-1018660 and CCF-1320953.

REFERENCES

[1] R. Rubinstein and M. Elad, in Proc. SPARS 2011; p. 73.
[2] S. Ravishankar and Y. Bresler, IEEE Trans Sig Proc 2013; 61: 1072-86.
[3] S. Ravishankar and Y. Bresler, IEEE Trans Image Proc 2013; 22: 4598-612.
[4] A. Beck and M. Teboulle, SIAM Journal Imag Sciences 2009; 2: 183-202.
[5] S. Ravishankar and Y. Bresler, IEEE Trans Med Imag 2011; 30: 1028-41.