Regularized Dictionary Learning for Sparse Approximation

Mehrdad Yaghoobi, Thomas Blumensath, Mike Davies

IDCOM, School of Engineering and Electronics
University of Edinburgh, UK

Eusipco08, Aug 26, 2008

Outline

Underdetermined Linear Systems and Sparse Approximation

$$\overbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}}^{\mathbf{y}} = \overbrace{\begin{bmatrix} d_{1,1} & & d_{1,k} & & d_{1,N} \\ d_{2,1} & & d_{2,k} & & d_{2,N} \\ \vdots & \dots & \vdots & \dots & \vdots \\ d_{d,1} & & d_{d,k} & & d_{d,N} \end{bmatrix}}^{\mathbf{D}} \overbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_N \end{bmatrix}}^{\mathbf{x}} + \overbrace{\begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_d \end{bmatrix}}^{\nu}$$

- Noisy sparse approximation: $\min_{\mathbf{x}} ||\mathbf{x}||_0 \ s.t. \ ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 < \epsilon$
- Convex sparse approximation:
$$\min_{\mathbf{x}} ||\mathbf{x}||_1 \ s.t. \ ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 < \epsilon$$
- Unconstrained convex sparse approximation:
$$\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 + \lambda ||\mathbf{x}||_1$$

- Dictionary learning:
  Finding a dictionary such that sparse approximations of the training samples are sparser (with the same representation error) or have less representation error (with the same sparsity) or both.
- Methods are mostly based on block relaxation: iterate between sparse approximation of the learning blocks and the dictionary updates.

$$min_{\{\mathbf{D}\in\mathscr{D},\mathbf{X}\}}\Phi(\mathbf{D},\mathbf{X})\ ;\ \Phi(\mathbf{D},\mathbf{X}) = ||\mathbf{Y}-\mathbf{D}\mathbf{X}||_F^2 + \lambda J_{0,0}(\mathbf{X})$$

$\mathbf{X} = [\mathbf{x}^{(1)}\ \mathbf{x}^{(2)} \ldots \mathbf{x}^{(L)}]$ $\quad$ $\mathscr{D}$ admissible dictionary set

$\mathbf{Y} = [\mathbf{y}^{(1)}\ \mathbf{y}^{(2)} \ldots \mathbf{y}^{(L)}]$ $\quad$ $J_{0,0}(\mathbf{X}) = \#\{x_{i,j} \neq 0\}$

$||\mathbf{A}||_F = (\sum_i \sum_j (a_{i,j})^2)^{\frac{1}{2}}$

Dictionary Learning: Admissible Sets

- Two important classes of dictionaries are:
  - Dictionaries with constrained Frobenius-norm,

$$\mathscr{D} = \{\mathbf{D}_{d \times N} : ||\mathbf{D}||_F \leq c_F^{1/2}\}$$

  - Dictionaries with constrained Column-norm,

$$\mathscr{D} = \{\mathbf{D}_{d \times N} : ||\mathbf{d}_i||_2 \leq c_C^{1/2}\},$$

- The dictionary learning algorithm turns to the following optimization problem,

$$min_{\{\mathbf{D} \in \mathscr{D}, \mathbf{X}\}} \Phi(\mathbf{D}, \mathbf{X}) \ ; \ \Phi(\mathbf{D}, \mathbf{X}) = ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 + \lambda J_{1,1}(\mathbf{X})$$

Majorization Method

Majorize minimization method: replacing the original objective functions with surrogate majorizing objective functions.

Original objective function

$$\min_{\omega \in \Omega} \phi(\omega)$$
$$c \leq \phi(\omega)$$

Majorizing objective function

$$\phi(\omega) \leq \psi(\omega, \xi) \quad \forall \omega, \xi \in \Omega$$
$$\phi(\omega) = \psi(\omega, \omega) \quad \forall \omega \in \Omega$$

Two-step optimization

1- $\omega_{new} = \arg\min_{\omega \in \Omega} \psi(\omega, \xi)$, *fixed* ξ

2- $\xi_{new} = \omega = \arg\min_{\xi \in \Omega} \psi(\omega, \xi)$, *fixed* ω

Sparse Matrix Approximation with the Majorization Method (Iterative Thresholding)

- Unconstrained convex objective function,

$$\min_{\mathbf{X}} \Phi(\mathbf{X}) \quad ; \quad \Phi(\mathbf{X}) = ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 + \lambda J_{1,1}(\mathbf{X})$$

- The majorizing function is made by adding a convex function $\Theta_{\mathbf{D}}(\mathbf{X}, \mathbf{X}^{[n]})$ to $\Phi(\mathbf{X})$.

$$\Psi(\mathbf{X}, \mathbf{X}^{[n]}) = \Phi(\mathbf{X}) + \Theta_{\mathbf{D}}(\mathbf{X}, \mathbf{X}^{[n]})$$

$$\Theta_{\mathbf{D}}(\mathbf{X}, \mathbf{X}^{[n]}) = c_x||\mathbf{X} - \mathbf{X}^{[n]}||_F^2 - ||\mathbf{D}\mathbf{X} - \mathbf{D}\mathbf{X}^{[n]}||_F^2$$

- Optimization based on $\mathbf{X}$: $\mathbf{X}^{[n+1]} = \min_{\mathbf{X}} \Psi(\mathbf{X}, \mathbf{X}^{[n]})$

$$\{\mathbf{X}^{[n+1]}\}_{i,j} = \begin{cases} a_{i,j} - \lambda/2 \; sign(a_{i,j}) & |a_{i,j}| > \lambda/2 \\ 0 & otherwise \end{cases},$$

$$\mathbf{A} := \frac{1}{c_X}(\mathbf{D}^T\mathbf{Y} + (c_x\mathbf{I} - \mathbf{D}^T\mathbf{D})\mathbf{X}^{[n]})$$

- Updating the current coefficient matrix: $\mathbf{X}^{[n+1]} \rightarrow \mathbf{X}^{[n]}$

- By using the Lagrangian multiplier method we get a cost function as follows,

$$\Phi_{\gamma}(\mathbf{D}, \mathbf{X}) = ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 + \lambda J_{1,1}(\mathbf{X}) + \gamma(||\mathbf{D}||_F^2 - c_F)$$

- It is a convex function and its minimum is in a point with zero gradient.

$$\mathbf{D} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I})^{-1}$$

- An appropriate $\gamma \geq 0$ should be selected such that $\mathbf{D}$ is admissible. If $\mathbf{D}|_{\gamma=0}$ is not admissible then it can be found by a line-search method.

- The cost function for this admissible set is,

$$\Phi_{\mathbf{G}}(\mathbf{D},\mathbf{X}) = ||\mathbf{Y}-\mathbf{D}\mathbf{X}||_F^2 + \lambda J_{1,1}(\mathbf{X}) + tr\{\mathbf{G}(\mathbf{D}^T\mathbf{D} - c_C\mathbf{I})\},$$

  where $\mathbf{G}$ is a diagonal matrix with the Lagrangian multipliers on its main diagonal.

- The majorizing function is made by adding a convex function to $\Phi_{\mathbf{G}}(\mathbf{D},\mathbf{X})$.

$$\Psi_{\mathbf{G}}(\mathbf{D},\mathbf{D}^{[n]}) = \Phi_{\mathbf{G}}(\mathbf{D},\mathbf{X}) + \Theta_D(\mathbf{D},\mathbf{D}^{[n]})$$

$$\Theta_D(\mathbf{D},\mathbf{D}^{[n]}) = c_D||\mathbf{D}-\mathbf{D}^{[n]}||_F^2 - ||\mathbf{D}\mathbf{X}-\mathbf{D}^{[n]}\mathbf{X}||_F^2$$

- Optimization based on $\mathbf{D}$: $\mathbf{D}^{[n+1]} = \min_{\mathbf{D}} \Psi_{\mathbf{G}}(\mathbf{D},\mathbf{D}^{[n]})$

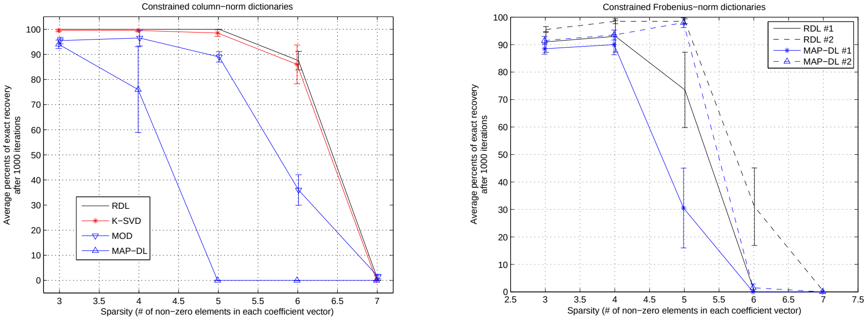
$$\{\mathbf{D}^{[n]}\}_j = \begin{cases} \mathbf{b}_j & ||\mathbf{b}_j||_2 \le c_C^{1/2} \\ \frac{c_C^{1/2}}{||\mathbf{b}_j||_2}\mathbf{b}_j & otherwise \end{cases}, \quad \mathbf{B} := \frac{1}{c_D}(\mathbf{Y}\mathbf{X}^T + \mathbf{D}^{[n]}(c_D\mathbf{I} - \mathbf{X}\mathbf{X}^T))$$

- Updating the current dictionary: $\mathbf{D}^{[n+1]} \rightarrow \mathbf{D}^{[n]}$

Simulations: Synthetic data

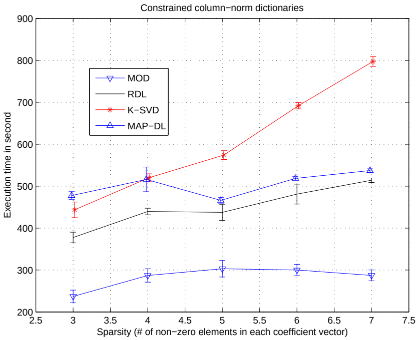
- Synthetic data was generated to test the ability of the algorithms to recover the dictionary exactly.
  - Random $\mathbf{D}_{20X40}$ were generated, followed by normalization of the dictionary to have fixed column or Frobenius-norm.
  - A set of 1280 uniformly random coefficient vectors was generated where the absolute values of the non-zero coefficients were between 0.2 and 1, followed by dividing by the norm of the corresponding atom (for the fixed Frobenius-norm dictionaries).
  - The locations of the non-zero values in the coefficient vectors were selected uniformly at random.
  - An atom was called “recovered“ when the inner-product between (Normalized) original atom and (Normalized) recovered atom was more than 0.99 .

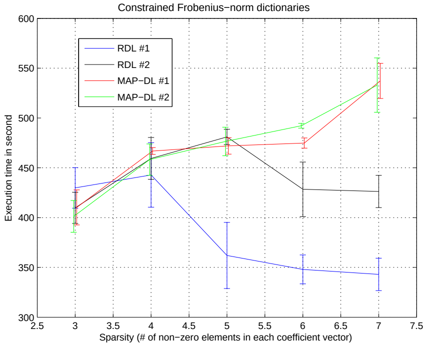
- Starting from random dictionaries
- 1000 iterations of alternating minimizations
- Majorization method was used for the sparse approximation followed by debiasing of the coefficient matrices.
- Results of the average percentages and standard deviations of the atom recovery of 5 trials are shown.



Atom recovery: computation costs

- Simulations ran on the Intel Xeon 2.6 GHz dual core processor machines.

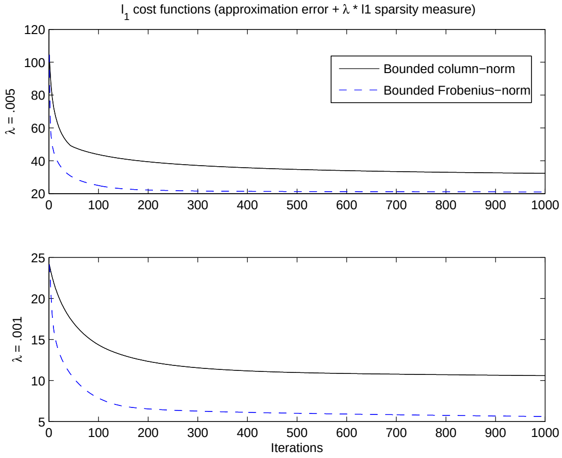




Sparse coding of audio signals

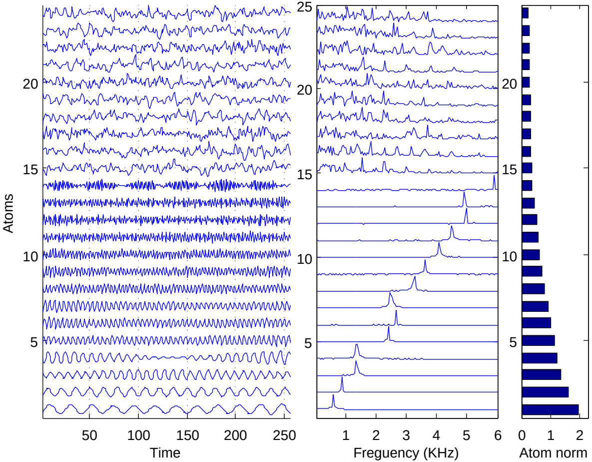
- Audio signals have been shown to have some sparse structures, which consist of sinusoids, transients and a noisy residual component.
- In this simulation audio samples were selected randomly from more than 8 hours recorded from BBC radio 3.
- The recorded 48kHz audio was down-sampled by a factor of 4.
- The window length of 256 was selected for all the simulations.
- Dictionary learning algorithms were started with a 2 times overcomplete dictionary as the initial point.

Constrained column and Frobenius-norm dictionary learning comparison

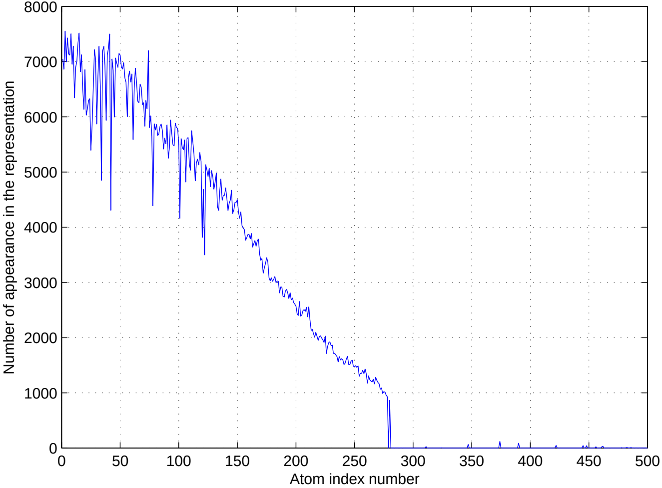


Audio dictionary learning

- Random initialized dictionary.
- Bounded Frobenius-norm dictionary admissible set
- 100,000 iterations of alternative updates.

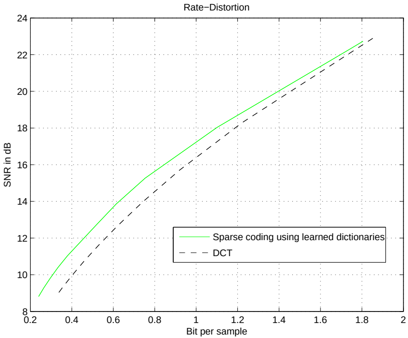


Number of appearances of the learned atoms in the sparse approximations



Entropy coding estimation for coding with the shrunk learned dictionary and DCT

- Starting with 2 times overcomplete DCT
- Ran for 250 iterations of alternative optimizations.
- Different λ were used to find optimal dictionaries for different bitrates
- Convex hull of the R-D curves has been plotted.



Conclusions

- New algorithms were presented for dictionary learning, using two norm constraints, which are fast and their performance are as good as currently available methods.
- The new algorithms can not only apply typically used constraints on the dictionaries, but are also flexible enough to use the corresponding convex relaxed admissible sets.
- In a recent work, we could show that the majorization method for the optimization of the joint objective function converges to a fixed point or gets as close as possible to a connected set of fixed points.
- Using a constraint on the Frobenius-norm of the dictionary, jointly with an ℓ_1 sparsity measure, was found to be a better choice.