Proximal alternating algorithms in dictionary learning

Bao Chenglong
Department of Mathematics
National University of Singapore

Joint work with Yuhui Quan, Hui Ji and Zuowei Shen.
Which kind of data we face?

Text data
Which kind of data we face?

Text data

Financial data
Which kind of data we face?

Text data

Healthy data

Financial data
Which kind of data we face?

- Text data
- Healthy data
- Financial data
- Vision data
Why focus on data representation?

- 4Vs in data science
  a) Volume
  b) Variety
  c) Velocity
  d) Veracity

- Key of the success
  • Efficient and effective data representation

Photo source: IBM
Data representation

• Main difficulty: curse of dimensionality (e.g. Hughes phenomenon)

• Sparsity prior
  • Regularity
  • Temporal and spatial coherence
  • Hierarchical organization of explanatory factors
  • Dictionary learning

  Non-convex models

• Goal: “good” sparse representation
  • Efficient and convergent numerical schemes and effective models

Sparse matrices

Photo source: Gallery of the University of Florida
Sparsity based dictionary learning


- Many variants
  - Redundant dictionary learning [Elad et al, TIP, 2006]
  - Data-driven tight frame construction [Cai et al, ACHA, 2014]
  - Task-driven dictionary learning [Mairal et al, TPAMI, 2012]
  - Hierarchical dictionary learning [Jenatton et al, JMLR, 2011]

- Success in applications
  - Image denoising [Vincent et al, JMLR, 2010]
  - Abnormal event detection [Zhao et al, CVPR 2011]
  - Natural language processing [Bagnell et al, NIPS, 2009]
  - Visual tracking [Mei et al, TPAMI, 2011]
• Data-driven sparse representation
  • Dictionary learning: find an adaptive dictionary $D = (d_1, d_2, \ldots, d_n)$ such that

\[ y = \sum_i c_i d_i, \]

and most $c_i$s are zero.
• Data-driven sparse representation
  • Dictionary learning: find an adaptive dictionary $D = (d_1, d_2, \ldots, d_n)$ such that

$$y = \sum_i c_i d_i,$$

and **most $c_i$s are zero**

How to solve the non-convex model?
The most popular dictionary learning method

- Data $Y = (y_1, y_2 \ldots y_p)$ and $D = (d_1, d_2, \ldots, d_m)$ and $C = (c_1, c_2, \ldots, c_p)$

- **K-SVD** [Elad et al, TIP, 2006]
  
  $\min_{D,C} ||Y - DC||_F^2 \text{ s.t. } ||c_i||_0 \leq k, ||d_j||_2 = 1, \forall i \in [p], \forall j \in [m].$

  - Alternating minimization: OMP and sequential SVD *(No convergence guarantee and high computational complexity)*

- **K-SVD based applications**
  - Image classification [Jiang et al, TPAMI, 2013]
  - Text corpora representation [Jenatton et al, JMLR, 2011]
  - Multi-task learning [Ruvolo et al, ICML, 2014]
  - Recommendation system [Gediminas et al, TKDE, 2012]
The numerical behavior of K-SVD

L2 norm of increments of the sequence generated by K-SVD
The numerical behavior of K-SVD

L2 norm of increments of the sequence generated by K-SVD

Goal: convergent and efficient numerical algorithm.
Mathematical formulation

- General dictionary learning model

\[
\min_{D,C} \sum_{i=1}^{p} f(y_i, Dc_i) + r_1(C) + r_2(D)
\]

- \( f \) is the loss function, e.g. \( \ell_2 \) loss,
- \( r_1, r_2 \) are constraint functions on \( C, D \), respectively.
Mathematical formulation

• General dictionary learning model

$$\min_{D,C} \sum_{i=1}^{p} f(y_i, Dc_i) + f_r(C) + r_2(D)$$

• $f$ is the loss function, e.g. $\ell_2$ loss,
• $r_1, r_2$ are constraint functions on $C, D$, respectively.

• Multi-block non-convex optimization

$$\min_{x=(x_1,\ldots,x_n)} H(x) := f(x) + \sum_{i=1}^{n} r_i(x_i)$$

• Assumptions

1. $\nabla f$ is Lipschitz on any bounded set and $\nabla_i f$ is $L_i$-Lipschitz
2. $r_i, i \in [n]$ are proper and lower semi-continuous
Existing schemes

• Define $f_i^k(\cdot) := f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}, \ldots, x_n^k)$

Alternating minimization

$$x_i^{k+1} \in \arg\min_{x_i} f_i^k(x_i) + r_i(x_i)$$

Proximal alternating minimization

$$x_i^{k+1} \in \arg\min_{x_i} f_i^k(x_i) + r_i(x_i) + \lambda_i \|x_i - x_i^k\|^2_F$$

Proximal alternating linearized minimization

$$x_i^{k+1} \in \arg\min_{x_i} \langle \nabla f_i^k(x_i^k), x_i - x_i^k \rangle + r_i(x_i) + \lambda_i \|x_i - x_i^k\|^2_F$$
Literature review

• Alternating minimization (AM)
  • No convergence guarantee

• Proximal alternating minimization (PAM)
  • Global convergence property [Attouch et al, MOR, 2010]

• Proximal alternating linearized minimization (PALM)
  • Global convergence property [Bolte et al, MP, 2014]

• Hybrid method
  • Multi-convex case [Xu et al, SIIMS, 2013]
Literature review

• Alternating minimization (AM)
  • No convergence guarantee

• Proximal alternating minimization (PAM)
  • Global convergence property [Attouch et al, MOR, 2010]

• Proximal alternating linearized minimization (PALM)
  • Global convergence property [Bolte et al, MP, 2014]

• Hybrid method
  • Multi-convex case [Xu et al, SIIMS, 2013]

Inner iter. No. : AM \approx PAM > PALM
Outer iter. No. : AM < PAM < PALM
Literature review

• Alternating minimization (AM)
  • No convergence guarantee

• Proximal alternating minimization (PAM)
  • Global convergence property [Attouch et al, MOR, 2010]

• Proximal alternating linearized minimization (PALM)
  • Global convergence property [Bolte et al, MP, 2014]

• Hybrid method
  • Multi-convex case [Xu et al, SIIMS, 2013]

Inner iter. No. : $AM \approx PAM > PALM$
Outer iter. No. : $AM < PAM < PALM$  Hybrid them for non-convex case.
Hybrid proximal alternating method

- HPAM [Bao et al, TPAMI, 2016]
  For $k = 1, 2, \ldots$, for $i = 1, 2, \ldots, n$

\[
x_{i}^{k+1} \in \begin{cases} 
\arg\min f_{i}^{k}(x_{i}) + r_{i}(x_{i}) + \lambda_{i} \|x_{i} - x_{i}^{k}\|_{F}^{2} \quad \text{or} \\
\arg\min \left\langle \nabla f_{i}^{k}(x_{i}^{k}), x_{i} - x_{i}^{k} \right\rangle + r_{i}(x_{i}) + \lambda_{i} \|x_{i} - x_{i}^{k}\|_{F}^{2} 
\end{cases}
\]

where $\lambda_{i}$ is the appropriate step size.
Hybrid proximal alternating method

- HPAM [Bao et al., TPAMI, 2016]
  
  For $k = 1, 2, \ldots$, for $i = 1, 2, \ldots, n$

  \[
x_{i}^{k+1} \in \begin{cases}
  \arg\min f_{i}^{k}(x_{i}) + r_{i}(x_{i}) + \lambda_{i}||x_{i} - x_{i}^{k}||_{F}^{2} \text{ or } \\
  \arg\min \langle \nabla f_{i}^{k}(x_{i}^{k}), x_{i} - x_{i}^{k} \rangle + r_{i}(x_{i}) + \lambda_{i}||x_{i} - x_{i}^{k}||_{F}^{2}
  \end{cases}
\]

  where $\lambda_{i}$ is the appropriate step size.

**Theorem (Global convergence):** Let \( \{x^{k}\} \) be the infinite sequence generated by the HPAM. If \( H \) is a KL function and \( \{x^{k}\} \) is bounded. Then, the sequence \( \{x^{k}\} \) converges to a point \( x^{*} \), which is a critical point of \( H \), i.e. \( 0 \in \partial H(x^{*}) \).

Remark: AM can be included under certain condition.
Sketch of the proof

• The proof is based on the result from [Attouch et al, MP, 2013].

• Four main steps:
  1. Sufficient decrease property

\[ H(x^k) - H(x^{k+1}) \geq \rho_1 ||x^k - x^{k+1}||^2_F, \text{ for some } \rho_1 > 0 \]

  2. Bounded the subgradient:

\[ \text{dist} \left( 0, \partial H(x^k) \right) \leq \rho_2 ||x^k - x^{k-1}||, \text{ for some } \rho_2 > 0 \]

  3. Subsequence continuity: \( \exists \) a subsequence \( \{x^{k_j}\} \) such that

\[ x^{k_j} \to \bar{x}, \quad \text{and} \quad H(x^{k_j}) \to H(\bar{x}). \]

  4. \( H \) satisfies Kurdyka-Lojasiewicz (KL) property.
HPAM in dictionary learning

• Plain dictionary learning model

\[
\min_{D, C} \|Y - DC^T\|_F^2 + \lambda \|C\|_0, \text{ s.t. } \|C\|_\infty \leq M, \|d_j\|_2 = 1, \forall j \in [m].
\]

• Numerical schemes

- **C1** \((x_1, x_2, ..., x_n) = (c_1, c_2, ..., c_q, d_1, d_2, ..., d_q)\)
- **C2** \((x_1, x_2, ..., x_n) = (C, D)\)
- **C3** \((x_1, x_2, ..., x_n) = (C, d_1, c_1, d_2, c_2, ..., d_q, c_q)\)
HPAM in dictionary learning

• Plain dictionary learning model

\[
\min_{D,C} ||Y - DC^T||_F^2 + \lambda ||C||_0, \text{ s.t. } ||C||_\infty \leq M, ||d_j||_2 = 1, \forall j \in [m].
\]

• Numerical schemes

  • C1 \((x_1, x_2, ..., x_n) = (c_1, c_2, ..., c_q, d_1, d_2, ..., d_q)\)
  • C2 \((x_1, x_2, ..., x_n) = (C, D)\)
  • C3 \((x_1, x_2, ..., x_n) = (C, d_1, c_1, d_2, c_2, ..., d_q, c_q)\)

<table>
<thead>
<tr>
<th>SCHEMES</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block choice</td>
<td>C1</td>
<td>C2</td>
<td>C3</td>
</tr>
<tr>
<td>Step I</td>
<td>PAM</td>
<td>PALM</td>
<td>PALM</td>
</tr>
<tr>
<td>Step II</td>
<td>PAM</td>
<td>PALM</td>
<td>PAM+AM</td>
</tr>
</tbody>
</table>
HPAM in dictionary learning

• Plain dictionary learning model

$$\min_{D, C} ||Y - DC^T||_F^2 + \lambda ||C||_0, \text{ s.t. } ||C||_\infty \leq M, ||d_j||_2 = 1, \forall j \in [m].$$

• Numerical schemes
  
  • C1 \((x_1, x_2, \ldots, x_n) = (c_1, c_2, \ldots, c_q, d_1, d_2, \ldots, d_q)\)
  
  • C2 \((x_1, x_2, \ldots, x_n) = (C, D)\)
  
  • C3 \((x_1, x_2, \ldots, x_n) = (C, d_1, c_1, d_2, c_2 \ldots, d_q, c_q)\)

<table>
<thead>
<tr>
<th>SCHEMES</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block choice</td>
<td>C1</td>
<td>C2</td>
<td>C3</td>
</tr>
<tr>
<td>Step I</td>
<td>PAM</td>
<td>PALM</td>
<td>PALM</td>
</tr>
<tr>
<td>Step II</td>
<td>PAM</td>
<td>PALM</td>
<td>PAM+AM</td>
</tr>
</tbody>
</table>

All the above schemes converge to a critical point.
Scheme 1 (S1)

Block choice: $C_1 (x_1, x_2, ..., x_n) = (c_1, c_2, ..., c_q, d_1, d_2, ..., d_q)$

Step 1. Sparse coding: for $i = 1, ..., m$,

$$c^k_i \in \arg\min_{|c|_\infty \leq M} \lambda ||c||_0 + ||J_i^k - d_i^k c||_F^2 + \lambda_c ||c - c_i^{k-1}||_F^2$$

where $J_i^k = Y - \sum_{j<i} d_j^k c_j^{k\top} + \sum_{j>i} d_j^k c_j^{k-1\top}$.

Step 2. Dictionary update: for $i = 1, ..., m$,

$$d_i^{k+1} \in \arg\min_{d} ||E_i^k - d c_i^{k\top}||_F^2 + \lambda_d ||d - d_i^k||_F^2 \text{ s.t. } ||d|| = 1$$

where $E_i^k = Y - \sum_{j<i} d_j^{k+1} c_j^{k\top} + \sum_{j>i} d_j^k c_j^{k\top}$.
Scheme 2 (S2)

Block choice: C2 \((x_1, x_2, \ldots, x_n) = (C, d_1, d_2, \ldots, d_q)\)

**Step 1. Sparse coding:**

\[
C^k \in \arg\min_{\|C\|_\infty \leq M} \lambda \|C\|_0 + \langle \nabla_C f(D^k, C^{k-1}), C - C^{k-1} \rangle + \lambda_C \|C - C^{k-1}\|_F^2
\]

**Step 2. Dictionary update:** for \(i = 1, \ldots, m\),

\[
d_i^{k+1} \in \arg\min_{d} \|E_i^k - d c_i^{kT}\|_F^2 + \lambda_d \|d - d_i^k\|_F^2 \text{ s.t. } \|d\| = 1
\]

where \(E_i^k = Y - \sum_{j<i} d_j^{k+1} c_j^{kT} + \sum_{j>i} d_j^k c_j^{kT} \).
Scheme 3 (S3)

Step 1. Sparse coding:

\[ C^k \in \text{argmin}_{\|C\|_\infty \leq M} \lambda \|C\|_0 + \langle \nabla_C f(D^k, C^{k-1}), C - C^{k-1} \rangle + \lambda_C \|C - C^{k-1}\|_F^2 \]

Step 2. Dictionary update: for \( i = 1, \ldots, m \),

1. Update the \( d_i \)

\[ d_i^{k+1} \in \text{argmin}_d \|E_i^k - d c_i^{k\top}\|_F^2 + \lambda_d \|d - d_i^k\|_F^2 \text{ s.t. } \|d\| = 1 \]

where \( E_i^k = Y - \sum_{j<i} d_j^{k+1} c_j^{k\top} + \sum_{j>i} d_j^k c_j^{k\top} \).

2. Reupdate the non-zero coefficients \( c_i \)

\[ c_i^{k+1} \in \text{argmin}_c \|E_i^k - d_i^{k+1} c\|_F^2 \text{ s.t. } c(j) = 0, \forall j \in I_j \]

where \( I_j = \{q: c_i^k(q) = 0\} \).

Further decrease the residual.
Numerical behavior

Increments of the sequence \( \{C^k\} \)
Numerical behavior

Objective value versus iteration

Different convergent algorithms achieve different critical points in DL.
Computational efficiency

Training time versus the atom dimensions

<table>
<thead>
<tr>
<th>Atom Dim.</th>
<th>6*6</th>
<th>8*8</th>
<th>10*10</th>
<th>12*12</th>
<th>14*14</th>
<th>16*16</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-SVD</td>
<td>39</td>
<td>70</td>
<td>114</td>
<td>164</td>
<td>228</td>
<td>308</td>
</tr>
<tr>
<td>S1</td>
<td>71</td>
<td>217</td>
<td>465</td>
<td>1011</td>
<td>1848</td>
<td>3094</td>
</tr>
<tr>
<td>S2</td>
<td>9</td>
<td>16</td>
<td>28</td>
<td>42</td>
<td>60</td>
<td>86</td>
</tr>
<tr>
<td>S3</td>
<td>10</td>
<td>18</td>
<td>30</td>
<td>45</td>
<td>66</td>
<td>96</td>
</tr>
</tbody>
</table>

Efficiency: S2 ≈ S3 > K-SVD > S1
Image denoising

<table>
<thead>
<tr>
<th>Image</th>
<th>Fingerprint</th>
<th>Lena</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise 5</td>
<td>36.59</td>
<td>38.59</td>
</tr>
<tr>
<td>K-SVD</td>
<td>32.39</td>
<td>35.47</td>
</tr>
<tr>
<td>S1</td>
<td>30.06</td>
<td>33.70</td>
</tr>
<tr>
<td>S2</td>
<td>28.47</td>
<td>32.38</td>
</tr>
<tr>
<td>S3</td>
<td>28.24</td>
<td>32.21</td>
</tr>
</tbody>
</table>

Performance: $S3 \approx S1 \approx K$-SVD $>$ S2
## Image denoising

<table>
<thead>
<tr>
<th>Image</th>
<th>Fingerprint</th>
<th>Lena</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noise 5 10 15 20</td>
<td>Noise 5 10 15 20</td>
</tr>
<tr>
<td>K-SVD</td>
<td>36.59 32.39 30.06 28.47</td>
<td>38.59 35.47 33.70 32.38</td>
</tr>
<tr>
<td>S1</td>
<td>36.58 32.27 29.87 28.24</td>
<td>38.48 35.37 33.55 32.21</td>
</tr>
<tr>
<td>S2</td>
<td>36.50 32.21 29.84 28.18</td>
<td>38.46 35.35 33.50 32.15</td>
</tr>
<tr>
<td>S3</td>
<td>36.59 32.35 30.03 28.44</td>
<td>38.49 35.41 33.57 32.25</td>
</tr>
</tbody>
</table>

Performance: $S3 \approx S1 \approx K\text{-SVD} > S2$

S3 is the most appropriate scheme.
Nonlinear dictionary learning

• Capturing the nonlinear data patterns

• Equiangular kernel learning [Quan et al, CVPR, 2016]

\[
\min_{D, C} \| \Phi(Y) - \Phi(D)C \|_F^2,
\]

s. t. \( \| c_i \|_0 \leq k, \forall i, D^\top D = I \)

where \( \Phi \) is a map associated with the kernel \( K \).
Nonlinear dictionary learning

• Capturing the nonlinear data patterns

• Equiangular kernel learning [Quan et al, CVPR, 2016]

\[
\min_{D,C} \|\Phi(Y) - \Phi(D)C\|_F^2, \quad \text{s.t.} \|c_i\|_0 \leq k, \forall i, D^\top D = I
\]

where \(\Phi\) is a map associated with the kernel \(K\).

• Properties:
  1. \(\langle \Phi(d_m), \Phi(d_n) \rangle = \mu_0, \forall m \neq n\) if \(K = \psi(||d_m - d_n||^2)\).
  2. More scalable than kernel K-SVD (free of \(K(Y,Y)\))
Dynamic texture classification

- Recognizing the moving textures with certain stationary temporal changes
- Nonlinear data patterns

Fountain  Flag  Candle
Experiments

- **UCLA-DT**
  - 50 categories, 200 videos
- **DynTex**
  - 10 categories, 275 videos
- **DynTex++**
  - 36 categories, 3600 videos

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DFS</th>
<th>DFS+</th>
<th>LBP-TOP</th>
<th>KGDL</th>
<th>OTF</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCLA</td>
<td>97.5</td>
<td>97.5</td>
<td>N.A.</td>
<td>N.A.</td>
<td>97.2</td>
<td>98.6</td>
</tr>
<tr>
<td>DynTex</td>
<td>74.5</td>
<td>74.8</td>
<td>72.0</td>
<td>75.1</td>
<td>73.5</td>
<td>75.6</td>
</tr>
<tr>
<td>DynTex++</td>
<td>89.9</td>
<td>91.7</td>
<td>89.2</td>
<td>92.8</td>
<td>89.8</td>
<td>93.4</td>
</tr>
</tbody>
</table>
Discussion

• The hybrid proximal alternating scheme has certain convergence guarantee for multi-block nonconvex problems

• Application based alternating scheme can lead to more efficient and effective algorithm

Future work:
1. Find better initialization
2. Connection between hierarchical dictionary learning and convolutional neural networks
Thank you!