Efficient numerical methods for $\ell_0$-norm related minimizations

Chenglong Bao

Department of mathematics
National University of Singapore

December 6, 2017

Workshop on spline approximation and its applications on Carl de Boor’s 80th birthday, IMS, NUS
Overview

1. Background
   - Wavelet tight frame
   - Sparsity based image restoration

2. Numerical methods for \( \ell_0 \) norm based balanced models
   - \( \ell_0 \) norm based balanced model
   - Data-driven balanced model
   - An example in dynamic texture classification

3. Conclusion
Wavelet tight frame

- Let $\phi$ be a B-spline function
- Denote $h_0[k]$ and $r$ wavelet masks $\{h_\ell[k]\}_{\ell=1}^r$ satisfy

\[
\phi = 2 \sum_k h_\ell[k] \phi(2 \cdot -k), \quad \psi_\ell = 2 \sum_k h_\ell[k] \phi(2 \cdot -k), \quad 1 \leq \ell \leq r
\]

- Define $H_\ell(v) = \downarrow (\sqrt{2} \cdot h_\ell(\cdot) \ast v)$ and $H_\ell^* = \sqrt{2} h_\ell \ast (\uparrow v)$

\[
\{h_\ell[k]\}_{\ell=0}^r \text{ satisfies the UEP}^1 \implies \sum_{\ell=0}^r H_\ell^* H_\ell = \text{Id.}
\]

- **Example:** $\phi$ is piecewise linear B-spline function

\[
h_0 = \frac{1}{4}[1, 2, 1], \quad h_1 = \frac{1}{4}[\,-1, 2, -1], \quad h_2 = \frac{\sqrt{2}}{4}[1, 0, -1].
\]

\(^1\text{Ron and Shen, 1997.}\)
Sparse representation of images

Let $W_L = H_0(\cdot)$ and $W_H = (H_1(\cdot); H_2(\cdot); \ldots; H_r(\cdot))$.

- The perfect reconstruction property

$$x = W_L^\top W_L x + W_H^\top W_H x = W^\top W x,$$

where $W = (W_L^\top, W_H^\top)^\top$.

- $W_H(x)$ is sparse
The $\ell_0$-norm related minimization

$$\min_c \frac{1}{2}\|y - AW^T c\|_2^2 + \frac{\kappa}{2}\|(Id - WW^T)c\|_2^2 + \|\lambda \cdot c\|_0$$

where $A$ is the corruption operator, $c = [c_L, c_H]$ and $\lambda = [0, \lambda]$.

- $\kappa = 0$, it is synthesis model; $\kappa = +\infty$, it is analysis model.
- **Special case:** $A = Id$ and $\kappa = 1$

\[c_L^* = W_L y \text{ and } c_H^* = T_{\sqrt{2\lambda}}(W_H y).\]

where $T_\lambda(x) = x$ if $|x| \geq \lambda$, and $T_\lambda(x) = 0$ otherwise.

- NP-hard in general.
Related work

- Convex relaxation.
  - Substitute $\| \cdot \|_0$ with $\| \cdot \|_1$ in (1).
  - Many numerical schemes: (accelerated) proximal gradient, primal-dual descent, coordinate descent, etc.
  - Convergence analysis is established
  - Bias estimation in large coefficients and loss of contrast.

- Nonconvex or $\ell_0$ norm minimization.
  - Better image recovery results.
  - Nonconvex relaxations: MCP, SCAD, $\ell_p$, $p \in (0, 1)$
  - Numerical schemes: adaptive penalty decomposition, proximal gradient descent, mean doubly augmented Lagrangian, inertial proximal algorithm.
  - The convergence analysis for solving model (1) is not clear.

Goal: design efficient algorithms for solving (1) with convergence guarantee.
Numerical scheme

The balance model (1) is equivalent to

\[
\min_{x} \quad G(x) := F(x) + \|\lambda \cdot x\|_0,
\]

where

\[
F(x) = \frac{1}{2}\|y - AW^\top x\|_2^2 + \frac{\kappa}{2}\|(Id - WW^\top)x\|_2^2.
\]

Define

\[
S_\tau(x, y) = F(y) + \langle \nabla F(y), x - y \rangle + \frac{\tau}{2}\|x - y\|_2^2 + \|\lambda \cdot x\|_0.
\]

**Algorithm 1 (EPIHT method)**

Given \( x_k, x_{k-1} \) and \( w_k \in (0, w] \) where \( w < 1 \).

1. \( y_{k+1} = x_k + w_k(x_k - x_{k-1}) \) \hspace{2cm} (S1)
2. \( z_{k+1} = \arg \min \{ G(y_{k+1}), G(x_k) \} \) \hspace{2cm} (S2)
3. \( x_{k+1} = \arg \min_x S_\tau(x, z_{k+1}) = \sqrt{\frac{2\lambda}{\tau}}(z_{k+1} - \nabla F(z_{k+1})/\tau) \) \hspace{2cm} (S3)

The step size \( \tau > \|\nabla^2 F\|_2^2 \) which might be very large.
Algorithm 2 (Line search for (S3))

Choose $\sigma > 0$ and $\tau_k^0 \in [\tau_{\text{min}}, \tau_{\text{max}}]$ and $\eta \in (0, 1)$

- **Find the smallest integer** $i_k \in \mathbb{N}$ **such that** with $\hat{\tau} = \frac{\tau_0^k}{\eta^i_k}$

$$H(\hat{x}) \leq S_{\hat{\tau}}(\hat{x}, z_{k+1}) - \frac{\sigma}{2} \| \hat{x} - z_{k+1} \|^2$$

where $\hat{x} = \mathcal{T} \sqrt{2\lambda/\hat{\tau}}(z_{k+1} - \nabla F(z_{k+1})/\hat{\tau})$.

- **Set** $\tau_k = \frac{\tau_0^k}{\eta^i_k}$ **and**

$$x_{k+1} \in \arg \min_x S_{\tau_k}(x, z_{k+1}).$$

1. The step size $\tau_0^k$ is initialized by Barzilai-Borwein (BB) method.
2. The choice of $\tau_{\text{max}}$ is bounded by $L + \sigma$ where $L = \| \nabla^2(F) \|_2$. 


Convergence results

Theorem 1

Let \( \{x_k\}_{k=1}^{+\infty} \) to be the sequence generated by (S1)-(S3). If \( \tau \) is \( L + \sigma \) or chosen by Alg.2.

1. If \( \{x_k\}_{k=1}^{+\infty} \) is bounded, then there exists some \( \bar{x} \) such that

\[
x_k \to \bar{x} \text{ as } k \to +\infty.
\]

and \( \bar{x} \) is a local minimizer of \( G(x) \).

2. If \( F \) is strongly convex, then \( x_k \) converges to \( \bar{x} \) linearly.

Remark:

- The boundedness \( \{x_k\} \) holds if \( F \) is coercive.
- \( \text{Ker}(W^\top) \cap \text{Ker}(A) \neq \emptyset \Rightarrow \) strong convexity of \( F \).
- Add \( t\|x\|^2/2 \) to \( F \) ensures the strong convexity of \( F \).
Experiments

- Compressed sensing

\[
\min_x \frac{1}{2} \| y - Ax \|_2^2 + \lambda \| x \|_0
\]

where \( A \in \mathbb{R}^{m \times n} \).

- The number of samples is \( m = 500 \).

- The signal dimension \( n \) various and the sparsity level is \( n/100 \).
Image restoration

- CT image reconstruction
  - Recovery of Shepp-Logan phantom from different number of projections.

<table>
<thead>
<tr>
<th>No. of Proj.</th>
<th>Split Bregman</th>
<th>FISTA</th>
<th>PIHT</th>
<th>Ours-LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>22.98</td>
<td>27.77</td>
<td>28.96</td>
<td>29.52</td>
</tr>
<tr>
<td>50</td>
<td>26.17</td>
<td>32.34</td>
<td>35.15</td>
<td>35.88</td>
</tr>
</tbody>
</table>

**Table:** The PSNR value of the recovered image.

- Image deblur
  - Gaussian kernel with size 9 and variance 1.5 + Gaussian noise with std. 3.

<table>
<thead>
<tr>
<th>Images</th>
<th>Split Bregman</th>
<th>FISTA</th>
<th>Ours-LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>House</td>
<td>32.37 (0.8593)</td>
<td>32.31 (0.8574)</td>
<td><strong>32.45 (0.8638)</strong></td>
</tr>
<tr>
<td>Lena</td>
<td>32.71 (0.9572)</td>
<td>32.81 (0.9584)</td>
<td><strong>33.15 (0.9618)</strong></td>
</tr>
<tr>
<td>Clock</td>
<td>29.49 (0.9159)</td>
<td>29.47 (0.9102)</td>
<td><strong>29.91 (0.9229)</strong></td>
</tr>
</tbody>
</table>

**Table:** The PSNR(SSIM) of the recovered image.
Data-driven balanced model

- Data-driven tight frame construction (Cai et al. ACHA, 2014)

$$\min_{C \in \mathbb{R}^{m \times p}, D \in \mathbb{R}^{m \times m}} \| Y - DC \|_F^2 + \lambda \| C \|_0, \text{ s.t. } DD^\top = I$$

where $Y = (y_1, y_2, \ldots, y_p) \in \mathbb{R}^{m \times p}$ is the collection of all $\sqrt{m} \times \sqrt{m}$ patches and $D = (a_1, a_2, \ldots, a_m) \in \mathbb{R}^{m \times m}$.

- $W(a_1, a_2, \ldots, a_m)$ forms a tight frame for $\ell_2(\mathbb{Z})$. 

<table>
<thead>
<tr>
<th>Image</th>
<th>Haar filters</th>
<th>Linear filters</th>
<th>Learned filters</th>
</tr>
</thead>
</table>

"Redundant dictionary learning"
Data-driven balanced model

- Data-driven tight frame construction (Cai et al. ACHA, 2014)

\[
\begin{align*}
\min_{C \in \mathbb{R}^{m \times p}, D \in \mathbb{R}^{m \times m}} & \quad \| Y - DC \|_F^2 + \lambda \| C \|_0, \quad \text{s.t.} \quad DD^\top = I \\
\text{where} \quad Y = (y_1, y_2, \ldots, y_p) \in \mathbb{R}^{m \times p} \text{ is the collection of all } \sqrt{m} \times \sqrt{m} \text{ patches and } D = (a_1, a_2, \ldots, a_m) \in \mathbb{R}^{m \times m}.
\end{align*}
\]

- \(W(a_1, a_2, \ldots, a_m)\) forms a tight frame for \(\ell_2(\mathbb{Z})\).

Redundant dictionary learning

\[
\begin{align*}
\min_{C \in \mathbb{R}^{q \times p}, D \in \mathbb{R}^{m \times q}} & \quad \| Y - DC \|_F^2 + \lambda \| C \|_0, \quad \text{s.t.} \quad \| d_i \|_2 = 1, \quad i = 1, 2, \ldots, q.
\end{align*}
\]
Classical methods

- Alternating minimization: OMP and sequential SVD.
- No convergence guarantee and high computational complexity.
- K-SVD does not converge numerically.

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

- Alternating minimization: OMP and sequential SVD.
- No convergence guarantee and high computational complexity.
- K-SVD does not converge numerically.

Goal: design efficient and convergent numerical algorithm for solving dictionary learning models.
Mathematical formulation

Multi-block non-convex optimization

\[
\min_{x=(x_1,x_2,\ldots,x_n)} H(x) = F(x) + \sum_{i=1}^{n} r_i(x_i).
\]

Assumption:

- \( F(x) \) is smooth and \( \nabla F \) has Lipschitz constant \( L > 0 \), i.e.

\[
\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|
\]

- \( r_i, i = 1, 2, \ldots, n \) is proper, lower semi-continuous

An example in dictionary learning:

\[
F(D, C) = \|Y - DC\|_F^2, \quad r_1(C) = \lambda \|C\|_0, \quad r_2 = \delta_D(D)
\]

where \( D = \{D||d_i|| = 1, i = 1, 2, \ldots, q\} \).
Related work

- Alternating minimization (AM) might not converge
  - Counter example constructed by M. Powell (Math. Program. 1973)

- Methods with global convergence property
  - The generated sequence converges to a critical point
  - Proximal alternating linearized method (PALM) (Bolte et al, Math. Program., 2014)

- Hybrid method for multi-convex function (Xu et al, SIAM J. Imaging Sci., 2013)
- Block stochastic gradient method (Xu et al, SIAM J. Optim., 2015)
Related work

- Alternating minimization (AM) might not converge
  - Counter example constructed by M. Powell (Math. Program. 1973)

- Methods with global convergence property
  - The generated sequence converges to a critical point
  - Proximal alternating linearized method (PALM) (Bolte et al, Math. Program., 2014)

- Hybrid method for multi-convex function (Xu et al, SIAM J. Imaging Sci., 2013)
- Block stochastic gradient method (Xu et al, SIAM J. Optim., 2015)

Hybrid them for solving non-convex problems.
Define

\[ \tilde{F}(x, y) = F(y) + \langle F(x), x - y \rangle, \]

the linearization of \( F(x) \) at the point \( y \).

**HPAM iterates:**

\[ x_i^{k+1} : \in \begin{cases} 
\arg \min_{x_i} r_i(x_i) + F(x_{<i}^{k+1}, x_i, x_{>i}^k) + \frac{1}{2} \| x_i - x_i^k \|^2_{T_i^k}, \text{(PAM)} \text{ or,} \\
\arg \min_{x_i} r_i(x_i) + \tilde{F}(x_{<i}^{k+1}, x_i, x_{>i}^k) + \frac{1}{2} \| x_i - x_i^k \|^2_{T_i^k}, \text{(PALM)}
\end{cases} \]

where \( T_i^k > 0 \) for \( i = 1, 2, \ldots, n \) and \( k = 1, 2, \ldots, n \).

**Advantages of HPAM:**

- Fast computation: (always) closed form solution for subproblems.
- Theoretical convergence guarantee: stable local behavior.

**Remark:** AM iteration can also be incorporated in some cases.
**Convergence analysis**

**Theorem 2 (Global convergence)**

Let \( \{x^k = (x^k_1, \ldots, x^k_n)\}_{k=1}^{\infty} \) to be the infinite sequence generated by the HPAM. If \( H \) is a KL function and \( \{x_k\} \) is bounded, then there exists an \( x^* \), such that

\[
x^k \to x^* \text{ as } k \to \infty, \text{ and } 0 \in \partial H(x^*),
\]

when \( T_i^k \) is chosen appropriately.

**Criterion of choice of \( T_i^k \):**

- **PAM iterates:** \( A \preceq T_i^k \preceq B \) for some \( A, B \succ 0 \) (more flexible).
- **PALM iterates:** \( L_i \preceq T_i^k \preceq B \) for some \( B \succ 0 \) where \( L_i \) is the Lipschitz constant of \( \nabla F(x) \) at the block \( x_i \), i.e. \( \nabla_i F(x) \)

**KL functions:** polynomial functions, indicator functions of polyhedral sets, Euclidean norm, TV semi-norm, \( \ell_0 \) norm and rank function, etc.
HPAM in dictionary learning

- Redundant dictionary learning

\[
\min_{C \in \mathbb{R}^{p \times q}, D \in \mathbb{R}^{m \times q}} \| Y - DC^\top \|_F^2 + \lambda \| C \|_0, \text{ s.t. } , \| C \|_\infty \leq M, \| d_i \|_2 = 1, \forall i.
\]

- Numerical schemes
  - Block choices
    - (a1) \((x_1, x_2, \ldots, x_n) = (c_1, c_2 \ldots, c_q, d_1, d_2, \ldots, d_q)\)
    - (a2) \((x_1, x_2, \ldots, x_n) = (C, D)\)
    - (a3) \((x_1, x_2, \ldots, x_n) = (C, d_1, c_1, d_2, c_2, \ldots, d_q, c_q)\)
  - Update schemes

<table>
<thead>
<tr>
<th>SCHEMES</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block choice</td>
<td>a1</td>
<td>a2</td>
<td>a3</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 schemes (S1),(S2),(S3) converge to a critical point.
**Closed solutions for Step I**

**Step I: sparse coding**

- The solution of \( \min_{\|x\|_\infty \leq M} (x - y)^2 + \lambda |x|_0 \) is

\[
x^* = \mathcal{P}_\lambda(y) := \text{sign}(y) \odot \min(|\mathcal{T}_\lambda(y)|, M)
\]

if \( M \geq \sqrt{\lambda} \) where \( \mathcal{T}_\lambda(\cdot) \) is the hard-thresholding operator.

- **Scheme 1:** for \( i = 1, 2, \ldots, q \)

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

\[
= \mathcal{P}_{\lambda/(1+\lambda_c)}((J_i^k d_i^k + \lambda_c c_i^{k-1})/(1 + \lambda_c)), \text{ if } M > \sqrt{\lambda/(1 + \lambda_c)},
\]

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

- **Scheme 2:**

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

\[
= \mathcal{P}_{\lambda/\lambda_C}(C^{k-1} - \nabla_C f(D^k, C^{k-1}/2\lambda_c)), \text{ if } M > \sqrt{\lambda/\lambda_C},
\]

where \( f(D, C) = \|Y - DC\|_F^2 \).

- **Scheme 3** is the same as scheme 2.
Closed solutions for Step II

Step II: Dictionary update

▶ Scheme 1: for $i = 1, 2, \ldots, q$

$$d_i^{k+1} \in \arg \min_{\|d\|_2=1} \|E_i^k - dc_i^{k\top}\|_F^2 + \lambda_d \|d - d_i^k\|_2^2$$

$$= \text{Proj}(d_i^k - E_i^k c_i^k / \lambda_d)$$

where $E_i^k = Y - \sum_{j<i} d_j^{k+1} c_j^{k\top} + \sum_{j>i} d_j^k c_j^{k\top}$ and $\text{Proj}(x) = x/\|x\|_2$.

▶ Scheme 2:

$$D \in \arg \min_D \langle \nabla_D f(D^k, C^k), D - D^k \rangle + \lambda_D \|D - D^k\|_F^2, \text{ s.t. } \|d_i\| = 1, \forall i,$$

$$= \text{Proj}(D^k - \nabla_D f(D^k, C^k) / \lambda_D)$$

▶ Scheme 3: Scheme 1 + re-update the nonzero coefficients of $c$

$$c_i^{k+1} = \arg \min_{\|c\|_\infty \leq M} \|E_i^k - d_i^{k+1} c^{\top}\|_F^2, \text{ s.t. } c_j = 0, \forall j \in I_j^k.$$
Experiments: convergence behavior

Our algorithm does generate a convergent sequence to a critical point.
**Experiments**

1. Different schemes converge to different critical points.

\[ \sigma = 10 \quad \text{and} \quad \sigma = 15 \]

2. Different initializations lead to different denoise results ($\sim 0.15$ dB).

**Figure:** The difference of denoising results based on random and DCT initializations.
Experiments

1. **Computational efficiency:** \( S_3 \approx S_2 \) > K-SVD > \( S_1 \)

<table>
<thead>
<tr>
<th>Atom Dim.</th>
<th>6x6</th>
<th>8x8</th>
<th>10x10</th>
<th>12x12</th>
<th>14x14</th>
<th>16x16</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>

2. **Image denoise:** \( S_3 \approx S_1 \approx K\text{-SVD} > S_2 \)

<table>
<thead>
<tr>
<th>Image</th>
<th>Fingerprint</th>
<th>Image</th>
<th>Lena</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>K-SVD</td>
<td>36.59</td>
<td>32.39</td>
<td>30.06</td>
</tr>
<tr>
<td>S1</td>
<td>36.58</td>
<td>32.27</td>
<td>29.87</td>
</tr>
<tr>
<td>S2</td>
<td>36.50</td>
<td>32.21</td>
<td>29.84</td>
</tr>
<tr>
<td>S3</td>
<td>36.59</td>
<td>32.35</td>
<td>30.03</td>
</tr>
</tbody>
</table>

3. Scheme 3 is the most appropriate scheme in dictionary learning.
Application: dynamic texture classification

- Recognizing the moving textures with certain stationary temporal changes

- Equiangular kernel dictionary learning

\[
\min_{D,C} \| \psi(Y) - \psi(D)C \|^2_2 + \lambda \| C \|_0, \text{ s.t. } D^\top D = I_d.
\]

where \( \psi \) is a feature map.

- \( \langle \psi(d_i), \psi(d_j) \rangle = \text{const.} \) for all \( i \neq j \) if \( \psi \) corresponds to Gaussian kernels.
Experiments

- Numerical algorithm: HPAM.
- More scalable for high dimensional data
- Datasets:
  - UCLA-DT (50 categories, 200 videos)
  - Dyn-Tex (10 categories, 275 videos)
  - Dyn-Tex++ (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>

Classification accuracy
Conclusion

1. The sparse property of the wavelet tight frame induced by the B-splines plays important role in image restoration.

2. Efficient numerical algorithms for solving $\ell_0$ norm based (data-driven) balance model are proposed.

3. The convergence analysis of these algorithms are established.

4. Experiments show the advantage of our proposed algorithms.
Thank you!