Augmented L1 and Nuclear-Norm Models with Globally Linearly Convergent Algorithms

Wotao Yin (CAAM @ Rice U.)

Joint work with Ming-Jun Lai (Math @ U.Georgia)
L1 versus LS (Least Squares)

Example: matrix $\mathbf{A} = \text{randn}(100, 200)$, vector $\mathbf{x}^0$ has 20 nonzeros, samples $\mathbf{b} := \mathbf{A}\mathbf{x}^0$

$$\min\{\|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

$$\min\{\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

L1 has a \textit{sparse} solution. LS has a \textit{dense} solution.
L1 versus LS (Least Squares)

Example: matrix $A = \text{randn}(100, 200)$, vector $x^0$ has 20 nonzeros, samples $b := Ax^0$

$$\min\{\|x\|_1 : Ax = b\}$$

$L1$ has a sparse solution. LS has a dense solution.

How about

$$(L1 + \alpha LS) \min\{\|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 : Ax = b\}?$$
To get a sparse solution, \((L1+\alpha LS)\) is seemingly a bad idea.

\[1\] Yin, Osher, Goldfarb, and Darbon [2008]
\[2\] Zou and Hastie [2005]
To get a sparse solution, \((L1+\alpha LS)\) is seemingly a bad idea.

However, we will see in this talk:

- Sufficiently small \(\alpha\) leads to an L1 minimizer, which is sparse
- Theoretical and numerical advantages of adding \(\frac{1}{2\alpha} \|x\|_2^2\)

---

1Yin, Osher, Goldfarb, and Darbon [2008]
2Zou and Hastie [2005]
To get a sparse solution, \((L1+\alpha LS)\) is seemingly a bad idea.

However, we will see in this talk:

- Sufficiently small \(\alpha\) leads to an L1 minimizer, which is sparse
- Theoretical and numerical advantages of adding \(\frac{1}{2\alpha}\|x\|_2^2\)

The model is related to

- Linearized Bregman algorithm\(^1\)
- Elastic net\(^2\) (it is a different purpose, looking for non-L1 minimizer)

\(^1\)Yin, Osher, Goldfarb, and Darbon [2008]
\(^2\)Zou and Hastie [2005]
Related problem: minimize nuclear-norm + LS

**Notation.** $\| \cdot \|_*$: nuclear norm; $\| \cdot \|_F$: Frobenius norm; $\| \cdot \|_1$: sum of entries’ absolute values.

---

\(^3\)Fazel [2002], Candes and Recht [2008]
Related problem: minimize nuclear-norm + LS

**Notation.** $\| \cdot \|_*$: nuclear norm; $\| \cdot \|_F$: Frobenius norm; $\| \cdot \|_1$: sum of entries’ absolute values.

- Low-rank matrix completion\(^3\) from $\Omega$–subsamples

\[(Nu + \alpha LS) \min_X \left\{ \|X\|_* + \frac{1}{2\alpha} \|X\|_F^2 : X_{ij} = M_{ij}, \forall (i,j) \in \Omega \right\}\]

Code: SVT by Cai, Candes, and Shen [2008]

---

\(^3\)Fazel [2002], Candes and Recht [2008]
Related problem: minimize nuclear-norm + LS

**Notation.** \( \| \cdot \|_* \): nuclear norm; \( \| \cdot \|_F \): Frobenius norm; \( \| \cdot \|_1 \): sum of entries’ absolute values.

- Low-rank matrix completion\(^3\) from \( \Omega \)–subsamples

\[
\begin{align*}
(Nu+\alpha LS) \quad \min_X \left\{ \|X\|_* + \frac{1}{2\alpha} \|X\|_F^2 : X_{ij} = M_{ij}, \forall (i,j) \in \Omega \right\}
\end{align*}
\]

Code: SVT by Cai, Candes, and Shen [2008]

- Robust PCA (low-rank + sparse decomposition)

\[
\begin{align*}
\min_{L,S} \left\{ \|L\|_* + \lambda \|S\|_1 + \frac{1}{2\alpha} (\|L\|_F^2 + \|S\|_F^2) : L + S = D \right\}
\end{align*}
\]

Code: IT by Wright, Ganesh, Rao, and Ma [2009]

---

\(^3\)Fazel [2002], Candes and Recht [2008]
Outline

1. Guaranteed sparse/low-rank solutions
2. Linearized Bregman algorithm and its global linear convergence
3. Numerical performance with 2nd-order information
Exact regularization

**Theorem (Friedlander and Tseng [2007], Yin [2010])**

There exists $\alpha^0 > 0$ such that whenever $\alpha > \alpha^0$, the unique solution to

\[(L1 + \alpha LS) \quad \min\{\|x\|_1 + \frac{1}{2\alpha}\|x\|_2^2 : Ax = b\}\]

is also a solution to

\[(L1) \quad \min\{\|x\|_1 : Ax = b\}\].
Exact regularization

Theorem (Friedlander and Tseng [2007], Yin [2010])

There exists $\alpha^0 > 0$ such that whenever $\alpha > \alpha^0$, the unique solution to

$$(L1+\alpha LS) \quad \min\{\|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 : Ax = b\}$$

is also a solution to

$$(L1) \quad \min\{\|x\|_1 : Ax = b\}.$$
Exact regularization

Theorem (Friedlander and Tseng [2007], Yin [2010])

There exists $\alpha^0 > 0$ such that whenever $\alpha > \alpha^0$, the unique solution to

$$(L1+\alpha LS) \quad \min \{ \|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 : Ax = b \}$$

is also a solution to

$$(L1) \quad \min \{ \|x\|_1 : Ax = b \}.$$
\[
\min \|x\|_1 \text{ v.s. } \min \|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 \text{ subject to } Ax = b
\]

\begin{align*}
\pm 1 \text{ sparse} & \quad \text{Gaussian sparse} & \quad \text{Power-law sparse}
\end{align*}

Level curves of relative-error $10^{-3}$. Higher is better.
\[ \min \|x\|_1 \quad \text{v.s.} \quad \min \|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 \quad \text{subject to} \quad Ax = b \]

\[ \pm 1 \text{ sparse} \quad \text{Gaussian sparse} \quad \text{Power-law sparse} \]

Level curves of relative-error $10^{-3}$. Higher is better.

**Conclusion:** $\alpha = 10\|x^0\|_\infty$ works well for compressive sensing!

Also, there are various ways to estimate $\|x^0\|_\infty$. 
Recovery Guarantees for \((L1+\alpha LS)\)

The following properties, which guarantee sparse recovery by L1 minimization, can also guarantee that by \((L1+\alpha LS)\) minimization:

- **Null-space property**\(^4\) (NSP). An “if and only if” property for uniform recovery.
- **Restricted isometry principle**\(^5\) (RIP). Widely used. An “if” property shared by many randomly generated matrices.
- **Spherical section property**\(^6\) (SSP). Invariant to left-multiplying nonsingular matrices, but more difficult to use than RIP.
- **“RIPless” analysis**\(^7\). Useful when RIP/SSP does not hold, gives non-uniform guarantees with \(O(k \log(n))\) measurements.

... more ...

---

\(^4\)Donoho and Huo [2001], Gribonval and Nielsen [2003], Zhang [2005]

\(^5\)Candes and Tao [2005]

\(^6\)Zhang [2008], Vavasis [2009]

\(^7\)Candes and Plan [2010]
Recovery Guarantees: Null-Space Condition

Theorem (exact recovery)

Assume $\|x^0\|_\infty$ is fixed. $(L1+\alpha LS)$ uniquely recovers all $k$-sparse vectors $x^0$ from measurements $b = Ax^0$ if and only if

$$\left(1 + \frac{\|x^0_S\|_\infty}{\alpha}\right) \|h_S\|_1 \leq \|h_{S^c}\|_1,$$

holds for $\forall h \in \text{Null}(A)$ and $\forall$ coordinate sets $S$ of cardinality $|S| \leq k$. 

Theorem (matrix exact recovery)

Assume that $\|X^0\|_2$ is fixed. $(Nu+\alpha Fr)$ uniquely recovers all matrices $X^0$ of rank $r$ or less from measurements $b = A(X^0)$ if and only if

$$(1 + \|X^0\|_2 \alpha)r \sum_{i=1}^{\sigma_r(H)} \leq m \sum_{i=r+1}^{\sigma_r(H)} \sigma_i(H)$$

holds for all matrices $H \in \text{Null}(A)$. 

Hints:

(1) suggests $\alpha \geq C \cdot \|x^0\|_\infty$; (2) suggests $\alpha \geq C \cdot \|X^0\|_2$. 

Recovery Guarantees: Null-Space Condition

**Theorem (exact recovery)**

Assume $\|x^0\|_\infty$ is fixed. $(L1+\alpha LS)$ uniquely recovers all $k$-sparse vectors $x^0$ from measurements $b = Ax^0$ if and only if

$$\left(1 + \frac{\|x^0_S\|_\infty}{\alpha}\right) \|h_S\|_1 \leq \|h_{S^c}\|_1,$$

(1)

holds for $\forall h \in \text{Null}(A)$ and $\forall$ coordinate sets $S$ of cardinality $|S| \leq k$.

**Theorem (matrix exact recovery)**

Assume that $\|X^0\|_2$ is fixed. $(Nu+\alpha Fr)$ uniquely recovers all matrices $X^0$ of rank $r$ or less from measurements $b = A(X^0)$ if and only if

$$\left(1 + \frac{\|X^0\|_2}{\alpha}\right) \sum_{i=1}^{r} \sigma_i(H) \leq \sum_{i=r+1}^{m} \sigma_i(H)$$

(2)

holds for all matrices $H \in \text{Null}(A)$. 

Hints:

(1) suggests $\alpha \geq C \cdot \|x^0\|_\infty$; (2) suggests $\alpha \geq C \cdot \|X^0\|_2$. 
Recovery Guarantees: Null-Space Condition

Theorem (exact recovery)
Assume \( \|x^0\|_{\infty} \) is fixed. \((L1+\alpha LS)\) uniquely recovers all \( k \)-sparse vectors \( x^0 \) from measurements \( b = Ax^0 \) if and only if

\[
\left(1 + \frac{\|x^0_S\|_{\infty}}{\alpha}\right) \|h_S\|_1 \leq \|h_{S^c}\|_1,
\]

holds for \( \forall h \in \text{Null}(A) \) and \( \forall \) coordinate sets \( S \) of cardinality \( |S| \leq k \).

Theorem (matrix exact recovery)
Assume that \( \|X^0\|_2 \) is fixed. \((Nu+\alpha Fr)\) uniquely recovers all matrices \( X^0 \) of rank \( r \) or less from measurements \( b = A(X^0) \) if and only if

\[
\left(1 + \frac{\|X^0\|_2}{\alpha}\right) \sum_{i=1}^{r} \sigma_i(H) \leq \sum_{i=r+1}^{m} \sigma_i(H)
\]

holds for all matrices \( H \in \text{Null}(A) \).

Hints: (1) suggests \( \alpha \geq C \cdot \|x^0\|_{\infty} \); (2) suggests \( \alpha \geq C \cdot \|X^0\|_2 \).
Recovery Guarantees: Restricted Isometry Principle (RIP)

Definition (Candes and Tao [2005])

The RIP constant $\delta_k$ is the smallest value such that

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

holds for all $k$-sparse vectors $x \in \mathbb{R}^n$.

Theorem (exact recovery)

Assume that $x^0 \in \mathbb{R}^n$ is $k$-sparse. If $A$ satisfies RIP with $\delta_{2k} \leq 0.4404$ and $\alpha \geq 10\|x^0\|_\infty$, then $x^0$ is the unique minimizer of $(L1+\alpha LS)$ given measurements $b := Ax^0$. 

• Work on RIP constants: Candes [2008], Foucart and Lai [2009], Foucart [2010], Cai, Wang, and Xu [2010], Mo and Li [2011]. We used proof techniques from Mo and Li [2011].
Recovery Guarantees: Restricted Isometry Principle (RIP)

**Definition (Candès and Tao [2005])**

The RIP constant $\delta_k$ is the smallest value such that

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

holds for all $k$-sparse vectors $x \in \mathbb{R}^n$.

**Theorem (exact recovery)**

Assume that $x^0 \in \mathbb{R}^n$ is $k$-sparse. If $A$ satisfies RIP with $\delta_{2k} \leq 0.4404$ and $\alpha \geq 10\|x^0\|_\infty$, then $x^0$ is the unique minimizer of $(L1 + \alpha LS)$ given measurements $b := Ax^0$.

**Theorem (matrix exact recovery)**

Let $X^0$ be a matrix with rank $r$ or less. If $A$ satisfies RIP with $\delta_{2r} \leq 0.4404$ and $\alpha \geq 10\|X^0\|_2$, then $X^0$ is the unique minimizer of $(Nu + \alpha Fr)$ given measurements $b := A(X^0)$. 

---

• Work on RIP constants: Candès [2008], Foucart and Lai [2009], Foucart [2010], Cai, Wang, and Xu [2010], Mo and Li [2011]. We used proof techniques from Mo and Li [2011].
Recovery Guarantees: Restricted Isometry Principle (RIP)

**Definition (Candes and Tao [2005])**

The RIP constant $\delta_k$ is the smallest value such that

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

holds for all $k$-sparse vectors $x \in \mathbb{R}^n$.

**Theorem (exact recovery)**

Assume that $x^0 \in \mathbb{R}^n$ is $k$-sparse. If $A$ satisfies RIP with $\delta_{2k} \leq 0.4404$ and $\alpha \geq 10\|x^0\|_{\infty}$, then $x^0$ is the unique minimizer of $(L1 + \alpha LS)$ given measurements $b := Ax^0$.

**Theorem (matrix exact recovery)**

Let $X^0$ be a matrix with rank $r$ or less. If $A$ satisfies RIP with $\delta_{2r} \leq 0.4404$ and $\alpha \geq 10\|X^0\|_2$, then $X^0$ is the unique minimizer of $(Nu + \alpha Fr)$ given measurements $b := A(X^0)$.

- Work on RIP constants: Candes [2008], Foucart and Lai [2009], Foucart [2010], Cai, Wang, and Xu [2010], Mo and Li [2011]. We used proof techniques from Mo and Li [2011].
Recovery Guarantees: Restricted Isometry Principle (RIP)

For approximately sparse signals and/or noisy measurements, solve the $\ell_2$-constrained model:

$$\min_{x} \left\{ \|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 : \|Ax - b\|_2 \leq \sigma \right\}$$  \hspace{1cm} (3)

Theorem (stable recovery)

Let $x_0 \in \mathbb{R}^n$ be an arbitrary vector, $S = \{\text{largest } k \text{ components of } x_0\}$, and $Z = S^C$. Let $b := Ax_0 + n$, where $n$ is an arbitrary noisy vector. If $A$ satisfies RIP with $\delta_{2k} \leq 0.3814$, then the solution $x^*$ of (3) with $\alpha \geq 10 \|x_0\|_\infty$ and $\sigma = \|n\|_2$ satisfies

$$\|x^* - x_0\|_1 \leq C_1 \cdot \sqrt{k} \|n\|_2 + C_2 \cdot \|x_0\|_Z \|_1,$$

$$\|x^* - x_0\|_2 \leq \bar{C}_1 \cdot \|n\|_2 + \bar{C}_2 \cdot \|x_0\|_Z \|_1 / \sqrt{k},$$

where $C_1$, $C_2$, $\bar{C}_1$, and $\bar{C}_2$ are constants depending on $\delta_{2k}$.

\footnote{One can use $\alpha / \|x^0\|_\infty$ and $\alpha / \|x^0_Z\|_\infty$ to improve the constants.}
Recovery Guarantees: Restricted Isometry Principle (RIP)

For approximately sparse signals and/or noisy measurements, solve the $\ell_2$-constrained model:

$$\min_x \left\{ \|x\|_1 + \frac{1}{2\alpha} \|x\|_2^2 : \|Ax - b\|_2 \leq \sigma \right\}$$  \hspace{1cm} (3)

**Theorem (stable recovery)**

Let $\mathbf{x}^0 \in \mathbb{R}^n$ be an arbitrary vector, $S = \{\text{largest } k \text{ components of } \mathbf{x}^0\}$, and $\mathcal{Z} = S^C$. Let $\mathbf{b} := A\mathbf{x}^0 + \mathbf{n}$, where $\mathbf{n}$ is an arbitrary noisy vector. If $A$ satisfies RIP with $\delta_{2k} \leq 0.3814$, then the solution $\mathbf{x}^*$ of (3) with $\alpha \geq 10\|\mathbf{x}^0\|_\infty$ and $\sigma = \|\mathbf{n}\|_2$ satisfies

$$\|\mathbf{x}^* - \mathbf{x}^0\|_1 \leq C_1 \cdot \sqrt{k}\|\mathbf{n}\|_2 + C_2 \cdot \|\mathbf{x}^0_\mathcal{Z}\|_1,$$

$$\|\mathbf{x}^* - \mathbf{x}^0\|_2 \leq \bar{C}_1 \cdot \|\mathbf{n}\|_2 + \bar{C}_2 \cdot \|\mathbf{x}^0_\mathcal{Z}\|_1 / \sqrt{k},$$

where $C_1$, $C_2$, $\bar{C}_1$, and $\bar{C}_2$ are constants depending$^8$ on $\delta_{2k}$.

---

$^8$One can use $\alpha / \|\mathbf{x}^0\|_\infty$ and $\alpha / \|\mathbf{x}^0_\mathcal{Z}\|_\infty$ to improve the constants.
Recovery Guarantees: Restricted Isometry Principle (RIP)

For approximately low-rank matrices and/or noisy measurements, solve $\ell_2$-constrained the model:

\[
\min_X \left\{ \|X\|_* + \frac{1}{2\alpha} \|X\|_F^2 : \|A(X) - b\|_2 \leq \sigma \right\}
\]

(4)

---

\[\text{One can use } \alpha/\sigma_1(X^0) \text{ and } \alpha/\sigma_{r+1}(X^0) \text{ to improve the constants.}\]
Recovery Guarantees: Restricted Isometry Principle (RIP)

For approximately low-rank matrices and/or noisy measurements, solve \( \ell_2 \)-constrained the model:

\[
\min_X \left\{ \|X\|_* + \frac{1}{2\alpha} \|X\|_F^2 : \|A(X) - b\|_2 \leq \sigma \right\}
\]

(4)

Theorem (matrix stable recovery)

Let \( X^0 \in \mathbb{R}^{n_1 \times n_2} \) be an arbitrary matrix, and \( \sigma_i(X^0) \) be its \( i \)-th largest singular value. Let \( b := A(X^0) + n \), where \( n \) is an arbitrary noisy vector. If linear operator \( A \) satisfies RIP with \( \delta_{2r} \leq 0.3814 \), then the solution \( X^* \) of (4) with \( \alpha \geq 10\|X^0\|_2 \) and \( \sigma = \|n\|_2 \) satisfies

\[
\|X^* - X^0\|_* \leq C_1 \cdot \sqrt{r} \|n\|_2 + C_2 \cdot \hat{\sigma}(X^0),
\]

\[
\|X^* - X^0\|_F \leq \tilde{C}_1 \cdot \|n\|_2 + (\tilde{C}_2 / \sqrt{r}) \cdot \hat{\sigma}(X^0),
\]

where \( \hat{\sigma}(X^0) = \sum_{i=r+1}^{\min\{n_1,n_2\}} \sigma_i(X^0) \), and \( C_1, C_2, \tilde{C}_1, \) and \( \tilde{C}_2 \) are constants depending\(^9\) on \( \delta_{2r} \).

---

\(^9\)One can use \( \alpha / \sigma_1(X^0) \) and \( \alpha / \sigma_{r+1}(X^0) \) to improve the constants.
Recovery Guarantees: Spherical Section Property

Definition (Vavasis [2009])
Assume $m > 0$, $n > 0$, and $m < n$. An $(n - m)$-dim subspace $\mathcal{V} \subset \mathbb{R}^n$ has the $\Delta$ spherical section property ($\Delta$-SSP) if

$$\frac{\|h\|_1}{\|h\|_2} \geq \sqrt{\frac{m}{\Delta}}, \quad \forall \ h \in \mathcal{V}.$$
Recovery Guarantees: Spherical Section Property

**Definition (Vavasis [2009])**

Assume $m > 0$, $n > 0$, and $m < n$. An $(n - m)$-dim subspace $V \subset \mathbb{R}^n$ has the $\Delta$ spherical section property ($\Delta$-SSP) if

$$\frac{\|h\|_1}{\|h\|_2} \geq \sqrt{\frac{m}{\Delta}}, \quad \forall h \in V.$$

**Significance:**

1. $\text{Null}(A)$ has $\Delta$-SSP and $\frac{m}{\Delta} \geq 4k \Rightarrow \ell_1$–NSP $\Rightarrow$ uniform recovery
2. A uniformly random $(n - m)$-dim subspace $V$ has $\Delta$-SSP\(^{11}\) for

$$\Delta = C_0(\log(n/m) + 1)$$

with prob $\geq 1 - e^{C_1(n-m)}$, where $C_0$ and $C_1$ are universal constants. Hence, uniformly random $\text{Null}(A)$ leads to exact recovery under $m = O(k \log(n/m))$ measurements by $\ell_1$ with overwhelming probability.

---

\(^{10}\)Zhang [2008]

\(^{11}\)Kashin [1977], Garnaev and Gluskin [1984]
Recovery Guarantees: Spherical Section Property

Theorem (exact recovery)

Suppose $\text{Null}(A)$ has $\Delta$-SSP. Fix $\|x^0\|_\infty$ and $\alpha > 0$. If

$$m \geq \left(2 + \frac{\|x^0\|_\infty}{\alpha}\right)^2 k\Delta,$$

(5)

then $(L1+\alpha LS)$ recovers all $k$-sparse $x^0$ from measurements $b = Ax^0$. 

• Similar results hold for matrix recovery under the $\Delta$-SSP of $\text{Null}(A)$.
Recovery Guarantees: Spherical Section Property

Theorem (exact recovery)

Suppose \( \text{Null}(A) \) has \( \Delta \)-SSP. Fix \( \|x^0\|_\infty \) and \( \alpha > 0 \). If

\[
m \geq \left( 2 + \frac{\|x^0\|_\infty}{\alpha} \right)^2 k\Delta, \tag{5}\]

then \( (L1+\alpha LS) \) recovers all \( k \)-sparse \( x^0 \) from measurements \( b = Ax^0 \).

Theorem (stable recovery)

Let \( x^0 \in \mathbb{R}^n \) be an arbitrary vector, \( S = \{ \text{largest } k \text{ components of } x^0 \} \), and \( Z = S^c \). Suppose \( \text{Null}(A) \) has \( \Delta \)-SSP. Let \( \alpha > 0 \). If

\[
m \geq 4 \left( 1 + \left( \frac{\alpha + \|x^0_S\|_\infty}{\alpha - \|x^0_Z\|_\infty} \right) \right)^2 k\Delta, \tag{6}\]

then the solution \( x^* \) of \( (L1+\alpha LS) \) satisfies

\[
\|x^* - x^0\|_1 \leq \frac{8\alpha}{\alpha - \|x^0_Z\|_\infty} \cdot \|x^0_Z\|_1. \tag{7}\]
Recovery Guarantees: Spherical Section Property

**Theorem (exact recovery)**

Suppose $\text{Null}(A)$ has $\Delta$-SSP. Fix $\|x^0\|_\infty$ and $\alpha > 0$. If

$$m \geq \left(2 + \frac{\|x^0\|_\infty}{\alpha}\right)^2 k \Delta,$$

then $(L1+\alpha LS)$ recovers all $k$-sparse $x^0$ from measurements $b = Ax^0$.

**Theorem (stable recovery)**

Let $x^0 \in \mathbb{R}^n$ be an arbitrary vector, $S = \{\text{largest } k \text{ components of } x^0\}$, and $Z = S^C$. Suppose $\text{Null}(A)$ has $\Delta$-SSP. Let $\alpha > 0$. If

$$m \geq 4 \left(1 + \left(\frac{\alpha + \|x^0_S\|_\infty}{\alpha - \|x^0_Z\|_\infty}\right)\right)^2 k \Delta,$$

then the solution $x^*$ of $(L1+\alpha LS)$ satisfies

$$\|x^* - x^0\|_1 \leq \frac{8\alpha}{\alpha - \|x^0_Z\|_\infty} \cdot \|x^0_Z\|_1.$$

- Similar results hold for matrix recovery under the $\Delta$-SSP of $\text{Null}(A)$.
Recovery Guarantees: an “RIPless” property\textsuperscript{12}

- Especially useful when NSP/RIP/SSP are difficult to check or do not hold (with good constants).
- Applications: orthogonal transform ensembles satisfying an incoherence condition, random Teoplitz/circulant ensembles, certain tight and continuous frame ensembles

**Theorem (exact recovery)**

Let $\mathbf{x}^0 \in \mathbb{R}^n$ be a fixed $k$-sparse vector. With prob $\geq 1 - 5/n - e^{-\beta}$, $\mathbf{x}^0$ is the unique solution to $(L_1 + \alpha LS)$ given $\mathbf{b} = \mathbf{A}\mathbf{x}^0$ and $\alpha \geq 8\|\mathbf{x}^0\|_2$ if

$$m \geq C_0(1 + \beta)\mu(\mathbf{A}) \cdot k \log n,$$

where $C_0$ is a constant and $\mu(\mathbf{A})$ is the incoherence parameter of $\mathbf{A}$.

\textsuperscript{12}Candes and Plan [2010]
Guarantee for matrix completion

Theorem (Zhang, Cai, Cheng, and Zhu [2012])

Consider matrix \( \mathbf{M} \in \mathbb{R}^{n_1 \times n_2} \) obeying the strong incoherence assumption\(^\text{13}\). With

\[
\alpha \geq \frac{4}{p} \| \text{Proj}_\Omega \mathbf{M} \|_F, \quad \text{where } p = \frac{m}{n_1 n_2}
\]

and probability \( \geq 1 - n^{-3} \), matrix \( \mathbf{M} \) is the unique solution to

\[
\min \{ \| \mathbf{X} \|_* + \frac{1}{2\alpha} \| \mathbf{X} \|_F^2 : \mathbf{X}_{ij} = \mathbf{M}_{ij}, \ \forall (i,j) \in \Omega \}.
\]

\(^{13}\)Candes and Tao [2010]
Outline

1. Guarantees for recovering sparse solutions

2. Linearized Bregman algorithm and its global geometric convergence

3. Numerical performance with 2nd-order information
Linearized Bregman

‖ Bregman distance

\[ D_J(x; y) = J(x) - [J(y) + \langle p, x - y \rangle], \quad p \in \partial J(y) \]

linearization of \( J \) at \( y \)

‖ Bregman iteration

\[ x^{k+1} \leftarrow \min D_J(x; x^k) + \frac{1}{2} \| Ax - b \|^2, \]
\[ p^{k+1} \leftarrow p^k + A^\top (b - Ax^{k+1}). \]

Equivalent to augmented Lagrangian after change of variables.

‖ Linearized Bregman iteration

\[ x^{k+1} \leftarrow \min D_J(x; x^k) + h \langle A^\top (Ax^k - b), x \rangle + \frac{1}{2\alpha} \| x - x^k \|^2, \]
\[ p^{k+1} \leftarrow p^k + hA^\top (b - Ax^k) - \frac{1}{\alpha} (x^{k+1} - x^k). \]
Compare Bregman and linearized Bregman Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Bregman</th>
<th>linearized Bregman</th>
</tr>
</thead>
<tbody>
<tr>
<td>sparse optimization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>subproblem</td>
<td>BPDN(^{14})</td>
<td>closed form</td>
</tr>
<tr>
<td>iterations</td>
<td>5 - 10</td>
<td>50 - 3000</td>
</tr>
<tr>
<td>minimizes ({J(x) : Ax = b})?</td>
<td>Yes(^{15})</td>
<td>(\alpha) exact regularization</td>
</tr>
<tr>
<td>truncation–error forgetting?</td>
<td>Yes(^{16})</td>
<td>n/a</td>
</tr>
</tbody>
</table>

\(^{14}\)\(\min_x \|x\|_1 + \frac{\mu}{2} \|Ax - b^k\|_2^2\)
\(^{15}\)Yin, Osher, Goldfarb, and Darbon [2008]
\(^{16}\)Yin and Osher [2012]
Linearized Bregman

The linearized Bregman iteration generates a sequence \( \{x^k\} \) converging to the solution of

\[
(L1 + \alpha LS) \min \left\{ J(x) + \frac{1}{2\alpha} \|x\|_2^2 : Ax = b \right\}.
\]

Clue: \( D_{J(\cdot)}(x; x^k) + \frac{1}{2\alpha} \|x - x^k\|_2^2 = D_{J(\cdot)} + \frac{1}{2\alpha} \|\cdot\|_2^2(x; x^k) \)

The linearized Bregman iteration = gradient descent to the Lagrange dual of \((L1 + \alpha LS):\)

\[
\min -b^\top y + \frac{\alpha}{2} \|A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y)\|_2^2.
\]
Lagrangian dual is unconstrained and $C^1$

Theorem (Convex Analysis by Rockafellar [1970])

If a convex program has a strictly convex objective, it has a unique solution and its Lagrangian dual program is differentiable.
Lagrangian dual is unconstrained and $C^1$

**Theorem (Convex Analysis by Rockafellar [1970])**

*If a convex program has a strictly convex objective, it has a unique solution and its Lagrangian dual program is differentiable.*

$$\min -b^\top y + \frac{\alpha}{2} \| A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y) \|_2^2.$$
Lagrangian dual is unconstrained and $C^1$

Theorem (Convex Analysis by Rockafellar [1970])

If a convex program has a strictly convex objective, it has a unique solution and its Lagrangian dual program is differentiable.

$$\min -b^T y + \frac{\alpha}{2} \| A^T y - \text{Proj}_{[-1,1]^n}(A^T y) \|^2_2.$$
Global linear (geometric) convergence

Theorem
Assume a solution \( x^* \neq 0 \) exists. Let \( \mathcal{Y}^* \) be the set of optimal dual solutions. Let \( f \) be the dual objective function, and \( f^* \) be the optimal dual objective value. The linearized Bregman iteration starting from any \( y^0 \) with step size \( 0 < h < 2\nu/(\alpha^2\|A\|^4) \) generates

- globally Q-linearly converging dual solutions \( \{y^k\} \):

\[
\text{dist}_{\ell^2}(y^k, \mathcal{Y}^*) \leq C^{k/2} \cdot \text{dist}_{\ell^2}(y^0, \mathcal{Y}^*),
\]

- globally R-linearly converging dual values \( \{f(y^k)\} \) and primal solutions \( \{x^k\} \):

\[
f(y^k) - f^* \leq (L/2)C^k \cdot (\text{dist}_{\ell^2}(y^0, \mathcal{Y}^*))^2,
\]

\[
\|x^{k+1} - x^*\|_2 \leq \alpha\|A\|_2 C^{k/2} \cdot \text{dist}_{\ell^2}(y^0, \mathcal{Y}^*),
\]

where \( \nu \) is a restricted strong convexity constant, and \( C := 1 - 2h\nu + h^2\alpha^2\|A\|_2^4 \) obeys \( 0 < C < 1 \).
Proof outline

Recall dual objective:

\[ f(y) := -b^\top y + \frac{\alpha}{2} \|A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y)\|_2^2 \]

• Dual solution set: \( \mathcal{Y}^* = \{ y' \in \mathbb{R}^m : \alpha \text{ shrink}(A^\top y') = x^* \} \)
Proof outline

Recall dual objective:

\[ f(y) := -b^T y + \frac{\alpha}{2} \| A^T y - \text{Proj}_{[-1,1]^n}(A^T y) \|_2^2 \]

- Dual solution set: \( \mathcal{Y}^* = \{ y' \in \mathbb{R}^m : \alpha \text{ shrink}(A^T y') = x^* \} \)
- (Key!) Restricted strong convexity (RSC): \( \exists \, \nu > 0 \) such that
  \[ \langle y - \text{Proj}_{\mathcal{Y}^*}(y), \nabla f(y) \rangle \geq \nu \| y - \text{Proj}_{\mathcal{Y}^*}(y) \|_2^2, \quad \forall \, y \in \mathbb{R}^m \]
Proof outline

Recall dual objective:

\[ f(y) := -b^T y + \frac{\alpha}{2} \| A^T y - \text{Proj}_{[-1,1]^n}(A^T y) \|_2^2 \]

- Dual solution set: \( \mathcal{Y}^* = \{ y' \in \mathbb{R}^m : \alpha \text{ shrink}(A^T y') = x^* \} \)
- (Key!) Restricted strong convexity (RSC): \( \exists \nu > 0 \) such that

\[ \langle y - \text{Proj}_{\mathcal{Y}^*}(y), \nabla f(y) \rangle \geq \nu \| y - \text{Proj}_{\mathcal{Y}^*}(y) \|_2^2, \quad \forall y \in \mathbb{R}^m \]

Compare with strong convexity (which does not hold in our case):

\[ \langle y - y', \nabla f(y) - \nabla f(y') \rangle \geq c \| y - y' \|_2^2, \quad \forall y, y' \in \mathbb{R}^m \]
Proof outline

Recall dual objective:

\[ f(y) := -b^T y + \frac{\alpha}{2} \| A^T y - \text{Proj}_{[-1,1]^n}(A^T y) \|^2 \]

- Dual solution set: \( \mathcal{Y}^* = \{ y' \in \mathbb{R}^m : \alpha \text{shrink}(A^T y') = x^* \} \)
- (Key!) Restricted strong convexity (RSC): \( \exists \nu > 0 \) such that

\[
\langle y - \text{Proj}_{\mathcal{Y}^*}(y), \nabla f(y) \rangle \geq \nu \| y - \text{Proj}_{\mathcal{Y}^*}(y) \|^2, \quad \forall y \in \mathbb{R}^m
\]

Compare with strong convexity (which does not hold in our case):

\[
\langle y - y', \nabla f(y) - \nabla f(y') \rangle \geq c \| y - y' \|^2, \quad \forall y, y' \in \mathbb{R}^m
\]

- From RSC to global linear convergence is standard
Proving RSC requires eigen-properties of $A$:

- Decompose $A = [“active cols.” “inactive cols.”] = [\tilde{A} \ \tilde{B}]$
- We need to bound RSC constant $\nu$ from zero, translating to proving

$$\min_{Ax \neq 0} \frac{(Ax)^\top (\tilde{A}D\tilde{A}^\top)(Ax)}{(Ax)^\top (Ax)} > 0, \quad \text{(where } \tilde{D} > 0 \text{ is fixed).}$$

It's true only if $\text{rank}(\tilde{A}) = \text{rank}(A)$, which is not the case since $x^*$ is sparse and thus active columns are very few!
Proving RSC requires eigen-properties of \( \mathbf{A} \):

- Decompose \( \mathbf{A} = ["active cols." "inactive cols." ] = [\bar{\mathbf{A}} \, \bar{\mathbf{B}}] \)
- We need to bound RSC constant \( \nu \) from zero, translating to proving

\[
\min_{\mathbf{Ax} \neq 0} \frac{(\mathbf{Ax})^\top (\bar{\mathbf{A}}\bar{\mathbf{D}}\bar{\mathbf{A}}^\top)(\mathbf{Ax})}{(\mathbf{Ax})^\top (\mathbf{Ax})} > 0, \quad (\text{where } \bar{\mathbf{D}} \succ 0 \text{ is fixed}).
\]

It’s true only if \( \text{rank}(\bar{\mathbf{A}}) = \text{rank}(\mathbf{A}) \), which is not the case since \( \mathbf{x}^* \) is sparse and thus active columns are very few!

- After finer analysis, we can instead bound \( \nu \) by

\[
\min \left\{ \frac{(\mathbf{Ax})^\top (\bar{\mathbf{A}}\bar{\mathbf{D}}\bar{\mathbf{A}}^\top)(\mathbf{Ax})}{(\mathbf{Ax})^\top (\mathbf{Ax})} : \mathbf{Ax} = \bar{\mathbf{A}}\bar{\mathbf{c}} + \bar{\mathbf{B}}\bar{\mathbf{d}} \neq 0, \bar{\mathbf{d}} \geq 0, \bar{\mathbf{B}}^\top(\bar{\mathbf{A}}\bar{\mathbf{c}} + \bar{\mathbf{B}}\bar{\mathbf{d}}) \leq 0 \right\}
\geq \min \{ \lambda^{++}_{\text{min}}(\bar{\mathbf{A}}\bar{\mathbf{D}}\bar{\mathbf{A}}^\top + \bar{\mathbf{C}}^\top \bar{\mathbf{C}}) : \bar{\mathbf{C}} \text{ is an } m\text{-by-}p \text{ submatrix of } \bar{\mathbf{B}}, \, p \geq 0 \}
Compare with convergence results of other algorithms

For sparse optimization:

- Iterative soft-thresholding (ISTA): asymptotic linear convergence\(^{17}\) (find support of \(x^*\) in finitely many steps; then converge linearly), no global linear convergence rate, but has global sublinear rate \(f(x^k) - f^* \approx O(1/k)\)

- FISTA: global sublinear convergence\(^{18}\) \(f(x^k) - f^* \approx O(1/k^2)\)

- Alternating-direction method /split Bregman: no known rate of convergence for \(\ell_1\) better than \(O(1/k)\)

- Accelerated linearized Bregman\(^{19}\): \(O(1/k^2)\)

- Linearized Bregman: \(O(\mu^k), \mu < 1\), for \(\|x^k - x^*\|\) and \(f(x^k) - f^*\)

---

\(^{17}\) Hale, Yin, and Zhang [2008]

\(^{18}\) Beck and Teboulle [2009]

\(^{19}\) Huang, Ma, and Goldfarb [2011]
Outline

1. Guarantees for recovering sparse solutions
2. Linearized Bregman algorithm and its global geometric convergence
3. **Numerical performance with 2nd-order information**
Much faster convergence

- Dual is differentiable, so we can apply gradient–based techniques

---

20 Osher, Mao, Dong, and Yin [2010]
21 Barzilai and Borwein [1988]
22 Zhang and Hager [2004]
23 Liu and Nocedal [1989]
Much faster convergence

- Dual is differentiable, so we can apply gradient–based techniques
- Primal $x^k$ can be recovered from dual $y^k$:

$$ x^k := A(A^\top y^k - \text{Proj}_{[-1,1]^n}(A^\top y^k)) $$

---

20 Osher, Mao, Dong, and Yin [2010]
21 Barzilai and Borwein [1988]
22 Zhang and Hager [2004]
23 Liu and Nocedal [1989]
Much faster convergence

- Dual is differentiable, so we can apply gradient–based techniques
- Primal $x^k$ can be recovered from dual $y^k$:

$$x^k := A(A^T y^k - \text{Proj}_{[-1,1]}(A^T y^k))$$

Numerically compare

1. Original linearized Bregman (dual grad descent)
2. (1) + Kicking\(^{20}\)
3. (2) + BB step\(^{21}\) + non-monotone\(^{22}\) line search
4. Limited-memory BFGS or L-BFGS\(^{23}\) (only use last 5 gradients)

The ADM approach by Yang, Moller, and Osher [2011] has impressive results but haven’t been compared yet.

\(^{20}\)Osher, Mao, Dong, and Yin [2010]
\(^{21}\)Barzilai and Borwein [1988]
\(^{22}\)Zhang and Hager [2004]
\(^{23}\)Liu and Nocedal [1989]
Test on a Gaussian Sparse Signal

Error of $x^{(k)}$

Error of $y^{(k)}$

Iteration

2-norm error

original
kicking
line search
L-BFGS
Test on a ±1 Sparse Signal

Error of $x^{(k)}$

Error of $y^{(k)}$
Conclusions

- (L1+\(\alpha\)LS) can still give sparse solutions (unlike the Huber-norm)
- Dual of (L1+\(\alpha\)LS) is differentiable, grad-descent has global linear convergence
- Using 2nd-order information significantly accelerates convergence

Current and Future Work

- More effective smoothing?
- Develop much faster algorithms using 2nd-order info for \(\ell_1\) and problems exploiting low-dimensional structures
- Upgrade existing codes with \(\alpha\)LS
- More ...
References:


Q. Mo and S. Li. New bounds on the restricted isometry constant $\delta_{2k}$. *Applied and Computational Harmonic Analysis*, 31(3):460–468, 2011.


