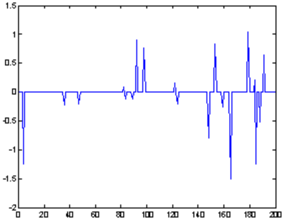
Augmented L1 and Nuclear-Norm Models with Globally Linearly Convergent Algorithms

Wotao Yin (CAAM @ Rice U.)

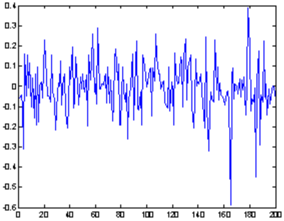
Joint work with Ming-Jun Lai (Math @ U.Georgia)

L1 versus LS (Least Squares)

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



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

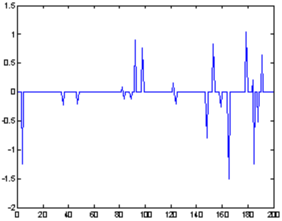


$\min\{\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$

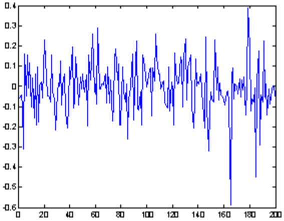
L1 has a *sparse* solution. LS has a *dense* solution.

L1 versus LS (Least Squares)

Example: matrix $\mathbf{A} = \texttt{randn(100,200)}$, 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 *sparse* solution. LS has a *dense* solution.

How about

$$(\text{L1}+\alpha\text{LS}) \qquad \min\{\|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b}\}?$$

To get a sparse solution, (L1+αLS) is seemingly a bad idea.

[1] Yin, Osher, Goldfarb, and Darbon [2008]
[2] Zou and Hastie [2005]

To get a sparse solution, (L1+αLS) is seemingly a bad idea.

However, we will see in this talk:

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

[1] Yin, Osher, Goldfarb, and Darbon [2008]
[2] Zou and Hastie [2005]

To get a sparse solution, (L1+αLS) is seemingly a bad idea.

However, we will see in this talk:

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

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 Ω–subsamples

$$(\text{Nu}+\alpha\text{LS}) \qquad \min_{\mathbf{X}} \left\{ \|\mathbf{X}\|_* + \frac{1}{2\alpha}\|\mathbf{X}\|_F^2 : \mathbf{X}_{ij} = \mathbf{M}_{ij}, \forall (i,j) \in \Omega \right\}$$

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 Ω–subsamples

$$(\text{Nu}+\alpha\text{LS}) \qquad \min_{\mathbf{X}} \left\{ \|\mathbf{X}\|_* + \frac{1}{2\alpha}\|\mathbf{X}\|_F^2 : \mathbf{X}_{ij} = \mathbf{M}_{ij}, \forall (i,j) \in \Omega \right\}$$

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

- Robust PCA (low-rank + sparse decomposition)

$$\min_{\mathbf{L},\mathbf{S}} \left\{ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \frac{1}{2\alpha}\left(\|\mathbf{L}\|_F^2 + \|\mathbf{S}\|_F^2\right) : \mathbf{L} + \mathbf{S} = \mathbf{D} \right\}$$

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

$$\text{(L1+}\alpha\text{LS)} \qquad \min\{\|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

is also a solution to

$$\text{(L1)} \qquad \min\{\|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}\}.$$

Exact regularization

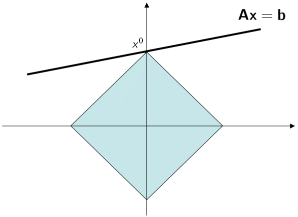
Theorem (Friedlander and Tseng [2007], Yin [2010])

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

$$\text{(L1+}\alpha\text{LS)} \qquad \min\{\|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

is also a solution to

$$\text{(L1)} \qquad \min\{\|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}\}.$$



L1 ball

Exact regularization

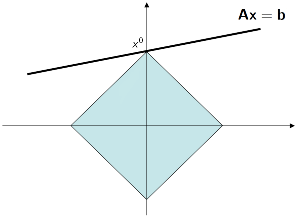
Theorem (Friedlander and Tseng [2007], Yin [2010])

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

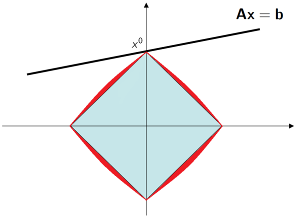
$$\text{(L1+}\alpha\text{LS)} \qquad \min\{\|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

is also a solution to

$$\text{(L1)} \qquad \min\{\|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}\}.$$



L1 ball



L1+LS ball

Recovery Guarantees

The following properties, which guarantee sparse recovery by L1 minimization, can also guarantee that by (L1+αLS) 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 $\|\mathbf{x}^0\|_\infty$ *is fixed. (L1+αLS) uniquely recovers all k-sparse vectors* $\mathbf{x}^0$ *from measurements* $\mathbf{b} = \mathbf{A}\mathbf{x}^0$ *if and only if*

$$\left(1 + \frac{\|\mathbf{x}^0_{\mathcal{S}}\|_\infty}{\alpha}\right)\|\mathbf{h}_{\mathcal{S}}\|_1 \leq \|\mathbf{h}_{\mathcal{S}^c}\|_1, \tag{1}$$

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

Recovery Guarantees: Null-Space Condition

Theorem (exact recovery)

Assume $\|\mathbf{x}^0\|_\infty$ is fixed. (L1+αLS) uniquely recovers all k-sparse vectors $\mathbf{x}^0$ from measurements $\mathbf{b} = \mathbf{A}\mathbf{x}^0$ if and only if

$$\left(1 + \frac{\|\mathbf{x}^0_{\mathcal{S}}\|_\infty}{\alpha}\right) \|\mathbf{h}_{\mathcal{S}}\|_1 \leq \|\mathbf{h}_{\mathcal{S}^c}\|_1, \tag{1}$$

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

Theorem (matrix exact recovery)

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

$$\left(1 + \frac{\|\mathbf{X}^0\|_2}{\alpha}\right) \sum_{i=1}^{r} \sigma_i(\mathbf{H}) \leq \sum_{i=r+1}^{m} \sigma_i(\mathbf{H}) \tag{2}$$

holds for all matrices $\mathbf{H} \in \mathrm{Null}(\mathcal{A})$.

Recovery Guarantees: Null-Space Condition

Theorem (exact recovery)

Assume $\|\mathbf{x}^0\|_\infty$ is fixed. (L1+αLS) uniquely recovers all k-sparse vectors $\mathbf{x}^0$ from measurements $\mathbf{b} = \mathbf{A}\mathbf{x}^0$ if and only if

$$\left(1 + \frac{\|\mathbf{x}^0_{\mathcal{S}}\|_\infty}{\alpha}\right)\|\mathbf{h}_{\mathcal{S}}\|_1 \le \|\mathbf{h}_{\mathcal{S}^c}\|_1, \tag{1}$$

holds for $\forall$ $\mathbf{h} \in \mathrm{Null}(\mathbf{A})$ and $\forall$ coordinate sets $\mathcal{S}$ of cardinality $|\mathcal{S}| \le k$.

Theorem (matrix exact recovery)

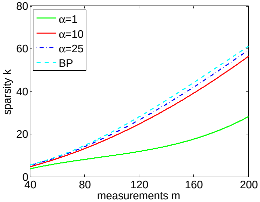
Assume that $\|\mathbf{X}^0\|_2$ is fixed. (Nu+αFr) uniquely recovers all matrices $\mathbf{X}^0$ of rank r or less from measurements $\mathbf{b} = \mathcal{A}(\mathbf{X}^0)$ if and only if

$$\left(1 + \frac{\|\mathbf{X}^0\|_2}{\alpha}\right)\sum_{i=1}^{r}\sigma_i(\mathbf{H}) \le \sum_{i=r+1}^{m}\sigma_i(\mathbf{H}) \tag{2}$$

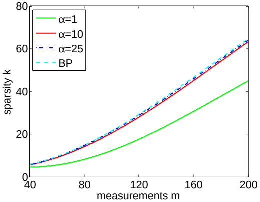
holds for all matrices $\mathbf{H} \in \mathrm{Null}(\mathcal{A})$.

Hints: (1) suggests $\alpha \ge C \cdot \|\mathbf{x}^0\|_\infty$; (2) suggests $\alpha \ge C \cdot \|\mathbf{X}^0\|_2$.

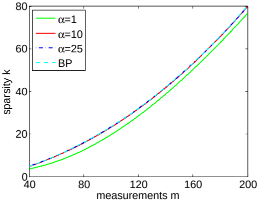
Level curves of relative-error 10^{-3}. Above each curve is where the relative error $>$1e-3. A higher curve means better performance. $\|\mathbf{x}^0\|_\infty = 1$ for all cases.



±1 sparse



Gaussian sparse



Power-law sparse

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

There are various ways to estimate $\|\mathbf{x}^0\|_\infty$.

Recovery Guarantees: Restricted Isometry Principle (RIP)

Definition (Candes and Tao [2005])

The RIP constant δ_k is the smallest value such that

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2$$

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

Theorem (exact recovery)

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

Recovery Guarantees: Restricted Isometry Principle (RIP)

Definition (Candes and Tao [2005])

The RIP constant δ_k is the smallest value such that

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2$$

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

Theorem (exact recovery)

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

Theorem (matrix exact recovery)

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

Recovery Guarantees: Restricted Isometry Principle (RIP)

Definition (Candes and Tao [2005])

The RIP constant δ_k is the smallest value such that

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \le \|\mathbf{A}\mathbf{x}\|_2^2 \le (1+\delta_k)\|\mathbf{x}\|_2^2$$

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

Theorem (exact recovery)

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

Theorem (matrix exact recovery)

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

• 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 ℓ_2-constrained model:

$$\min_{\mathbf{x}} \left\{ \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \sigma \right\} \tag{3}$$

[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 sparse signals and/or noisy measurements, solve the ℓ_2-constrained model:

$$\min_{\mathbf{x}} \left\{ \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \sigma \right\} \tag{3}$$

Theorem (stable recovery)

Let $\mathbf{x}^0 \in \mathbb{R}^n$ be an arbitrary vector, $\mathcal{S} = \{$largest k components of $\mathbf{x}^0\}$, and $\mathcal{Z} = \mathcal{S}^C$. Let $\mathbf{b} := A\mathbf{x}^0 + \mathbf{n}$, where $\mathbf{n}$ is an arbitrary noisy vector. If $\mathbf{A}$ 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

$$\begin{aligned} \|\mathbf{x}^* - \mathbf{x}^0\|_1 &\leq C_1 \cdot \sqrt{k}\|\mathbf{n}\|_2 + C_2 \cdot \|\mathbf{x}^0_{\mathcal{Z}}\|_1, \\ \|\mathbf{x}^* - \mathbf{x}^0\|_2 &\leq \bar{C}_1 \cdot \|\mathbf{n}\|_2 + \bar{C}_2 \cdot \|\mathbf{x}^0_{\mathcal{Z}}\|_1/\sqrt{k}, \end{aligned}$$

where C_1, C_2, $\bar{C}_1$, and $\bar{C}_2$ are constants depending[8] *on δ_{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 ℓ_2-constrained the model:

$$\min_{\mathbf{X}} \left\{ \|\mathbf{X}\|_* + \frac{1}{2\alpha}\|\mathbf{X}\|_F^2 : \|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|_2 \leq \sigma \right\} \tag{4}$$

[9] One can use $\alpha/\sigma_1(\mathbf{X}^0)$ and $\alpha/\sigma_{r+1}(\mathbf{X}^0)$ to improve the constants.

Recovery Guarantees: Restricted Isometry Principle (RIP)

For approximately low-rank matrices and/or noisy measurements, solve ℓ_2-constrained the model:

$$\min_{\mathbf{X}} \left\{ \|\mathbf{X}\|_* + \frac{1}{2\alpha}\|\mathbf{X}\|_F^2 : \|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|_2 \le \sigma \right\} \tag{4}$$

Theorem (matrix stable recovery)

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

$$\begin{aligned} \|\mathbf{X}^* - \mathbf{X}^0\|_* &\le C_1 \cdot \sqrt{r}\|\mathbf{n}\|_2 + C_2 \cdot \hat{\sigma}(\mathbf{X}^0), \\ \|\mathbf{X}^* - \mathbf{X}^0\|_F &\le \bar{C}_1 \cdot \|\mathbf{n}\|_2 + \left(\bar{C}_2/\sqrt{r}\right) \cdot \hat{\sigma}(\mathbf{X}^0), \end{aligned}$$

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

[9] One can use $\alpha/\sigma_1(\mathbf{X}^0)$ and $\alpha/\sigma_{r+1}(\mathbf{X}^0)$ 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 Δ spherical section property (Δ-SSP) if

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

[10] Zhang [2008]

[11] Kashin [1977], Garnaev and Gluskin [1984]

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 Δ spherical section property (Δ-SSP) if

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

Significance:[10]

1. $\mathrm{Null}(\mathbf{A})$ has Δ-SSP and $\frac{m}{\Delta} \geq 4k \ \Rightarrow \ell_1$–NSP $\Rightarrow$ uniform recovery
2. A uniformly random $(n-m)$-dim subspace $\mathcal{V}$ has Δ-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 $\mathrm{Null}(\mathbf{A})$ leads to exact recovery under $m = O(k\log(n/m))$ measurements by ℓ_1 with overwhelming probability.

[10] Zhang [2008]

[11] Kashin [1977], Garnaev and Gluskin [1984]

Recovery Guarantees: Spherical Section Property

Theorem (exact recovery)

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

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

then (L1+α*LS) recovers all* k*-sparse* $\mathbf{x}^0$ *from measurements* $\mathbf{b} = \mathbf{A}\mathbf{x}^0$*.*

Recovery Guarantees: Spherical Section Property

Theorem (exact recovery)

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

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

then (L1+α*LS) recovers all* k*-sparse* $\mathbf{x}^0$ *from measurements* $\mathbf{b} = \mathbf{A}\mathbf{x}^0$*.*

Theorem (stable recovery)

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

$$m \geq 4\left(1 + \left(\frac{\alpha + \|\mathbf{x}^0_{\mathcal{S}}\|_\infty}{\alpha - \|\mathbf{x}^0_{\mathcal{Z}}\|_\infty}\right)\right)^2 k\Delta, \tag{6}$$

then the solution $\mathbf{x}^*$ *of (L1+*α*LS) satisfies*

$$\|\mathbf{x}^* - \mathbf{x}^0\|_1 \leq \frac{8\alpha}{\alpha - \|\mathbf{x}^0_{\mathcal{Z}}\|_\infty} \cdot \|\mathbf{x}^0_{\mathcal{Z}}\|_1. \tag{7}$$

Recovery Guarantees: Spherical Section Property

Theorem (exact recovery)

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

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

then (L1+α*LS) recovers all* k*-sparse* $\mathbf{x}^0$ *from measurements* $\mathbf{b} = \mathbf{A}\mathbf{x}^0$*.*

Theorem (stable recovery)

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

$$m \geq 4\left(1 + \left(\frac{\alpha + \|\mathbf{x}^0_{\mathcal{S}}\|_\infty}{\alpha - \|\mathbf{x}^0_{\mathcal{Z}}\|_\infty}\right)\right)^2 k\Delta, \tag{6}$$

then the solution $\mathbf{x}^*$ *of (L1+*α*LS) satisfies*

$$\|\mathbf{x}^* - \mathbf{x}^0\|_1 \leq \frac{8\alpha}{\alpha - \|\mathbf{x}^0_{\mathcal{Z}}\|_\infty} \cdot \|\mathbf{x}^0_{\mathcal{Z}}\|_1. \tag{7}$$

- Similar results hold for matrix recovery under the Δ-SSP of $\mathrm{Null}(\mathcal{A})$

Recovery Guarantees: an "RIPless" property[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 (L1+*α*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* $\mathbf{A}$.

[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*[13]. *With*

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

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

$$D_J(\mathbf{x};\mathbf{y}) = J(\mathbf{x}) - [\underbrace{J(\mathbf{y}) + \langle \mathbf{p}, \mathbf{x}-\mathbf{y}\rangle}_{\text{linearization of } J \text{ at } \mathbf{y}}], \quad \mathbf{p} \in \partial J(\mathbf{y})$$

- Bregman iteration

$$\mathbf{x}^{k+1} \leftarrow \min D_J(\mathbf{x};\mathbf{x}^k) + \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2,$$
$$\mathbf{p}^{k+1} \leftarrow \mathbf{p}^k + \mathbf{A}^\top(\mathbf{b}-\mathbf{A}\mathbf{x}^{k+1}).$$

  Equivalent to augmented Lagrangian after change of variables.

- Linearized Bregman iteration

$$\mathbf{x}^{k+1} \leftarrow \min D_J(\mathbf{x};\mathbf{x}^k) + h\langle \mathbf{A}^\top(\mathbf{A}\mathbf{x}^k-\mathbf{b}), \mathbf{x}\rangle + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{x}^k\|_2^2,$$
$$\mathbf{p}^{k+1} \leftarrow \mathbf{p}^k + h\mathbf{A}^\top(\mathbf{b}-\mathbf{A}\mathbf{x}^k) - \frac{1}{\alpha}(\mathbf{x}^{k+1}-\mathbf{x}^k).$$

Compare Bregman and linearized Bregman Algorithms

sparse optimization	Bregman	linearized Bregman
subproblem	BPDN[14]	closed form
iterations	5 - 10	50 - 3000
minimizes $\{J(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}$?	Yes[15]	α exact regularization
truncation–error forgetting?	Yes[16]	n/a

[14] $\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}^k\|_2^2$

[15] Yin, Osher, Goldfarb, and Darbon [2008]

[16] Yin and Osher [2012]

Linearized Bregman

Theorem (Cai, Osher, and Shen [2009], Yin [2010])

The linearized Bregman iteration generates a sequence $\{\mathbf{x}^k\}$ *converging to the solution of*

$$\text{(L1+}\alpha\text{LS)} \qquad \min \left\{ J(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}.$$

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

Theorem (Yin [2010])

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

$$\min -\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2.$$

Lagrangian dual is unconstrained and C^1

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

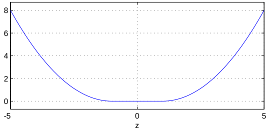
If a convex program has a strictly convex objective, it has a unique solution and its Lagrangian dual program is differentiable.

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.

$$\min -\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2.$$



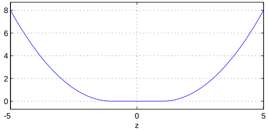
$\frac{1}{2}(z - \mathrm{Proj}_{[-1,1]}(z))^2$

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.

$$\min -\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2.$$



$\frac{1}{2}(z - \mathrm{Proj}_{[-1,1]}(z))^2$

Lead to important consequences:

- Grad descent has global linear (geometric) convergence
- Much faster algorithms using (approximate) 2nd-order information

Global linear (geometric) convergence

Theorem

Assume a solution $\mathbf{x}^* \neq \mathbf{0}$ *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* $\mathbf{y}^0$ *with step size* $0 < h < 2\nu/(\alpha^2\|\mathbf{A}\|^4)$ *generates*

- *globally Q-linearly converging dual solutions* $\{\mathbf{y}^k\}$*:*

$$\mathrm{dist}_{\ell_2}(\mathbf{y}^k, \mathcal{Y}^*) \leq C^{k/2} \cdot \mathrm{dist}_{\ell_2}(\mathbf{y}^0, \mathcal{Y}^*),$$

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

$$f(\mathbf{y}^k) - f^* \leq (L/2)C^k \cdot \left(\mathrm{dist}_{\ell_2}(\mathbf{y}^0, \mathcal{Y}^*)\right)^2,$$

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2 \leq \alpha\|\mathbf{A}\|_2 C^{k/2} \cdot \mathrm{dist}_{\ell_2}(\mathbf{y}^0, \mathcal{Y}^*),$$

where ν *is a restricted strong convexity constant, and* $C := 1 - 2h\nu + h^2\alpha^2\|\mathbf{A}\|_2^4$ *obeys* $0 < C < 1$.

Proof outline

Recall dual objective:

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

- Dual solution set: $\mathcal{Y}^* = \{\mathbf{y}' \in \mathbb{R}^m : \alpha\,\mathrm{shrink}(\mathbf{A}^\top \mathbf{y}') = \mathbf{x}^*\}$

Proof outline

Recall dual objective:

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

- Dual solution set: $\mathcal{Y}^* = \{\mathbf{y}' \in \mathbb{R}^m : \alpha\,\mathrm{shrink}(\mathbf{A}^\top \mathbf{y}') = \mathbf{x}^*\}$
- (Key!) Restricted strong convexity (RSC): $\exists\ \nu > 0$ such that

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

Proof outline

Recall dual objective:

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

- Dual solution set: $\mathcal{Y}^* = \{\mathbf{y}' \in \mathbb{R}^m : \alpha\,\mathrm{shrink}(\mathbf{A}^\top \mathbf{y}') = \mathbf{x}^*\}$
- (Key!) Restricted strong convexity (RSC): $\exists\ \nu > 0$ such that

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

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

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

Proof outline

Recall dual objective:

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

- Dual solution set: $\mathcal{Y}^* = \{\mathbf{y}' \in \mathbb{R}^m : \alpha\,\mathsf{shrink}(\mathbf{A}^\top \mathbf{y}') = \mathbf{x}^*\}$
- (Key!) Restricted strong convexity (RSC): $\exists\ \nu > 0$ such that

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

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

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

- From RSC to global linear convergence is standard

Proving RSC requires eigen-properties of $\mathbf{A}$:

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

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

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

Proving RSC requires eigen-properties of $\mathbf{A}$:

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

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

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

- After finer analysis, we can instead bound ν by

$$\min \left\{ \frac{(\mathbf{Ax})^\top (\bar{\mathbf{A}}\bar{\mathbf{D}}\bar{\mathbf{A}}^\top)(\mathbf{Ax})}{(\mathbf{Ax})^\top (\mathbf{Ax})} : \mathbf{Ax} = \bar{\mathbf{A}}\bar{\mathbf{c}} + \bar{\mathbf{B}}\bar{\mathbf{d}} \neq \mathbf{0}, \bar{\mathbf{d}} \geq \mathbf{0}, \bar{\mathbf{B}}^\top (\bar{\mathbf{A}}\bar{\mathbf{c}} + \bar{\mathbf{B}}\bar{\mathbf{d}}) \leq \mathbf{0} \right\}$$
$$\geq \min\{\lambda_{\min}^{++}(\bar{\mathbf{A}}\bar{\mathbf{D}}\bar{\mathbf{A}}^\top + \bar{\mathbf{C}}^\top \bar{\mathbf{C}}) : \bar{\mathbf{C}} \text{ is an } m\text{-by-}p \text{ submatrix of } \bar{\mathbf{B}}, \ p \geq 0\}$$

Compare with convergence results of other algorithms

For sparse optimization:

- Iterative soft-thresholding (ISTA): *asymptotic* linear convergence[17] (find support of $\mathbf{x}^*$ in finitely many steps; then converge linearly), no global linear convergence rate, but has global sublinear rate $f(\mathbf{x}^k) - f^* \approx O(1/k)$
- FISTA: global *sublinear* convergence[18] $f(\mathbf{x}^k) - f^* \approx O(1/k^2)$
- Alternating-direction method /split Bregman: no known rate of convergence for ℓ_1 better than $O(1/k)$
- Accelerated linearized Bregman[19]: $O(1/k^2)$
- Linearized Bregman: $O(\mu^k)$, $\mu < 1$, for $\|\mathbf{x}^k - \mathbf{x}^*\|$ and $f(\mathbf{x}^k) - f^*$

---

[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 $\mathbf{x}^k$ can be recovered from dual $\mathbf{y}^k$:

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

[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 $\mathbf{x}^k$ can be recovered from dual $\mathbf{y}^k$:

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

Numerically compare

(1) Original linearized Bregman (dual grad descent)

(2) (1) + Kicking[20]

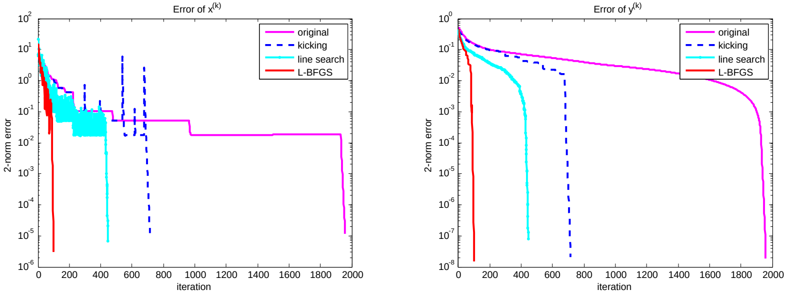
(3) (2) + BB step[21] + non-monotone[22] line search

(4) Limited-memory BFGS or L-BFGS[23] (only use last 5 gradients)

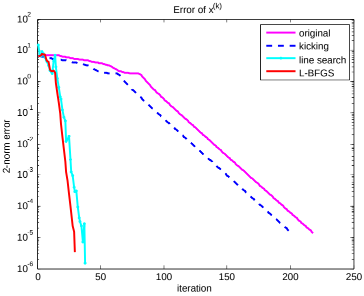
The ADM approach by Yang, Moller, and Osher [2011] has impressive results but haven't been compared yet.

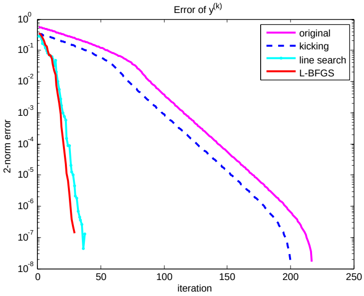
[20] Osher, Mao, Dong, and Yin [2010]
[21] Barzilai and Borwein [1988]
[22] Zhang and Hager [2004]
[23] Liu and Nocedal [1989]

Test on a Gaussian Sparse Signal



Test on a ± 1 Sparse Signal





Summary

Conclusions

- (L1+αLS) can still give sparse solutions (unlike the Huber-norm)
- Dual of (L1+αLS) 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 ℓ_1 and problems exploiting low-dimensional structures
- Upgrade existing codes with αLS
- More ...

References:

W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

M. Fazel. Matrix rank minimization with applications. *Department of Electrical Engineering, Stanford University*, Doctoral Thesis, 2002.

E. Candes and B. Recht. Exact matrix completion via convex optimization. 2008.

J.-F. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2008.

J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *arXiv:0905.0233*, 2009.

M.P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4):1326–1350, 2007.

W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM Journal on Imaging Sciences*, 3(4):856–877, 2010.

D. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001.

R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

Y. Zhang. A simple proof for recoverability of ℓ_1-minimization: go over or under? *Rice University CAAM Technical Report TR05-09*, 2005.

E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

Y. Zhang. Theory of compressive sensing via l1-minimization: a non-RIP analysis and extensions. *Rice University CAAM Technical Report TR08-11*, 2008.

S.A. Vavasis. Derivation of compressive sensing theorems from the spherical section property, www.student.math.uwaterloo.ca/~co769/simplif.pdf, 2009.

E. Candes and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2010.

E. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear systems via lq-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

S. Foucart. A note on guaranteed sparse recovery via l1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

T.T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *Signal Processing, IEEE Transactions on*, 58(3):1300–1308, 2010.

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

B. S. Kashin. Diameters of certain finite-dimensional sets in classes of smooth functions. *Izv. Akad. Nauk SSSR, Ser. Mat.*, 41(2):334–351, 1977.

A. Garnaev and E. D. Gluskin. The widths of a euclidean ball. *Dokl. Akad. Nauk SSSR*, 277(5):1048–1052, 1984.

H. Zhang, J.F. Cai, L. Cheng, and J. Zhu. Strongly convex programming for exact matrix completion and robust principal component analysis. *Preprint*, 2012.

E. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

W. Yin and S. Osher. Error forgetting of bregman iteration. *Rice University CAAM Technical Report TR12-03*, 2012. [pdf].

J.-F. Cai, S. Osher, and Z. Shen. Convergence of the linearized Bregman iteration for ℓ_1-norm minimization. *Mathematics of Computation*, 78(268): 2127–2136, 2009.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

B. Huang, S. Ma, and D. Goldfarb. Accelerated linearized Bregman method. *Arxiv preprint arXiv:1106.5413*, 2011.

S. Osher, Y. Mao, B. Dong, and W. Yin. Fast linearized Bregman iteration for compressive sensing and sparse denoising. *Communications in Mathematical Sciences*, 8(1):93–111, 2010. [pdf].

J. Barzilai and J. Borwein. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

H. Zhang and W. Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM Journal on Optimization*, 14:1043–1056, 2004.

D. C. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming, Series B*, 45(3):503–528, 1989.

Y. Yang, M. Moller, and S. Osher. A dual split Bregman method for fast l1 minimization. *UCLA CAM Report 11-57*, 2011.