The Bregman Methods: Reviews and New Error Cancellation Results

Wotao Yin
(joint work with Stan Osher)

Department of Computational and Applied Mathematics
Rice University

(Work supported by NSF, ONR, and Sloan Foundation)

February 14, 2010

RICE

Bregman iteration has been unreasonably successful in

1. Better regularization quality over ℓ_1, total variation, ...
2. Fast, accurate iterations for *constrained* ℓ_1-like minimization.

Points 1 and 2 are different!

Bregman iteration has been unreasonably successful in

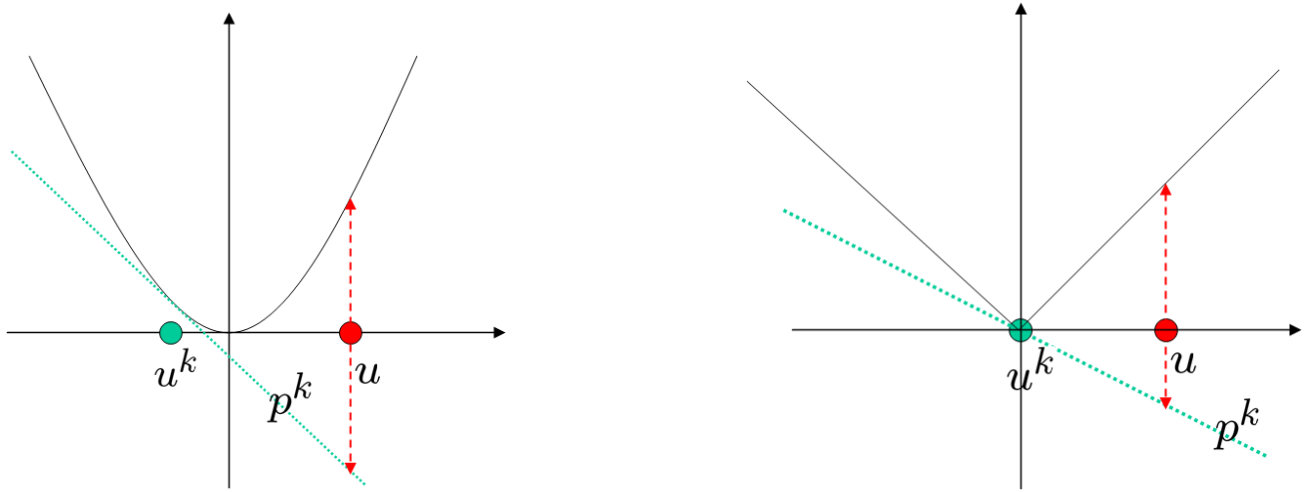
1. Better regularization quality over ℓ_1, total variation, ...
2. Fast, accurate iterations for *constrained* ℓ_1-like minimization.

Points 1 and 2 are different!

Bregman Distance

- Original model: $\min J(u) + f(u)$. Regularizer $J(\cdot)$
- Given u^k, $p^k \in \partial J(u^k)$
- Bregman distance:

$$D(u, u^k) := J(u) - \left(J(u^k) + \langle p^k, u - u^k \rangle\right)$$



- New model: $u^{k+1} \leftarrow \min \alpha D(u, u^k) + f(u)$. E.g.: $\alpha = 5$. p^k is obtainable from previous iteration.

RICE

Bregman = Add Back Residuals

- Original model:

$$u \leftarrow \min \ \mu J(u) + \frac{1}{2}\|Au - b\|_2^2.$$

- Bregman original form:

$$\begin{aligned} u^{k+1} &\leftarrow \min \ \mu \left[J(u) - (J(u^k) + \langle p^k, u - u^k \rangle)\right] + \frac{1}{2}\|Au - b\|_2^2 \\ p^{k+1} &\leftarrow p^k + A^\top (b - Au^{k+1}). \end{aligned}$$

- Add-Back form:

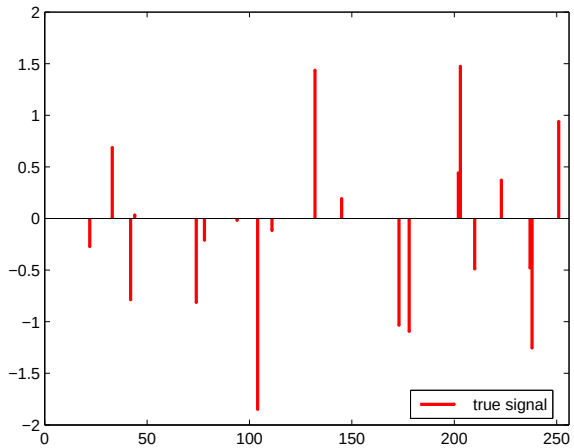
$$\begin{aligned} u^{k+1} &\leftarrow \min \ \mu J(u) + \frac{1}{2}\|Au - b^k\|_2^2 \\ b^{k+1} &\leftarrow b + (b^k - Au^{k+1}). \end{aligned}$$

Each subproblem has the same form of the original problem.

Bregman Regularization: Better Solution Quality

Example: Compressive Sensing Reconstruction from Noise Input

Original signal u, sparse



Noisy Gaussian measurements: $b = Au + \omega$, where A: 100×250.

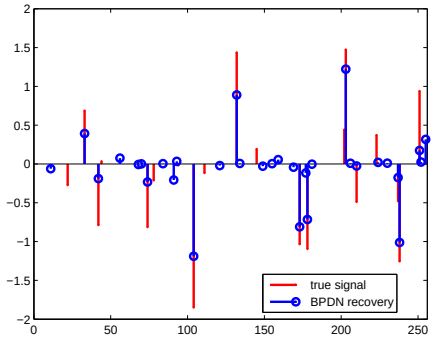
Compare:

1. Basis pursuit: $u \leftarrow \min \mu\|u\|_1 + \frac{1}{2}\|Au - b\|_2^2$
2. Bregman: $u^{k+1} \leftarrow \min \bar{\mu}\|u\|_1 + \frac{1}{2}\|Au - b^k\|_2^2$, $b^{k+1} \leftarrow$ add back

RICE

Basis Pursuit (Non-Bregman) vs Bregman

1. Recover u by: $\quad \min \ \mu\|u\|_1 + \frac{1}{2}\|Au - b\|_2^2$



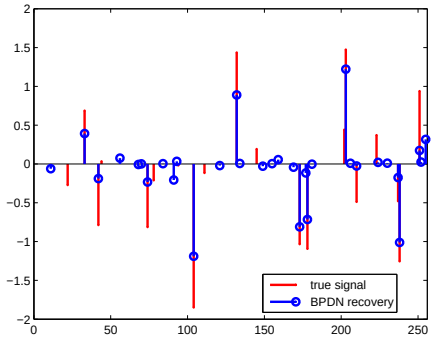
$\mu = 48.5$

Not sparse

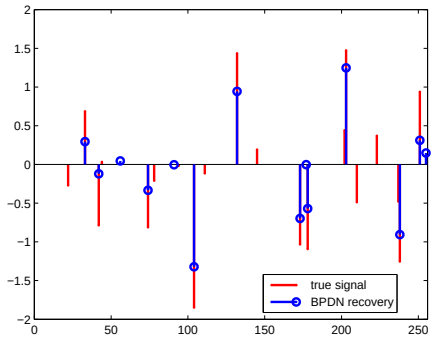
μ too small

Basis Pursuit (Non-Bregman) vs Bregman

1. Recover u by: $\min \ \mu\|u\|_1 + \frac{1}{2}\|Au - b\|_2^2$



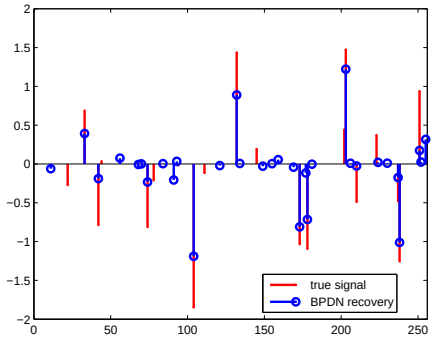
$\mu = 48.5$
Not sparse
μ too small



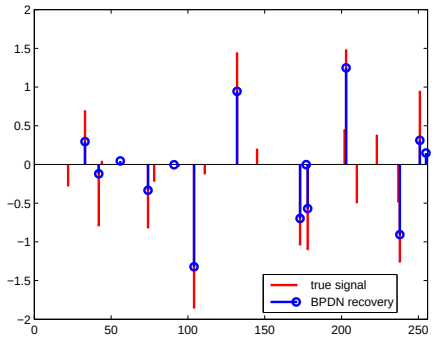
$\mu = 49$
Sparse but poor fitting

Basis Pursuit (Non-Bregman) vs Bregman

1. Recover u by: $\min \ \mu\|u\|_1 + \frac{1}{2}\|Au - b\|_2^2$

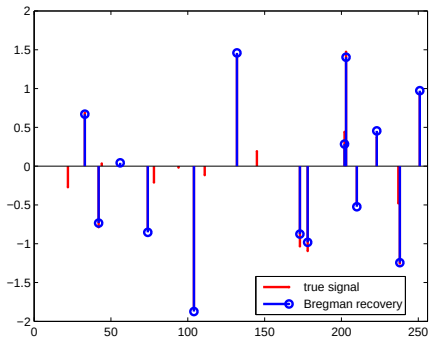


$\mu = 48.5$
Not sparse
μ too small



$\mu = 49$
Sparse but poor fitting

2. Recover u by Bregman: set $\bar{\mu} = 150$, after 5 iterations

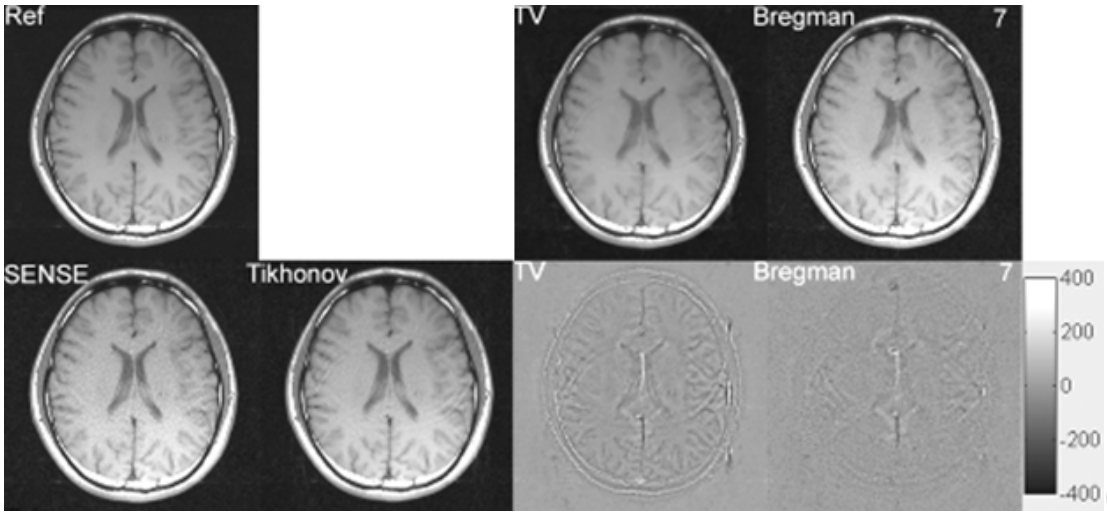


Sparse, better fitting

RICE

Example: image deblurring and/or denoising

- $J(u) = \mu TV(u)$
- $f(u) = \frac{1}{2}\|Au - b\|_2^2$
- Stop when $\|Au^k - b\|_2^2 \approx$ est.$\|Au^{true} - b\|_2^2$



(UWM-CMRI Lab)

Less signal in the residual.

- For ℓ_1, Bregman gives *sparser, better fitted* signals
- For TV, Bregman gives *less staircasing, higher contrast*

- For ℓ_1, Bregman gives *sparser, better fitted* signals
- For TV, Bregman gives *less staircasing, higher contrast*
- Reasons: *iterative boosting*
  1. For small k, u^k is over-regularized yet correctly captures larger nonzeros/edges.
  2. Minimizing $D(u, u^k)$ does not penalize the nonzeros/edges in u^k.

RICE

Bregman Regularization vs Prox-Point

The proximal point algorithm: (Rockafellar'76, Censor-Zenios'92, Kiwiel'97, etc.)

$$u^{k+1} \leftarrow \min_{u} \; f(u) + D(u, u^k) \tag{1}$$

Differences (Bregman Regularization vs Prox-Point):

1. Goal: Bregman – not to minimize $f(u)$; Prox – to minimize $f(u)$;
2. J: Bregman – nonsmooth, convex; Prox – smooth, strongly convex;
3. Stop: Bregman – prior convergence; Prox – near convergence.

1. Better regularization quality over ℓ_1, total variation, ...
   - Work for noisy data
   - Start with over-regularization
   - $f(u^k) \downarrow$, stop $f(u^k) \approx f(\textit{real } u)$ est.

1. Better regularization quality over ℓ_1, total variation, ...
   - Work for noisy data
   - Start with over-regularization
   - $f(u^k) \downarrow$, stop $f(u^k) \approx f(\textit{real } u)$ est.

2. Fast, accurate methods for constrained ℓ_1 and TV minimization.
   - Work for *noiseless* data
   - $f(u^k) \downarrow$, stop $f(u^k) = 0$.

1. Better regularization quality over ℓ_1, total variation, ...
   - Work for noisy data
   - Start with over-regularization
   - $f(u^k) \downarrow$, stop $f(u^k) \approx f(\textit{true } u)$ est.

2. Giving fast, accurate methods for constrained ℓ_1 and TV minimization.
   - Work for *noiseless* data
   - $f(u^k) \downarrow$, stop $f(u^k) = 0$.

Applied to Constrained Minimization

Y.-Osher-Goldfarb-Burger 07

- Purpose: $u_{real} \leftarrow \min\{J(u) : Au = b\}$, constrained
- Bregman: $u^{k+1} \leftarrow \min D_J(u, u^k) + \frac{1}{2}\|Au - b\|_2^2$, unconstrained
- Properties:
  - $u^k \rightarrow u_{real}$
  - Fast, finite convergence for ℓ_1–like $J(u)$
  - Even if subproblems are solved inexactly, under some conditions, solution is accurate and error converges to machine precision.

Convergence results

Assumption: J is convex, f is convex & differentiable, subproblem solutions exist; p^k starts from 0 (equivalent, b^k starts from b).

Theorem (General convergence)

Under the Assumption, $\{u^k\}$ of (**??**) *satisfies*

1. *Monotonic decrease:* $f(u^{k+1}) \leq D(u^{k+1}) + f(u^{k+1}) \leq f(u^k)$.
2. *Convergence: if u^* minimizes f and $J(u) < \infty$, then $f(u^k) \leq f(u^*) + J(u^*)/k$ and, thus, $f(u^k) \to f(u^*)$.*
3. *Denoise b, noise reduction: let $f(\cdot) = f(\cdot; b)$ (e.g., $f(\cdot) = \frac{1}{2}\|A \cdot - b\|_2^2$) and suppose $f(\bar{u}, \bar{b}) = 0$ ($\bar{b}$ and $\bar{u}$ are noiseless input and signal, resp); then $D(\bar{u}, u^{k+1}) < D(\bar{u}, u^k)$ as long as $f(u^{k+1}; b) > f(\bar{u}; b)$.*

RICE

Convergence results

Lemma

If $f(\cdot) = \frac{1}{2}\|A \cdot -b\|_2^2$ *and an* u^k *satisfies* $f(u^k) = 0$, *then* u^k *is a solution of* $\min\{J(u) : f(x) = 0\}$. *This holds even if subproblems are inexactly solved.*

Theorem (Finite convergence for ℓ_1)

Let $J(\cdot) = \mu\|\cdot\|_1$ *and* $f(\cdot) = \frac{1}{2}\|A \cdot -b\|_2^2$. *If* $Ax = b$ *is consistent, then there exists* K *such that any* u^k, $k > K$, *is a solution of* $\min\{J(u) : f(x) = 0\}$.

The theorem extends to any piece-wise linear J.

RICE

Error Cancellation

- Error cancellation is a happy result due to *adding back*!

$$b^k \leftarrow b + (b^{k-1} - Au^k) \tag{2}$$

$$u^{k+1} \leftarrow \min J(u) + \frac{1}{2}\|Au - b^k\|_2^2. \tag{3}$$

- Suppose we computed $u^k_{inexact} = u^k + w^k$. w^k is error.
- (**??**) becomes $b^k_{inexact} \leftarrow b + (b^{k-1} - Au^k_{inexact}) = b^k - Aw^k$.
- (**??**) becomes

$$u^{k+1} \leftarrow \min J(u) + \frac{1}{2}\|A(u + w^k) - b^k\|_2^2$$

which includes the *model error* w^k.

Let w be a model error, and consider

$$\min J(u) + f(u + w). \tag{4}$$

Define

- u_{exact}: exact solution of (**??**)
- $u_{inexact} = u_{exact} + \epsilon$: computed inexact sol of (**??**)
- u_{real}: exact solution of $\min J(u) + f(u)$.

Let w be a model error, and consider

$$\min J(u) + f(u + w). \tag{4}$$

Define

- u_{exact}: exact solution of (**??**)
- $u_{inexact} = u_{exact} + \epsilon$: computed inexact sol of (**??**)
- u_{real}: exact solution of $\min J(u) + f(u)$.

Theorem (General case)

If u_{exact} and $u_{exact} - w$ are on the same face of $\mathrm{graph}(J)$, then

$$u_{inexact} - u_{real} = \epsilon - w.$$

Let w be a model error, and consider

$$\min J(u) + f(u + w). \tag{4}$$

Define

- u_{exact}: exact solution of (**??**)
- $u_{inexact} = u_{exact} + \epsilon$: computed inexact sol of (**??**)
- u_{real}: exact solution of $\min J(u) + f(u)$.

Theorem (General case)

If u_{exact} and $u_{exact} - w$ are on the same face of $\mathrm{graph}(J)$, then

$$u_{inexact} - u_{real} = \epsilon - w.$$

Corollary (J is ℓ_1)

If u_{exact}^{k+1} and $u_{exact}^{k+1} - w^k$ have no opposite signs, then

$$u_{inexact}^{k+1} - u_{real}^{k+1} = \epsilon^{k+1} - w^k.$$

RICE

Theorem Proof:

$$
\begin{aligned}
u_{inexact} &= u_{exact} + \epsilon \\
&= [\operatorname*{argmin}_{u} J(u) + f(u+w)] + \epsilon \\
&\quad \text{(introduce } v := u + w\text{)} \\
&= [\operatorname*{argmin}_{v} J(v-w) + f(v)] - w + \epsilon \\
&\quad \text{(use the condition)} \\
&= [\operatorname*{argmin}_{v} J(v) + f(v)] - w + \epsilon \\
&= u_{real} - w + \epsilon.
\end{aligned}
$$

In other words, $u_{inexact} - u_{real} = \epsilon - w$.

Error Cancellation Example

- u_{real}: 500 entries, 25 nonzero, sparse
- $b = Au_{real}$: 250 linear projections, A has Gaussian random entries
- Recover u_{real} by solving $\min\{\|u\|_1 : Au = b\}$
- Run Bregman, each subproblem inexactly solved with the same tol $\equiv$ 1e-6

  by: FPC, FPC-BB, GPSR, GPSR–BB, or SpaRSA

Error Cancellation Example

- u_{real}: 500 entries, 25 nonzero, sparse
- $b = Au_{real}$: 250 linear projections, A has Gaussian random entries
- Recover u_{real} by solving $\min\{\|u\|_1 : Au = b\}$
- Run Bregman, each subproblem inexactly solved with the same tol $\equiv$ 1e-6

  by: FPC, FPC-BB, GPSR, GPSR–BB, or SpaRSA

ltr k	1	2	3	4	5
$\frac{\|u_{real} - u^k_{inexact}\|}{\|u_{real}\|}$	6.5e-2	2.3e-7	6.2e-14	7.9e-16	5.6e-16.

Relative error converged to the machine precision!

Error Cancellation Example

- u_{real}: 500 entries, 25 nonzero, sparse
- $b = Au_{real}$: 250 linear projections, A has Gaussian random entries
- Recover u_{real} by solving $\min\{\|u\|_1 : Au = b\}$
- Run Bregman, each subproblem inexactly solved with the same tol $\equiv$ 1e-6

  by: FPC, FPC-BB, GPSR, GPSR–BB, or SpaRSA

ltr k	1	2	3	4	5
$\frac{\|u_{real} - u^k_{inexact}\|}{\|u_{real}\|}$	6.5e-2	2.3e-7	6.2e-14	7.9e-16	5.6e-16.

  Relative error converged to the machine precision!

- Classical results require *diminishing tolerances* for convergence, but they are *not* needed for ℓ_1 and above solvers. Why?

RICE

Short Answer:

In $u_{inexact}^{k+1} - u_{real}^{k+1} = \epsilon^{k+1} - w^k$, ϵ^{k+1} almost cancels w^k for all k large.

A Slightly Long Answer:

1. **Finiteness.** With enough accuracy, Bregman u^k reaches the optimal face in finitely many iterations (denoted by K) and stays.
   - The Theorem applies for $k \geq K$.
2. **Error forgetting.** For any $k > K$, two *exact* Bregman iterations yields the global solution. In other words, errors before K can be forgotten.
   - For any $k \geq K$, $u_{real}^{k+1} = u_{real}$, the global solution.
   - $w^k = \epsilon^k$ and, thus, $u_{inexact}^{k+1} - u_{real} = \epsilon^{k+1} - \epsilon^k$.
3. **Convergence.** Use a first-order solver with a fixed tol. Given $\epsilon^{k+1} - \epsilon^k$ is small enough, $\|\epsilon^{k+1} - \epsilon^k\| \to 0$ geometrically in k.
   - $u_{inexact}^k$ converges to u_{real}, the global solution, geometrically in k.

RICE

Generalizations

- Inverse scale space (Burger, Gilboa, Osher, Xu, etc.)
- Linearized Bregman (Yin, Osher, Mao, etc.)
- Logistic Regression (Shi, et al. Rice CAAM 08-08)
- Split Bregman (Goldstein, Osher, UCLA CAM08-29)
- A unified primal-dual framework, BOS (X.Zhang, et al. UCLA CAM09-99)
- More ... People use the words "Bregmanize"

Linearized Bregman

Idea: Linearize the fidelity term at u^k

Work: Y.-Osher-Goldfarb-Darbon 07, Osher-Mao-Dong-Y. 08, Cai-Osher-Shen 08, Y. 09

Linearized Bregman

Idea: Linearize the fidelity term at u^k

Work: Y.-Osher-Goldfarb-Darbon 07, Osher-Mao-Dong-Y. 08, Cai-Osher-Shen 08, Y. 09

- Example: data fitting $= \frac{1}{2}\|Au - b\|_2^2$

$$u^{k+1} \leftarrow \min_u D(u, u^k) + \langle A^\top (Au^k - b), u\rangle + \frac{1}{2\delta}\left\|u - u^k\right\|_2^2$$

Linearized Bregman

Idea: Linearize the fidelity term at u^k

Work: Y.-Osher-Goldfarb-Darbon 07, Osher-Mao-Dong-Y. 08, Cai-Osher-Shen 08, Y. 09

- Example: data fitting $= \frac{1}{2}\|Au - b\|_2^2$

$$u^{k+1} \leftarrow \min_u D(u, u^k) + \langle A^\top(Au^k - b), u\rangle + \frac{1}{2\delta}\left\|u - u^k\right\|_2^2$$

- For $D(u, u^k)$ induced by $J(u) = \mu\|u\|_1$, iterations become

$$\begin{aligned} u^{k+1} &\leftarrow \delta \ \mathrm{shrink}(v^k, \mu) \\ v^{k+1} &\leftarrow v^k + A^\top(b - Au^{k+1}). \end{aligned}$$

RICE

Linearized Bregman

Idea: Linearize the fidelity term at u^k

Work: Y.-Osher-Goldfarb-Darbon 07, Osher-Mao-Dong-Y. 08, Cai-Osher-Shen 08, Y. 09

- Example: data fitting $= \frac{1}{2}\|Au - b\|_2^2$

$$u^{k+1} \leftarrow \min_u D(u, u^k) + \langle A^\top(Au^k - b), u\rangle + \frac{1}{2\delta}\left\|u - u^k\right\|_2^2$$

- For $D(u, u^k)$ induced by $J(u) = \mu\|u\|_1$, iterations become

$$\begin{aligned} u^{k+1} &\leftarrow \delta \ \mathrm{shrink}(v^k, \mu) \\ v^{k+1} &\leftarrow v^k + A^\top(b - Au^{k+1}). \end{aligned}$$

- Application: non-negative least–squares, matrix completion

RICE

Linearized Bregman, Cont'd

Properties:

- gradient-ascend the dual of $\min\{\mu\|u\|_1 + \frac{1}{2\delta}\|u\|^2 : Au = b\}$
- Exact regularization: $\exists\bar{\delta}$: if $\delta > \bar{\delta}$, then solves $\min\{\|u\|_1 : Au = b\}$
- Empirically, # nonzeros of u^k grows monotonically in k

Yin, W. Analysis and Generalizations of the Linearized Bregman Method, Rice CAAM Report TR09-02. [link]

RICE

Operator Splitting ADM

Operator splitting + ADM gives Split Bregman (Goldstein–Osher 08)

- Operator splitting by (Wang–Yang–Y.–Zhang 07,08): Split $TV(u)$ to Du and $\sum_i \|(\cdot)_i\|$. Great payoff for many imaging problems.
- Apply the alternating direction of multipliers (ADM) method to above splitting.

Operator Splitting ADM

Alternating direction method: (Douglas–Rachford 60s, Glowinski–Marocco, Gabay–Mercier, 70s)

1. fix u, minimize w.r.t. v
2. fix v, minimize w.r.t. u
3. update λ

Example (Wang–Yang–Y.–Zhang 07,08) Compressed MRI, image deblr

$$\min_u \mu TV(u) + \frac{1}{2}\|Au - b\|_2^2 \Leftrightarrow \min_u \{\mu\|w\|_1 + \frac{1}{2}\|Au - b\|_2^2 : w = Du\}$$

where A is partial Fourier or convolution. ADM extends to color images, duals, rank-minimization

Summary

1. Bregman improves ℓ_1-like regularization quality for noisy data
2. Bregman applied to constrained ($Au = b$) minimization is not new but is fast and accurate due *adding back*
3. Various extensions take advantages of model structures

More details and solvers at *Rice L_1-Related Optimization Project*

RICE