The Abdus Salam International Centre for Theoretical Physics
50th Anniversary 1964-2014

2585-28

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications

2 - 20 June 2014

Numerical algorithms for sparse recovery contd.

I. Loris
ULB, Brussels
Belgium

Numerical algorithms for sparse recovery (part 2)

Ignace Loris

Université Libre de Bruxelles

Coherent state transforms,
time-frequency and time-scale analysis, applications

Trieste, Italy, June 2–21, 2014

Summary of part 1

- Sparse (approximate) solution to linear equations $Ku = y$ by solving

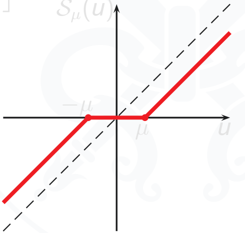
$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1$$

- Iterative soft-thresholding algorithm:

$$u_{n+1} = \mathcal{S}_{\alpha\mu}\left[u_n + \alpha K^T(y - Ku_n)\right] \qquad (1)$$

  with $\mathcal{S}_\mu$ = component-wise soft-thresholding:

$$\mathcal{S}_\mu(u) = \begin{cases} u - \lambda & u \geq \mu \\ 0 & |u| \leq \mu \\ u + \lambda & u \leq -\mu \end{cases}$$



  converges to $\hat{u}$ if $\alpha < 2/\|K\|^2$.

- u_n and $\hat{u}$ will have many components equal to zero ('sparsity')
- soft-thresholding is the 'proximal operator' of $\|u\|_1$

Summary of part 1

- Sparse (approximate) solution to linear equations $Ku = y$ by solving

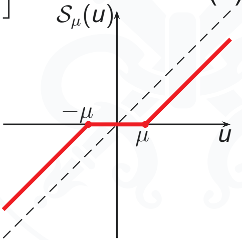
$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1$$

- Iterative soft-thresholding algorithm:

$$u_{n+1} = \mathcal{S}_{\alpha\mu}\left[u_n + \alpha K^T(y - Ku_n)\right] \qquad (1)$$

with $\mathcal{S}_\mu =$ component-wise soft-thresholding:

$$\mathcal{S}_\mu(u) = \begin{cases} u - \lambda & u \geq \mu \\ 0 & |u| \leq \mu \\ u + \lambda & u \leq -\mu \end{cases}$$



converges to $\hat{u}$ if $\alpha < 2/\|K\|^2$.

- u_n and $\hat{u}$ will have many components equal to zero ('sparsity')
- soft-thresholding is the 'proximal operator' of $\|u\|_1$

Summary of part 1

- Sparse (approximate) solution to linear equations $Ku = y$ by solving

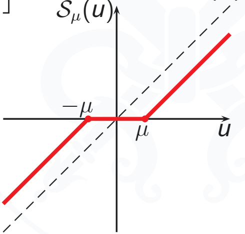
$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1$$

- Iterative soft-thresholding algorithm:

$$u_{n+1} = \mathcal{S}_{\alpha\mu}\left[u_n + \alpha K^T(y - Ku_n)\right] \qquad (1)$$

with $\mathcal{S}_\mu$ = component-wise soft-thresholding:

$$\mathcal{S}_\mu(u) = \begin{cases} u - \lambda & u \geq \mu \\ 0 & |u| \leq \mu \\ u + \lambda & u \leq -\mu \end{cases}$$



converges to $\hat{u}$ if $\alpha < 2/\|K\|^2$.

- u_n and $\hat{u}$ will have many components equal to zero ('sparsity')
- soft-thresholding is the 'proximal operator' of $\|u\|_1$

Summary of part 1

- Sparse (approximate) solution to linear equations $Ku = y$ by solving

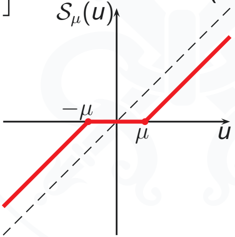
$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1$$

- Iterative soft-thresholding algorithm:

$$u_{n+1} = \mathcal{S}_{\alpha\mu}\left[u_n + \alpha K^T(y - Ku_n)\right] \qquad (1)$$

with $\mathcal{S}_\mu =$ component-wise soft-thresholding:

$$\mathcal{S}_\mu(u) = \begin{cases} u - \lambda & u \geq \mu \\ 0 & |u| \leq \mu \\ u + \lambda & u \leq -\mu \end{cases}$$



converges to $\hat{u}$ if $\alpha < 2/\|K\|^2$.

- u_n and $\hat{u}$ will have many components equal to zero ('sparsity')
- soft-thresholding is the 'proximal operator' of $\|u\|_1$

Problem statement

- Before: Sparse (approximate) solution to linear equations $Ku = y$ by solving

$$\arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1 \qquad (2)$$

- Now: instead of u sparse, we require Au sparse: many $(Au)_i = 0$, where A is a linear operator ('analysis style sparsity')
- (approximate) solution $\hat{u}$ to linear equations $Ku = y$ with Au sparse by solving:

$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|Au\|_1 \qquad (3)$$

- A is not necessarily invertible
- (2) is a special case of (3): $A = \mathrm{Id}$ or change of variables if $\exists A^{-1}$
- proximal operator of $\mu\|Au\|_1$ does not have a simple expression in general, otherwise one could use:

$$u_{n+1} = \mathrm{prox}_{\alpha\mu\|A\cdot\|_1}\left[u_n + \alpha K^T(y - Ku_n)\right]$$

Problem statement

- Before: Sparse (approximate) solution to linear equations $Ku = y$ by solving
$$\arg\min_u \frac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1 \quad (2)$$
- Now: instead of u sparse, we require Au sparse: many $(Au)_i = 0$, where A is a linear operator ('analysis style sparsity')
- (approximate) solution $\hat{u}$ to linear equations $Ku = y$ with Au sparse by solving:
$$\hat{u} = \arg\min_u \frac{1}{2}\|Ku - y\|_2^2 + \mu\|Au\|_1 \quad (3)$$
- A is not necessarily invertible
- (2) is a special case of (3): $A = \mathrm{Id}$ or change of variables if $\exists A^{-1}$
- proximal operator of $\mu\|Au\|_1$ does not have a simple expression in general, otherwise one could use:
$$u_{n+1} = \mathrm{prox}_{\alpha\mu\|A\cdot\|_1}\left[u_n + \alpha K^T(y - Ku_n)\right]$$

Problem statement

- Before: Sparse (approximate) solution to linear equations $Ku = y$ by solving
$$\arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1 \qquad (2)$$

- Now: instead of u sparse, we require Au sparse: many $(Au)_i = 0$, where A is a linear operator ('analysis style sparsity')
- (approximate) solution $\hat{u}$ to linear equations $Ku = y$ with Au sparse by solving:
$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|Au\|_1 \qquad (3)$$
- A is not necessarily invertible
- (2) is a special case of (3): $A = \mathrm{Id}$ or change of variables if $\exists A^{-1}$
- proximal operator of $\mu\|Au\|_1$ does not have a simple expression in general, otherwise one could use:
$$u_{n+1} = \mathrm{prox}_{\alpha\mu\|A\cdot\|_1}\left[u_n + \alpha K^T(y - Ku_n)\right]$$

Problem statement

- Before: Sparse (approximate) solution to linear equations $Ku = y$ by solving
$$\arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1 \tag{2}$$
- Now: instead of u sparse, we require Au sparse: many $(Au)_i = 0$, where A is a linear operator ('analysis style sparsity')
- (approximate) solution $\hat{u}$ to linear equations $Ku = y$ with Au sparse by solving:
$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \mu\|Au\|_1 \tag{3}$$
- A is not necessarily invertible
- (2) is a special case of (3): $A = \mathrm{Id}$ or change of variables if $\exists A^{-1}$
- proximal operator of $\mu\|Au\|_1$ does not have a simple expression in general, otherwise one could use:
$$u_{n+1} = \mathrm{prox}_{\alpha\mu\|A\cdot\|_1}\left[u_n + \alpha K^T(y - Ku_n)\right]$$

Goal (1)

- Want an explicit iterative algorithm for:

$$\arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1$$

- for a generic A, i.e. **no** A^{-1}
- explicit: uses only K, K^T, A and A^T at every step,
  No equation solving at every step
  No non-trivial sub-problem at every step
  No smoothing of type : $\|Au\|_1 \approx \sum_i \sqrt{(Au)_i^2 + \epsilon^2}$,
- Guide: new algorithm should reduce to (1) if $A = 1$

  This is a choice, not a mathematical requirement

Goal (2)

- More generally: we want an explicit iterative algorithm for:

$$\arg\min_u f(u) + g(Au)$$

  where f is convex, differentiable with Lipschitz continuous gradient (L), g is convex, A is a linear map

- in terms of ∇f, prox_g (as before) and A, A^T,
- WITHOUT using $\text{prox}_{g(A\,\cdot)}$,
- and that reduces to the proximal gradient algorithm:

$$u_{n+1} = \text{prox}_{\alpha_n g}\left(u_n - \alpha_n \nabla f(u_n)\right) \tag{4}$$

  when $A = \text{Id}$.

Variational equations for $\hat{u} = \arg\min_u f(u) + g(Au)$

- Introducing $\tilde{g} = \alpha g(\beta^{-1}\cdot)$, we can equivalently write:

$$\arg\min_u \alpha f(u) + \tilde{g}(\beta Au)$$

- The variational equations are:

$$\alpha \nabla f(u) + \beta A^T w = 0 \quad \text{with} \quad w \in \partial \tilde{g}(\beta Au)$$

- This can be expressed as:

$$\alpha \nabla f(u) + \beta A^T w = 0 \quad \text{and} \quad \beta Au = \operatorname{prox}_{\tilde{g}}(w + \beta Au)$$

  Would like to write this as a fixed point equation.

- Introduce the function $\operatorname{prox}_{\tilde{g}} \equiv \mathrm{Id} - \operatorname{prox}_{\tilde{g}^*}$:

$$u = u - \alpha \nabla f(u) - \beta A^T w \quad \text{and} \quad \beta Au = w + \beta Au - \operatorname{prox}_{\tilde{g}^*}(w + \beta Au)$$

$$\Updownarrow$$

$$u = u - \alpha \nabla f(u) - \beta A^T w \quad \text{and} \quad w = \operatorname{prox}_{\tilde{g}^*}(w + \beta Au)$$

- NB: $\mathrm{Id} - \operatorname{prox}_{\tilde{g}}$ really is the proximal operator of a convex function $\tilde{g}^*$ ($\tilde{g}^*$ is the “convex conjugate” of $\tilde{g}$: $\tilde{g}^*(u) = \sup_v \langle u, v\rangle - \tilde{g}(v)$).

Iterative algorithms

- Variational equations are:

$$\hat{u} = \arg\min_u f(u) + g(Au) \Leftrightarrow \begin{cases} \hat{u} & = & \hat{u} - \alpha \nabla f(\hat{u}) - \beta A^T \hat{w} \\ \hat{w} & = & \mathrm{prox}_{\tilde{g}^*}(\hat{w} + \beta A \hat{u}) \end{cases} \quad (5)$$

  Need iterative algorithm for solving these equations.

- Now: make a choice and prove convergence.

- For example, one could set:

$$\begin{cases} w_{n+1} & = & \mathrm{prox}_{\tilde{g}^*}(w_n + \beta A u_n) \\ u_{n+1} & = & u_n - \alpha \nabla f(u_n) - \beta A^T w_{n+1} \end{cases}$$

  and study convergence.

- However this algorithm does not reduce to the proximal gradient algorithm (4) when $A = 1$ (variable w remains).

Generalized proximal gradient algorithm (1)

- Choice: Use a predict-correct step on the u variable
- By introducing $\bar{u}_{n+1}$:

$$\begin{cases} \bar{u}_{n+1} &= u_n - \alpha \nabla f(u_n) - \beta A^T w_n \\ w_{n+1} &= \mathrm{prox}_{\tilde{g}^*}(w_n + \beta A \bar{u}_{n+1}) \\ u_{n+1} &= u_n - \alpha \nabla f(u_n) - \beta A^T w_{n+1} \end{cases}$$

- Now, if $A = \mathrm{Id}$ and $\beta = 1$:

$$\begin{aligned} u_{n+1} &= u_n - \alpha \nabla f(u_n) - w_{n+1} \\ &= u_n - \alpha \nabla f(u_n) - \mathrm{prox}_{\tilde{g}^*}(w_n + \bar{u}_{n+1}) \\ &= u_n - \alpha \nabla f(u_n) - \mathrm{prox}_{\tilde{g}^*}(w_n + u_n - \alpha \nabla f(u_n) - w_n) \\ &= u_n - \alpha \nabla f(u_n) - \mathrm{prox}_{\tilde{g}^*}(u_n - \alpha \nabla f(u_n)) \\ &= \mathrm{prox}_{\tilde{g}}(u_n - \alpha \nabla f(u_n)) = \mathrm{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n)) \end{aligned}$$

  where we used $\mathrm{Id} - \mathrm{prox}_{\tilde{g}^*} = \mathrm{prox}_{\tilde{g}} = \mathrm{prox}_{\alpha g}$
- Similar reduction in case A orthogonal.

Generalized proximal gradient algorithm (1)

- Choice: Use a predict-correct step on the u variable
- By introducing $\bar{u}_{n+1}$:

$$\begin{cases} \bar{u}_{n+1} & = & u_n - \alpha \nabla f(u_n) - \beta A^T w_n \\ w_{n+1} & = & \operatorname{prox}_{\tilde{g}^*}(w_n + \beta A \bar{u}_{n+1}) \\ u_{n+1} & = & u_n - \alpha \nabla f(u_n) - \beta A^T w_{n+1} \end{cases}$$

- Now, if $A = \mathrm{Id}$ and $\beta = 1$:

$$\begin{aligned} u_{n+1} &= u_n - \alpha \nabla f(u_n) - w_{n+1} \\ &= u_n - \alpha \nabla f(u_n) - \operatorname{prox}_{\tilde{g}^*}(w_n + \bar{u}_{n+1}) \\ &= u_n - \alpha \nabla f(u_n) - \operatorname{prox}_{\tilde{g}^*}(w_n + u_n - \alpha \nabla f(u_n) - w_n) \\ &= u_n - \alpha \nabla f(u_n) - \operatorname{prox}_{\tilde{g}^*}(u_n - \alpha \nabla f(u_n)) \\ &= \operatorname{prox}_{\tilde{g}}(u_n - \alpha \nabla f(u_n)) = \operatorname{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n)) \end{aligned}$$

where we used $\mathrm{Id} - \operatorname{prox}_{\tilde{g}^*} = \operatorname{prox}_{\tilde{g}} = \operatorname{prox}_{\alpha g}$

- Similar reduction in case A orthogonal.

Generalized proximal gradient algorithm (1)

- Choice: Use a predict-correct step on the u variable
- By introducing $\bar{u}_{n+1}$:

$$\begin{cases} \bar{u}_{n+1} & = & u_n - \alpha \nabla f(u_n) - \beta A^T w_n \\ w_{n+1} & = & \text{prox}_{\tilde{g}^*}(w_n + \beta A \bar{u}_{n+1}) \\ u_{n+1} & = & u_n - \alpha \nabla f(u_n) - \beta A^T w_{n+1} \end{cases}$$

- Now, if $A = \text{Id}$ and $\beta = 1$:

$$\begin{aligned} u_{n+1} &= u_n - \alpha \nabla f(u_n) - w_{n+1} \\ &= u_n - \alpha \nabla f(u_n) - \text{prox}_{\tilde{g}^*}(w_n + \bar{u}_{n+1}) \\ &= u_n - \alpha \nabla f(u_n) - \text{prox}_{\tilde{g}^*}(w_n + u_n - \alpha \nabla f(u_n) - w_n) \\ &= u_n - \alpha \nabla f(u_n) - \text{prox}_{\tilde{g}^*}(u_n - \alpha \nabla f(u_n)) \\ &= \text{prox}_{\tilde{g}}(u_n - \alpha \nabla f(u_n)) = \text{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n)) \end{aligned}$$

  where we used $\text{Id} - \text{prox}_{\tilde{g}^*} = \text{prox}_{\tilde{g}} = \text{prox}_{\alpha g}$

- Similar reduction in case A orthogonal.

Generalized proximal gradient algorithm (2)

- Iterative algorithm with variable step lengths:

$$\begin{cases} \bar{u}_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_n \\ w_{n+1} &= \text{prox}_{\tilde{g}_n^*} \left[w_n + \beta_n A \bar{u}_{n+1} \right] \\ u_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_{n+1} \end{cases}$$

where $\tilde{g}_n(\cdot) = \alpha_n g(\beta_n^{-1} \cdot)$

- We will study convergence of:

$$\begin{cases} \bar{u}_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_n \\ \tilde{w}_{n+1} &= \text{prox}_{\tilde{g}_n^*} \left[w_n + \beta_n A \bar{u}_{n+1} \right] \\ \tilde{u}_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T \tilde{w}_{n+1} \\ w_{n+1} &= (1-\lambda_n) w_n + \lambda_n \tilde{w}_{n+1} \\ u_{n+1} &= (1-\lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{cases}$$

Generalized proximal gradient algorithm (2)

- Iterative algorithm with variable step lengths:

$$\begin{cases} \bar{u}_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_n \\ w_{n+1} &= \operatorname{prox}_{\tilde{g}_n^*} [w_n + \beta_n A \bar{u}_{n+1}] \\ u_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_{n+1} \end{cases}$$

  where $\tilde{g}_n(\cdot) = \alpha_n g(\beta_n^{-1} \cdot)$

- We will study convergence of:

$$\begin{cases} \bar{u}_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_n \\ \tilde{w}_{n+1} &= \operatorname{prox}_{\tilde{g}_n^*} [w_n + \beta_n A \bar{u}_{n+1}] \\ \tilde{u}_{n+1} &= u_n - \alpha_n \nabla f(u_n) - \beta_n A^T \tilde{w}_{n+1} \\ w_{n+1} &= (1 - \lambda_n) w_n + \lambda_n \tilde{w}_{n+1} \\ u_{n+1} &= (1 - \lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{cases}$$

Generalized proximal gradient algorithm

Theorem (generalized proximal gradient algorithm [1, 3, 6])

Let $\epsilon > 0$. *IF* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex with Lipschitz continuous gradient (L),* $g : \mathbb{R}^{d'} \to \mathbb{R}$ *is convex, proper, lower semi-continuous,* $A : \mathbb{R}^d \to \mathbb{R}^{d'}$ *is a linear map, and a minimizer of* $F(u) = f(u) + g(Au)$ *exists, THEN the iterative algorithm:*

$$\left\{ \begin{array}{lcl} \bar{u}_{n+1} & = & u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_n \\ \tilde{w}_{n+1} & = & \mathrm{prox}_{\tilde{g}_n^*}\left[w_n + \beta_n A \bar{u}_{n+1}\right] \\ \tilde{u}_{n+1} & = & u_n - \alpha_n \nabla f(u_n) - \beta_n A^T \tilde{w}_{n+1} \\ w_{n+1} & = & (1-\lambda_n) w_n + \lambda_n \tilde{w}_{n+1} \\ u_{n+1} & = & (1-\lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{array} \right. \tag{6}$$

with $u_0 =$ *arbitrary,* $\epsilon \le \alpha_n \le 2/L - \epsilon$, $\epsilon \le \beta_1 \le \beta_2 \le \ldots \le 1/\|A\| - \epsilon$ *and* $\epsilon \le \lambda_n \le 1$ *converges to a minimizer of* $F(u)$.

Proof of convergence

Let $\hat{u} \in \arg\min_u f(u) + g(Au)$, i.e.:

$$\begin{cases} \hat{u} &= \hat{u} - \alpha \nabla f(\hat{u}) - \beta A^T \hat{w} \\ \hat{w} &= \operatorname{prox}_{\tilde{g}^*} (\hat{w} + \beta A \hat{u}) \end{cases}$$

for some $\alpha, \beta > 0$ and where $\tilde{g}(\cdot) = \alpha g(\beta^{-1} \cdot)$.

Convexity of $\|w - \hat{w}\|_2^2$ and $w_{n+1} = (1 - \lambda_n) w_n + \lambda_n \tilde{w}_{n+1}$ implies:

$$\|w_{n+1} - \hat{w}\|_2^2 \leq (1 - \lambda_n) \|w_n - \hat{w}\|_2^2 + \lambda_n \|\tilde{w}_{n+1} - \hat{w}\|_2^2 \quad (7)$$

Convexity of $\|u - \hat{u}\|_2^2$ and $u_{n+1} = (1 - \lambda_n) u_n + \lambda_n \tilde{u}_{n+1}$ implies:

$$\|u_{n+1} - \hat{u}\|_2^2 = (1 - \lambda_n) \|u_n - \hat{u}\|_2^2 + \lambda_n \|\tilde{u}_{n+1} - \hat{u}\|_2^2 - \lambda_n (1 - \lambda_n) \|u_n - \tilde{u}_{n+1}\|_2^2 \quad (8)$$

Recall property of proximal operators

If $t^+ = \text{prox}_g(t^- + \Delta)$ then:

$$\|t^+ - t\|_2^2 \leq \|t^- - t\|_2^2 - \|t^+ - t^-\|_2^2 + 2\langle t^+ - t, \Delta\rangle + 2g(t) - 2g(t^+) \quad (9)$$

for all t.

We will use this property on the (iteration) relation:

$$\tilde{w}_{n+1} = \text{prox}_{\tilde{g}_n^*}\left[w_n + \beta_n A\bar{u}_{n+1}\right]$$

with $t^+ = \tilde{w}_{n+1}, t^- = w_n, t = \hat{w}, \Delta = \beta_n A\bar{u}_n$, and for $\tilde{g}_n^*$ instead of g

and on the fixed-point relation:

$$\hat{w} = \text{prox}_{\tilde{g}_n^*}\left[\hat{w} + \beta_n A\hat{u}\right]$$

with $t^+ = \hat{w}, t^- = \hat{w}, t = \tilde{w}_{n+1}, \Delta = \beta_n A\hat{u}$, and for $\tilde{g}_n^*$ instead of g.

Proof of convergence

It follows from $\tilde{w}_{n+1} = \text{prox}_{\tilde{g}_n^*}[w_n + \beta_n A\bar{u}_{n+1}]$ and eq. (9) with $t^+ = \tilde{w}_{n+1}, t^- = w_n, t = \hat{w}, \Delta = \beta_n A\bar{u}_n$ that:

$$\begin{aligned} \|\tilde{w}_{n+1} - \hat{w}\|_2^2 &\leq \|w_n - \hat{w}\|_2^2 - \|\tilde{w}_{n+1} - w_n\|_2^2 + 2\langle \tilde{w}_{n+1} - \hat{w}, \beta_n A\bar{u}_{n+1}\rangle \\ &\quad +2\tilde{g}_n^*(\hat{w}) - 2\tilde{g}_n^*(\tilde{w}_{n+1}) \end{aligned}$$

It follows from $\hat{w} = \text{prox}_{\tilde{g}_n^*}[\hat{w} + \beta_n A\hat{u}]$ and eq. (9) with $t^+ = \hat{w}, t^- = \hat{w}, t = \tilde{w}_{n+1}, \Delta = \beta_n A\hat{u}$ that:

$$\begin{aligned} \cancel{\|\hat{w} - \tilde{w}_{n+1}\|_2^2} &\leq \cancel{\|\hat{w} - \tilde{w}_{n+1}\|_2^2} - \cancel{\|\hat{w} - \hat{w}\|_2^2} + 2\langle \hat{w} - \tilde{w}_{n+1}, \beta_n A\hat{u}\rangle \\ &\quad +2\tilde{g}_n^*(\tilde{w}_{n+1}) - 2\tilde{g}_n^*(\hat{w}) \end{aligned}$$

Together (and using $\bar{u}_{n+1} = \tilde{u}_{n+1} - \beta_n A^T(w_n - \tilde{w}_{n+1})$):

$$\begin{aligned} \|\tilde{w}_{n+1} - \hat{w}\|_2^2 &\leq \|w_n - \hat{w}\|_2^2 - \|\tilde{w}_{n+1} - w_n\|_2^2 \\ &\quad +2\beta_n\langle \tilde{w}_{n+1} - \hat{w}, A(\bar{u}_{n+1} - \hat{u})\rangle \\ &= \|w_n - \hat{w}\|_2^2 - \|\tilde{w}_{n+1} - w_n\|_2^2 \\ &\quad +2\beta_n\langle \tilde{w}_{n+1} - \hat{w}, A(\tilde{u}_{n+1} - \hat{u} - \beta_n A^T(w_n - \tilde{w}_{n+1}))\rangle \end{aligned} \tag{10}$$

Proof of convergence

It follows from $\tilde{u}_{n+1} = u_n - \alpha_n \nabla f(u_n) - \beta_n A^T \tilde{w}_{n+1}$ that:

$$\|\tilde{u}_{n+1} - \hat{u}\|_2^2 \overset{(=)}{\leq} \|u_n - \hat{u}\|_2^2 - \|\tilde{u}_{n+1} - u_n\|_2^2 + 2\langle \tilde{u}_{n+1} - \hat{u}, -\alpha_n \nabla f(u_n) - \beta_n A^T \tilde{w}_{n+1}\rangle$$

It follows from $\hat{u} = \hat{u} - \alpha_n \nabla f(\hat{u}) - \beta_n A^T \hat{w}$ that:

$$\|\hat{u} - \tilde{u}_{n+1}\|_2^2 \overset{(=)}{\leq} \|\hat{u} - \tilde{u}_{n+1}\|_2^2 - \|\hat{u} - \hat{u}\|_2^2 + 2\langle \hat{u} - \tilde{u}_{n+1}, -\alpha_n \nabla f(\hat{u}) - \beta_n A^T \hat{w}\rangle$$

Together:

$$\begin{aligned}\|\tilde{u}_{n+1} - \hat{u}\|_2^2 &\leq \|u_n - \hat{u}\|_2^2 - \|\tilde{u}_{n+1} - u_n\|_2^2 \\ &\quad + 2\alpha_n \langle \hat{u} - \tilde{u}_{n+1}, \nabla f(u_n) - \nabla f(\hat{u})\rangle \\ &\quad - 2\beta_n \langle \tilde{u}_{n+1} - \hat{u}, A^T(\tilde{w}_{n+1} - \hat{w})\rangle \\ &\overset{\text{(see part 1)}}{\leq} \|u_n - \hat{u}\|_2^2 - \|\tilde{u}_{n+1} - u_n\|_2^2 + 2\alpha_n \tfrac{L}{4} \|u_n - \tilde{u}_{n+1}\|_2^2 \\ &\quad - 2\beta_n \langle \tilde{u}_{n+1} - \hat{u}, A^T(\tilde{w}_{n+1} - \hat{w})\rangle \end{aligned} \tag{11}$$

Proof of convergence

Adding inequalities (10) and (11) yields:

$$
\begin{aligned}
\|\tilde{u}_{n+1} - \hat{u}\|_2^2 + \|\tilde{w}_{n+1} - \hat{w}\|_2^2 \quad \leq \quad & \|u_n - \hat{u}\|_2^2 + \|w_n - \hat{w}\|_2^2 \\
& -(1 - \tfrac{\alpha_n L}{2})\|\tilde{u}_{n+1} - u_n\|_2^2 - \|\tilde{w}_{n+1} - w_n\|_2^2 \\
& -2\beta_n^2 \langle A^T(\tilde{w}_{n+1} - \hat{w}), A^T(w_n - \tilde{w}_{n+1})\rangle
\end{aligned}
$$

The inner product $2\langle A^T(\tilde{w}_{n+1} - \hat{w}), A^T(w_n - \tilde{w}_{n+1})\rangle$ equals $\|A^T(w_n - \hat{w})\|_2^2 - \|A^T(\tilde{w}_{n+1} - \hat{w})\|_2^2 - \|A^T(w_n - \tilde{w}_{n+1})\|_2^2$ and hence:

$$
\begin{aligned}
& \|\tilde{u}_{n+1} - \hat{u}\|_2^2 + \|\tilde{w}_{n+1} - \hat{w}\|_2^2 - \beta_n^2\|A^T(\tilde{w}_{n+1} - \hat{w})\|_2^2 \\
& \qquad \leq \|u_n - \hat{u}\|_2^2 + \|w_n - \hat{w}\|_2^2 - \beta_n^2\|A^T(w_n - \hat{w})\|_2^2 \\
& \qquad -(1 - \tfrac{\alpha_n L}{2})\|\tilde{u}_{n+1} - u_n\|_2^2 - \|\tilde{w}_{n+1} - w_n\|_2^2 + \beta_n^2\|A^T(w_n - \tilde{w}_{n+1})\|_2^2
\end{aligned}
$$

If $\beta_n \leq \beta_{n+1}$ then $-\beta_{n+1}^2\|A^T(\tilde{w}_{n+1} - \hat{w})\|_2^2 \leq -\beta_n^2\|A^T(\tilde{w}_{n+1} - \hat{w})\|_2^2$ such that:

Proof of convergence

$$\|\tilde{u}_{n+1} - \hat{u}\|_2^2 + \|\tilde{w}_{n+1} - \hat{w}\|_{\beta_{n+1}A}^2 \leq \|u_n - \hat{u}\|_2^2 + \|w_n - \hat{w}\|_{\beta_n A}^2 - (1 - \frac{\alpha_n L}{2})\|\tilde{u}_{n+1} - u_n\|_2^2 - \|\tilde{w}_{n+1} - w_n\|_{\beta_n A}^2 \quad (12)$$

where $\|w\|_{\beta_n A}^2 = \|w\|_2^2 - \beta_n^2 \|A^T w\|_2^2$ (a norm).
Combining inequalities (7), (8) and (12) yields:

$$\|u_{n+1} - \hat{u}\|_2^2 + \|w_{n+1} - \hat{w}\|_{\beta_{n+1}A}^2 \leq \|u_n - \hat{u}\|_2^2 + \|w_n - \hat{w}\|_{\beta_n A}^2 - \lambda_n \left[(1 - \frac{\alpha_n L}{2}) + (1 - \lambda_n)\right] \|\tilde{u}_{n+1} - u_n\|_2^2 - \lambda_n \|\tilde{w}_{n+1} - w_n\|_{\beta_n A}^2$$

With $\epsilon \leq \lambda_n \leq 1$ and $\alpha_n \leq 2/L - \epsilon$, it follows that:

$$\|u_{n+1} - \hat{u}\|_2^2 + \|w_{n+1} - \hat{w}\|_{\beta_{n+1}A}^2 \leq \|u_n - \hat{u}\|_2^2 + \|w_n - \hat{w}\|_{\beta_n A}^2 - c_1 \|\tilde{u}_{n+1} - u_n\|_2^2 - c_2 \|\tilde{w}_{n+1} - w_n\|_{\beta_n A}^2 \quad (13)$$

for some $c_1, c_2 > 0$.

Proof of convergence

It follows that:

$$\begin{aligned} \|u_{n+1}-\hat{u}\|_2^2 + \|w_{n+1}-\hat{w}\|^2_{\beta_{n+1}A} &\leq \|u_n-\hat{u}\|_2^2 + \|w_n-\hat{w}\|^2_{\beta_n A} \\ &\leq \dots \\ &\leq \|u_0-\hat{u}\|_2^2 + \|w_0-\hat{w}\|^2_{\beta_0 A} = C \end{aligned}$$

With $\bar{\beta} = \sup_n \beta_n < 1/||A||$ and $\|w\|^2_{\bar{\beta}A} = \|w\|_2^2 - \bar{\beta}^2\|A^T w\|_2^2$, one finds:

$$\|u_{n+1}-\hat{u}\|_2^2 + \|w_{n+1}-\hat{w}\|^2_{\bar{\beta}A} \overset{\beta_{n+1}\leq\bar{\beta}}{\leq} \|u_{n+1}-\hat{u}\|_2^2 + \|w_{n+1}-\hat{w}\|^2_{\beta_{n+1}A} \leq C$$

i.e. (u_n, w_n) is bounded.
As $(\alpha_n)_n$ and $(\beta_n)_n$ are also bounded, there exists a common converging subsequence:

$$u_{n_j} \overset{j\to\infty}{\longrightarrow} u^\dagger, \quad w_{n_j} \overset{j\to\infty}{\longrightarrow} w^\dagger, \quad \alpha_{n_j} \overset{j\to\infty}{\longrightarrow} \alpha > 0, \quad \beta_{n_j} \overset{j\to\infty}{\longrightarrow} \beta > 0$$

Proof of convergence

Inequality (13) also implies ($N \geq M$):

$$\begin{aligned}
&\sum_{n=M}^{N-1} c_1 \|\tilde{u}_{n+1} - u_n\|_2^2 + c_2 \|\tilde{w}_{n+1} - w_n\|_{\beta_n A}^2 \\
&\leq \sum_{n=M}^{N-1} \|u_n - \hat{u}\|_2^2 + \|w_n - \hat{w}\|_{\beta_n A}^2 - \|u_{n+1} - \hat{u}\|_2^2 - \|w_{n+1} - \hat{w}\|_{\beta_{n+1} A}^2 \\
&= \|u_M - \hat{u}\|_2^2 + \|w_M - \hat{w}\|_{\beta_M A}^2 - \|u_N - \hat{u}\|_2^2 - \|w_N - \hat{w}\|_{\beta_N A}^2 \\
&\leq \|u_M - \hat{u}\|_2^2 + \|w_M - \hat{w}\|_{\beta_M A}^2
\end{aligned} \tag{14}$$

This means that $\|\tilde{u}_{n+1} - u_n\|_2 \xrightarrow{n\to\infty} 0$ and thus: $\tilde{u}_{n_j+1} \xrightarrow{j\to\infty} u^\dagger$.

It also implies that $\|\tilde{w}_{n+1} - w_n\|_{\bar{\beta} A}^2 \leq \|\tilde{w}_{n+1} - w_n\|_{\beta_n A}^2 \xrightarrow{n\to\infty} 0$, and thus: $\tilde{w}_{n_j+1} \xrightarrow{j\to\infty} w^\dagger$.

Using $\bar{u}_{n+1} = \tilde{u}_{n+1} - \beta_n A^T(w_n - \tilde{w}_{n+1}))$, one also finds that $\bar{u}_{n_j+1} \xrightarrow{j\to\infty} u^\dagger$.

Proof of convergence

But as

$$\begin{cases} \bar{u}_{n_j+1} & = & u_{n_j} - \alpha_{n_j} \nabla f(u_{n_j}) - \beta_{n_j} A^T w_{n_j} \\ \tilde{w}_{n_j+1} & = & \text{prox}_{\tilde{g}^*_{n_j}} \left[w_{n_j} + \beta_{n_j} A \bar{u}_{n_j+1} \right] \end{cases}$$

one finds ($j \to \infty$) by continuity of ∇f and prox (2x !) that:

$$\begin{cases} u^\dagger & = & u^\dagger - \alpha \nabla f(u^\dagger) - \beta A^T w^\dagger \\ w^\dagger & = & \text{prox}_{\tilde{g}^*} \left[w^\dagger + \beta A u^\dagger \right] \end{cases}$$

i.e. $u^\dagger$ is a minimizer of $f(u) + g(Au)$.

Proof of convergence

Finally, choosing $\hat{u} = u^\dagger$ and $\hat{w} = w^\dagger$, inequality (14) implies that:

$$\begin{aligned}
\|u_N - u^\dagger\|_2^2 + \|w_N - w^\dagger\|_{\bar{\beta}A}^2 &\overset{\bar{\beta}\geq\beta_N}{\leq} \|u_N - u^\dagger\|_2^2 + \|w_N - w^\dagger\|_{\beta_N A}^2 \\
&\overset{(14)}{\leq} \|u_M - u^\dagger\|_2^2 + \|w_M - w^\dagger\|_{\beta_M A}^2 \\
&\leq \|u_M - u^\dagger\|_2^2 + \|w_M - w^\dagger\|_2^2
\end{aligned}$$

($N \geq M$). As there is a subsequence (u_{n_j}, w_{n_j}) that converges to $(u^\dagger, w^\dagger)$, the rhs can be made as small as one likes (choice of M). This shows that the whole sequence $(u_n, w_n)_n$ converges to $(u^\dagger, w^\dagger)$.
One also shows that $\tilde{u}_n$ converges to $u^\dagger$ and that $\tilde{w}_n$ converges to $w^\dagger$. $\square$

Generalized proximal gradient algorithm with error terms

Theorem (generalized proximal gradient algorithm [1])

Let $\epsilon > 0$. *IF* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex with Lipschitz continuous gradient (L),* $g : \mathbb{R}^{d'} \to \mathbb{R}$ *is convex, proper, lower semi-continuous,* $A : \mathbb{R}^d \to \mathbb{R}^{d'}$ *is a linear map, and a minimizer of* $F(u) = f(u) + g(Au)$ *exists, THEN the iterative algorithm:*

$$\left\{ \begin{array}{lcl} \bar{u}_{n+1} & = & u_n - \alpha_n \nabla f(u_n) - \beta_n A^T w_n + \epsilon_n \\ \tilde{w}_{n+1} & = & \mathrm{prox}_{\tilde{g}_n^*} [w_n + \beta_n A \bar{u}_{n+1}] + \delta_n \\ \tilde{u}_{n+1} & = & u_n - \alpha_n \nabla f(u_n) - \beta_n A^T \tilde{w}_{n+1} + \eta_n \\ w_{n+1} & = & (1-\lambda_n) w_n + \lambda_n \tilde{w}_{n+1} \\ u_{n+1} & = & (1-\lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{array} \right. \quad (15)$$

with $u_0 =$ *arbitrary,* $\epsilon \le \alpha_n \le 2/L - \epsilon$, $\epsilon \le \beta_1 \le \beta_2 \le \ldots \le 1/\|A\| - \epsilon$, $\epsilon \le \lambda_n \le 1$, $\sum_n \|\epsilon_n\|_2 < +\infty$, $\sum_n \|\delta_n\|_2 < +\infty$, $\sum_n \|\eta_n\|_2 < +\infty$ *converges to a minimizer of* $F(u)$.

Generalized iterative soft-thresholding algorithm (GISTA)

- For $f = \frac{1}{2}\|Ku - y\|_2^2$ and $g(u) = \mu\|u\|_1$, the problem reduces to:

$$\hat{u} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1$$

  and the algorithm (6) reduces to:

$$\left\{\begin{array}{rcl} \bar{u}_{n+1} & = & u_n - \alpha\nabla f(u_n) - \beta A^T w_n \\ w_{n+1} & = & P_{\alpha\mu/\beta}\left[w_n + \beta A\bar{u}_{n+1}\right] \\ u_{n+1} & = & u_n - \alpha\nabla f(u_n) - \beta A^T \tilde{w}_{n+1} \end{array}\right. \quad (16)$$

  with $u_0 =$ arbitrary, $0 < \alpha < 2/\|K\|^2$, $0 < \beta < 1/\|A\|$ and $P_\mu = \mathrm{Id} - \mathcal{S}_\mu$ (= projection on ℓ_∞ ball).
- Convenience: conditions on α and β do not mix A and K
- Coefficients of Au_n are not sparse in every step (only as $n \to \infty$).
- One also shows:

$$F(\tilde{u}_N) - F(\hat{u}) \leq C/N$$

  for the Césaro means: $\tilde{u}_N = \frac{1}{N}\sum_{n=1}^{N} u_n$

Applications

- In image restoration:
  - 'Total Variation' (TV) penalty
  - $A = \begin{pmatrix} \Delta_x \\ \Delta_y \end{pmatrix}$ is the local gradient of image
  - $\|Au\|_1 = \sum_{\text{pixels}} \sqrt{(\Delta_x u)^2 + (\Delta_y u)^2}$
  - P_λ =component-wise:

$$P_\lambda(w_x, w_y) = \begin{cases} \frac{\lambda}{\sqrt{w_x^2+w_y^2}}(w_x, w_y) & \sqrt{w_x^2 + w_y^2} > \lambda \\ (w_x, w_y) & \sqrt{w_x^2 + w_y^2} \leq \lambda \end{cases}$$

  - promotes images with sparse gradients (=piecewise constant images)
- Group sparsity (possibly with overlapping groups):
  - A defines the groups (a single 1 on each row)
  - Columns of A may have more than a single nonzero entry (overlapping groups).
  - $\|Au\|_1 = \sum_{k \in \text{groups}} \|(u_i)_{i \in \text{group } k}\|_p$ with $p = 2$ or $p = \infty$

Applications

- In image restoration:
  - 'Total Variation' (TV) penalty
  - $A = \begin{pmatrix} \Delta_x \\ \Delta_y \end{pmatrix}$ is the local gradient of image
  - $\|Au\|_1 = \sum_{\text{pixels}} \sqrt{(\Delta_x u)^2 + (\Delta_y u)^2}$
  - P_λ =component-wise:

$$P_\lambda(w_x, w_y) = \begin{cases} \frac{\lambda}{\sqrt{w_x^2+w_y^2}}(w_x, w_y) & \sqrt{w_x^2 + w_y^2} > \lambda \\ (w_x, w_y) & \sqrt{w_x^2 + w_y^2} \leq \lambda \end{cases}$$

  - promotes images with sparse gradients (=piecewise constant images)

- Group sparsity (possibly with overlapping groups):
  - A defines the groups (a single 1 on each row)
  - Columns of A may have more than a single nonzero entry (overlapping groups).
  - $\|Au\|_1 = \sum_{k \in \text{groups}} \|(u_i)_{i \in \text{group } k}\|_p$ with $p = 2$ or $p = \infty$

Other algorithms

- Many other algorithms exist
- See also [2, 7, 4, 5] and references therein

Acknowledgements

- Thanks to organizers.
- Thanks to Hoan-Phung Bui, Federica Porta and Caroline Verhoeven.

Bibliography I

[1] Hoan-Phung Bui.
Algorithme itératif permettant la résolution de problèmes inverses.
Master's thesis, Université Libre de Bruxelles, 2014.

[2] Antonin Chambolle and Thomas Pock.
A first-order primal-dual algorithm for convex problems with applications to imaging.
J Math Imaging Vis, 40:120–145, 2011.
hal-00490826.

[3] Peijun Chen, Jianguo Huang, and Xiaoqun Zhang.
A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration.
Inverse Problems, 29(2):025011–, 2013.

[4] P. L. Combettes and J.-C. Pesquet.
Proximal splitting methods in signal processing," in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Editors), pp. 185-212. , New York, 2011, chapter 1, pages 185–212.
Springer-Verlag, 2011.

[5] Ignace Loris.
L1Packv2: A Mathematica package for minimizing an ℓ_1-penalized functional.
Computer Physics Communications, 179(12):895–902, 2008.

[6] Ignace Loris and Caroline Verhoeven.
On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty.
Inverse Problems, 27(12):125007, 2011.

[7] Xiaoqun Zhang, Martin Burger, and Stanley Osher.
A unified primal-dual algorithm framework based on bregman iteration.
J Sci Comput, 46:20–46, 2011.