Convergence rates of nested accelerated inexact proximal methods

Silvia Villa

Joint work with S. Salzo, L. Baldassarre and A. Verri

Laboratory for Computational and Statistical Learning, IIT and MIT
http://lcsl.mit.edu

OPT2012, NIPS

Problem setting

Given X a Hilbert space, we consider the problem of computing

$$\min_{x \in X} \quad \underbrace{f(x)}_{\text{data term}} + \underbrace{g(x)}_{\text{regularization}} \quad =: \quad F(x)$$

Problem setting

Given X a Hilbert space, we consider the problem of computing

$$\min_{x \in X} \underbrace{f(x)}_{\text{data term}} + \underbrace{g(x)}_{\text{regularization}} =: F(x)$$

with

- $f : X \to \mathbb{R}$ convex and continuously differentiable, with Lipschitz continuous gradient, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Problem setting

Given X a Hilbert space, we consider the problem of computing

$$\min_{x \in X} \underbrace{f(x)}_{\text{data term}} + \underbrace{g(x)}_{\text{regularization}} =: F(x)$$

with

- $f : X \to \mathbb{R}$ convex and continuously differentiable, with Lipschitz continuous gradient, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

- $g : X \to \mathbb{R} \cup \{+\infty\}$ of the form $g(x) = \omega(Bx)$ with

  $B : X \to H$ bounded linear, H Hilbert space
  $\omega : H \to \mathbb{R} \cup \{+\infty\}$ convex and lsc

Accelerated forward-backward splitting algorithm

Given $y_0 = x_0 \in X$, and $\lambda_k \in (0, 2/L]$, consider

$$x_{k+1} = \text{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$$
$$y_{k+1} = a_k x_k + b_k x_{k+1} + c_k y_k$$

with a_k, b_k, c_k appropriately defined, then

Accelerated forward-backward splitting algorithm

Given $y_0 = x_0 \in X$, and $\lambda_k \in (0, 2/L]$, consider

$$x_{k+1} = \text{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$$
$$y_{k+1} = a_k x_k + b_k x_{k+1} + c_k y_k$$

with a_k, b_k, c_k appropriately defined, then

$$F(x_k) - \min F \leq \frac{C}{k^2}$$

if the minimum is attained.

Nesterov '05; Tseng '08
Beck-Teboulle '09: FISTA ($c_k = 0$)

Proximity operator of g

Definition (Moreau '65)

$$\text{prox}_{\lambda g}(y) := \underset{x \in X}{\text{argmin}} \left\{ g(x) + \frac{1}{2\lambda}\|x - y\|^2 \right\} = \Phi_\lambda(x).$$

Proximity operator of g

Definition (Moreau '65)

$$\text{prox}_{\lambda g}(y) := \underset{x \in X}{\text{argmin}} \left\{ g(x) + \frac{1}{2\lambda}\|x - y\|^2 \right\} = \Phi_\lambda(x).$$

Example

$$g(x) = \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases} \quad \Longrightarrow \quad \text{prox}_{\lambda \iota_C} = P_C$$

Proximity operator of g

Definition (Moreau '65)

$$\text{prox}_{\lambda g}(y) := \underset{x \in X}{\text{argmin}} \left\{ g(x) + \frac{1}{2\lambda}\|x - y\|^2 \right\} = \Phi_\lambda(x).$$

Example

$$g(x) = \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases} \quad \Longrightarrow \quad \text{prox}_{\lambda \iota_C} = P_C$$

Problem : in general $\text{prox}_{\lambda g}(y)$ is not available in closed form.

Main Results

Consider the algorithm

$$x_{k+1} \approx_{\epsilon_k} \text{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$$
$$y_{k+1} = a_k x_k + b_k x_{k+1} + c_k y_k$$

Our contributions

1. convergence rates for inexact FB (without considering the computation of prox)
2. definition of inexact proximal points and algorithms for computing them
3. convergence rates for the nested procedure (considering the prox)

Main Results

Consider the algorithm

$$x_{k+1} \approx_{\epsilon_k} \operatorname{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$$
$$y_{k+1} = a_k x_k + b_k x_{k+1} + c_k y_k$$

Our contributions

1. **convergence rates for inexact FB (without considering the computation of prox)**
2. definition of inexact proximal points and algorithms for computing them
3. convergence rates for the nested procedure (considering the prox)

Convergence rate for accelerated inexact FB

Theorem (V. Salzo-Baldassarre-Verri)

If $\epsilon_k = O\left(\frac{1}{k^q}\right)$, *with* $q > 1/2$ *then*

$$F(x_k) - F_* = \begin{cases} O(\frac{1}{k^2}) & \text{if } q > 3/2 \\ O(\frac{\log k}{k^2}) & \text{if } q = 3/2 \\ O(\frac{1}{k^{2q-1}}) & \text{if } q < 3/2. \end{cases}$$

See also:
Schmidt - Le Roux - Bach '11
Guler '92, Salzo - Villa '12 ($f = 0$)
Combettes-Wajs '05 (non accelerated)
Rockafellar '76 (non accelerated, $f = 0$)

Main Results

Consider the algorithm

$$x_{k+1} \approx_{\epsilon_k} \text{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$$
$$y_{k+1} = a_k x_k + b_k x_{k+1} + c_k y_k$$

Our contributions

1. convergence rates for inexact FB (without considering the computation of prox) convergence rates for inexact FB (without considering the computation of prox)
2. **definition of inexact proximal points and algorithms for computing them**
3. convergence rates for the nested procedure (considering the prox)

Inexact computation of the proximity operator

$$\text{prox}_{\lambda g}(y) = \underset{x \in X}{\text{argmin}} \{ g(x) + \frac{1}{2\lambda} \|x - y\|^2 \}$$

Definition (Lemaire '91)

$$z = \text{prox}_{\lambda g}(y) \quad \Longleftrightarrow \quad 0 \in \partial g(z) + \frac{z - y}{\lambda} \quad \Longleftrightarrow \quad \frac{y - z}{\lambda} \in \partial g(z)$$

Inexact computation of the proximity operator

$$\text{prox}_{\lambda g}(y) = \underset{x \in X}{\text{argmin}} \{ g(x) + \frac{1}{2\lambda} \|x - y\|^2 \}$$

Definition (Lemaire '91)

$$z \approx_\epsilon \text{prox}_{\lambda g}(y) \iff 0 \in \partial_{\frac{\epsilon^2}{2\lambda}} g(z) + \frac{z-y}{\lambda} \iff \frac{y-z}{\lambda} \in \partial_{\frac{\epsilon^2}{2\lambda}} g(z)$$

Recall that the ϵ-subdifferential of a function Φ is

$$\partial_\epsilon \Phi(z) = \{ \eta \in X \, : \, \Phi(x) \geq \Phi(z) + \langle \eta, x - z \rangle - \epsilon, \, \forall x \in X \}.$$

Admissible approximations of $g = \omega \circ B$

$$\Phi_\lambda(x) = \omega(Bx) + \frac{1}{2\lambda}\|x - y\|^2$$

$B : X \to H$ $\quad H$ Hilbert space, B bounded linear operator

$\omega : H \to \mathbb{R} \cup \{+\infty\}$ $\quad$ convex and lower semicontinuous

The dual problem of $\min \Phi_\lambda$ is

$$\min_{v \in H} \frac{1}{2\lambda}\|\lambda B^* v - y\|^2 + \omega^*(v) - \frac{1}{2\lambda}\|y\|^2 =: \Psi_\lambda(v)$$

and if strong duality holds the duality gap $G(x, v) := \Phi_\lambda(x) + \Psi_\lambda(v)$ satisfies

$$\min_{(x,v) \in X \times H} G(x, v) = 0\,.$$

Inexact proximal points of $g = \omega \circ B$: duality gap characterization

Theorem (Duality and proximal points)

The following are equivalent:

- $B^*\bar{v} = \text{prox}_{g^*/\lambda}(y/\lambda)$ *(equiv. $\bar{v}$ is a minimizer of Ψ_λ)*
- $y - \lambda B^*\bar{v} = \text{prox}_{\lambda g}(y)$ *(equiv. $y - \lambda B^*\bar{v}$ is a minimizer of Φ_λ)*
- $G(y - \lambda B^*\bar{v}, \bar{v}) = 0$

Inexact proximal points of $g = \omega \circ B$: duality gap characterization

Theorem (Duality and proximal points)

The following are equivalent:

- $B^*\bar{v} = \text{prox}_{g^*/\lambda}(y/\lambda)$ *(equiv. $\bar{v}$ is a minimizer of Ψ_λ)*
- $y - \lambda B^*\bar{v} = \text{prox}_{\lambda g}(y)$ *(equiv. $y - \lambda B^*\bar{v}$ is a minimizer of Φ_λ)*
- $G(y - \lambda B^*\bar{v}, \bar{v}) = 0$

Theorem (Duality and inexact proximal points) (V.-Salzo-Baldassarre-Verri)

The following are equivalent:

- $B^*v \approx_{\epsilon/\lambda} \text{prox}_{g^*/\lambda}(y/\lambda)$
- $y - \lambda B^*v \approx_{\epsilon} \text{prox}_{\lambda g}(y)$
- $G(y - \lambda B^*v, v) \leq \frac{\epsilon^2}{2\lambda}$

Algorithms for computing inexact proximal points

Idea: find a minimizing sequence for the duality gap.

Theorem (V. Salzo-Baldassarre-Verri)

Assume ω continuous. Suppose that $\Psi_\lambda(v_n) - \Psi_\lambda(\bar{v}) = O(1/n^{2p})$. Then

$$G(y - \lambda B^* v_n, v_n) = O(1/n^p)$$

Main Results

Consider the algorithm

$$x_{k+1} \approx_{\epsilon_k} \operatorname{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$$
$$y_{k+1} = a_k x_k + b_k x_{k+1} + c_k y_k$$

Our contributions

1. convergence rates for inexact FB (without considering the computation of prox) convergence rates for inexact FB (without considering the computation of prox)
2. definition of inexact proximal points and algorithms for computing them
3. **convergence rates for the nested procedure (considering the prox)**

Global convergence rate

Suppose that $\epsilon_k = 1/k^q$ and consider an inner algorithm solving the prox subproblem in at most $O(1/\epsilon_k^{2/p})$ iterations.

$$\mathcal{C}_g(q,p) = \begin{cases} O\left(1/\varepsilon^{\frac{2q/p+1}{2q-1}}\right) + O\left(1/\varepsilon^{\frac{1}{2q-1}}\right) & \text{if } 1/2 < q < 3/2 \\ O\left(1/\varepsilon^{\frac{2q/p+1}{2}}\right) + O\left(1/\varepsilon^{\frac{1}{2}}\right) & \text{if } q > 3/2\,. \end{cases}$$

Global convergence rate

Suppose that $\epsilon_k = 1/k^q$ and consider an inner algorithm solving the prox subproblem in at most $O(1/\epsilon_k^{2/p})$ iterations.

$$\mathcal{C}_g(q,p) = \begin{cases} O\left(1/\varepsilon^{\frac{2q/p+1}{2q-1}}\right) + O\left(1/\varepsilon^{\frac{1}{2q-1}}\right) & \text{if } 1/2 < q < 3/2 \\ O\left(1/\varepsilon^{\frac{2q/p+1}{2}}\right) + O\left(1/\varepsilon^{\frac{1}{2}}\right) & \text{if } q > 3/2\,. \end{cases}$$

- The lower global complexity is reached for $q \to 3/2$ and it is

$$\mathcal{C}_g(p) = O(1/\varepsilon^{\frac{p+3}{2p}+\delta}), \quad \text{for any } \delta > 0$$

Global convergence rate

Suppose that $\epsilon_k = 1/k^q$ and consider an inner algorithm solving the prox subproblem in at most $O(1/\epsilon_k^{2/p})$ iterations.

$$\mathcal{C}_g(q,p) = \begin{cases} O\left(1/\varepsilon^{\frac{2q/p+1}{2q-1}}\right) + O\left(1/\varepsilon^{\frac{1}{2q-1}}\right) & \text{if } 1/2 < q < 3/2 \\ O\left(1/\varepsilon^{\frac{2q/p+1}{2}}\right) + O\left(1/\varepsilon^{\frac{1}{2}}\right) & \text{if } q > 3/2\,. \end{cases}$$

- The lower global complexity is reached for $q \to 3/2$ and it is

$$\mathcal{C}_g(p) = O(1/\varepsilon^{\frac{p+3}{2p}+\delta}), \quad \text{for any } \delta > 0$$

- As $p \to +\infty$, $\quad \mathcal{C}_g(p) \to O(1/\varepsilon^{\frac{1}{2}+\delta})$

Global convergence rate

Suppose that $\epsilon_k = 1/k^q$ and consider an inner algorithm solving the prox subproblem in at most $O(1/\epsilon_k^{2/p})$ iterations.

$$\mathcal{C}_g(q,p) = \begin{cases} O\left(1/\varepsilon^{\frac{2q/p+1}{2q-1}}\right) + O\left(1/\varepsilon^{\frac{1}{2q-1}}\right) & \text{if } 1/2 < q < 3/2 \\ O\left(1/\varepsilon^{\frac{2q/p+1}{2}}\right) + O\left(1/\varepsilon^{\frac{1}{2}}\right) & \text{if } q > 3/2\,. \end{cases}$$

- The lower global complexity is reached for $q \to 3/2$ and it is

$$\mathcal{C}_g(p) = O(1/\varepsilon^{\frac{p+3}{2p}+\delta}), \quad \text{for any } \delta > 0$$

- As $p \to +\infty$, $\quad \mathcal{C}_g(p) \to O(1/\varepsilon^{\frac{1}{2}+\delta})$
- If FISTA is used, $\quad \mathcal{C}_g(p) = O(1/\varepsilon^{2+\delta}), \quad$ for any $\delta > 0$

Impact of the errors on the global iteration complexity

Precision	10^{-4}			10^{-6}			10^{-8}		
Algo	Time	# Ext	# Int	Time	# Ext	# Int	Time	# Ext	# Int
AIFB									
$q = 1$	11.8	137	1062	124.2	905	12313	1750	8776	182006
$q = 1.3$	16.2	118	1600	**63.6**	387	6437	**272.1**	1300	28350
$q = 1.5$	26.0	117	2734	98.7	373	10540	414.5	1085	45297
ISTA									
$q = 0.1$	36.9	1341	1341	147.2	5346	5346	635.4	23031	23031
$q = 0.8$	36.9	1341	1341	147.2	5346	5346	635.4	23031	23031
$q = 1.0$	63.2	1337	4533	189.9	5226	11126	745.1	18224	48333
PRIDU									
$\sigma = 10$	7.4	362	-	165.7	8186	-	4684	231848	-
$\sigma = 12.5$	**6.2**	310	-	132.2	6609	-	3715	185588	-

Deblurring with Total Variation regularization.

PRIDU: Algorithm 1 in Chambolle - Pock '11

Impact of the errors on the global iteration complexity

Precision	10^{-4}			10^{-6}			10^{-8}		
Algo	Time	# Ext	# Int	Time	# Ext	# Int	Time	# Ext	# Int
AIFB									
$q = 1$	3.9	104	3985	41.5	983	42239	414.1	9748	421769
$q = 1.3$	**2.1**	51	2103	11.2	247	11389	60.4	1179	61915
$q = 1.5$	2.8	50	2857	16.2	199	16945	61.3	548	64518
ISTA									
$q = 0.3$	5.2	1613	1730	10.3	3246	3363	15.9	5065	5182
$q = 0.5$	4.4	1217	1827	**9.5**	2850	3460	**14.9**	4603	5213
$q = 0.8$	7.0	585	6092	15.5	2218	11264	19.8	3599	12645
PRIDU									
$\sigma = 10$	10.5	2901	-	25.4	7040	-	47.4	13141	-
$\sigma = 1.07$	5.8	1602	-	11.0	3026	-	16.1	4452	-

Breast cancer dataset: overlapping group lasso

PRIDU: Algorithm 1 in Chambolle - Pock '11

Contributions

- convergence rates for inexact and accelerated forward-backward algorithm
- characterization of the notion of inexactness in terms of duality gap
- convergence rates for nested procedures, where the prox is computed via duality
- numerical comparison with non accelerated methods and with a primal-dual algorithm

References


S. Villa, S. Salzo, L. Baldassarre and A. Verri,
Accelerated and inexact forward-backward splitting algorithms,
http://www.optimization-online.org/DB_HTML/2011/08/3132.html
(submitted)


S. Salzo and S. Villa,
Accelerated and inexact proximal point algorithms, J. Convex Anal. 2012


L. Rosasco, S. Mosci, M. Santoro, A. Verri, and S. Villa
Proximal Methods for Structured Sparsity Regularization,
(Proceedings of ECML, LNAI Springer, 2010)