Stochastic forward-backward splitting algorithm

Silvia Villa

joint work with L. Rosasco and B. C. Vũ

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

Ottimizzazione e processi dinamici in apprendimento statistico e problemi inversi
Sestri Levante — September 2014

Problem setting

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

$$\min_{w \in H} T(w), \qquad T(w) = F(w) + R(w),$$

with

- $F \colon H \to \mathbb{R}$ convex and continuously differentiable, with Lipschitz continuous gradient, i.e.,

$$\|\nabla F(w) - \nabla F(w')\| \leq \beta \|w - w'\|$$

- $R \colon H \to \mathbb{R} \cup \{+\infty\}$ proper, convex, and lower semicontinuous

Problem setting

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

$$\min_{w \in H} T(w), \qquad T(w) = \underbrace{F(w)}_{\text{data fitting term}} + \underbrace{R(w)}_{\text{regularization term}}$$

with

- $F \colon H \to \mathbb{R}$ convex and continuously differentiable, with Lipschitz continuous gradient, i.e.,

$$\|\nabla F(w) - \nabla F(w')\| \leq \beta \|w - w'\|$$

- $R \colon H \to \mathbb{R} \cup \{+\infty\}$ proper, convex, and lower semicontinuous

Statistical learning with regularization

Given a training set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ of data points in an input/output space $H \times Y$, $\ell : \mathbb{R} \times Y \to \mathbb{R}_+$, the goal is to approximate the infimum of

$$F(w) = \int_{H \times Y} \ell(\langle w, x \rangle, y) dP(x, y) + \lambda R(w),$$

If, for every y, $\nabla \ell(\cdot, y)$ is Lipschitz continuous, then ∇F is Lipschitz continuous.

Statistical learning with regularization cont'd

A common strategy is to minimize

$$T(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(\langle w, x_i \rangle, y_i) + \lambda R(w)$$

Forward-backward splitting algorithm

Given $w_0 \in H$ and $\gamma_n \in]0, 2/\beta[$, define

FB

$$w_{n+1} = \text{prox}_{\gamma_n R}(w_n - \gamma_n \nabla F(w_n))$$

See Bausckhe-Combettes, Convex analysis and monotone operator theory in Hilbert spaces, 2011.

Proximity operator of R

Definition (Moreau '65)

$$\operatorname{prox}_{\lambda R}(w) = \operatorname*{argmin}_{v \in H} \underbrace{\left(R(v) + \frac{1}{2\lambda} \|v - w\|^2 \right)}_{=\Phi_\lambda(w)}$$

Proximity operator of R

Definition (Moreau '65)

$$\operatorname{prox}_{\lambda R}(w) = \operatorname*{argmin}_{v \in H} \underbrace{\left(R(v) + \frac{1}{2\lambda}\|v - w\|^2 \right)}_{=\Phi_\lambda(w)}$$

Example

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

Stochastic forward-backward splitting algorithm

Given w_0 such that $E[\|w_0\|^2] < +\infty$, $\gamma_n > 0$, define

SFB

$$w_{n+1} = \text{prox}_{\gamma_n R}(w_n - \gamma_n \nabla F(w_n))$$

Stochastic forward-backward splitting algorithm

Given w_0 such that $E[\|w_0\|^2] < +\infty$, $\gamma_n > 0$, define

SFB

$$w_{n+1} = \text{prox}_{\gamma_n R}(w_n - \gamma_n \mathbf{G_n})$$

where G_n is a stochastic estimate of the gradient.

Stochastic forward-backward splitting algorithm

Given w_0 such that $E[\|w_0\|^2] < +\infty$, $\gamma_n > 0$, $\lambda_n \in [0, 1]$, define

SFB

$$y_n = \mathrm{prox}_{\gamma_n R}(w_n - \gamma_n \mathbf{G_n})$$
$$w_{n+1} = (1 - \lambda_n) w_n + \lambda_n y_n.$$

where G_n is a stochastic estimate of the gradient.

Assumptions

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Define the filtration $\mathcal{F}_n = \sigma(\{w_0, \dots, w_n\})$ and assume

- $E[\|G_n\|^2] < +\infty$
- $E[G_n|\mathcal{F}_n] = \nabla F(w_n)$
- $E[\|G_n - \nabla F(w_n)\|^2|\mathcal{F}_n] \leq \sigma^2(1 + \|\nabla F(w_n)\|^2)$,

The case of statistical learning

Given a Hilbert space, the objective is to minimize

$$T(w) = \int_{H \times Y} \ell(\langle w, x \rangle, y) dP(x, y) + \lambda R(w)$$

given a sequence of i.i.d. samples $(x_i, y_i)_{i \in \mathbb{N}}$.

Then, $\mathcal{F}_n = \sigma((x_1, y_1), \dots, (x_n, y_n))$ and we can choose

$$G_n = \nabla \ell(\cdot, y_n)(\langle w_n, x_n \rangle) x_n \implies E[G_n | \mathcal{F}_n] = \nabla F(w_n)$$

$E[\|G_n - \nabla F(w_n)\|^2 | \mathcal{F}_n] \leq \sigma^2(1 + \|\nabla F(w_n)\|^2)$ is a condition on the variance of the random variable

$$(x, y) \in H \times Y \mapsto \nabla \ell(\cdot, y)(\langle w, x \rangle) x$$

Statistical learning - Incremental FB algorithm

The goal is to minimize

$$T(w) = \frac{1}{n} \sum_{i=1}^{m} \ell(\langle w, x_i \rangle, y_i) + \lambda R(w)$$

Let $i_n : \Omega \to \{1, \dots, m\}$ be a sequence of independent r.v. such that, for every n and i, $P[i_n = i] = 1/m$. Then

$$G_n = \nabla \ell(\cdot, y_{i_n}))(\langle w, x_{i_n} \rangle) x_{i_n}$$

is such that $E[G_n | \mathcal{F}_n] = E[G_n] = \frac{1}{n} \sum_{i=1}^{m} \ell(\langle w, x_i \rangle, y_i)$.

Comparison between FB and SFB algorithm

The stochastic incremental FB algorithm becomes

SFB

$$w_{n+1} = \mathrm{prox}_{\gamma_n R}(w_n - \gamma_n \nabla \ell(\langle w_{i_n}, x_{i_n} \rangle, y_{i_n}) x_{i_n})$$

The FB algorithm is

FB

$$w_{n+1} = \mathrm{prox}_{\gamma_n R}\left(w_n - \gamma_n \frac{1}{m} \sum_{i=1}^{m} \nabla \ell(\langle w_n, x_i \rangle, y_i) x_i\right)$$

Main Results

Assume that a solution $\overline{w}$ exists. Given w_0 such that $E[\|w_0\|^2] < +\infty$, γ_n, $(\lambda_n)_n$ in $(0, 1]$, and G_n.

SFBA

$$y_n = \text{prox}_{\gamma_n R}(w_n - \gamma_n G_n)$$
$$w_{n+1} = (1 - \lambda_n)w_n + \lambda_n y_n.$$

Our contributions

1. Convergence rates for $(E[\|w_n - \overline{w}\|^2]$
2. Almost sure convergence of w_n

Main Results

Assume that a solution $\overline{w}$ exists. Given w_0 such that $E[\|w_0\|^2] < +\infty$, γ_n and G_n.

SFBA

$$w_{n+1} = \text{prox}_{\gamma_n R}(w_n - \gamma_n G_n)$$

Our contributions

1. **convergence rates for** $E[\|w_n - \overline{w}\|^2]$
2. Almost sure convergence of w_n

Convergence rates for $E[\|w_n - \overline{w}\|^2]$

Main assumption

F is μ strongly convex and R is ν strongly convex with $\mu + \nu > 0$.

Theorem

Let $\alpha > 0$ and $\theta \in]0, 1]$. Assume that $\gamma_n = \alpha / n^\theta$ and suppose that there exists $\epsilon > 0$ s.t. $\gamma_n < \dfrac{(1-\epsilon)}{(1+2\sigma^2)\beta}$ (β is the Lipschitz constant of ∇F).
Then, setting $c = 2\alpha(\nu + \mu\epsilon)/(1+\nu)^2$,

$$E[\|w_n - \overline{w}\|^2] \leq \begin{cases} O(1/n^\theta) & \text{if } \theta \in]0, 1[\\ O(1/n^c) + O(1/n) & \text{if } \theta = 1 \end{cases}$$

Remarks and related work

- c can be made greater than 1 by properly choosing α (knowledge of μ and ν is required)

Remarks and related work

- c can be made greater than 1 by properly choosing α (knowledge of μ and ν is required)
- the obtained rate of convergence is the same that can be obtained using "accelerated" methods (see e. g. Kwok-Hu-Pan, NIPS 2009 and Ghadimi-Lan 2012, Li-Chen-Peña 2014)

Remarks and related work

- c can be made greater than 1 by properly choosing α (knowledge of μ and ν is required)
- the obtained rate of convergence is the same that can be obtained using "accelerated" methods (see e. g. Kwok-Hu-Pan, NIPS 2009 and Ghadimi-Lan 2012, Li-Chen-Peña 2014)
- the result is not asymptotic. An explicit estimate of the constants in the O terms is available (Chung Lemma)

Remarks and related work

- c can be made greater than 1 by properly choosing α (knowledge of μ and ν is required)
- the obtained rate of convergence is the same that can be obtained using "accelerated" methods (see e. g. Kwok-Hu-Pan, NIPS 2009 and Ghadimi-Lan 2012, Li-Chen-Peña 2014)
- the result is not asymptotic. An explicit estimate of the constants in the O terms is available (Chung Lemma)
- extends to the non smooth case results that were known only in the smooth case (for stochastic gradient algorithm, Bach-Moulines 2011)

Comparison with FOBOS (Duchi-Singer,2009)

A closely related algorithm is

FOBOS

$$w_{n+1} = \mathrm{prox}_{\gamma_n R}(w_n - \gamma_n G_n)$$
$$\overline{w}_{n+1} = \frac{\sum_{k=0}^{n+1} \gamma_k w_k}{\sum_{k=0}^{n+1} \gamma_k}$$

- no averages are computed in SFB algorithm
- convergence rate of SFB is faster $(1/n)$ w.r.t. the one of FOBOS $(\log n/n)$
- the convergence analysis of FOBOS does not require F differentiable and relies on boundedness of ∂F (square loss excluded) and ∂R
- this answers a question posed by Rakhlin-Shamir-Sridaran, 2012 (Making gradient descent optimal for strongly convex stochastic optimization).

Open problems

1) High probability bounds.
2) Assume that $\mu = \nu = 0$ (no strong convexity). If $R = 0$ (see Bach and Moulines 2011), a non asymptotic bound for the sequence

$$E[T(w_n) - \min_{w \in H} T(w)]$$

can be proved. Also, setting,

$$\overline{w}_n = \frac{\sum_{i=0}^{n} \gamma_i w_i}{\sum_{i=0}^{n} \gamma_i}$$

nonasymptotic bounds for

$$E[T(\overline{w}_n) - \min_{w \in H} T(w)]$$

are available. Can we extend this result to the case $R \neq 0$?

Main Results

Assume that a solution $\overline{w}$ exists. Given $w_0 \in H$, γ_n and G_n.

SFB

$$w_{n+1} = \text{prox}_{\gamma_n R}(w_n - \gamma_n G_n)$$

Our contributions

1. Convergence rates for $E[\|w_n - \overline{w}\|^2]$
2. **Almost sure convergence of** w_n

Almost sure convergence with uniform convexity

Theorem

Suppose that F is **uniformly convex** at $\overline{w}$ and

$$\sum \gamma_n = +\infty \qquad \sum \gamma_n^2 < +\infty.$$

Then

$$\|w_n - \overline{w}\| \to 0 \ \text{ almost surely}$$

Recall that a function $F\colon H \to \mathbb{R}$ is uniformly convex at a point $\overline{w}$ if $\exists\, \phi\colon \mathbb{R}_+ \to \mathbb{R}_+$ vanishing only at 0 such that

$$(\forall w \in H) \qquad \langle \nabla F(w) - \nabla F(\overline{w}), w - \overline{w} \rangle \geq \Phi(\|w - \overline{w}\|)$$

Almost sure convergence

Theorem

Suppose that F is **strictly convex** at $\overline{w}$ and

$$\sum \gamma_n = +\infty \qquad \sum \gamma_n^2 < +\infty.$$

- If ∇F is weakly continuous, then there exists a subsequence t_n

$$\|w_{t_n} - \overline{w}\| \to 0 \ \text{ almost surely}$$

- If $R = \iota_C$ for some closed convex set C, then

$$\|w_n - \overline{w}\| \to 0 \ \text{ almost surely}$$

Recall that a function $F\colon H \to \mathbb{R}$ is strictly convex at a point $\overline{w}$ if

$$(\forall w \neq \overline{w}) \qquad \langle \nabla F(w) - \nabla F(\overline{w}), w - \overline{w} \rangle > 0$$

Remarks and related work

- ∇F is weakly continuous in finite dimensions, or if it is linear

Remarks and related work

- ∇F is weakly continuous in finite dimensions, or if it is linear
- We extend results in Barty-Roy-Strugarek 2007

Remarks and related work

- ∇F is weakly continuous in finite dimensions, or if it is linear
- We extend results in Barty-Roy-Strugarek 2007
- the proof is based on the concept of stochastic quasi Fejér sequences.

Extensions: monotone inclusions framework

Given a Hilbert space H, we consider the problem

$$\text{Find } \overline{w} \in H \quad \text{such that} \quad 0 \in A\overline{w} + B\overline{w},$$

where

- $A\colon H \to 2^H$ is maximally monotone, i.e. for every w and w' in H and for every $u \in Aw$ and $u' \in Aw'$

  $$\langle w - w', u - u' \rangle \geq 0$$

  and there exists no monotone operator whose graph properly contains the graph of A.
- $B\colon H \to H$ is single valued and cocoercive, i.e. for every v and w in H

  $$\langle v - w, Bv - Bw \rangle \geq (1/\beta)\|Bv - Bw\|^2$$

Extensions: monotone inclusions framework

Given a Hilbert space H, we consider the problem

$$\text{Find } \overline{w} \in H \quad \text{such that} \quad 0 \in A\overline{w} + B\overline{w},$$

Given w_0 such that $E[\|w_0\|^2] < +\infty$, $\gamma_n > 0$, λ_n in $(0, 1]$, and G_n.

SFBA

$$w_{n+1} = J_{\gamma_n A}(w_n - \gamma_n G_n)$$
$$w_{n+1} = (1 - \lambda_n)w_n + \lambda_n y_n$$

with $J_{\gamma_n A} = (I + \gamma_n A)^{-1}$ and $E[G_n|\mathcal{F}_n] = Bw_n$.

Open problems

Without unform monotonicity assumptions, the analysis in Combettes and Pesquet (2014) implies that

$$w_n \rightharpoonup \overline{w} \text{ almost surely}$$

under the assumption

$$\sum_{n\in\mathbb{N}} \sqrt{E[\|G_n - Bw_n\|^2|\mathcal{F}_n]} < +\infty.$$

It would be interesting to understand if almost sure convergence holds under the error conditions:

$$E[\|G_n - Bw_n\|^2|\mathcal{F}_n] \leq \sigma^2(1 + \|Bw_n\|^2)$$

Averaging could help (Passty, 1979) and could allow to consider the case where B is only maximally monotone.

Behavior with respect to the choice of parameters

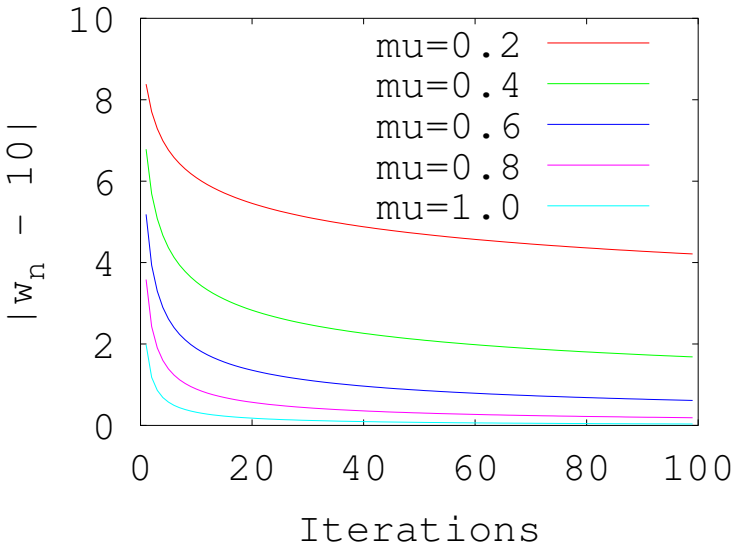
One dimensional toy example

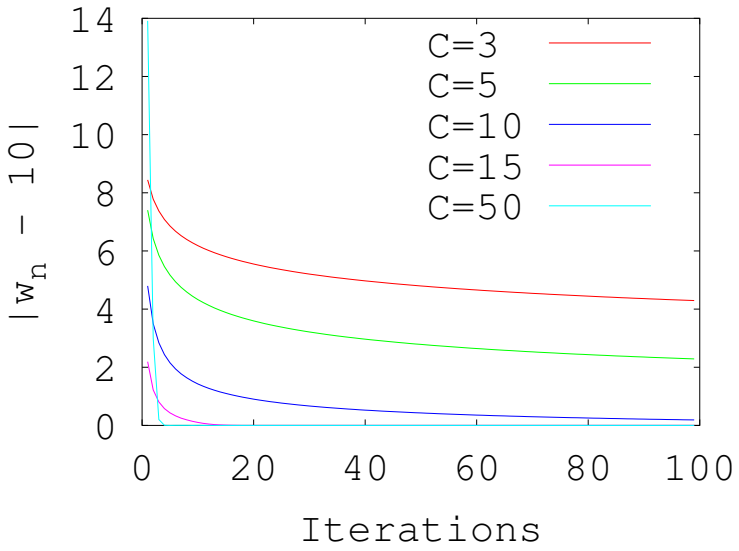
$$\phi(w) := \frac{\mu}{2}(w - 10)^2 + 0.02|w - 10|. \tag{1}$$

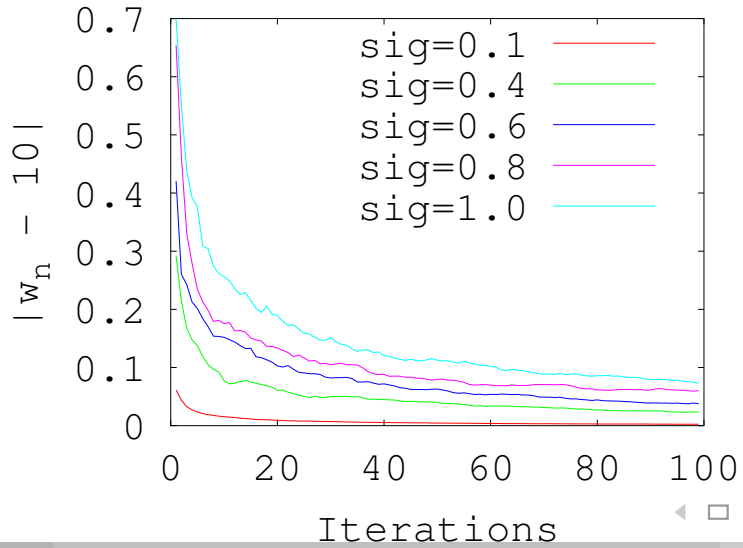
- We know the solution and the gradients exactly
- We add gaussian noise to the gradients
- The prox is the soft thresholding operator
- One hundred simulations for each experiment

$$\text{prox}_{\gamma R}(w) = 10 + \textit{sign}(w - 10)\max\{|w - 10| - \gamma, 0\}$$

Behavior with respect to the choice of parameters





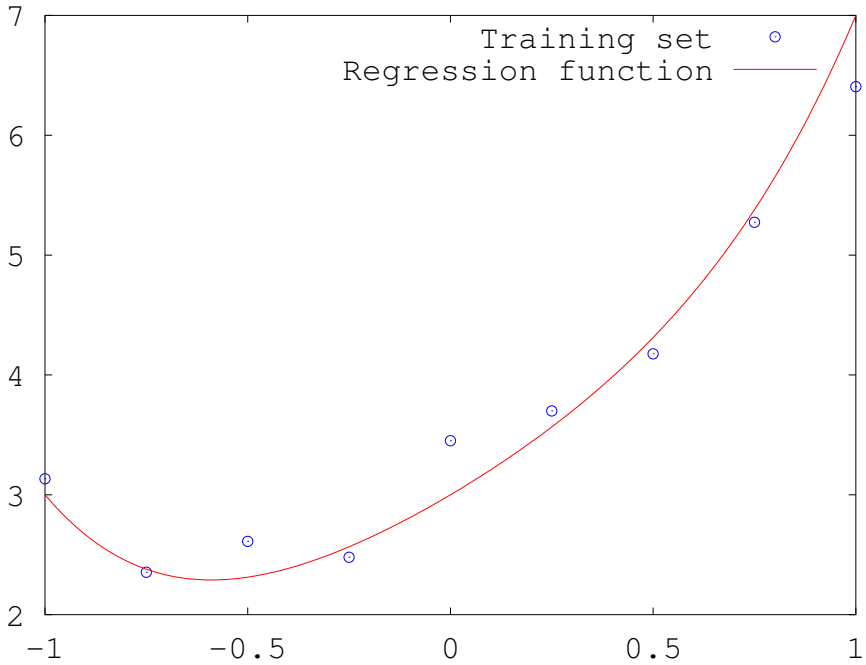


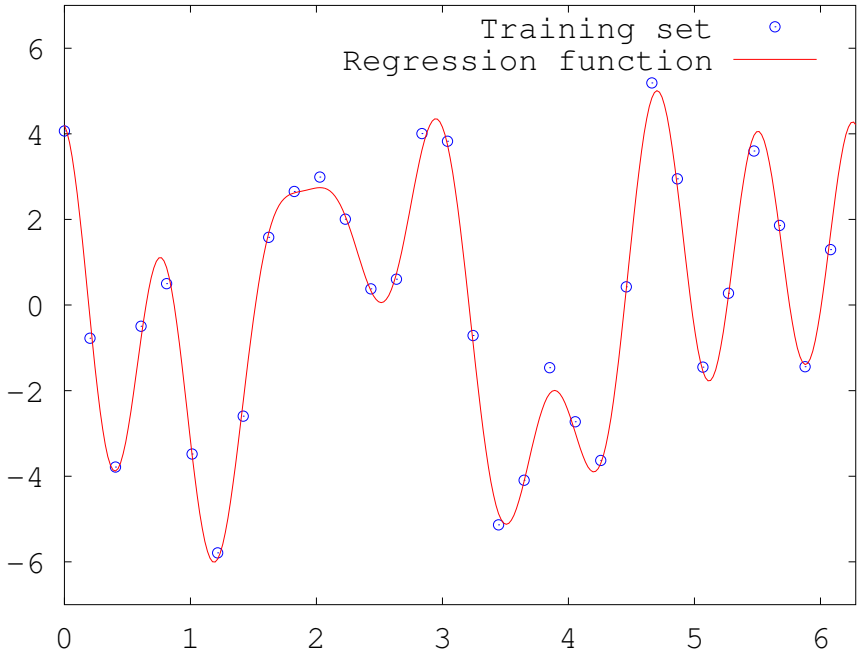
Comparison with other stochastic first order methods

We compare three algorithms

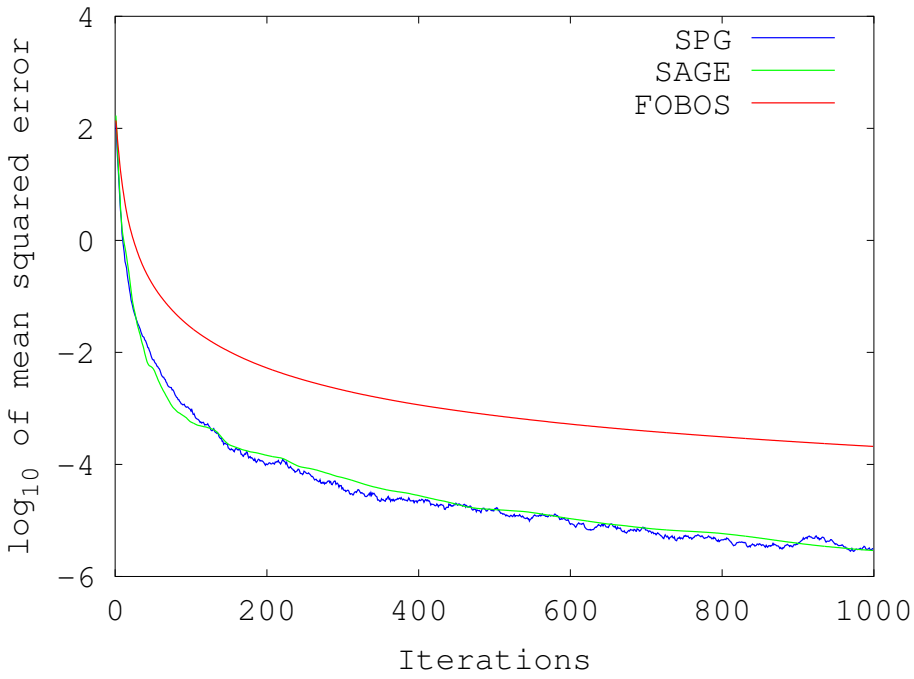
- Stochastic forward backward algorithm (SPG)
- Stochastic forward-backward with averages (FOBOS, Duchi-Singer 2009)
- Accelerated stochastic forward-backward (SAGE, Hu-Kwok-Pan 2011)

Comparison with other methods: regression problems





Comparison with other methods: regression problems



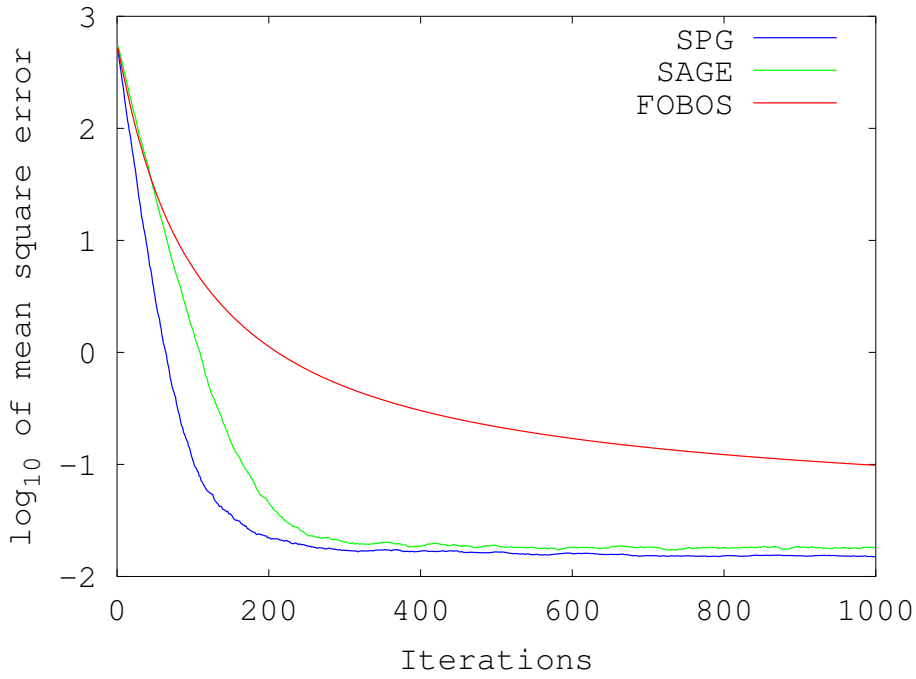


Figure: Convergence of the iterations to the optimal solution with polynomial basis (left) and with Fourier basis (right).

Averaging could be a bad idea

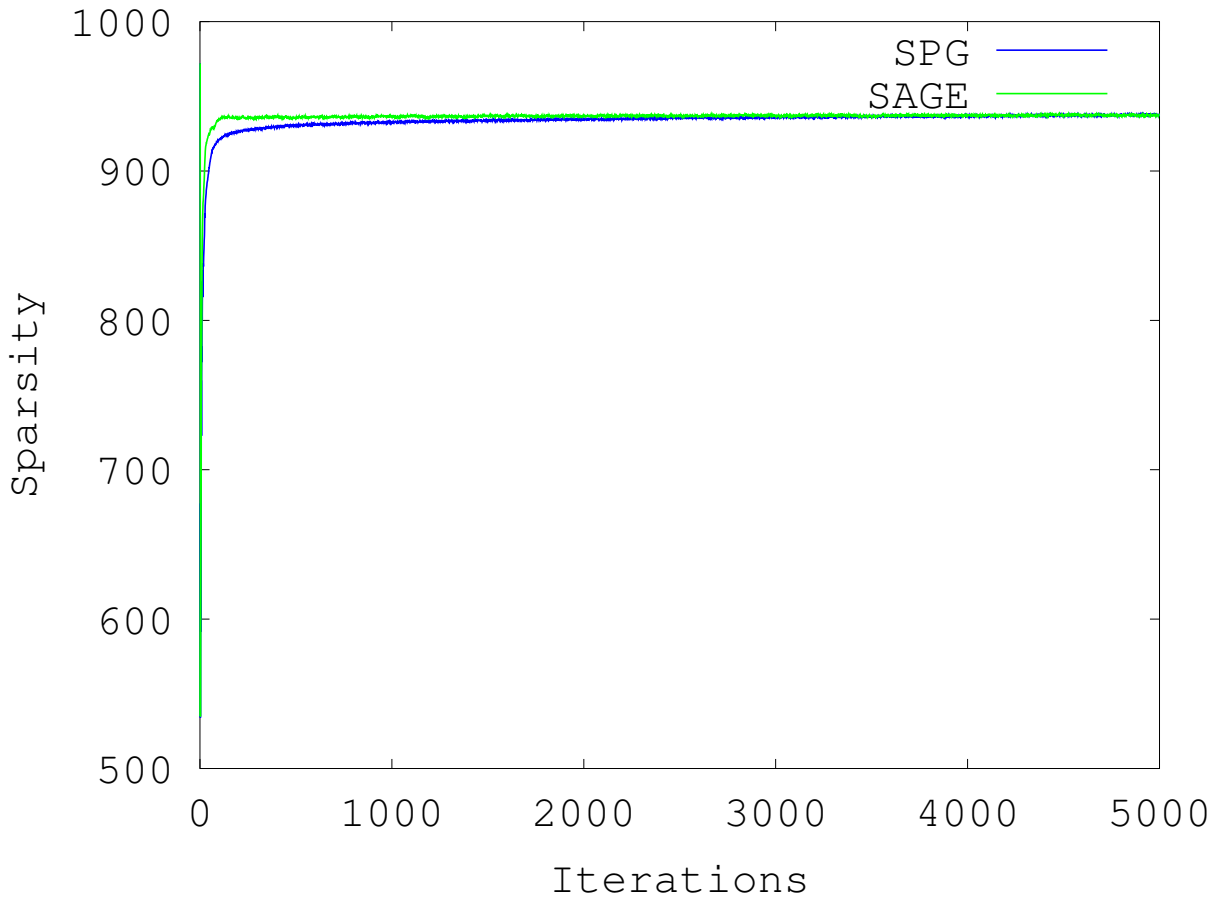


Figure: Number of zero components of the vector $w_n - \overline{w}$ with the same initial point for SPG and SAGE. The number of zero components of FOBOS is decreasing with the iterations, and close to 400.

Conclusions

- convergence rates and almost sure convergence for stochastic forward-backward algorithm under general error conditions
- extend the study to the monotone case
- averaging: yes or no?

References


L. Rosasco, S. Villa, and B. C. Vũ,
Convergence of stochastic proximal gradient algorithm, arxiv 1403.5074


L. Rosasco, S. Villa, and B. C. Vũ,
Stochastic forward-backward splitting for monotone inclusions, arxiv 1403.7999