Proximal Stochastic Gradient Method with Variance Reduction

Lin Xiao (Microsoft Research)

Joint work with
Tong Zhang (Rutgers University, Baidu)

IPAM Workshop on *Stochastic Gradient Methods*
UCLA, February 27, 2013

Minimizing finite average of convex functions

problem

$$\text{minimize} \quad F(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

stochastic gradient method:

$$x_{k+1} = x_k - \eta_k \nabla f_{i_k}(x_k)$$

Minimizing finite average of convex functions

problem

$$\text{minimize} \quad F(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

stochastic gradient method:

$$x_{k+1} = x_k - \eta_k \nabla f_{i_k}(x_k)$$

two perspectives:

- *stochastic optimization:* a special case of minimizing $\mathbb{E}_\xi f(x,\xi)$
- *deterministic optimization:* a randomized incremental gradient method for a structured convex problem

Mind the problem structure

stochastic optimization perspective:

- a general method used to solve a special case
- complexity theory: $O(\frac{1}{\epsilon^2})$ or $O(\frac{1}{\epsilon})$ with strong convexity
- recent improvements by Bach & Moulines

Mind the problem structure

stochastic optimization perspective:

- a general method used to solve a special case
- complexity theory: $O(\frac{1}{\epsilon^2})$ or $O(\frac{1}{\epsilon})$ with strong convexity
- recent improvements by Bach & Moulines

deterministic optimization perspective:

- a special method for solving a structured problem
- sanity test: should at least beat full gradient methods: complexity $O(n\frac{L}{\mu}\log\frac{1}{\epsilon})$ or $O(n\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$
- recent progresses: SAG and SVRG

Stochastic average gradient (SAG)

- SAG method (Le Roux, Schmidt, Bach 2012)

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} g_k^{(i)}$$

where

$$g_k^{(i)} = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \\ g_{k-1}^{(i)} & \text{otherwise} \end{cases}$$

- a randomized variant of incremental aggregated gradient (IAG) of Blatt, Hero, & Gauchman (2007)
- complexity (gradient evaluations): $O(\max\{n, \frac{L}{\mu}\} \log \frac{1}{\epsilon})$
- need to store most recent gradient of each component, but can be avoided for some structured problems

Stochastic variance reduced gradient (SVRG)

- SVRG (Johnson & Zhang 2013, Mahdavi, Zhang & Jin 2013)
  - update form

$$x_{k+1} = x_k - \eta(\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x}))$$

  - update $\tilde{x}$ periodically (every few passes)
- still a stochastic gradient method

$$\begin{aligned} & \mathbb{E}[\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x})] \\ = \ & \nabla F(x_k) - \nabla F(\tilde{x}) + \nabla F(\tilde{x}) \\ = \ & \nabla F(x_k) \end{aligned}$$

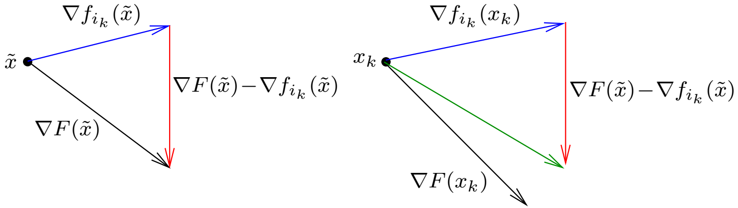
  - expected update direction is the same as $\mathbb{E} f_{i_k}(x_k)$
  - variance can be diminishing if $\tilde{x}$ updated periodically
- complexity: $O\left((n + \frac{L}{\mu}) \log \frac{1}{\epsilon}\right)$, cf. SAG: $O(\max\{n, \frac{L}{\mu}\} \log \frac{1}{\epsilon})$

Stochastic variance reduced gradient (SVRG)

- computational cost per iteration:
  - unlike SAG, no need to store gradients for each components
  - need to compute two gradients at each iteration, and also full gradient periodically (no more than three per epoch)
  - for many structured problems, two gradients at each iteration can be reduced to only one

Stochastic variance reduced gradient (SVRG)

- computational cost per iteration:
  - unlike SAG, no need to store gradients for each components
  - need to compute two gradients at each iteration, and also full gradient periodically (no more than three per epoch)
  - for many structured problems, two gradients at each iteration can be reduced to only one
- intuition of variance reduction



SAG vs SVRG

- SAG: more like a full gradient method?

$$x_{k+1} = x_k - \frac{\eta}{n} \sum_{i=1}^{n} g_k^{(i)}, \quad \text{where} \quad g_k^{(i)} = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \\ g_{k-1}^{(i)} & \text{otherwise} \end{cases}$$

  each new stochastic gradient is weighted by η/n, but re-used in many iterations

- SVRG: more like a stochastic gradient method?

$$x_{k+1} = x_k - \eta(\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x}))$$

  each new stochastic gradient is weighted by η, but only used once and then discarded

Contributions of this talk

- extend SVRG to minimization of composite objective functions

$$\min_{x\in\mathbb{R}^d} P(x) \stackrel{\text{def}}{=} F(x) + R(x), \quad \text{where} \;\; F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

  - each $f_i(x)$ is convex and smooth
  - $P(x)$ strongly convex
  - $R(x)$ convex and possibly nondifferentiable
- prove same complexity $O\left((n + \frac{L}{\mu}) \log \frac{1}{\epsilon}\right)$
- weighted sampling strategy to achieve $O\left((n + \frac{L_{\text{avg}}}{\mu}) \log \frac{1}{\epsilon}\right)$

Problem statement and assumptions

$$\min_{x \in \mathbb{R}^d} P(x) \stackrel{\text{def}}{=} F(x) + R(x), \quad \text{where} \;\; F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Problem statement and assumptions

$$\min_{x\in\mathbb{R}^d} P(x) \overset{\text{def}}{=} F(x) + R(x), \quad \text{where} \ \ F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

assumptions:

- each $f_i(x)$ and $R(x)$ are convex; $f_i(x)$ differentiable on $\mathrm{dom}(R)$
- each $f_i(x)$ is smooth with Lipschitz constant L

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|$$

  (which implies that $\nabla F(x)$ also has Lipschitz constant L)
- $P(x)$ strongly convex: for all $x \in \mathrm{dom}(R)$ and $y \in \mathbb{R}^d$,

$$P(y) \ge P(x) + \xi^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \quad \forall\, \xi \in \partial P(x)$$

Proximal stochastic gradient (Prox-SG) method

- Prox-SG: for $k = 1, 2, \ldots$, draw i_k randomly from $\{1, \ldots, n\}$,

$$x_k = \arg\min_{x \in \mathbb{R}^d} \left\{ \nabla f_{i_k}(x_{k-1})^T x + \frac{1}{2\eta} \|x - x_{k-1}\|^2 + R(x) \right\}$$

- with definition of *proximal mapping*

$$\mathrm{prox}_R(y) = \arg\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|x - y\|^2 + R(x) \right\}$$

  Prox-SG can be written as

$$x_k = \mathrm{prox}_{\eta R}\big(x_{k-1} - \eta \nabla f_{i_k}(x_{k-1})\big)$$

- complexity $O(\frac{1}{\mu\epsilon})$ (Duchi & Singer 2009, Langford et al. 2009)

Prox-SVRG

- proceed in stages:
  - update $\tilde{x}$ at beginning of each stage (every few passes)
  - each iteration takes the form

$$x_k = \mathrm{prox}_{\eta R}(x_{k-1} - \eta v_k)$$

    where

$$v_k = \nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x})$$

- still a Prox-SG method, since

$$\mathbb{E} v_k = \nabla F(x_{k-1}) - \nabla F(\tilde{x}) + \nabla F(\tilde{x}) = \nabla F(x_{k-1})$$

  but with correction from gradients computed at $\tilde{x}$

Prox-SVRG

input: $\tilde{x}_0$, η, m
iterate: for $s = 1, 2, \dots$

$\tilde{x} = \tilde{x}_{s-1}$

$\tilde{v} = \nabla F(\tilde{x})$

$x_0 = \tilde{x}$

iterate: for $k = 1, 2, \dots, m$

pick $i_k \in \{1, \dots, n\}$ uniformly at random

$v_k = \nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\tilde{x}) + \tilde{v}$

$x_k = \mathrm{prox}_{\eta R}(x_{k-1} - \eta v_k)$

end

set $\tilde{x}_s = \frac{1}{m} \sum_{k=1}^{m} x_k$

end

Convergence analysis of Prox-SVRG

- **theorem:** suppose $0 < \eta \leq 1/4L$ and m sufficiently large so that
$$\rho = \frac{1}{\mu\eta(1-4L\eta)m} + \frac{4L\eta(m+1)}{(1-4L\eta)m} < 1$$
then we have geometric convergence in expectation:
$$\mathbb{E}P(\tilde{x}_s) - P(x_\star) \leq \rho^s[P(\tilde{x}_0) - P(x_\star)]$$

- *more concretely,* if $\eta = \theta/L$, then
$$\rho \approx \frac{L/\mu}{\theta(1-4\theta)m} + \frac{4\theta}{1-4\theta}$$
choosing $\theta = 0.1$ and $m = 100(L/\mu)$ results in $\rho = 5/6$

- overall complexity: $O\left(\left(\frac{L}{\mu} + n\right)\log\left(\frac{1}{\epsilon}\right)\right)$

Proof ideas

- define *stochastic gradient mapping*

$$g_k = \frac{1}{\eta}(x_{k-1} - x_k) = \frac{1}{\eta}\left(x_{k-1} - \mathrm{prox}_{\eta R}(x_{k-1} - \eta v_k)\right)$$

so that $x_k = x_{k-1} - \eta g_k$

- similar as in classical analysis of stochastic gradient methods

$$\begin{aligned}
\|x_k - x_\star\|^2 &= \|x_{k-1} - \eta g_k - x_\star\|^2 \\
&= \|x_{k-1} - x_\star\|^2 - 2\eta g_k^T(x_{k-1} - x_\star) + \eta^2\|g_k\|^2 \\
&\vdots \\
\mathbb{E}\|x_k - x_\star\|^2 &\leq \|x_{k-1} - x_\star\|^2 - 2\eta\left(\mathbb{E}P(x_k) - P(x_\star)\right) + 2\eta^2\|\Delta_k\|^2
\end{aligned}$$

where $\Delta_k = v_k - \nabla F(x_{k-1})$

Assumptions for weighted sampling

$$\min_{x \in \mathbb{R}^d} P(x) \stackrel{\text{def}}{=} F(x) + R(x), \quad \text{where} \;\; F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

assumptions: f_i and R are convex, and

- each $f_i(x)$ is smooth with Lipschitz constant L_i

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$$

- the average $F(x)$ has Lipschitz constant L ($\leq \frac{1}{n}\sum_{i=1}^{n} L_i$)

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

- $P(x)$ strongly convex: for all $x \in \mathrm{dom}(R)$ and $y \in \mathbb{R}^d$,

$$P(y) \geq P(x) + \xi^T(y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall\, \xi \in \partial P(x)$$

Prox-SVRG

input: $\tilde{x}_0$, η, m

iterate: for $s = 1, 2, \ldots$

$\tilde{x} = \tilde{x}_{s-1}$

$\tilde{v} = \nabla F(\tilde{x})$

$x_0 = \tilde{x}$

probability $Q = \{q_1, \ldots, q_n\}$ on $\{1, \ldots, n\}$

iterate: for $k = 1, 2, \ldots, m$

pick $i_k \in \{1, \ldots, n\}$ randomly according to Q

$v_k = (\nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\tilde{x}))/(q_{i_k} n) + \tilde{v}$

$x_k = \mathrm{prox}_{\eta R}(x_{k-1} - \eta v_k)$

end

set $\tilde{x}_s = \frac{1}{m} \sum_{k=1}^{m} x_k$

end

Convergence analysis

- **theorem:** let $L_Q = \max_i\{L_i/(q_i n)\}$, suppose $0 < \eta \leq 1/4L_Q$ and m sufficiently large so that

$$\rho = \frac{1}{\mu\eta(1-4L_Q\eta)m} + \frac{4L_Q\eta(m+1)}{(1-4L_Q\eta)m} < 1$$

  then we have geometric convergence in expectation:

$$\mathbb{E}P(\tilde{x}_s) - P(x_\star) \leq \rho^s[P(\tilde{x}_0) - P(x_\star)]$$

- we always have $L \leq L_Q$ and smallest possible L_Q is

$$L_Q = \frac{1}{n}\sum_{i=1}^{n} L_i, \quad \text{when} \quad q_i = \frac{L_i}{\sum_{j=1}^{n} L_j}$$

- overall complexity: $O\left(\left(\frac{L_Q}{\mu} + n\right)\log\left(\frac{1}{\epsilon}\right)\right)$

Numerical experiments

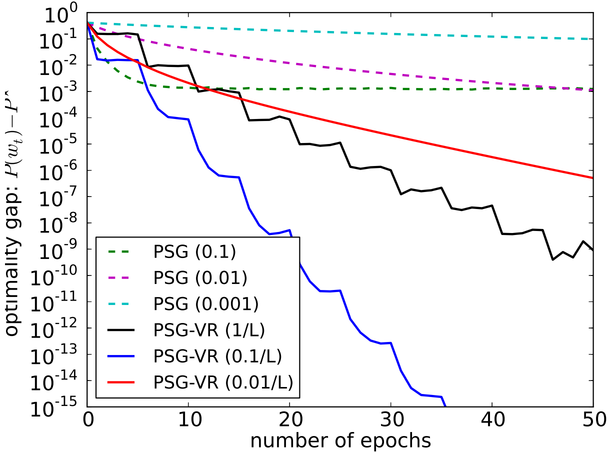
- data $(a_1, b_1), \ldots, (a_n, b_n)$ with $a_i \in \mathbb{R}^d$ and $b_i \in \{+1, -1\}$
- regularized logistic regression

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \log\big(1 + \exp(-b_i a_i^T x)\big) + \lambda_2 \|x\|_2^2 + \lambda_1 \|x\|_1$$

- data set: RCV1 binary training (LIBSVM website)
  - $n = 20,242$
  - $d = 47,236$
  - sparsity: 0.16%
  - $\lambda_2 = 0.0001$
  - $\lambda_1 = 0.00001$

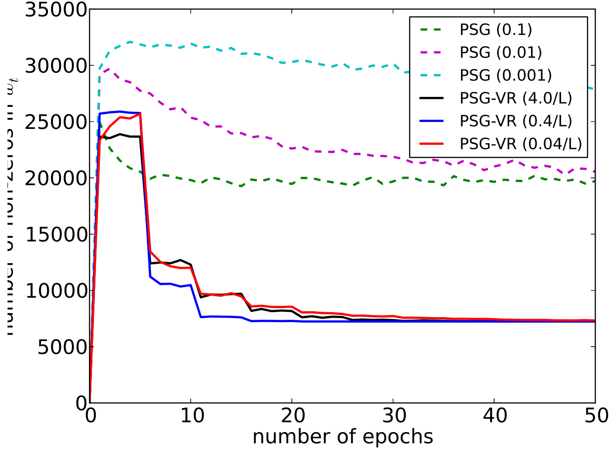
Comparison with Prox-SG: objective value

$m = 5$



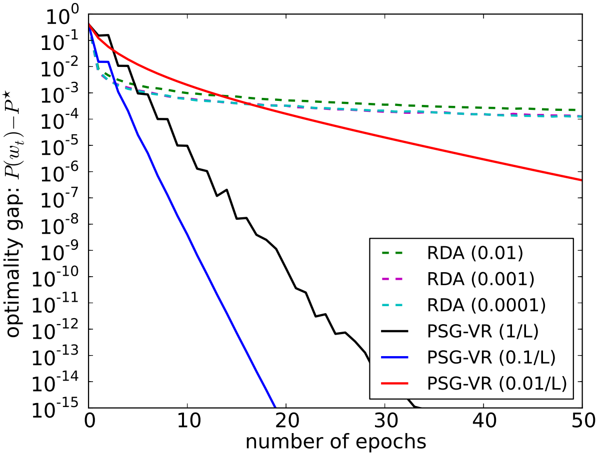
Comparison with Prox-SG: sparsity

$$m = 5$$



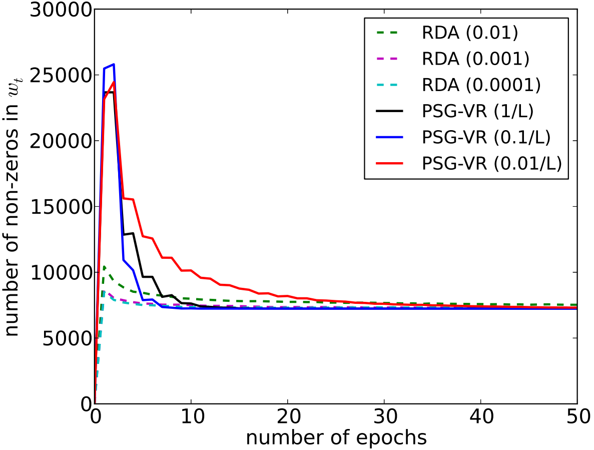
Comparison with RDA: objective value

$m = 2$



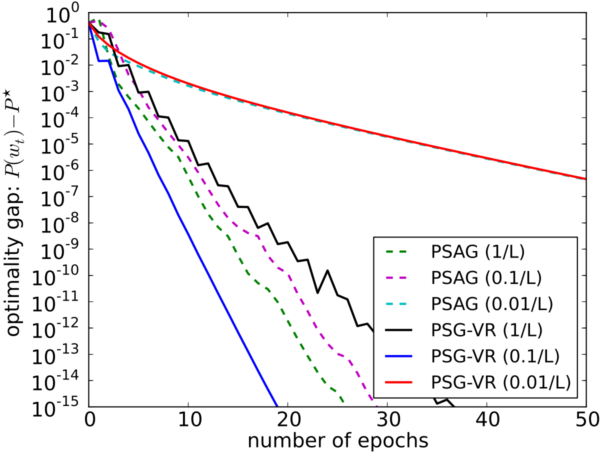
Comparison with RDA: sparsity

$$m = 2$$



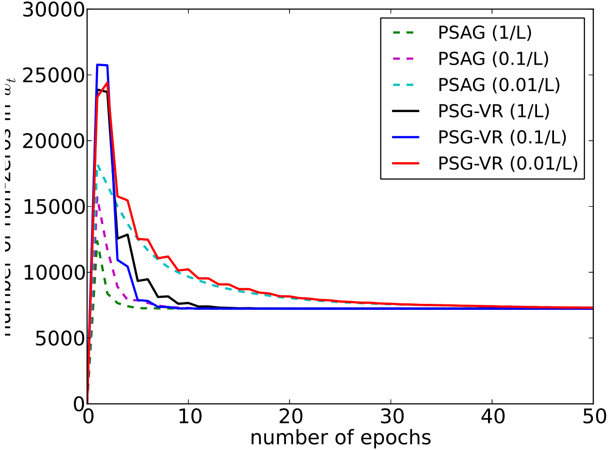
Comparison with Prox-SAG: objective value

$$m = 2$$



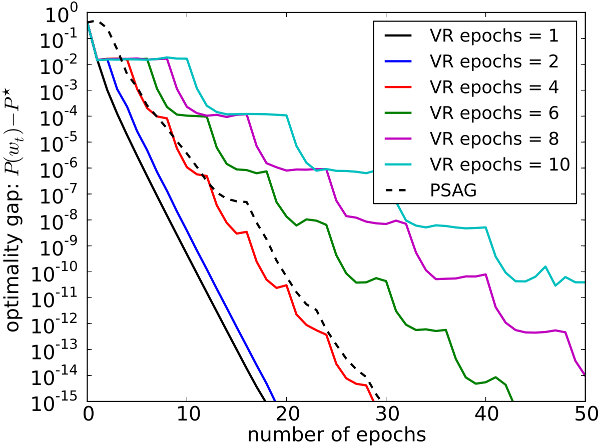
Comparison with Prox-SAG: sparsity

$m = 2$



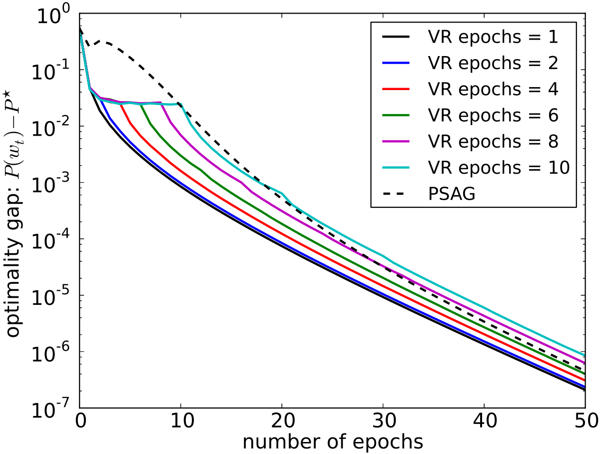
Different variance reduction periods

$\lambda_2 = 0.0001$



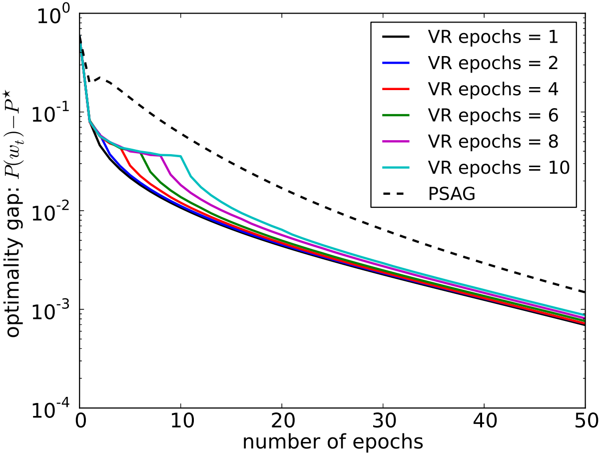
Different variance reduction periods

$$\lambda_2 = 0.00001$$



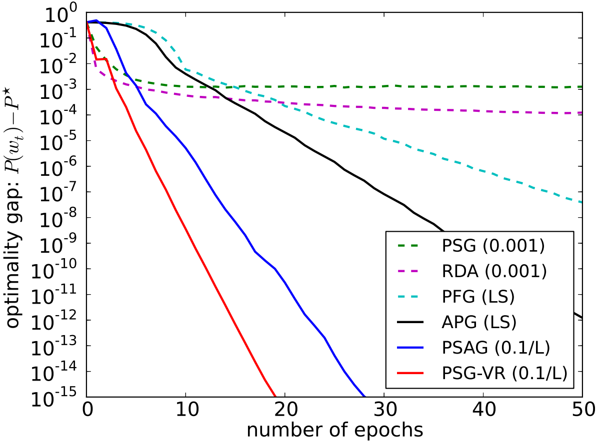
Different variance reduction periods

$$\lambda_2 = 0.000001$$



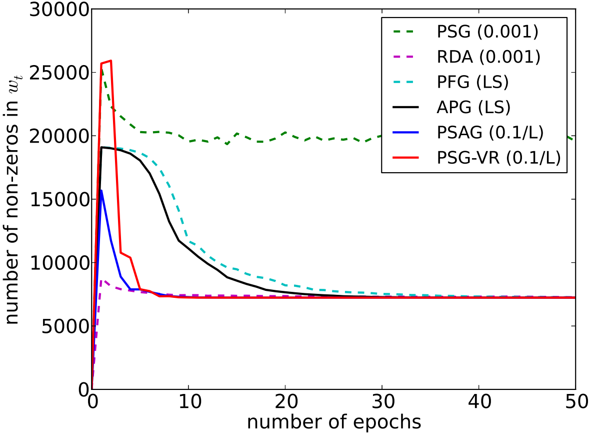
Comparison of different methods: objective value

$$\lambda_2 = 0.0001 \qquad m = 2$$



Comparison of different methods: sparsity

$$\lambda_2 = 0.0001 \qquad m = 2$$



Summary: Prox-SVRG

- exploit finite average structure to obtain faster convergence rate
- extended SVRG to proximal setting, established same complexity
- developed weighted sampling scheme for Prox-SVRG
- preliminary numerical experiments comparable with (Prox-) SAG

Summary: Prox-SVRG

- exploit finite average structure to obtain faster convergence rate
- extended SVRG to proximal setting, established same complexity
- developed weighted sampling scheme for Prox-SVRG
- preliminary numerical experiments comparable with (Prox-) SAG