Sparse Asynchronous Distributed Learning

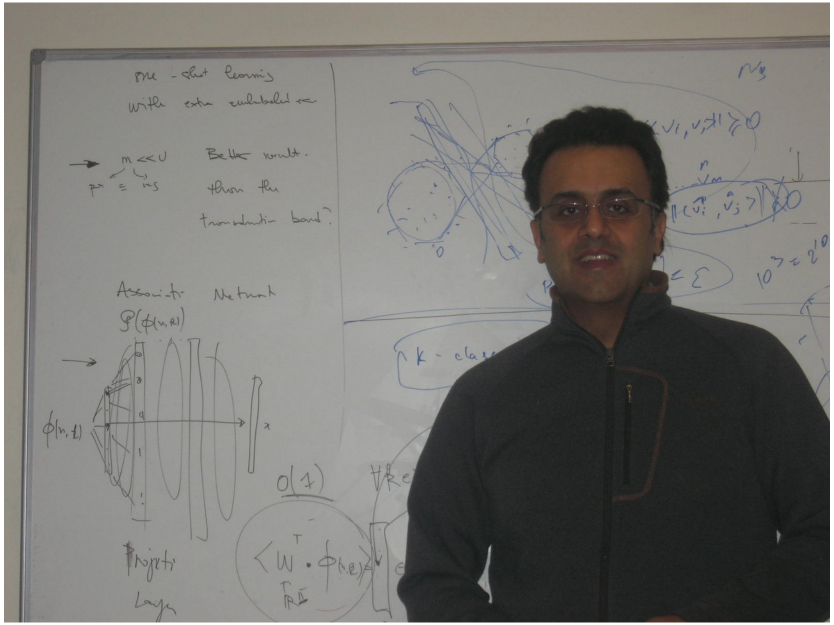
Dmitry Grishchenko

Université Grenoble Alpes

18 November 2020

Collaborators

Frank Iutzeler



Massih-Reza Amini

Problem

$$\min_{x \in \mathbb{R}^n} \underbrace{\frac{1}{m} \sum_{i=1}^{m} \ell(b_i, h(a_i, x))}_{f(x)} + \lambda_1 \|x\|_1$$

Problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{M} \alpha_i \underbrace{\left[\frac{1}{|\mathcal{D}_i|} \sum_{j \in \mathcal{D}_i} \ell(b_j, h(a_j, x)) \right]}_{f_i} + \lambda_1 \|x\|_1,$$

where the full dataset $\mathcal{D}$ is split onto M nonintersecting subsets $\mathcal{D}_i$ and α_i is the proportion of examples $\frac{|\mathcal{D}_i|}{m}$.

Problem

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{M} \alpha_i \underbrace{\left[\frac{1}{|\mathcal{D}_i|}\sum_{j\in\mathcal{D}_i} \ell(b_j, h(a_j, x))\right]}_{f_i} + \lambda_1 \|x\|_1,$$

L-smooth
μ-strongly convex

where the full dataset $\mathcal{D}$ is split onto M nonintersecting subsets $\mathcal{D}_i$ and α_i is the proportion of examples $\frac{|\mathcal{D}_i|}{m}$.

Problem

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{M} \alpha_i \underbrace{\left[\frac{1}{|\mathcal{D}_i|}\sum_{j\in\mathcal{D}_i} \ell(b_j, h(a_j, x))\right]}_{f_i} + \lambda_1\|x\|_1,$$

L-smooth
μ-strongly convex

where the full dataset $\mathcal{D}$ is split onto M nonintersecting subsets $\mathcal{D}_i$ and α_i is the proportion of examples $\frac{|\mathcal{D}_i|}{m}$.

These subsets $\mathcal{D}_i$ can be split over machines.

Worker

Communication

Master

Problem

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{M} \alpha_i \underbrace{\left[\frac{1}{|\mathcal{D}_i|}\sum_{j\in\mathcal{D}_i} \ell(b_j, h(a_j, x))\right]}_{f_i} + \lambda_1 \|x\|_1,$$

L-smooth
μ-strongly convex

where the full dataset $\mathcal{D}$ is split onto M nonintersecting subsets $\mathcal{D}_i$ and α_i is the proportion of examples $\frac{|\mathcal{D}_i|}{m}$.

These subsets $\mathcal{D}_i$ can be split over machines.

Worker

Communication

Bottleneck

Master

Algorithm: DAve-PG

Master

$$\bar{x}^k \leftarrow \bar{x}^{k-1} + \alpha_i \Delta^k$$
$$x^k \leftarrow \mathbf{prox}_{\gamma\lambda_1\|\cdot\|_1}(\bar{x}^k)$$

x^k $\quad$ Δ^k

Worker 1

Computation of

$\nabla f_1(x^{k-d_1^k})$

in process.

•••

Worker i

$$x_i^+ = x^{k-d_i^k} - \gamma \nabla f_i(x^{k-d_i^k})$$
$$\Delta^k \leftarrow x_i^+ - x_i$$
$$x_i \leftarrow x_i + \Delta^k$$

•••

Worker M

Computation of

$\nabla f_M(x^{k-d_M^k})$

in process.

Konstantin Mishchenko, Franck Iutzeler, Jérôme Malick, and Massih-Reza Amini. *A Delay-tolerant Proximal-Gradient Algorithm for Distributed Learning,* International Conference on Machine Learning, 3584-3592

Algorithm: SPY

Master

$$\bar{x}^k \leftarrow \bar{x}^{k-1} + \alpha_i \Delta^k$$
$$x^k \leftarrow \mathbf{prox}_{\gamma\lambda_1\|\cdot\|_1}(\bar{x}^k)$$

x^k

$[\Delta^k]_{\mathbf{S}^{k-d_i^k}}$

Worker 1

Computation of

$\nabla f_1(x^{k-d_1^k})$

in process.

Worker i

$$x_i^+ = x^{k-d_i^k} - \gamma \nabla f_i(x^{k-d_i^k})$$
$$\Delta^k \leftarrow x_i^+ - x_i$$
$$x_i \leftarrow x_i + [\Delta^k]_{\mathbf{S}^{k-D_i^k}}$$

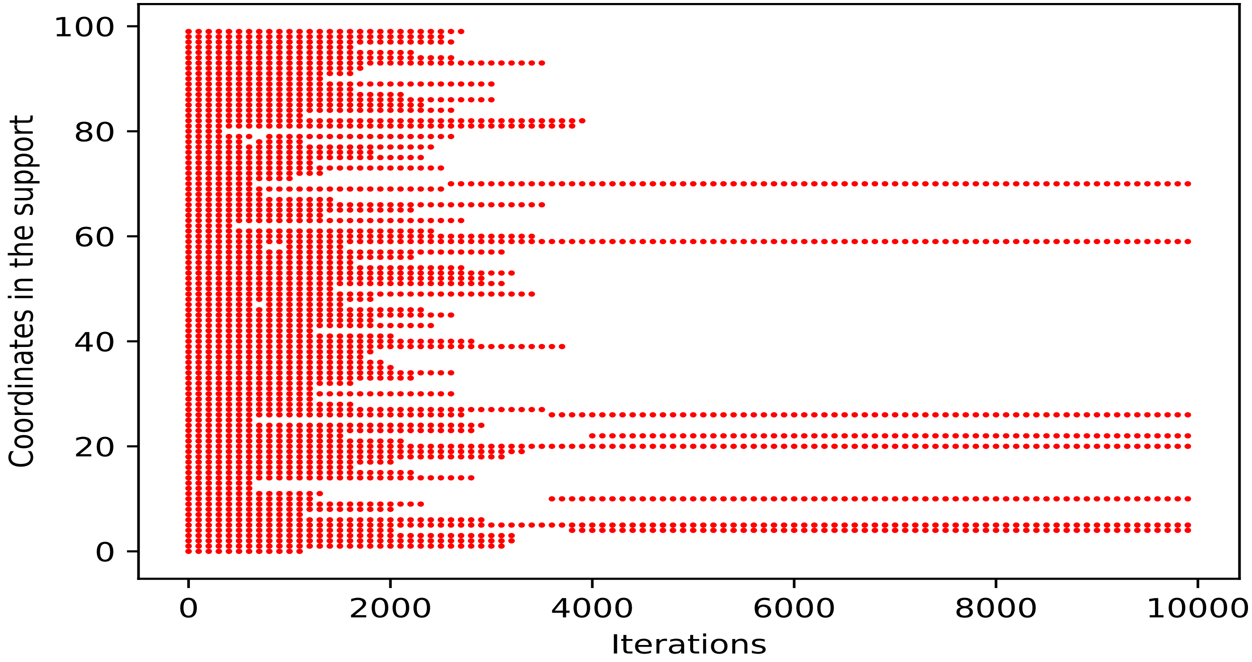
Worker M

Computation of

$\nabla f_M(x^{k-d_M^k})$

in process.

Identification (Example)



Synthetic LASSO problem min $\frac{1}{2}\|Ax - b\|_2^2 + \lambda_1\|x\|_1$ for random generated matrix $A \in \mathbb{R}^{100\times 100}$ and vector $b \in \mathbb{R}^{100}$ and hyperparameter λ_1 chosen to reach 8% of density (amount of non-zero coordinates) of the final solution.

Adaptive Selection

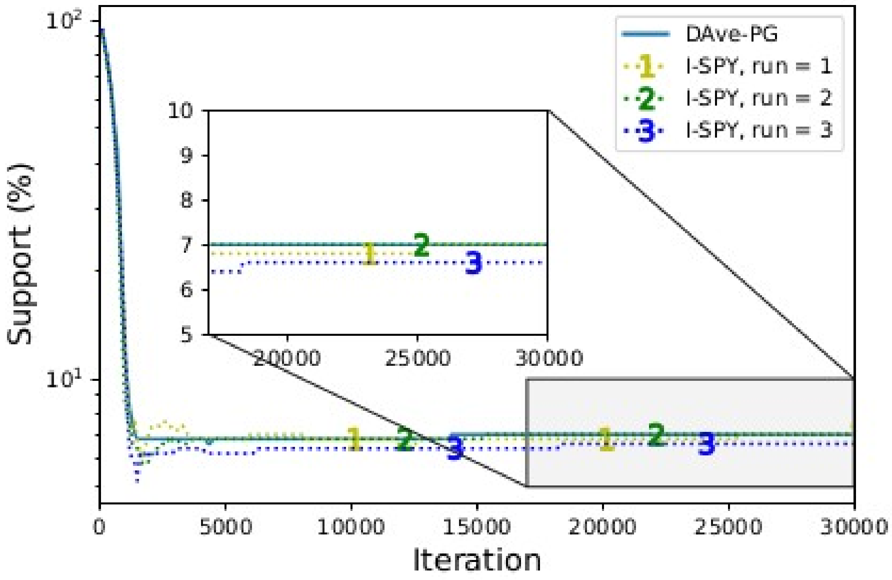
p is π-priority random vector w.r.t. the current iterate point x^k

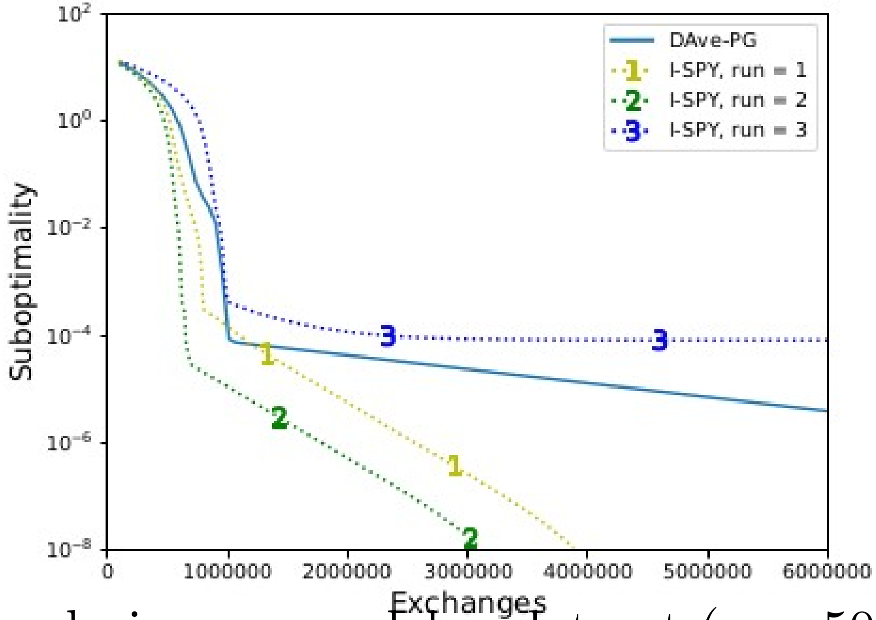
$$\mathbb{P}\left[j \in \mathbf{S}_\pi^k\right] = \begin{cases} 1 & \text{if j} \in \text{supp}(x^k), \\ \pi & \text{otherwise.} \end{cases}$$

This selection is not i.i.d.!

If support is fixed the selection is i.i.d.!

Experiments: Adaptive Selection

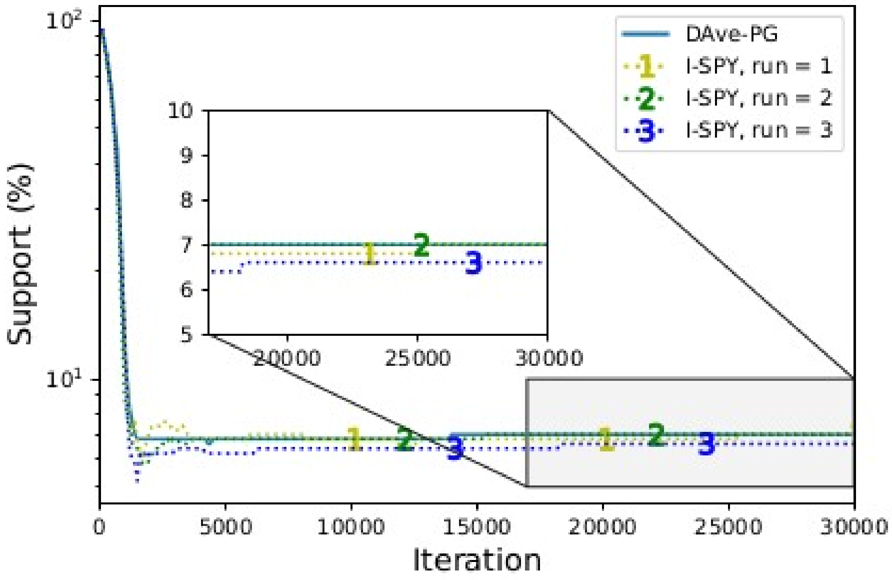


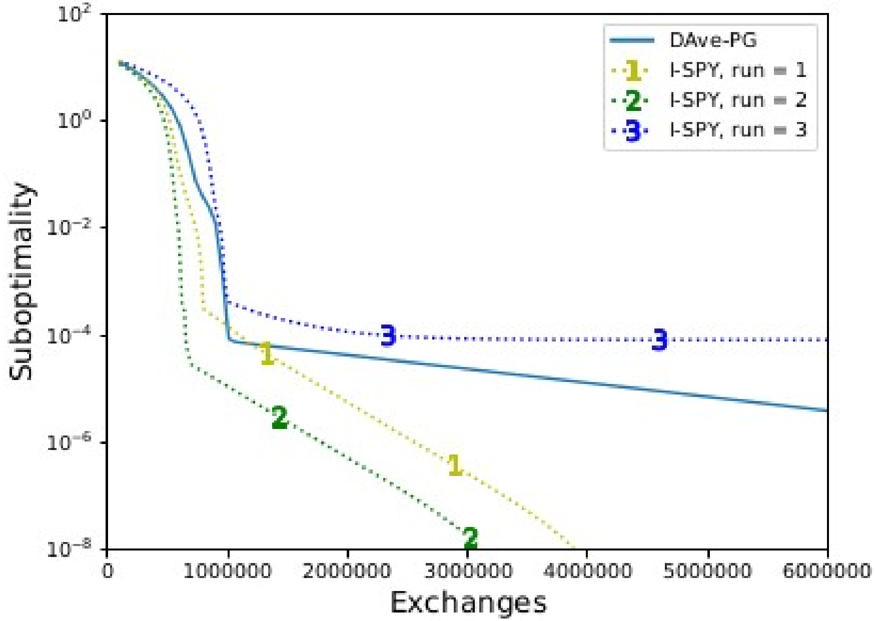


Logistic regression with elastic net regularizer on madelon dataset ($n = 500$ $m = 2000$) and $M = 10$ machines.

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{m}\sum_{j=1}^{m} \log(1+\exp(-y_j z_j^\top x)) + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|_2^2$$

Experiments: Adaptive Selection





It is better if it converges, but it can diverge!

Mask Selection

Mask Selection

$$\left[\Delta^k\right]_{\mathbf{S}^k}$$

Mask Selection

$$\left[\Delta^k\right]_{\mathbf{S}^k}$$

Random sparsificaiton with $p = (p_1, .., p_n) \in (0, 1]^n$.

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\}.$$

Mask Selection

$$\left[\Delta^k\right]_{\mathbf{S}^k}$$

Random sparsificaiton with $p = (p_1, .., p_n) \in (0,1]^n$.

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\}.$$

- p is an arbitrary probability vector.

Mask Selection

$$\left[\Delta^k\right]_{\mathbf{S}^k}$$

Random sparsificaiton with $p = (p_1, .., p_n) \in (0,1]^n$.

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\}.$$

- p is an arbitrary probability vector.

- p is a π-uniform probability vector.

Mask Selection

$$\left[\Delta^k\right]_{\mathbf{S}^k}$$

Random sparsificaiton with $p = (p_1, .., p_n) \in (0, 1]^n$.

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\}.$$

- p is an arbitrary probability vector.
- p is a π-uniform probability vector.
- p is a π-priority random vector w.r.t. some point x

$$\mathbb{P}\left[j \in \mathbf{S}_\pi^k\right] = \begin{cases} 1 & \text{if j} \in \text{supp}(x), \\ \pi & \text{otherwise.} \end{cases}$$

General Theoretical Result

General Theoretical Result

Assumption (on randomness)

The sparsity mask selectors $(\mathbf{S}_p^k)$ are independent and identically distributed random variables. We select a coordinate in the mask as follows:

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\},$$

with $p = (p_1, .., p_n) \in (0, 1]^n$.

General Theoretical Result

Assumption (on randomness)

The sparsity mask selectors $(\mathbf{S}_p^k)$ are independent and identically distributed random variables. We select a coordinate in the mask as follows:

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\},$$

with $p = (p_1, .., p_n) \in (0, 1]^n$.

Theorem (Limits of sparsification)

Take $\gamma = \frac{2}{\mu+L}$, then `SPY` verifies for all $k \in [k_m, k_{m+1})$

$$\mathbb{E}\left\|x^k - x^\star\right\|^2 \leq \left(p_{\max}\left(\frac{1-\kappa_{\mathrm{P}}}{1+\kappa_{\mathrm{P}}}\right)^2 + 1 - p_{\min}\right)^m \max_i \left\|x_i^0 - x_i^\star\right\|^2.$$

with the shifted local solutions $x_i^\star = x^\star - \gamma_i \nabla f_i(x^\star)$.

General Theoretical Result

Assumption (on randomness)

The sparsity mask selectors $(\mathbf{S}_p^k)$ are independent and identically distributed random variables. We select a coordinate in the mask as follows:

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\},$$

with $p = (p_1, .., p_n) \in (0, 1]^n$.

Theorem (Limits of sparsification)

Take $\gamma = \frac{2}{\mu+L}$, then $\texttt{SPY}$ verifies for all $k \in [k_m, k_{m+1})$

$$\mathbb{E}\left\|x^k - x^\star\right\|^2 \leq \underbrace{\left(p_{\max}\left(\frac{1-\kappa_{\mathrm{P}}}{1+\kappa_{\mathrm{P}}}\right)^2 + 1 - p_{\min}\right)}_{\in (0,1)}{}^{m} \max_i \left\|x_i^0 - x_i^\star\right\|^2.$$

with the shifted local solutions $x_i^\star = x^\star - \gamma_i \nabla f_i(x^\star)$.

General Theoretical Result

Assumption (on randomness)

The sparsity mask selectors $(\mathbf{S}_p^k)$ are independent and identically distributed random variables. We select a coordinate in the mask as follows:

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\},$$

with $p = (p_1, .., p_n) \in (0, 1]^n$.

Theorem (Limits of sparsification)

Take $\gamma = \frac{2}{\mu+L}$, then SPY with π-uniform sampling verifies for all $k \in [k_m, k_{m+1})$

$$\mathbb{E}\left\|x^k - x^\star\right\|^2 \leq \left(1 - \pi\frac{4\mu L}{(\mu+L)^2}\right)^m \max_i \left\|x_i^0 - x_i^\star\right\|^2.$$

with the shifted local solutions $x_i^\star = x^\star - \gamma_i \nabla f_i(x^\star)$.

General Theoretical Result

Assumption (on randomness)

The sparsity mask selectors $(\mathbf{S}_p^k)$ are independent and identically distributed random variables. We select a coordinate in the mask as follows:

$$\mathbb{P}[j \in \mathbf{S}_p^k] = p_j > 0 \qquad \text{for all } j \in \{1, .., n\},$$

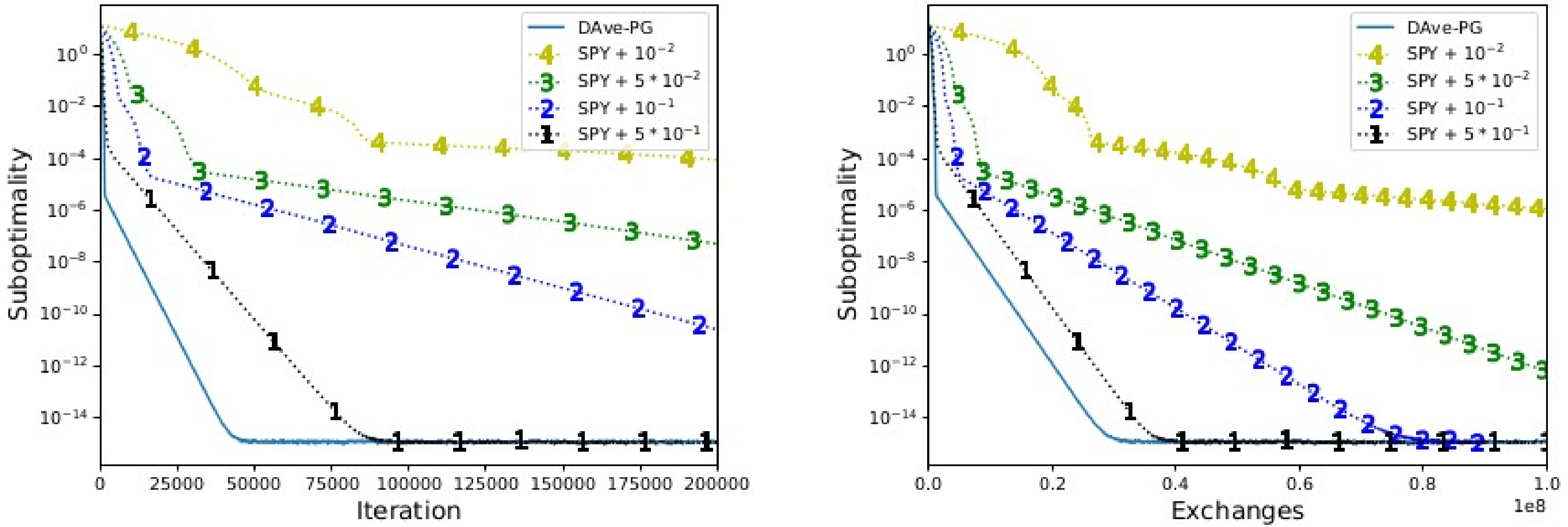
with $p = (p_1, .., p_n) \in (0, 1]^n$.

Limits of sparsification

`SPY` reaches linear convergence of the mean squared error in terms of epochs if

$$\frac{p_{\min}}{p_{\max}} > (1 - \gamma\mu)^2 \overset{\gamma = \frac{2}{\mu + L}}{\geq} \left(\frac{1 - \kappa_{\mathsf{P}}}{1 + \kappa_{\mathsf{P}}}\right)^2.$$

Uniform Sampling



Logistic regression with elastic net regularizer on madelon dataset ($n = 500$ $m = 2000$) and $M = 10$ machines.

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{m}\sum_{j=1}^{m}\log(1+\exp(-y_j z_j^\top x)) + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|_2^2$$

Identification (Example)

$$x^\star = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{i=1}^{M} \alpha_i f_i(x) + \lambda_1 \|x\|_1$$

Theorem (Enlarged identification)

Let (u^k) be an $\mathbb{R}^n$-valued sequence converging almost surely to $u^\star$ and define sequence (x^k) as $x^k = \mathbf{prox}_{\gamma r}(u^k)$ and $x^\star = \mathbf{prox}_{\gamma r}(u^\star)$. Then (x^k) identifies some subspaces with probability one; more precisely for any $\varepsilon > 0$, with probability one, after some finite time,

$$\operatorname{supp}(x^\star) \subseteq \operatorname{supp}(x^k) \subseteq \max_{u\in\mathcal{B}(u^\star,\varepsilon)} \left\{\operatorname{supp}(\mathbf{prox}_{\gamma r}(u))\right\}.$$

Non-degeneracy

Another way to define the non-degeneracy for the problem

$$\min_{x \in \mathbb{R}^n} f(x) + r(x)$$

is the following:

$$\nabla f(x^\star) \in \operatorname{ri} \partial r(x^\star).$$

In case of ℓ_1 regularizer $r(x) = \lambda_1 \|x\|_1$ this can be written explicitly as

$$\left| \nabla f(x^\star)_{[j]} \right| < \lambda_1 \qquad \text{for all } j \in \operatorname{supp}(x^\star).$$

Better rate

Assumption (Convergence)

Let us assume that for any $\varepsilon > 0$ there exists iterate number K such that for any $k > K$, the average point $\|\bar{x}^k - \bar{x}^\star\|_2^2 < \varepsilon$ is ε-close to the final solution.

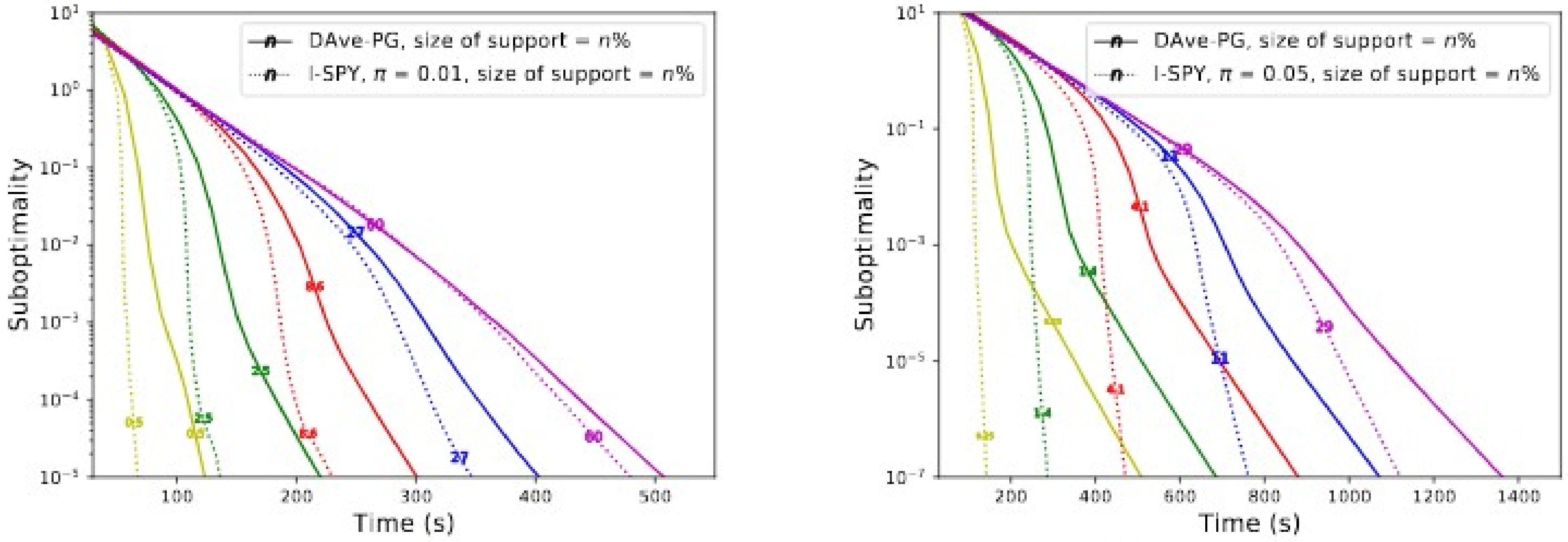
Theorem (Better rate)

Suppose that Assumption holds. For any $\gamma \in (0, 2/(\mu + L)]$ and for any $k \in [k_s, k_{s+1})$ we have:

$$\|x^k - x^\star\|^2 = \mathcal{O}_p\left(\left(1 - \frac{2\gamma\mu L}{\mu + L}\right)^s\right),$$

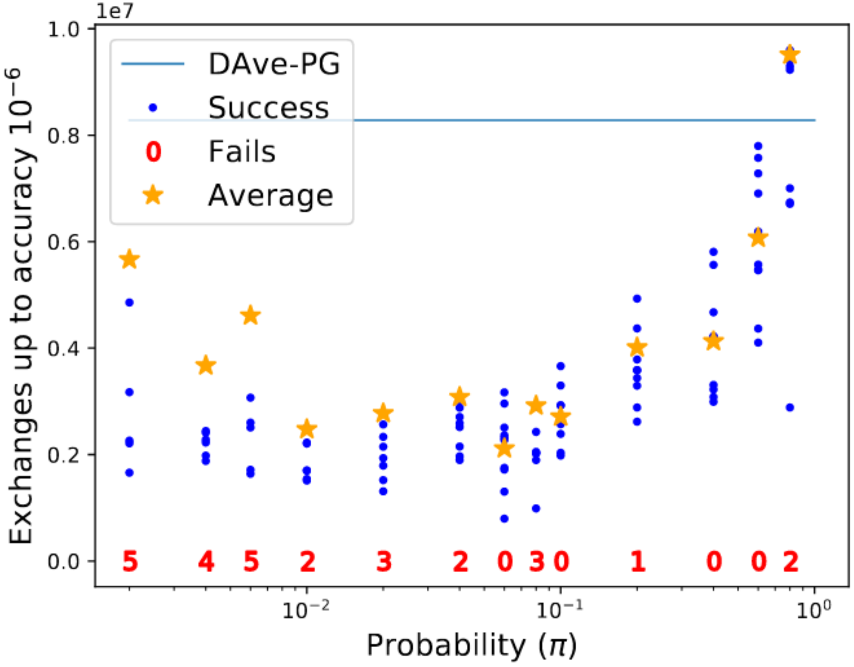
where $\mathcal{O}_p$ denotes big O in probability.

Time performance



(a) real-sim (b) rcv1_train

Stability



Logistic regression with elastic net regularizer on madelon dataset ($n = 500$ $m = 2000$) and $M = 10$ machines.

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{m} \sum_{j=1}^{m} \log(1+\exp(-y_j z_j^\top x)) + \lambda_1 \|x\|_1 + \frac{\lambda_2}{2} \|x\|_2^2$$


Thank You For
Your Attention!