Distributed Optimization with Sparse Communications

D. GRISHCHENKO[1,2], F. IUTZELER[1], J. MALICK[1] and M.-R. AMINI[2]

1. DAO team, LJK, Université Grenoble Alpes
2. AMA team, LIG, Université Grenoble Alpes

Correspondence: name.lastname@univ-grenoble-alpes.fr

Overview

Context

- **Optimization algorithms:** find minimizer of convex functions
- **Distributed setting:** several machines, without shared data
- **Communications between machines:** bottleneck

Results

We present a **distributed** version of **proximal gradient decent** with constant stepsize and **two-way sparse communications** with **linear convergence**.

Model

- **Distributed learning:**
  n observations are split over M machines
  machine i has a private examples subset $\mathcal{S}_i$ $\sum \mathcal{S}_i = \mathcal{S}$ – full set of examples
- **Shared prediction without moving data:**
  decoupling the ability to learn from the need to store the data in a centralized way.

Problem: $\boxed{\min_{x\in\mathbb{R}^d} \sum_{i=1}^{M} \pi_i f_i(x) + r(x)}$

$\pi_i = n_i/n$ the proportion of observations locally stored in machine i

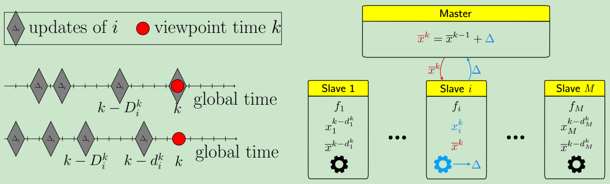
$f_i(x) = \frac{1}{n_i}\sum_{j\in\mathcal{S}_i} \ell_j(x)$ the local empirical risk estimated on machine i

Assumptions

- **On functions:** all f_i are $L-$smooth and $\mu-$strongly convex
- **On regularizer:** r is convex
  $x^\star$ – unique minimizer

Notations

- **For the master:**
  k = number of updates the master receives from any of the slaves
  $k_{m+1} = \min\left\{k : \text{each machine made at least 2 updates on the interval } [k_m, k]\right\}$
- **For slave i:**
  d_i^k = time elapsed from the last update slave i to the master
  D_i^k = time of the penultimate update



Sparsification of local updates

Master machine asynchronously gathers *sparsified* delayed gradient updates from slaves and sends them back the current point. At iteration k, this randomly drawn subset of entries of the gradient to be computed by agent i^k is called **mask** and is denoted by $\mathbf{S}^k$. Using $x_{[j]}$ is the j-th coordinate of $x \in \mathbb{R}^d$

$$x^k_{i[j]} = \begin{cases} \left(x^{k-D_i^k} - \gamma\nabla f_i(x^{k-D_i^k})\right)_{[j]} & \text{if } i = i^k \text{ and } j \in \mathbf{S}^{k-D_i^k} \\ x^{k-1}_{i[j]} & \text{otherwise} \end{cases}$$

$$x^k = \mathrm{prox}_{\gamma r}\Big(\underbrace{\sum_{i=1}^{M}\pi_i x_i^k}_{:=\bar{x}^k}\Big) = \arg\min_z \left\{ r(z) + \frac{1}{2\gamma}\|x - \bar{x}^k\|^2 \right\}$$

Assumption on sparsification

The sparsity mask selectors $(\mathbf{S}^k)$ are the only random variables:

$\mathbb{P}[j \in \mathbf{S}^k] = 1$ if $j \in \mathrm{supp}(x^k)$

$\mathbb{P}[j' \in \mathbf{S}^k] = p > 0$ for all $j' \notin \mathrm{supp}(x^k)$

Delays $(D_i^k)_{i=1,..,M}$ are independent of the future mask selectors $\{\mathbf{S}^\ell\}_{\ell\geq k}$.

Master

Initialize $\bar{x}^0$
while *not converged* **do**
  Receive $[\Delta^k]_{\mathbf{S}^{k-D_i^k}}$ from agent $i = i^k$
  $\bar{x}^k \leftarrow \bar{x}^{k-1} + \pi_i[\Delta^k]_{\mathbf{S}^{k-D_i^k}}$
  $x^k \leftarrow \mathrm{prox}_{\gamma r}(\bar{x}^k)$
  Choose sparsity mask $\mathbf{S}^k$
  Send $x^k, \mathbf{S}^k$ to agent $i = i^k$
end

Slave i

Initialize $x_i = x_i^+ = x = \bar{x}^0$
while *not interrupted by master* **do**
  $[x^+]_{\mathbf{S}\setminus\mathrm{supp}(x)} \leftarrow [x - \gamma\nabla f_i(x)]_{\mathbf{S}\setminus\mathrm{supp}(x)}$
  $[x^+]_{\mathrm{supp}(x)} \leftarrow p[x - \gamma\nabla f_i(x))]_{\mathrm{supp}(x)} + (1-p)[x_i]_{\mathrm{supp}(x)}$
  $\Delta \leftarrow x^+ - x$
  Send $[\Delta]_{\mathbf{S}}$ to master
  $[x_i]_{\mathbf{S}} \leftarrow [x_i^+]_{\mathbf{S}}$
  Receive x and $\mathbf{S}$ from master
end

Convergence rate

Take $\gamma \in (0, 2/(\mu+L)]$. **Then, for all** $k \in [k_m, k_{m+1})$,

$$\mathbb{E}\|x^k - x^\star\|^2 \leq \left(1 - 2\frac{\gamma p\mu L}{\mu + L}\right)^m \max_{i=1,..,M}\|x_i^0 - x_i^\star\|^2$$

- Linear convergence
- Same step-size as in standard proximal gradient
- If $M = 1$ and $p = 1$ – usual convergence rate

Identification

Assumptions

- **On regularizer:** let $r(x) = \lambda_1\|x\|_1$, **then**

$$\mathrm{prox}_{\gamma r}(x) = \begin{cases} x - \gamma\lambda_1 & \text{if } x > \gamma\lambda_1 \\ x + \gamma\lambda_1 & \text{if } x < -\gamma\lambda_1 \\ 0 & \text{otherwise} \end{cases}$$

- **On delays:** $\exists C : \forall\varepsilon > 0, m \in \mathbb{Z}_+, k_{m+1} - k_m \leq C(1+\varepsilon)^m$

Identification result The algorithm identifies a near-optimal support in finite time with probability one:

$$\exists K : \forall k \geq K, \quad \mathrm{supp}(x^\star) \subseteq \mathrm{supp}(x^k) \subseteq \mathrm{supp}(y_\varepsilon^\star)$$

where $y_\varepsilon^\star = \mathrm{prox}_{\gamma(1-\varepsilon)r}(\bar{x}^\star - x^\star)$ for any $\varepsilon > 0$. Furthermore, if the problem is non-degenerate, i.e. $-\sum_{i=1}^{M}\pi_i\nabla f_i(x^\star) \in \mathrm{ri}\,\partial r(x^\star)$ then, the algorithm identifies the optimal support with probability one:

$$\exists K : \forall k \geq K, \quad \mathrm{supp}(x^k) = \mathrm{supp}(x^\star)$$

Sparsity This identification result gives us **two-way sparsity** of algorithm in terms of communications.

Numerical experiments

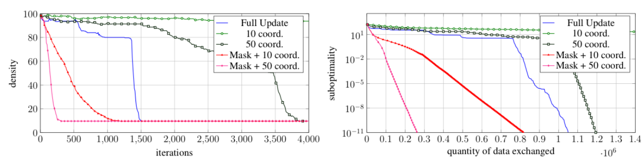


Figure 1: Evolution of the iterates density and functional suboptimality versus quantity of exchanged data on the lasso problem.

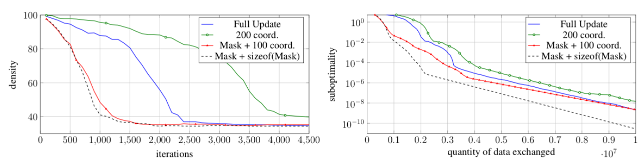


Figure 2: Evolution of the iterates density and functional suboptimality versus quantity of exchanged data on the logistic regression problem.

References

1. Jalal Fadili, Jérôme Malick, & Gabriel Peyré. *Sensitivity analysis for mirror-stratifiable convex functions.* to appear in SIAM Journal on Optimization, 2018.
2. Jianqiao Wangni, Jialei Wang, Ji Liu, & Tong Zhang. *Gradient sparsification for communication-efficient distributed optimization.* arXiv:1710.09854 , 2017.
3. Rémi Leblond, Fabian Pedregosa, & Simon Lacoste-Julien. *Asaga: Asynchronous parallel saga.* AISTATS 2017..
4. Bikash Joshi, Franck Iutzeler, & Massih-Reza Amini. *An Asynchronous Distributed Framework for Large-scale Learning Based on Parameter Exchanges,* to appear in International Journal of Data Science and Analytics, 2018.