A Delay-tolerant Proximal-Gradient Algorithm for Distributed Learning

SUPPLEMENTARY MATERIAL

A. Main technical lemma

The improvement of one iteration (with potentially inner multiple steps) at slave i is given by Lemma 1 of the main text which is the core technical lemma of the paper. We give its proof in the supplementary document. For convenience, we restate it below. Let us also recall the rule used to produce x_i^k.

Algorithm 1 Update rule for x_i^k.

Input: $\overline{x} = \overline{x}^{k-D_i^k}$
Take $x = x_i^{k-D_i^k}$ from the previous iteration
Select a number of repetitions p
Initialize $\Delta = 0$
for $q \leftarrow 1$ to p **do**
  $z \leftarrow \mathrm{prox}_{\gamma r}(\overline{x} + \Delta)$
  $x^+ \leftarrow z - \gamma \frac{1}{n_i} \sum_{j \in \mathcal{S}_i} \nabla \ell_j(z)$
  $\Delta \leftarrow \Delta + \frac{1}{M}(x^+ - x)$
  $x \leftarrow x^+$
end for
Output: $x_i^k = x$, $\Delta_i^k = \Delta$

Lemma 1. *Let Assumption 1 hold. For any i, define $x_i^\star = x^\star - \gamma \frac{1}{n_i} \sum_{j \in \mathcal{S}_i} \nabla \ell_j(x^\star)$. Then the algorithm's local iterates at slave i satisfy, for all k,*

$$\left\|x_i^k - x_i^\star\right\|^2 \leqslant (1-\gamma\mu)^2 c_{k-D_i^k}$$

where for all moment k we define

$$c_k = \max\left(\left\|\overline{x}^k - \overline{x}^\star\right\|^2, \left\|\overline{x}_{-i(k)}^k - \overline{x}_{-i(k)}^\star\right\|^2\right)$$

with $i(k)$ being the slave making the update k, and

$$\overline{x}^\star = \sum_{i=1}^M \pi_i x_i^\star, \qquad \overline{x}_{-i}^\star = \Big(\sum_{j\neq i} \pi_j\Big)^{-1} \sum_{j\neq i} \pi_j x_j^\star = \sum_{j\neq i} \frac{n_j}{n-n_i} x_j^\star, \qquad \overline{x}_{-i}^k = \Big(\sum_{j\neq i} \pi_j\Big)^{-1} \sum_{j\neq i} \pi_j x_j^k = \sum_{j\neq i} \frac{n_j}{n-n_i} x_j^k.$$

Proof. **Part 1.** First, we are going to prove a related result that gives a contraction result for the the slave machine that is updating at time k. By definition, we have $d_i^k = 0$, and the machine i last time started computing at moment $k - D_i^k$. Removing the global time index k for better readability, Let us also consider x_p, x_p^+, z_p are the local variables x, x^+, z at the last (the p-th) inner loop of slave i.

First, we notice that $f_i = \frac{1}{n_i}\sum_{j\in\mathcal{S}_i} \ell_j$ is a μ-strongly convex and L-smooth function, so that we write

$$\begin{aligned}
&\left\|x_i^k - x_i^\star\right\|^2 = \left\|x_p^+ - x_i^\star\right\|^2 \\
&= \left\|z_p - \gamma\nabla f_i(z_p) - (x^\star - \gamma \nabla f_i(x^\star))\right\|^2 \\
&= \left\|z_p - x^\star\right\|^2 + \gamma^2 \left\|\nabla f_i(z_p) - \nabla f_i(x^\star)\right\|^2 - 2\gamma\langle z_p - x^\star, \nabla f_i(z_p) - \nabla f_i(x^\star)\rangle \\
&\leqslant \left\|z_p - x^\star\right\|^2 + \gamma^2 \left\|\nabla f_i(z_p) - \nabla f_i(x^\star)\right\|^2 - 2\frac{\gamma}{\mu+L}\left\|\nabla f_i(z_p) - \nabla f_i(x^\star)\right\|^2 - 2\frac{\gamma\mu L}{\mu+L}\left\|z_p - x^\star\right\|^2 \\
&= \left(1 - \frac{2\gamma\mu L}{\mu+L}\right)\left\|z_p - x^\star\right\|^2 + \gamma\left(\gamma - \frac{2}{\mu+L}\right)\left\|\nabla f_i(z_p) - \nabla f_i(x^\star)\right\|^2. \qquad (1)
\end{aligned}$$

Moreover, since $\gamma \leqslant \frac{2}{\mu+L}$, the second term in (1) is negative and we can use strong convexity to further bound it as

$$\begin{aligned}\left\|x_i^k - x_i^\star\right\|^2 = \left\|x_p^+ - x_i^\star\right\|^2 &\leqslant \left(1 - \frac{2\gamma\mu L}{\mu+L}\right)\|z_p - x^\star\|^2 + \gamma\left(\gamma - \frac{2}{\mu+L}\right)\|\nabla f_i(z_p) - \nabla f_i(x^\star)\|^2 \\ &\leqslant \left(1 - \frac{2\gamma\mu L}{\mu+L}\right)\|z_p - x^\star\|^2 + \gamma\left(\gamma - \frac{2}{\mu+L}\right)\mu^2\|z_p - x^\star\|^2 \\ &= (1-\gamma\mu)^2\|z_p - x^\star\|^2. \end{aligned} \tag{2}$$

Using the non-expansiveness of the proximal operator, we have

$$\begin{aligned}\|z_p - x^\star\|^2 &= \left\|\mathrm{prox}_{\gamma r}(\overline{x} + \Delta_{p-1}) - \mathrm{prox}_{\gamma r}(\overline{x}^\star)\right\|^2 \\ &\leqslant \|\overline{x} + \Delta_{p-1} - \overline{x}^\star\|^2 \\ &= \left\|\sum_{j\neq i}\pi_j\left(x_j^{k-D_i^k} - x_j^\star\right) + \pi_i\left(x_{p-1} - x_i^\star\right)\right\|^2\end{aligned}$$

Note now that, as $\Delta_p = \pi_i \sum_{l=1}^p (x_l - x_{l-1})$, we have

$$\overline{x} + \Delta_p = \overline{x} + \pi_i(x_p - x_0) = \pi_i\left(\sum_{j\neq i} x_j^{k-D_i^k} + x_p\right).$$

We then get the following inequalities

$$\begin{aligned}\|z_p - x^\star\|^2 &= \left\|(1-\pi_i)\left(\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right) + \pi_i\left(x_{p-1} - x_i^\star\right)\right\|^2 \\ &\leqslant (1-\pi_i)\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\|^2 + \pi_i\|x_{p-1} - x_i^\star\|^2 \\ &\leqslant (1-\pi_i)\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\|^2 + \pi_i(1-\gamma\mu)^2\|z_{p-1} - x_i^\star\|^2. \end{aligned} \tag{3}$$

We now bound this expression, as follows

$$\begin{aligned}\|z_p - x^\star\|^2 &\leqslant \max\left(\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\|^2; \|z_{p-1} - x^\star\|^2\right) \\ &\leqslant \max\left(\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\|^2; \|z_1 - x^\star\|^2\right) \\ &= \max\left(\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\|^2; \left\|\mathrm{prox}_{\gamma r}(\overline{x}^{k-D_i^k}) - \mathrm{prox}_{\gamma r}(\overline{x})\right\|^2\right) \\ &\leqslant \max\left(\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\|^2; \left\|\overline{x}^{k-D_i^k} - \overline{x}^\star\right\|^2\right) \\ &= c_{k-D_i^k}\end{aligned} \tag{4}$$

Putting together Eqs. (2) and (4), we get that for time k and agent $i(k)$ that is updating at time k, we have

$$\left\|x_i^k - x_i^\star\right\|^2 \leqslant (1-\gamma\mu)^2 c_{k-D_i^k}.$$

Part 2. Now let us use this result to prove the full lemma. For a generic slave j, not necessarily finishing at moment k, we have $x_j^k = x_j^{k-d_j^k}$ as $k - d_j^k$ is the last time it was updated. In addition, we can apply the equation above to $x_j^{k-d_j^k}$. We, thus, obtain the claimed result that for any agent j and any time k

$$\left\|x_j^k - x_j^\star\right\|^2 = \left\|x_j^{k-d_j^k} - x_j^\star\right\|^2 \leqslant (1-\gamma\mu)^2 c_{k-D_j^k}$$

using the fact that $k - D_j^k = k - d_j^k - D_j^{k-d_j^k}$ is the penultimate update time of agent j. □

B. Resilience to infinite delays

As explained in the text, a unique feature of our algorithm is that the stepsize and convergence rate do not depend either on the delays or the computing system. The algorithm then shows of resilience to long delays as illustrated in the numerical section. Here we consider the extreme case of infinite delays, which correspond to a crash with lost of data. We show that the algorithm is still able to to converge to a point with guarantees depending of the part of the data lost and the known information at the moment of the crash

We recall our objective optimization problem

$$\min_{x\in\mathbb{R}^d} \quad \sum_{i=1}^{m} \pi_i f_i(x) + r(x). \tag{5}$$

Theorem 3. *Suppose that some machines, indexed by J, are unresponsive after moment K; so that the proposition $\pi := \sum_{j\in J}\pi_j$ of data is considered as missing after this moment. Then, the algorithm can still get an approximate solution $\hat{x}$ of the full Problem* (5) *with precision*

$$\|\hat{x} - x^\star\|^2 \leqslant \frac{\pi\delta}{1-(1-\rho)^2}$$

with $\rho = \gamma\mu$ is the convergence rate of the algorithm and $\delta \coloneqq \|\overline{x}_J^K - \overline{x}_J^\star\|^2$ measures the quality of the information available at the moment K. We use here the notation $x_J := \sum_{j\in j}\pi_j x_j$.

Proof. In this proof, we consider the case where $p = 1$ for all $k > K$; the general case follows similarly at the price of slightly heaviest notation. Take $k > K$. Let us split $\overline{x}^k$ between the weighted average iterate over the available data and the one about lost data: by Jensen's inequality

$$\|\overline{x}^k - \overline{x}^\star\|^2 \leqslant (1-\pi)\|\overline{x}_{\overline{J}}^k - \overline{x}_{\overline{J}}^\star\|^2 + \pi\|\overline{x}_J^k - \overline{x}_J^\star\|^2 = (1-\pi)\|\overline{x}_{\overline{J}}^k - \overline{x}_{\overline{J}}^\star\|^2 + \pi\delta. \tag{6}$$

From (2) in the proof of Lemma 1 and the contraction of the prox operator, we get

$$\|x_j^k - x_j^\star\|^2 \leqslant (1-\rho)^2\|\overline{x}^{k-D_j^k} - \overline{x}^\star\|^2,$$

from which we can obtain

$$\|\overline{x}_{\overline{J}}^k - \overline{x}_{\overline{J}}^\star\| \leqslant \sum_{j\in\overline{J}}\pi_j\|x_j^k - x_j^\star\| \leqslant \sum_{j\in\overline{J}}\pi_j(1-\rho)\|\overline{x}^{k-D_j^k} - \overline{x}^\star\| \leqslant (1-\rho)(1-\pi)\max_{j\in\overline{J}}\|\overline{x}^{k-D_j^k} - \overline{x}^\star\|. \tag{7}$$

Combining it with (6) yields

$$\|\overline{x}^k - \overline{x}^\star\|^2 \leqslant (1-\rho)^2(1-\pi)^3\max_{j\in\overline{J}}\|\overline{x}^{k-D_j^k} - \overline{x}^*\|^2 + \pi\delta. \tag{8}$$

From (6), observe now that $\|\overline{x}^k - \overline{x}^\star\|^2 \leqslant \max(\|\overline{x}_{\overline{J}}^k - \overline{x}_{\overline{J}}^\star\|^2; \delta) := d^k$, then (7) implies that $d^k \leqslant \max_{j\in\overline{J}} d^{k-D_j^k}$. Using the same reasoning than in the proof of Theorem 1 in the main text[1], we can prove that

$$d_k \leqslant \max_{\ell\in[k_{M-1};k_M)} d_\ell < \infty \quad \text{where } M = \max\{m : k_m \leqslant K\}.$$

Thus $\{\|\overline{x}^k - \overline{x}^*\|^2\}_k$ is bounded. Consider now c^* be the limit superior of this sequence and consider the sequence $\{l_k\}_k$ of indices such that

$$\|\overline{x}^{l_k} - \overline{x}^*\|^2 \to c^*.$$

Using (8) with l_k and taking limsup, we obtain

$$c^* = \lim_k \|\overline{x}^{l_k} - \overline{x}^*\|^2 \leqslant (1-\rho)^2(1-\pi)c^* + \pi\delta,$$

$$c^* \leqslant \frac{\pi\delta}{1-(1-\rho)^2(1-\pi)^3} \leqslant \frac{\pi\delta}{1-(1-\rho)^2}.$$

[1]The epoch sequence (k_m) just has to be slightly reformulated to take into account only the active slaves.

This means that all partial limits of $\{\|\overline{x}^k - \overline{x}^*\|^2\}_k$ are upper bounded by $\frac{\pi\delta}{1-(1-\rho)^2}$. Since $\hat{x}^k = \operatorname{prox}_{\gamma r}(\overline{x}^k)$ and the proximal operator contracts distances, we get the same result for the sequence $\{\|\hat{x}^k - x^*\|^2\}_k$. We can conclude with introducing $\hat{x}$ the limit of a converging subsequence. □

C. Refined rate with more than p_0 inner iterations

We provide here the proof of Theorem 2 of the paper.

Theorem 2. *In addition to the assumptions of Theorem 1, assume that every local loop in* `DAve-RPG` *uses $p \geqslant p_0$. Then,*

$$\left\|\hat{x}^k - x^\star\right\|^2 \leqslant [\eta(p_0)]^{2m} \max_i \left\|x_i^0 - x_i^\star\right\|^2$$

where the rate is defined from $\rho = \gamma\mu$ as

$$\eta(p_0) = (1-\rho)\left(1 - \frac{\rho}{M} - \cdots - \frac{\rho(1-\rho)^{p_0-2}}{M^{p_0-1}}\right)$$

Proof. The case $p_0 = 1$ reduces to Theorem 1, so we assume $p_0 \geqslant 2$. We use the notation of Lemma 1, and refine its argumentation with the extra information that $p \geq p_0$. Let us resume from (3): using the triangle inegality rather that the convexity inequality, we get similarly

$$\|z_p - x^\star\| \leqslant (1-\pi_i)\left\|\overline{x}_{-i}^{k-D_i^k} - \overline{x}_{-i}^\star\right\| + \pi_i(1-\rho)\|z_{p-1} - x_i^\star\|.$$

Note that we have this inequality of p but also recursively for $p-1$, $p-2$,... Encapsulating these inequalities, we obtain:

$$\begin{aligned}
\|z_p - \overline{x}^\star\| &\leqslant (1-\pi_i)\sqrt{c_{k-D_i^k}} + \pi_i(1-\rho)\|z_{p-1} - x_i^\star\| \\
&\leqslant (1-\pi_i)\sqrt{c_{k-D_i^k}} + \pi_i(1-\rho)\left((1-\pi_i)\sqrt{c_{k-D_i^k}} + \pi_i(1-\rho)\left\|t_i^k - x^\star\right\|\right) \\
&\leqslant \left(1 - \rho\pi_i - (1-\rho)\rho\pi_i^2 - \cdots - (1-\rho)^{p_0-2}\rho\pi_i^{p_0-1}\right)\sqrt{c_{k-D_i^k}}.
\end{aligned}$$

We now work for all machines at a fixed time k as follows

$$\begin{aligned}
\left\|\overline{x}^k - \overline{x}^\star\right\| &\leqslant \sum_{i=1}^M \left\|\pi_i(x_i^k - x_i^\star)\right\| \leqslant \sum_{i=1}^M \pi_i \left\|x_i^k - x_i^\star\right\| \leqslant (1-\rho)\sum_{i=1}^M \pi_i \left\|z_i^k - x^\star\right\| \\
&\leqslant (1-\rho)\max_i \sqrt{c_{k-D_i^k}} \sum_{i=1}^M \pi_i \left(1 - \rho\pi_i - (1-\rho)\rho\pi_i^2 - \cdots - (1-\rho)^{p_0-2}\rho\pi_i^{p_0-1}\right) \\
&\leqslant (1-\rho)\max_i \sqrt{c_{k-D_i^k}} \left(1 - \rho\sum_{i=1}^M \pi_i^2 - (1-\rho)\rho\sum_{i=1}^M \pi_i^3 - \cdots - (1-\rho)^{p_0-2}\rho\sum_{i=1}^M \pi_i^{p_0}\right).
\end{aligned}$$

Note finally that by Jensen's inequality for any p we have

$$\frac{1}{M}\sum_{i=1}^M \pi_i^p \geqslant (\frac{1}{M}\sum_{i=1}^M \pi_i)^p = \frac{1}{M^p}, \quad \text{so that} \quad \sum_{i=1}^M \pi_i^p \geqslant \frac{1}{M^{p-1}}.$$

This allows us to bound all the terms in the above expression and then proves the result. □

D. Additional numerical results

We run our numerical experiments on three standard datasets. To save room, we have not included in the main text the results for News20 dataset nor the suboptimality plots for different number of machines. They are provided here with the same setups as described in the main text for Fig. 3 and 6 respectively.

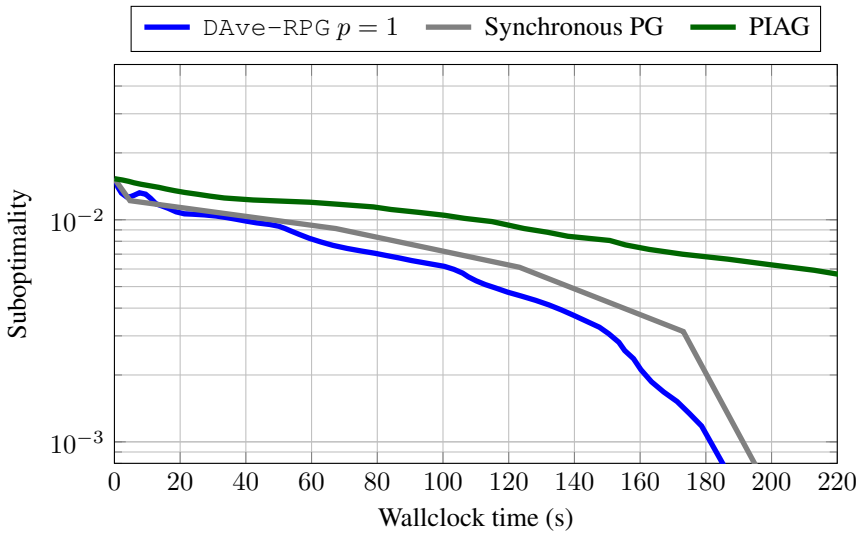

Figure 1. Regularized loss suboptimality on the training set versus wall clock time. News20 dataset, 30% of the data on the 1st machine.

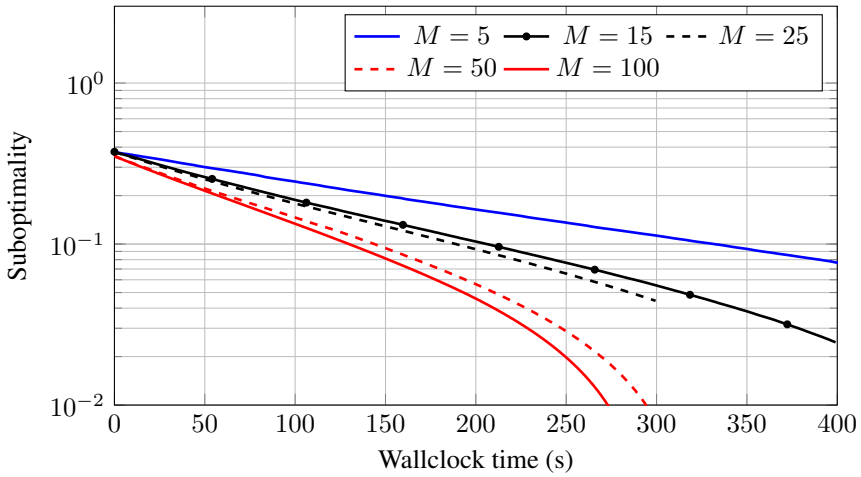

Figure 2. Scalability with respect to the number of slave machines for URL dataset.

E. Comparison with SAGA

As explained in introduction, parallel stochastic optimization methods are not adapted to the setting of distributed data. We go beyond this and we implement a direct extension of one of most popular stochastic algorithms, Prox-ASAGA (Leblond et al., 2017),(Pedregosa et al., 2017), that handles the considered optimization problems using only local data.

We run some numerical experiments in order to have a comparison between this algorithm and the batch algorithms (including ours). We observe on Figure 3 that Prox-ASAGA does not manage to reach high precision.

This bad behaviour was expected as our setting breaks the uniform sampling assumption under which Prox-ASAGA is proved to work well. In the experiment, we use indeed 100 workers, which implies heterogeneous delays and highly non-uniform sampling of the data. On top of this, Prox-ASAGA does not accept the efficient (fixed) stepsize.

References

Leblond, Rémi, Pedregosa, Fabian, and Lacoste-Julien, Simon. ASAGA: Asynchronous Parallel SAGA. In *20th International Conference on Artificial Intelligence and Statistics*, pp. 46–54, 2017.

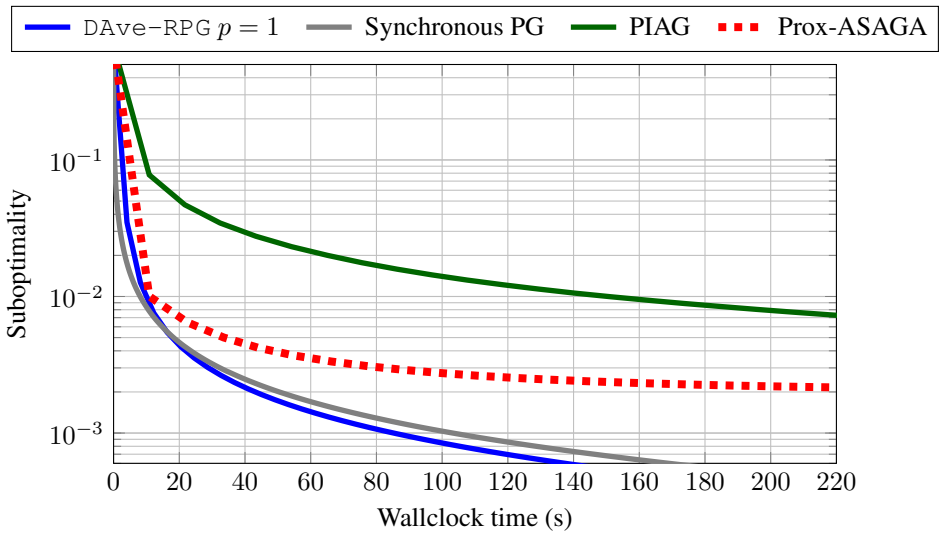

Figure 3. Regularized loss suboptimality on the training set versus wall clock time, for Covtype dataset, 100 workers.

Pedregosa, Fabian, Leblond, Rémi, and Lacoste-Julien, Simon. Breaking the nonsmooth barrier: A scalable parallel method for composite optimization. *Advances in Neural Information Processing System 30 (NIPS)*, 2017.