Optimal Algorithms for Smooth and Strongly Convex Distributed Optimization in Networks

SUPPLEMENTARY MATERIAL

Kevin Scaman [1] **Francis Bach** [2] **Sébastien Bubeck** [3] **Yin Tat Lee** [3] **Laurent Massoulié** [1]

Abstract

This supplementary document contains complete proofs of the theorems presented in the article, as well as an extension of our algorithm to composite problems particularly relevent for machine learning applications.

1. Optimal Convergence Rates

1.1. Centralized Algorithms

Proof of Theorem 1. This proof relies on splitting the function used by Nesterov to prove oracle complexities for strongly convex and smooth optimization (Nesterov, 2004; Bubeck, 2015). Let $\beta \geq \alpha > 0$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a graph and $A \subset \mathcal{V}$ a set of nodes of $\mathcal{G}$. For all $d > 0$, we denote as $A_d^c = \{v \in \mathcal{V} \,:\, d(A, v) \geq d\}$ the set of nodes at distance at least d from A, and let, for all $i \in \mathcal{V}$, $f_i^A : \ell_2 \to \mathbb{R}$ be the functions defined as:

$$f_i^A(\theta) = \begin{cases} \frac{\alpha}{2n}\|\theta\|_2^2 + \frac{\beta-\alpha}{8|A|}(\theta^\top M_1 \theta - 2\theta_1) & \text{if } i \in A \\ \frac{\alpha}{2n}\|\theta\|_2^2 + \frac{\beta-\alpha}{8|A_d^c|}\theta^\top M_2 \theta & \text{if } i \in A_d^c \\ \frac{\alpha}{2n}\|\theta\|_2^2 & \text{otherwise} \end{cases} \quad (1)$$

where $M_2 : \ell_2 \to \ell_2$ is the infinite block diagonal matrix with $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ on the diagonal, and $M_1 = \begin{pmatrix} 1 & 0 \\ 0 & M_2 \end{pmatrix}$. First, note that, since $0 \preceq M_1 + M_2 \preceq 4I$, $\bar{f}^A = \frac{1}{n}\sum_{i=1}^n f_i^A$ is α-strongly convex and β-smooth. Then, Theorem 1 is a direct consequence of the following lemma:

Lemma 1. *If $A_d^c \neq \emptyset$, then for any $t \geq 0$ and any black-box procedure one has, for all $i \in \{1, ..., n\}$,*

$$\bar{f}^A(\theta_{i,t}) - \bar{f}^A(\theta^*) \geq \frac{\alpha}{2}\left(\frac{\sqrt{\kappa_g}-1}{\sqrt{\kappa_g}+1}\right)^{2\left(1+\frac{t}{1+d\tau}\right)} \|\theta_{i,0} - \theta^*\|^2, \quad (2)$$

where $\kappa_g = \beta/\alpha$.

Proof. This lemma relies on the fact that most of the coordinates of the vectors in the memory of any node will remain equal to 0. More precisely, let $k_{i,t} = \max\{k \in \mathbb{N} \,:\, \exists \theta \in \mathcal{M}_{i,t} \text{ s.t. } \theta_k \neq 0\}$ be the last non-zero coordinate of a vector in the memory of node i at time t. Then, under any black-box procedure, we have, for any local computation step,

$$k_{i,t+1} \leq \begin{cases} k_{i,t} + \mathbb{1}\{k_{i,t} \equiv 0 \bmod 2\} & \text{if } i \in A \\ k_{i,t} + \mathbb{1}\{k_{i,t} \equiv 1 \bmod 2\} & \text{if } i \in A_d^c \\ k_{i,t} & \text{otherwise} \end{cases}. \quad (3)$$

Indeed, local gradients can only increase even dimensions for nodes in A and odd dimensions for nodes in A_d^c. The same holds for gradients of the dual functions, since these have the same block structure as their convex conjugates. Thus, in order to reach the third coordinate, algorithms must first perform one local computation in A, then d communication steps in order for a node in A_d^c to have a non-zero second coordinate, and finally, one local computation in A_d^c. Accordingly, one must perform at least k local computation steps and $(k-1)d$ communication steps to achieve $k_{i,t} \geq k$ for at least one node $i \in \mathcal{V}$, and thus, for any $k \in \mathbb{N}^*$,

$$\forall t < 1 + (k-1)(1 + d\tau), k_{i,t} \leq k - 1. \quad (4)$$

This implies in particular:

$$\forall i \in \mathcal{V}, k_{i,t} \leq \left\lfloor \frac{t-1}{1+d\tau} \right\rfloor + 1 \leq \frac{t}{1+d\tau} + 1. \quad (5)$$

Furthermore, by definition of $k_{i,t}$, one has $\theta_{i,k} = 0$ for all $k > k_{i,t}$, and thus

$$\|\theta_{i,t} - \theta^*\|_2^2 \geq \sum_{k=k_{i,t}+1}^{+\infty} \theta_k^{*2}. \quad (6)$$

[1]MSR-INRIA Joint Center, Palaiseau, France [2]INRIA, Ecole Normale Supérieure, Paris, France [3]Theory group, Microsoft Research, Redmond, United States. Correspondence to: Kevin Scaman <kevin.scaman@gmail.com>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

and, since $\bar{f}^A$ is α-strongly convex,

$$\bar{f}^A(\theta_{i,t}) - \bar{f}^A(\theta^*) \geq \frac{\alpha}{2}\|\theta_{i,t} - \theta^*\|_2^2. \tag{7}$$

Finally, the solution of the global problem $\min_\theta \bar{f}^A(\theta)$ is $\theta_k^* = \left(\frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}}\right)^k$. Combining this result with Eqs. (5), (6) and (7) leads to the desired inequality. □

Using the previous lemma with $d = \Delta$ the diameter of $\mathcal{G}$ and $A = \{v\}$ one of the pair of nodes at distance Δ returns the desired result. □

1.2. Decentralized Algorithms

Proof of Theorem 2. Let $\gamma_n = \frac{1-\cos(\frac{\pi}{n})}{1+\cos(\frac{\pi}{n})}$ be a decreasing sequence of positive numbers. Since $\gamma_2 = 1$ and $\lim_n \gamma_n = 0$, there exists $n \geq 2$ such that $\gamma_n \geq \gamma > \gamma_{n+1}$. The cases $n = 2$ and $n \geq 3$ are treated separately. If $n \geq 3$, let $\mathcal{G}$ be the linear graph of size n ordered from node v_1 to v_n, and weighted with $w_{i,i+1} = 1 - a\mathbb{1}\{i = 1\}$. Then, if $A = \{v_1, ..., v_{\lceil n/32 \rceil}\}$ and $d = (1 - 1/16)n - 1$, we have $|A_d^c| \geq |A|$ and Lemma 1 implies:

$$\bar{f}^A(\theta_{i,t}) - \bar{f}^A(\theta^*) \geq \frac{n\alpha}{2}\left(\frac{\sqrt{\kappa_g}-1}{\sqrt{\kappa_g}+1}\right)^{2\left(1+\frac{t}{1+d\tau}\right)}\|\theta_{i,0} - \theta^*\|^2. \tag{8}$$

A simple calculation gives $\kappa_l = 1 + (\kappa_g - 1)\frac{n}{2|A|}$, and thus $\kappa_g \geq \kappa_l/16$. Finally, if we take W_a as the Laplacian of the weighted graph $\mathcal{G}$, a simple calculation gives that, if $a = 0$, $\gamma(W_a) = \gamma_n$ and, if $a = 1$, the network is disconnected and $\gamma(W_a) = 0$. Thus, by continuity of the eigenvalues of a matrix, there exists a value $a \in [0, 1]$ such that $\gamma(W_a) = \gamma$. Finally, by definition of n, one has $\gamma > \gamma_{n+1} \geq \frac{2}{(n+1)^2}$, and $d \geq \frac{15}{16}(\sqrt{\frac{2}{\gamma}} - 1) - 1 \geq \frac{1}{5\sqrt{\gamma}}$ when $\gamma \leq \gamma_3 = \frac{1}{3}$.

For the case $n = 2$, we consider the totally connected network of 3 nodes, reweight only the edge (v_1, v_3) by $a \in [0, 1]$, and let W_a be its Laplacian matrix. If $a = 1$, then the network is totally connected and $\gamma(W_a) = 1$. If, on the contrary, $a = 0$, then the network is a linear graph and $\gamma(W_a) = \gamma_3$. Thus, there exists a value $a \in [0, 1]$ such that $\gamma(W_a) = \gamma$, and applying Lemma 1 with $A = \{v_1\}$ and $d = 1$ returns the desired result, since then $\kappa_g \geq 2\kappa_l/3$ and $d = 1 \geq \frac{1}{\sqrt{3\gamma}}$. □

2. Optimal Decentralized Algorithms

2.1. Single-Step Dual Accelerated Method

Proof of Theorem 3. Each step of the algorithm can be decomposed in first computing gradients, and then communicating these gradients across all neighborhoods. Thus, one step takes a time $1 + \tau$. Moreover, the Hessian of the dual function $F^*(\lambda\sqrt{W})$ is

$$(\sqrt{W} \otimes I_d)\nabla^2 F^*(\lambda\sqrt{W})(\sqrt{W} \otimes I_d), \tag{9}$$

where $\otimes$ is the Kronecker product and I_d is the identity matrix of size d. Also, note that, in Alg.(2), the current values x_t and y_t are always in the image of $\sqrt{W} \otimes I_d$ (i.e. the set of matrices x such that $x^\top \mathbb{1} = 0$). The condition number (in the image of $\sqrt{W} \otimes I_d$) can thus be upper bounded by $\frac{\kappa_l}{\gamma}$, and Nesterov's acceleration requires $\sqrt{\frac{\kappa_l}{\gamma}}$ steps to achieve any given precision (Bubeck, 2015). □

2.2. Multi-Step Dual Accelerated Method

Proof of Theorem 4. First, since $P_K(W)$ is a gossip matrix, Theorem 3 implies the convergence of Alg.(3). In order to simplify the analysis, we multiply W by $\frac{2}{(1+\gamma)\lambda_1(W)}$, so that the resulting gossip matrix has a spectrum in $[1 - c_2^{-1}, 1 + c_2^{-1}]$. Applying Theorem 6.2 in (Auzinger, 2011) with $\alpha = 1 - c_2^{-1}$, $\beta = 1 + c_2^{-1}$ and $\gamma = 0$ implies that the minimum

$$\min_{p \in \mathbb{P}_K, p(0)=0} \max_{x \in [1-c_2^{-1}, 1+c_2^{-1}]} |p(t) - 1| \tag{10}$$

is attained by $P_K(x) = 1 - \frac{T_K(c_2(1-x))}{T_K(c_2)}$. Finally, Corollary 6.3 of (Auzinger, 2011) leads to

$$\gamma(P_K(W)) \geq \frac{1 - 2\frac{c_1^K}{1+c_1^{2K}}}{1 + 2\frac{c_1^K}{1+c_1^{2K}}} = \left(\frac{1 - c_1^K}{1 + c_1^K}\right)^2, \tag{11}$$

where $c_1 = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, and taking $K = \lfloor \frac{1}{\sqrt{\gamma}} \rfloor$ implies

$$\frac{1}{\sqrt{\gamma(P_K(W))}} \leq \frac{1 + c_1^{\frac{1}{\sqrt{\gamma}}+1}}{1 - c_1^{\frac{1}{\sqrt{\gamma}}+1}} \leq 2. \tag{12}$$

The time required to reach an $\varepsilon > 0$ precision using Alg.(3) is thus $O\left((1 + K\tau)\sqrt{\frac{\kappa_l}{\gamma(P_K(W))}}\ln(1/\varepsilon)\right) = O\left(\sqrt{\kappa_l}(1 + \frac{1}{\sqrt{\gamma}}\tau)\ln(1/\varepsilon)\right)$. □

3. Composite Problems for Machine Learning

When the local functions are of the form

$$f_i(\theta) = g_i(B_i\theta) + c\|\theta\|^2, \tag{13}$$

where $B_i \in \mathbb{R}^{m_i \times d}$ and g_i is smooth and has proximal operator which is easy to compute (and hence also g_i^*), an additional Lagrange multiplier ν can be used to make the Fenchel conjugate of g_i appear in the dual optimization problem. More specifically, from the primal problem of Eq. (12), one has, with $\rho > 0$ an arbitrary parameter:

$$\begin{aligned}
&\inf_{\Theta\sqrt{W}=0} F(\Theta) \\
= &\inf_{\Theta\sqrt{W}=0,\ \forall i, x_i = B_i\theta_i} \frac{1}{n}\sum_{i=1}^n g_i(x_i) + c\|\theta_i\|_2^2 \\
= &\inf_{\Theta}\sup_{\lambda,\nu} \frac{1}{n}\sum_{i=1}^n \Big\{\nu_i^\top B_i\theta_i - g_i^*(\nu_i) + c\|\theta_i\|_2^2\Big\} \\
&+\frac{\rho}{n}\operatorname{tr}(\lambda^\top \Theta\sqrt{W}) \\
= &\sup_{\nu\in\prod_{i=1}^n \mathbb{R}^{m_i},\ \lambda\in\mathbb{R}^{d\times n}} -\frac{1}{n}\sum_{i=1}^n g_i^*(\nu_i) \\
&-\frac{1}{4cn}\sum_{i=1}^n \|B_i^\top\nu_i + \rho\lambda\sqrt{W}_i\|_2^2.
\end{aligned}$$

To maximize the dual problem, we can use (accelerated) proximal gradient, with the updates:

$$\begin{aligned}
\nu_{i,t+1} &= \inf_{\nu\in\mathbb{R}^{m_i}} g_i^*(\nu) \\
&\quad +\frac{1}{2\eta}\Big\|\nu - \nu_{i,t} + \frac{\eta}{2c}B_i(B_i^\top\nu_{i,t} + \rho\lambda_t\sqrt{W}_i)\Big\|_2^2 \\
\lambda_{t+1} &= \lambda_t - \eta\frac{\rho}{2cn}\sum_{i=1}^n (B_i^\top\nu_{i,t} + \rho\lambda_t\sqrt{W}_i)\sqrt{W}_i^\top.
\end{aligned}$$

We can rewrite all updates in terms of $z_t = \lambda_t\sqrt{W} \in \mathbb{R}^{d\times n}$, as

$$\begin{aligned}
\nu_{i,t+1} &= \inf_{\nu\in\mathbb{R}^{m_i}} g_i^*(\nu) \\
&\quad +\frac{1}{2\eta}\Big\|\nu - \nu_{i,t} + \frac{\eta}{2c}B_i(B_i^\top\nu_{i,t} + \rho z_{i,t})\Big\|_2^2 \\
z_{t+1} &= z_t - \eta\frac{\rho}{2cn}\sum_{i=1}^n (B_i^\top\nu_{i,t} + \rho z_i)W_i^\top.
\end{aligned}$$

In order to compute the convergence rate of such an algorithm, if we assume that:

- each g_i is μ-smooth,
- the largest singular value of each B_i is less than M,

then we simply need to compute the condition number of the quadratic function

$$Q(\nu,\lambda) = \frac{1}{2\mu}\sum_{i=1}^n \|\nu_i\|_2^2 + \frac{1}{4c}\sum_{i=1}^n \|B_i^\top\nu_i + \rho\lambda\sqrt{W}_i\|_2^2.$$

With the choice $\rho^2 = \frac{1}{\lambda_{\max}(W)}\big(\frac{c}{\mu} + M^2\big)$, it is lower bounded by $\big(1+\mu\frac{M^2}{c}\big)\frac{4}{\gamma}$, which is a natural upper bound on κ_l/γ. Thus this essentially leads to the same convergence rate than the non-composite case with the Nesterov and Chebyshev accelerations, i.e. $\sqrt{\kappa_l/\gamma}$.

The bound on the conditional number may be shown through the two inequalities:

$$\begin{aligned}
Q(\nu,\lambda) &\leqslant \frac{1}{2\mu}\sum_{i=1}^n \|\nu_i\|^2 + \frac{1}{2c}\sum_{i=1}^n \|\rho\lambda\sqrt{W}_i\|_2^2 \\
&\quad +\frac{1}{2c}\sum_{i=1}^n \|B_i^\top\nu_i\|_2^2, \\
Q(\nu,\lambda) &\geqslant \frac{1}{2\mu}\sum_{i=1}^n \|\nu_i\|^2 + \frac{1}{1+\eta}\frac{1}{4c}\sum_{i=1}^n \|\rho\lambda\sqrt{W}_i\|_2^2 \\
&\quad -\frac{1}{\eta}\frac{1}{4c}\sum_{i=1}^n \|B_i^\top\nu_i\|_2^2,
\end{aligned}$$

with $\eta = M^2\mu/c$.

References

Auzinger, W. *Iterative Solution of Large Linear Systems*. Lecture notes, TU Wien, 2011.

Bubeck, Sébastien. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Nesterov, Yurii. *Introductory lectures on convex optimization : a basic course*. Kluwer Academic Publishers, 2004.