Asynchronous Newton-Raphson Consensus for Distributed Convex Optimization ⋆

Filippo Zanella * **Damiano Varagnolo** ** **Angelo Cenedese** *
Gianluigi Pillonetto * **Luca Schenato** *

* *Department of Information Engineering, Università di Padova, Via Gradenigo 6/a, Padova, Italy. Emails:* {fzanella | angelo.cenedese | giapi | schenato }@dei.unipd.it
** *School of Electrical Engineering, Royal Institute of Technology, Osquldas väg 10, Stockholm, Sweden. Email:* damiano@kth.se

Abstract: We consider the distributed unconstrained minimization of separable convex cost functions, where the global cost is given by the sum of several local and private costs, each associated to a specific agent of a given communication network. We specifically address an asynchronous distributed optimization technique called Newton-Raphson Consensus. Beside having low computational complexity, low communication requirements and being interpretable as a distributed Newton-Raphson algorithm, the technique has also the beneficial properties of requiring very little coordination and naturally supporting time-varying topologies. In this work we analytically prove that under some assumptions it shows either local or global convergence properties, and corroborate this result by the means of numerical simulations.

Keywords: second order methods, separable cost functions, distributed minimization

1. INTRODUCTION

The interests on systems where agents collaborate to achieve a common goal are driven by the possibility of synergies, i.e., coordinated actions whose total effects are bigger than the ones achievable without coordination. Thus it does not surprise that *distributed optimization* received in the past years an increasing attention from various research communities, being it a pervasive building block for all the decision making processes, including estimation and control.

We can trace back the roots of this topic to the seminal work of Tsitsiklis (1984), and classify the literature into: methods based on primal decompositions, methods based on dual decompositions, and heuristic or ad-hoc methods.

Primal methods exploit suitable decompositions of the primal problem and operate on the values of the primal variables explicitly. The best known ones are Distributed Subgradient Methods (DSMs) (Shor, 1985, Chap. 2) (Nedić and Ozdaglar, 2009, and references therein). Their advantages can be summarized in easy implementability and wide applicability, while the drawbacks may lie in rather slow convergence rates in practical applications (Johansson, 2008, Chap. 6).

Dual methods exploit instead decompositions of the dual problem into simpler tasks. The most widely known approach, the Alternating Direction Method of Multipliers (ADMM), alternates local dual ascents and communication steps (Bertsekas and Tsitsiklis, 1997, pp. 253-261). It usually has faster convergence rates than DSMs, still maintaining a wide applicability (Boyd et al., 2010; Erseghe et al., 2011).

Heuristic or ad-hoc methods may use several different techniques, e.g., swarm optimization (Van Ast et al., 2008) or genetic algorithms. Other approaches are instead tailored for suitable classes of cost functions, e.g., the Fast-Lipschitz methods (Fischione, 2011; Fischione and Jönsson, 2011), and may have convergences faster than the ones of ADMMs or DSMs.

Primarily, these algorithms require different degrees of coordination among the agents. E.g., DSMs may be implemented without requiring synchronized communications (Nedić and Ozdaglar, 2009), while ADMM generally requires the preservation of the order of the operations.

Considering that the applicability of distributed algorithms relates to how much agents must coordinate, our aim is to consider the Newton-Raphson Consensus (NRC), a promising primal-based distributed optimizer originally proposed in Zanella et al. (2011), and lessen its coordination requirements. More specifically, we propose an asynchronous version of it and prove its convergence properties.

The interest on the NRC technique can be motivated as follows. First of all, at the best of our knowledge it is the unique primal-based distributed algorithm whose estimates evolve as driven by a Newton-Raphson optimization scheme, and that can be implemented without requiring a-priori knowledge about the topology of the network (see, e.g., Jadbabaie et al. (2009)). Secondly, the NRC exploits average consensus algorithms (Fagnani and

⋆ The research leading to these results was supported by European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n°257462 HYCON2 Network of excellence, by Progetto di Ateneo CPDA090135/09 funded by the University of Padova, and by the Italian Ing. Aldo Gini foundation.

Zampieri, 2008, and references therein). Thus it inherits all their favorable properties, like immediate adaptation to time-varying topologies and extreme simplicity of implementation.

The paper is organized as follows: we formulate the problem in Sec. 2 and summarize the notation in Sec. 3. Then introduce the original NRC algorithm in Sec. 4, and describe its asynchronous version in Sec. 5. In Sec. 6 we perform some numerical comparisons between the NRC and a DSM, and eventually draw some concluding remarks and future research topics in Sec. 7. All the proofs are collected in the technical report deposited in arXiv.

2. PROBLEM FORMULATION

Let the N agents of a network be endowed with local strictly convex cost functions $f_i : \mathbb{R} \mapsto \mathbb{R}$, so that the following function

$$\overline{f} : \mathbb{R} \mapsto \mathbb{R} \qquad \overline{f}\,(x) := \frac{1}{N}\sum_{i=1}^{N} f_i\,(x) \tag{1}$$

is a well-defined global cost. The aim of the agents is to cooperate and distributedly compute the minimizer of $\overline{f}$, namely

$$x^* := \arg\min_x \overline{f}\,(x)\ . \tag{2}$$

Due to memory / computational / communication constraints, the distributed optimization schemes must have low-complexities and be based only on local communications.

The network underlying the information exchange process is modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose vertexes $\{1, \dots, N\} \in \mathcal{V}$ and edges $(i, j) \in \mathcal{E}$ represent respectively the agents and the available communication links. We pose the following assumptions, some of which may be relaxed to the detriment of notational simplicity: the graph is undirected, connected and not time-varying. Moreover, we assume the following:

Assumption 1. (Convexity). The function $\overline{f} : \mathbb{R} \to \mathbb{R}$ defined in (1) is smooth, closed, proper, and strictly convex, i.e., $\overline{f}''(x) := \dfrac{d^2\overline{f}(x)}{dx^2} > 0$, $\forall x \in \mathbb{R}$.

Ass. 1 implies that x^* in (2) exists and is unique. Moreover, the positive second derivative is a mild sufficient condition to guarantee that the minimum x^* defined in (2) will be exponentially stable under the continuous Newton-Raphson dynamics described in the following Thm. 2. We notice that in principle just the average function $\overline{f}$ needs to have specific properties, and thus no conditions for the single f_i's are required: in fact they might even be non convex. We also notice that Ass. 1 allows us to apply standard singular perturbation analysis techniques (Khalil, 2001, Chap. 11) Kokotović et al. (1999). The following theorem provides a preliminary results to apply such tools.

Theorem 2. Let

$$\dot{x} = -\frac{\overline{f}'(x)}{\overline{f}''(x)} =: \psi(x + x^*), \quad x(0) \in D_r \tag{3}$$

describe a continuous Newton-Raphson algorithm with $\overline{f}$ satisfying Ass. 1. Let $D_r := \{x \in \mathbb{R} \,|\, \overline{f}(x) \le r\}$ be s.t. $r > \overline{f}(x^*)$. Then x^* is an exponentially stable equilibrium, i.e., $|x(t)| \le ce^{-\gamma t}|x(0)|$, $\forall\ t$'s and $x(0) \in D_r$, for suitable positive constants c and γ possibly depending on r.

We notice that the previous theorem can be used to show that the point x^* is actually globally stable. In fact the properness of $\overline{f}$ implies that for any $x \in \mathbb{R}$ there exists r such that $x \in D_r$. Thus one can start from any point and have an exponential convergence, although a uniform convergence rate in $\mathbb{R}$ might not exists. Nonetheless we can notice that, locally and around the optimum, the rate of convergence of the Newton-Raphson dynamics is $\gamma = 1$ independently of the convex function $\overline{f}$. In fact, if we linearize ψ around 0 (i.e., the dynamics around x^*) we obtain

$$\begin{aligned}\psi(x) &= \psi(0) + \psi'(0)x + o(x)\\ &= -\frac{\overline{f}'(x^*)}{\overline{f}''(x^*)} - \frac{\overline{f}''(x^*)\overline{f}''(x^*) - \overline{f}'(x^*)\overline{f}'''(x^*)}{\left(\overline{f}''(x^*)\right)^2}x + o(x)\\ &= -x + o(x)\end{aligned}$$

since $\overline{f}'(x^*) = 0$ and $\overline{f}''(x^*) \neq 0$.

3. NOTATION

In the following we use bold fonts to indicate vectors or functions whose range is vectorial, and plain italic fonts to indicate scalars or functions whose range is scalar. Furthermore we use the following shorthands:

$$\begin{aligned} f'(x) &:= \frac{df(x)}{dx} \qquad f''(x) := \frac{d^2 f(x)}{dx^2}\\ g_i\big(x_i\,(k)\,\big) &:= f_i''\big(x_i\,(k)\,\big)x_i\,(k) - f_i'\big(x_i\,(k)\,\big)\\ h_i\big(x_i\,(k)\,\big) &:= f_i''\big(x_i\,(k)\,\big)\\ \boldsymbol{x}\,(k) &:= [x_1\,(k)\ \cdots\ x_N\,(k)]^T\\ \boldsymbol{g}\big(\boldsymbol{x}(k)\big) &:= \big[g_1\big(x_1\,(k)\,\big)\ \cdots\ g_N\big(x_N\,(k)\,\big)\big]^T\\ \boldsymbol{h}\big(\boldsymbol{x}(k)\big) &:= \big[h_1\big(x_1\,(k)\,\big)\ \cdots\ h_N\big(x_N\,(k)\,\big)\big]^T\\ \boldsymbol{f}'(\boldsymbol{x}) &:= [f_1'\,(x_1)\ \cdots\ f_N'\,(x_N)]^T. \end{aligned}$$

We use the division bar to indicate also the componentwise division, e.g.,

$$\frac{\boldsymbol{g}\big(\boldsymbol{x}(k)\big)}{\boldsymbol{h}\big(\boldsymbol{x}(k)\big)} := \left[\frac{g_1\big(x_1\,(k)\,\big)}{h_1\big(x_1\,(k)\,\big)}\ \cdots\ \frac{g_N\big(x_N\,(k)\,\big)}{h_N\big(x_N\,(k)\,\big)}\right]^T.$$

4. DISTRIBUTED NEWTON-RAPHSON CONSENSUS: THE SYNCHRONOUS CASE

Consider a global cost function like in (1). Assume that it is given by a sum of quadratic costs $f_i(x) = a_i\big(x - b_i\big)^2$. Then x^* in (2) is computable by means of two average consensus algorithms in parallel (Xiao et al., 2005; Bolognani et al., 2010), since

$$\arg\min_{x\in\mathbb{R}} \sum_i a_i\big(x - b_i\big)^2 = \frac{\sum_{i=1}^N a_i b_i}{\sum_{i=1}^N a_i} = \frac{\frac{1}{N}\sum_{i=1}^N a_i b_i}{\frac{1}{N}\sum_{i=1}^N a_i}\ . \tag{4}$$

Notice now that $a_i b_i = f_i''(x)x - f_i'(x)$ and $a_i = f_i''(x)$. The intuition then goes as follows: let each agent pick a local x_i, and compute the corresponding local quantities

$$g_i(x_i) := f_i''(x_i)x_i - f_i'(x_i), \qquad h_i(x_i) := h_i''(x_i). \tag{5}$$

We can ask whether the guess

$$\widehat{x}^* = \frac{\frac{1}{N}\sum_{i=1}^N g_i(x_i)}{\frac{1}{N}\sum_{i=1}^N h_i(x_i)}, \tag{6}$$

that can be distributedly computed by the agents as a parallel of two average consensus, is equal to the global optimum x^* even if the f_i's are *not* quadratic.

In general, this is not. Nonetheless, $\widehat{x}^*$ corresponds to a guess of the global optimum computed through quadratic approximations of the local cost functions around the local estimates x_i. The consequent intuition is that alternating steps that compute the averages of the various g_i's and h_i's and steps that update the local x_i's will eventually lead to get the global optimum.

A synchronous algorithm implementing this procedure has been proposed in Zanella et al. (2011, 2012), and is summarized in Alg. 1. Its synchronous communication steps 6 and 7 rely on a symmetric consensus communication matrix $P = P^T$, i.e., a matrix whose elements are non-negative, where $P_{ij} > 0$ only if $(i,j) \in \mathcal{E}$, and with the properties that $P\mathbb{1} = \mathbb{1}$ ($\mathbb{1} := [1\,1\,\cdots\,1]^T$) and $\lim_{k\to\infty} P^k = \frac{1}{N}\mathbb{1}\mathbb{1}^T$. As a result, each k in Alg. 1 should be treated as an event when all the agents synchronously communicate and then update their local values.

Algorithm 1 Synchronous Newton-Raphson Consensus (SNRC)

(storage allocation and constraints on parameters)
1: $\boldsymbol{x}(k), \boldsymbol{y}(k), \boldsymbol{z}(k) \in \mathbb{R}^N; k = 0, 1, \ldots$
2: $P \in \mathbb{R}^{N\times N}$, positive and doubly stochastic
3: $\varepsilon \in (0,1)$

(initialization)
4: $\boldsymbol{x}(0) = \boldsymbol{x_0},\ \boldsymbol{y}(0) = \boldsymbol{z}(0) = \boldsymbol{g}(\boldsymbol{x}(-1)) = \boldsymbol{h}(\boldsymbol{x}(-1)) = \boldsymbol{0}$

(main algorithm)
5: **for** $k = 1, 2, \ldots$ **do**
(update of the auxiliary variables and consensus)
6: $\quad \boldsymbol{y}(k) = P\left[\boldsymbol{y}(k-1) + \boldsymbol{g}(\boldsymbol{x}(k-1)) - \boldsymbol{g}(\boldsymbol{x}(k-2))\right]$
7: $\quad \boldsymbol{z}(k) = P\left[\boldsymbol{z}(k-1) + \boldsymbol{h}(\boldsymbol{x}(k-1)) - \boldsymbol{h}(\boldsymbol{x}(k-2))\right]$
(update of the local guesses)
8: $\quad \boldsymbol{x}(k) = (1-\varepsilon)\boldsymbol{x}(k-1) + \varepsilon\dfrac{\boldsymbol{y}(k)}{\boldsymbol{z}(k)}$
9: **end for**

Alg. 1 has two fundamental features: the first is that the consensus steps on local variables y_i and z_i track the changing values of the functions $g_i\big(x_i(k)\big)$ and $h_i\big(x_i(k)\big)$. The second is the presence of the parameter ε that acts as a low pass filter. In fact it can be considered a forgetting factor which regulates, in the update of the estimate of the global minimum, the relative importance between the past estimate $x_i(k)$ and the current estimate $y_i(k)/z_i(k)$. Additionally, the parameter ε slows the dynamics of $x_i(k)$ and lets the consensus on $y_i(k)$, $z_i(k)$ take place.

As explained in Zanella et al. (2011, 2012), Alg. 1 can be approximated with the continuous-time system

$$\begin{cases} \varepsilon\dot{\boldsymbol{v}}(t) = -\boldsymbol{v}(t) + \boldsymbol{g}(\boldsymbol{x}(t)) \\ \varepsilon\dot{\boldsymbol{w}}(t) = -\boldsymbol{w}(t) + \boldsymbol{h}(\boldsymbol{x}(t)) \\ \varepsilon\dot{\boldsymbol{y}}(t) = -K\boldsymbol{y}(t) + (I-K)\left[\boldsymbol{g}(\boldsymbol{x}(t)) - \boldsymbol{v}(t)\right] \\ \varepsilon\dot{\boldsymbol{z}}(t) = -K\boldsymbol{z}(t) + (I-K)\left[\boldsymbol{h}(\boldsymbol{x}(t)) - \boldsymbol{w}(t)\right] \\ \dot{\boldsymbol{x}}(t) = -\boldsymbol{x}(t) + \dfrac{\boldsymbol{y}(t)}{\boldsymbol{z}(t)} \end{cases} \tag{7}$$

where K is a positive semidefinite matrix with kernel generated by the vector $\mathbb{1}$ and with eigenvalues $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N < 2$. (7) is thus a two-time scales dynamical system combining fast dynamics, that make the g_i's and h_i's converge to their averages, and slow dynamics, that make the local x_i's evolve approximatively as the scaled continuous Newton-Raphson algorithm

$$\dot{\overline{x}}(t) = -\varepsilon\frac{\overline{f}'\big(\overline{x}(t)\big)}{\overline{f}''\big(\overline{x}(t)\big)}\ . \tag{8}$$

Under Ass. 1, Alg. 1 (SNRC) has global convergence properties (Zanella et al., 2011):

Theorem 3. Consider Alg. 1 (SNRC), with Ass. 1 holding true and the initialization performed as in step 4. Then for every open ball $B_r^{x^*} := \{\boldsymbol{x} \mid \|\boldsymbol{x} - x^*\mathbb{1}\| < r\}$ there exist two positive constants $\overline{\varepsilon}_r$, c_r such that if $\varepsilon < \overline{\varepsilon}_r$, then there exists $\gamma_\varepsilon > 0$ such that, for all $\boldsymbol{x}_0 \in B_r^{x^*}$,

$$\|\boldsymbol{x}(k) - x^*\mathbb{1}\| \leq c_r e^{-\gamma_\varepsilon k}\|\boldsymbol{x}_0 - x^*\mathbb{1}\|.$$

The proof of the previous theorem is based on singular perturbation theory for continuous time systems and consensus tracking. The critical value for the parameter $\overline{\varepsilon}_r$ depends on r and the function $\overline{f}$. The explicit computation of the critical value based on Lyapunov theory is in general very pessimistic and therefore of no practical use. However, the proof shows that if r and ε are sufficiently small, the rate of convergence of the algorithm tends to $\gamma_\varepsilon = \varepsilon$, and the dynamics of the local estimates, if $\overline{x}(0) := \frac{1}{N}\sum_{i=1}^N x_i(0)$, is approximately given by

$$x_i(k) \approx x^* + \left(\overline{x}(0) - x^*\right)e^{-\varepsilon k}\ .$$

In the following we extend these ideas to be amenable to more realistic asynchronous implementations.

5. ASYNCHRONOUS NEWTON-RAPHSON CONSENSUS

As noticed before, steps 6 and 7 of Alg. 1 rely on synchronous communications and on the updates of the various y_i's and z_i's. Thus this implementation requires a high degree of coordination among the agents, being consequently of limited practical applicability.

Here we propose an asynchronous version of the NRC that is built upon the standard symmetric gossip consensus: at every time a single agent is activated, then this agent selects one of its neighbors and communicates with it. To describe precisely this process we use the following notation: $k = 1, 2, \ldots$ correspond to the time instants $t_1, t_2, \ldots$ where a generic agent i activates and communicates with one of its neighbors $j \in \mathcal{N}_i$. $v(k) : \mathbb{N} \mapsto \mathcal{V}$ and $e(k) : \mathbb{N} \mapsto \mathcal{E}$ indicate which agent and which edge have been activated at time k, respectively. $w_i(k)$ / $u_{(i,j)}(k)$ are instead flags indicating whether agent i / edge (i,j) have been activated at time k or not. Thus $w_i(k) = 1$ if $v(k) = i$, $w_i(k) = 0$

otherwise, and $u_{(i,j)}(k) = 1$ if $e(k) = (i, j)$, $u_{(i,j)}(k) = 0$ otherwise. Notice that we thus allow the activation of just a single agent and single edge for each time instant k.

As for the agent activation process, we exploit either uniform or persistent agent activation hypotheses:

Assumption 4. (uniform activation). There exist a strictly decreasing function σ and a positive integer B s.t.

$$\left|\frac{1}{T}\sum_{k=h}^{h+T-1} w_i(k) - \frac{1}{N}\right| \leq \sigma(T), \quad \forall i \in \mathcal{V}, \forall h \in \mathbb{N} \quad (9)$$

$$\sum_{k=h}^{B+k-1} u_{(i,j)}(k) \geq 1, \quad \forall (i,j) \in \mathcal{E}, \forall h \in \mathbb{N}. \quad (10)$$

(9) basically states that, on the long run, all the agents are activated the same number of times. (10) instead states that every edge is activated at least once in any window of length B, which can be arbitrarily large but finite.

Assumption 5. (persistent activation). There exists a positive integer B s.t.

$$\sum_{k=h}^{B+h-1} w_i(k) \geq 1, \quad \forall i \in \mathcal{V}, \forall h \in \mathbb{N} \quad (11)$$

and (10) simultaneously hold.

(11) is weaker than (9) in the sense that the former states just that each agent activates at least once in every sufficiently large time window.

Exploiting the previous definitions we introduce the *agent selection matrix* $S(k) \in \mathbb{R}^{N\times N}$, the *edge selection matrix* $E(k) \in \mathbb{R}^{N\times N}$ and the *symmetric gossip consensus matrix* $P(k) \in \mathbb{R}^{N\times N}$ as follows:

$$b_{(i,j)} := [0 \cdots 0 \overset{(i)}{1} 0 \cdots 0 \overset{(j)}{-1} 0 \cdots 0]^T \in \mathbb{R}^N \quad (12)$$

$$S(k) := \mathrm{diag}\big(w_1(k), \ldots, w_N(k)\big) \quad (13)$$

$$E(k) := \big(\mathrm{diag}(b_{u(k)})\big)^2 \quad (14)$$

$$P(k) := I - \alpha b_{u(k)} b_{u(k)}^T, \quad \alpha \in (0,1). \quad (15)$$

Basically, $S(k)$ is zero everywhere except for a one in the diagonal element (i, i) corresponding to the activated agent i. $E(k)$ is zero everywhere except for two ones in the diagonal elements (i, i) and (j, j), corresponding to the agents of the activated edge (i, j). $P(k)$ is the standard symmetric gossip consensus matrix with weight $1 - \alpha$ on the diagonals elements (i, i) and (j, j), and α on the (i, j) and (j, i) elements. With this notation it is possible to derive the Asynchronous Newton-Raphson Consensus (ANRC), presented in Alg. 2, as a straightforward modification of the SNRC.

Lines 7, 8 and 9 in Alg. 2 compactly represent the fact that all agents do not perform any action except for the selected ones i, j, updating their local variables y_i, y_j as

$$y_i(k+1) = (1-\alpha)\Big(y_i(k) + g_i\big(x_i(k)\big) - g_i\big(x_i(k-1)\big)\Big) + \alpha y_j(k)$$

$$y_j(k+1) = (1-\alpha)\Big(y_j(k) + g_j\big(x_j(k)\big) - g_j\big(x_j(k-1)\big)\Big) + \alpha y_i(k)$$

and z_i, z_j in a similar way. We notice that, among x_i and x_j, just the former is updated: in the proposed version we require the local guess to be updated just for the agent that initiates the communication. Line 9 in Alg. 2 thus reads as

$$x_i(k+1) = x_i(k) + \varepsilon\left(-x_i + \frac{y_i(k+1)}{z_i(k+1)}\right).$$

Algorithm 2 Asynchronous Newton-Raphson Consensus (ANRC)

(storage allocation and constraints on parameters)

1: $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^N$

2: $\alpha, \varepsilon \in (0, 1)$

(initialization)

3: $\boldsymbol{x}(0) = \boldsymbol{x_0}$

4: $\boldsymbol{y}(0) = \boldsymbol{g}\left(\boldsymbol{x}(-1)\right) = \boldsymbol{g}\left(\boldsymbol{x}(0)\right)$

5: $\boldsymbol{z}(0) = \boldsymbol{h}\left(\boldsymbol{x}(-1)\right) = \boldsymbol{h}\left(\boldsymbol{x}(0)\right)$

(main algorithm)

6: **for** $k = 1, 2, \ldots$ **do**

(update of the auxiliary variables and consensus)

7: $\boldsymbol{y}(k) = P(k)\Big[\ \boldsymbol{y}(k-1) + E(k)\Big(\boldsymbol{g}\big(\boldsymbol{x}(k-1)\big) - \boldsymbol{g}\big(\boldsymbol{x}(k-2)\big)\Big)\Big]$

8: $\boldsymbol{z}(k) = P(k)\Big[\ \boldsymbol{z}(k-1) + E(k)\Big(\boldsymbol{h}\big(\boldsymbol{x}(k-1)\big) - \boldsymbol{h}\big(\boldsymbol{x}(k-2)\big)\Big)\Big]$

(update of the local guesses)

9: $\boldsymbol{x}(k) = \boldsymbol{x}(k-1) + \varepsilon S(k)\left(-\boldsymbol{x}(k-1) + \dfrac{\boldsymbol{y}(k)}{\boldsymbol{z}(k)}\right)$

10: **end for**

The convergence properties of Alg. 2 are summarized by the two following theorems:

Theorem 6. (global stability). Consider Alg. 2 and let Assumptions 1 and 4 hold true. Then for every open ball $B_r^{x^*} := \{\boldsymbol{x} \mid \|\boldsymbol{x} - x^*\mathbb{1}\| < r\}$ there exist two positive constants $\bar{\varepsilon}_r$, c_r such that if $\varepsilon < \bar{\varepsilon}_r$, then there exists $\gamma_\varepsilon > 0$ such that

$$\|\boldsymbol{x}(k) - x^*\mathbb{1}\| \leq c_r e^{-\gamma_\varepsilon k}\|\boldsymbol{x}_0 - x^*\mathbb{1}\|$$

for all $\boldsymbol{x}_0 \in B_r^{x^*}$.

Theorem 7. (local stability). Consider Alg. 2 and let Assumptions 1 and 5 hold true. Then there exist an open ball $B_0^{x^*} := \{\boldsymbol{x} \mid \|\boldsymbol{x} - x^*\mathbb{1}\| < r_0\}$ and two positive constants $\bar{\varepsilon}$, c such that if $\varepsilon < \bar{\varepsilon}_r$ then there exists $\gamma_\varepsilon > 0$ s.t.

$$\|\boldsymbol{x}(k) - x^*\mathbb{1}\| \leq c e^{-\gamma_\varepsilon k}\|\boldsymbol{x}_0 - x^*\mathbb{1}\|$$

for all $\boldsymbol{x}_0 \in B_0^{x^*}$.

Under the hypothesis of both the previous theorems, for sufficiently small ε and initial points $x_i(0)$ sufficiently close to the equilibrium point x^*, the dynamics can be summarized in

$$(x_i(k) - x^*) \approx (x_i(0) - x^*)\, e^{-\varepsilon \sum_{t=0}^{k-1} w_i(t)}$$

for all $i \in \mathcal{V}$, i.e., all the local estimates show a linear convergence to the global optimum, with rates depending both on ε and on the number of local updates. As a consequence, one would like to increase ε as much as possible, however large ε might lead the system to instability if the initial conditions are not sufficiently close to the global optimum x^*.

Remark 8. In Alg. 1 we considered linear iterative average consensus schemes for notational simplicity reasons. However, *the proof of Prop. 6 does not rely on linear consensus updates on a undirected graph.* The only requirement is to update the y_i's and z_i's with an algorithm that achieves average consensus exponentially uniformly fast. Therefore it is possible to exploit also convergence acceleration methods (see, e.g., Aysal et al. (2009)) or average consensus algorithms for directed graphs (Franceschelli et al., 2011; Cai and Ishii, 2012).

Remark 9. The previous theorems are based on Assumptions 4 and 5, that are deterministic worst-case hypotheses on the agent and edge activations processes. We conjecture that substituting the previous deterministic assumptions with randomized ones where i.i.d. agents and edge activations satisfy

$$\mathbb{E}[w_i(k)] \geq \overline{w} > 0, \quad \forall i \in \mathcal{V}, \forall k \in \mathbb{N} \tag{16}$$

$$\mathbb{E}[u_{(i,j)}] \geq \overline{u} > 0, \quad \forall (i,j) \in \mathcal{E}, \forall k \in \mathbb{N} \tag{17}$$

for some positive constants $\overline{w}$, $\overline{u}$ will lead to exponential bounds holding almost surely.

6. NUMERICAL EXAMPLES

We compare the performance of the ANRC with the asynchronous DSM summarized in Alg. 3.

Algorithm 3 DSM (Nedić and Ozdaglar, 2009)

(storage allocation and constraints on parameters)

$\mathbf{x} \in \mathbb{R}^N$ for $k = 0, 1, \ldots$

c_i = counter associated to agent i, $i = 1, \ldots, N$ ($\boldsymbol{c} := [c_1, \ldots, c_N]^T$)

$\rho \in \mathbb{R}_+$

(initialization)

$\mathbf{x}(0) = \mathbf{0}$, $\boldsymbol{c}(0) = \mathbf{0}$

(main algorithm)

for $k = 0, 1, \ldots$ **do**

$$\mathbf{x}(k+1) = P(k)\left[\boldsymbol{x}(k) - \rho E(k)\frac{\boldsymbol{f}'(\boldsymbol{x}(k))}{\boldsymbol{c}(k)}\right]$$

$$\boldsymbol{c}(k+1) = \boldsymbol{c}(k) + E(k)\mathbb{1}$$

end for

The aim is to show that, for the considered experiments, the convergence rates of Alg. 2 are faster than the ones of Alg. 3. We now present the quantities involved in the simulations and the kind of experiments performed. Then we describe the results in Sec. 6.1.

The tuning parameters of Algorithms 2 and Alg. 3 have been manually selected in order to achieve the fastest convergence rates possible while preventing divergence effects. The empirically selected parameters are $\varepsilon = 0.15$ for the NRC and $\rho = 100$ for the DSM.

We also consider two particular graphs, both of $N = 25$ agents: the random geometric graph of Fig. 1 and a complete one. We generate the local costs as

$$f_i(x) = c_i e^{a_i x} + d_i e^{-b_i x}, \quad i = 1, \ldots, N \tag{18}$$

where $a_i, b_i \sim \mathcal{U}[0, 0.2]$ and $c_i, d_i \sim \mathcal{U}[0, 1]$. Some examples are shown in Fig. 1.

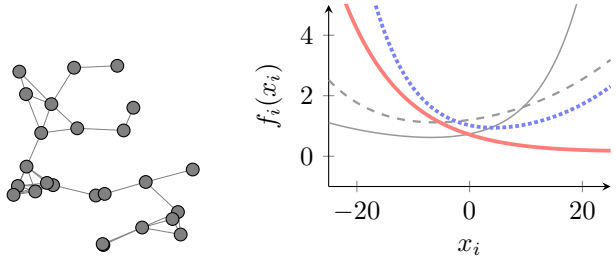


Fig. 1. The random geometric graph used in the simulations and some examples of local cost functions (18).

We use symmetric gossip for the consensus protocol as in (15) with $\alpha = 0.5$. The agents activation sequence is obtained concatenating independent permutations of the elements of $\mathcal{V}$. Similarly, the edges activation sequence is obtained concatenating independent permutations of the elements of $j \in \mathcal{N}_i$ once a agent i is selected, so that Ass. 4 is ensured with $B = Nd_{max}$ where d_{max} is the largest agent degree of the network.

We then consider the following Monte-Carlo experiment. Run M independent trials, where in each trial the network and the set of cost functions (18) are fixed, $\boldsymbol{x}_0 = 0$, and $S(k), E(k), P(k)$'s and generated based on the activation strategy just described and are used in both Alg. 2 and 3.

6.1 Experimental results

Let $e(k)$ and $v(k)$ be the average temporal evolution of the mean error and of its dispersion, i.e.,

$$e(k) := \frac{1}{MN}\sum_{m=1}^{M} ||\boldsymbol{x}_m(k) - x^*\mathbb{1}|| \tag{19}$$

$$v(k) := \frac{1}{MN}\sum_{m=1}^{M} \Big(||\boldsymbol{x}_m(k) - x^*\mathbb{1}|| - e(k)\Big)^2. \tag{20}$$

Fig. 3 plots the evolutions of the $e(k)$'s and $v(k)$'s relative to the outcomes of Algorithms 2 and 3 considering respectively the graph of Fig. 1 and the complete graph (some of the local costs are shown in Fig. 1). We can notice that both Algorithms 2 and 3 always converge to x^*, that the ANRC performs statistically better than the DSM, and that the effect of the topology of the network can play a crucial role on the convergence properties of these algorithms (the intuition being that the faster the consensus is, the faster the optimization converges).

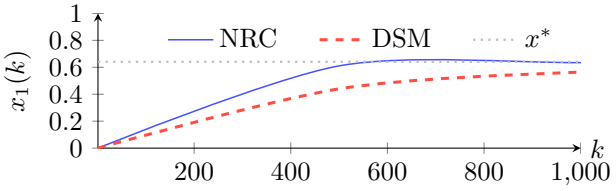
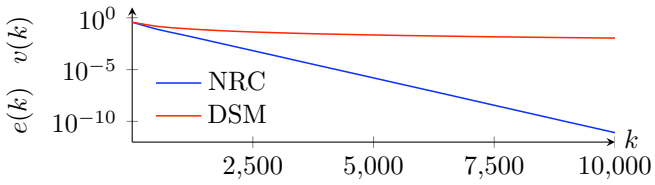
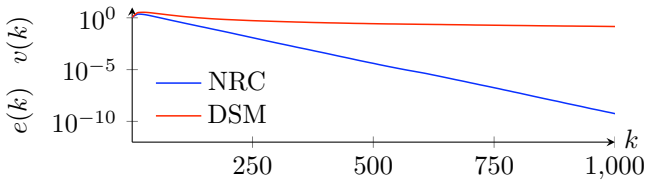


Fig. 2. Temporal evolution of the local guess $x_1(k)$.

Fig. 2 considers instead the first Monte-Carlo simulation used to generate Fig. 3-(b), and compares the temporal evolutions of the local guess $x_1(k)$ for the ANRC and DSM, supporting the claim that ANRC has faster convergence rates capabilities than DSM.



(a) graph of Fig. 1



(b) complete graph

Fig. 3. Temporal evolution of the performance indicators $e(k)$ (solid lines) and $e(k) \pm v(t)$ (dashed lines).

7. CONCLUSIONS AND FUTURE WORKS

We proposed an Asynchronous Newton-Raphson Consensus (ANRC) algorithm, a second-order distributed convex optimization technique with low computational and communication requirements.

By proposing this extension we showed that NRC may play an important role among the distributed optimization algorithms. It has in fact a natural niche, composed by the situations where the network topology is unknown and possibly time-varying (for which ADMMs may suffer of extremely complex implementations), and where the local cost functions are sufficiently smooth (for which ANRCs converge faster than DSMs due to the fact that the former uses also second-order information).

Remarkably, we showed that the proposed algorithm uses average-consensus as a building block. Thus it naturally supports the use of accelerated consensus techniques, that can further improve its convergence properties.

Besides stating that NRC can be asynchronous and it can use accelerated consensus strategies, this paper offers just preliminary results. In fact it provides convergence proofs assuming deterministic communication protocols. Nonetheless numerical investigations lead to conjecture that the algorithm preserves convergence properties for certain opportune stochastic protocols.

Future works, that need firstly to address this issue, should also analyze the effects of numerical errors and packet losses, the convergence speed under specific graphs and local cost functions scenarios, and also extend the technique to constrained problems.

REFERENCES

Aysal, T., Oreshkin, B., and Coates, M. (2009). Accelerated Distributed Average Consensus via Localized Node State Prediction. *IEEE Transactions on Signal Processing*, 57(4), 1563–1576.

Bertsekas, D.P. and Tsitsiklis, J.N. (1997). *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific.

Bolognani, S., Del Favero, S., Schenato, L., and Varagnolo, D. (2010). Consensus-based distributed sensor calibration and least-square parameter identification in WSNs. *International Journal of Robust and Nonlinear Control*, 20(2), 176–193.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Technical report, Stanford Statistics Dept.

Cai, K. and Ishii, H. (2012). Average Consensus on General Strongly Connected Digraphs. *Automatica (accepted for publication)*, 1–32.

Erseghe, T., Zennaro, D., Dall'Anese, E., and Vangelista, L. (2011). Fast Consensus by the Alternating Direction Multipliers Method. *IEEE Transactions on Signal Processing*, 59(11), 5523–5537.

Fagnani, F. and Zampieri, S. (2008). Randomized consensus algorithms over large scale networks. *IEEE Journal on Selected Areas in Communications*, 26(4), 634–649.

Fischione, C. (2011). F-Lipschitz Optimization with Wireless Sensor Networks Applications. *IEEE TAC*, 56(10), 2319 – 2331.

Fischione, C. and Jönsson, U. (2011). Fast-Lipschitz Optimization with Wireless Sensor Networks Applications. In *Int. symp. on Inf. Proc. in sensor networks*.

Franceschelli, M., Giua, A., and Seatzu, C. (2011). Distributed Averaging in Sensor Networks Based on Broadcast Gossip Algorithms. *IEEE Sensors Journal*, 11(3), 808–817.

Jadbabaie, A., Ozdaglar, A., and Zargham, M. (2009). A Distributed Newton Method for Network Optimization. In *CDC*, 2736–2741. IEEE.

Johansson, B. (2008). *On Distributed Optimization in Networked Systems*. Ph.D. thesis, KTH Royal Institute of Technology.

Khalil, H.K. (2001). *Nonlinear Systems*. Prentice Hall, third edition.

Kokotović, P., Khalil, H.K., and O'Reilly, J. (1999). *Singular Perturbation Methods in Control: Analysis and Design*. Number 25 in Classics in applied mathematics. SIAM.

Nedić, A. and Ozdaglar, A. (2009). Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE TAC*, 54(1), 48–61.

Shor, N.Z. (1985). *Minimization Methods for Non-Differentiable Functions*. Springer-Verlag.

Tsitsiklis, J.N. (1984). *Problems in decentralized decision making and computation*. Ph.D. thesis, MIT.

Van Ast, J., Babška, R., and Schutter, B.D. (2008). Particle swarms in optimization and control. In *IFAC World Congress*.

Xiao, L., Boyd, S., and Lall, S. (2005). A Scheme for Robust Distributed Sensor Fusion Based on Average Consensus. In *Int. symp. on Inf. Proc. in sensor networks*, 63–70.

Zanella, F., Varagnolo, D., Cenedese, A., Pillonetto, G., and Schenato, L. (2011). Newton-Raphson consensus for distributed convex optimization. In *CDC and European Control Conference*, 5917–5922.

Zanella, F., Varagnolo, D., Cenedese, A., Pillonetto, G., and Schenato, L. (2012). Multidimensional Newton-Raphson consensus for distributed convex optimization. In *ACC*.