# Spectral-like gradient method for distributed optimization

Dusan Jakovetić*, Nataša Krejić*, Nataša Krklec Jerinkić*
*Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad
{dusan.jakovetic, natasa, natasa.krklec}@dmi.uns.ac.rs

***Abstract*—We consider a standard distributed multi-agent optimization setting where $n$ nodes (agents) in a network minimize the aggregate sum of their local convex cost functions. We present a distributed spectral-like gradient method, wherein step-sizes are node- and iteration-varying, and they are inspired by classical spectral methods from centralized optimization. Simulation examples illustrate the performance of the presented method.**

***Index Terms*—Distributed optimization, Consensus optimization, Spectral gradient method, Barzilai-Borwein method.**

## I. Introduction

We consider distributed optimization problems where $n$ nodes in a generic network cooperate to minimize the sum of their local convex costs. Such problems find applications in many relevant fields, including distributed inference, e.g., [1], [2], distributed control, e.g., [3], and parallel and distributed machine learning, e.g., [14]. In this paper, to solve the described class of problems, we present a novel distributed first order method based on the class of spectral gradient methods from centralized optimization.

*Spectral gradient methods* are a popular class of methods in the context of conventional, centralized optimization, due to their simplicity and efficiency. The class originated with the Barzilei-Borwein method [6] and its analysis therein for convex quadratic functions, while the method has been subsequently extended to more general optimization problems, both unconstrained and constrained, by Raydan and Birgin et al, [4], [5], [8]. Spectral gradient methods incorporate second-order information in a computationally efficient manner into gradient descent methods, achieving in practice significantly faster per-iteration convergence than standard gradient methods while the additional computational overhead per iteration is very small. Roughly, the main idea behind spectral gradient methods is to approximate the Hessian at each iteration with a scalar matrix (the leading scalar of the matrix is called the spectral coefficient) that approximately fits the secant equation. Reference [5] demonstrates that the spectral gradient method can be more efficient than the conjugate gradient method for certain classes of optimization problems. R-linear convergence of the method was established in [9], while extensions to constrained optimization in the form of Spectral Projected Gradient (SPG) methods are developed, e.g., in [7]. A vast number of applications is available in the literature and a comprehensive overview is presented in [8]. In the context of distributed systems and algorithms, reference [17] considers a method with spectral-like step-sizes for localization problems.

In this paper, we present a distributed method that is a generalization of spectral gradient methods for centralized optimization. Extension of spectral gradient methods to a distributed setting is a highly nontrivial task. We present an *exact* method (converging to the exact solution) that utilizes step-sizes that are akin to those of conventional (centralized) spectral methods, where the spectral-like step-sizes are "embedded" into the exact distributed first order method in [11]. We utilize the primal-dual interpretation of the method in [11], as developed in [13], and the corresponding form of the error recursion equation. We then exploit an analogy with the error recursion of the conventional (centralized) spectral method [4] to define the time-varying, node dependent, algorithm driven step-sizes. This analogy also allows for an intuitive interpretation of the presented method. R-linear convergence of the presented method under appropriate conditions on the cost functions and with appropriate safeguarding on the step-sizes is proved in [16]. Convergence proofs and further analytical and numerical studies can be found in [16].

The paper is organized as follows. Section 2 describes the problem and gives preliminaries. The novel distributed spectral gradient method is presented in Section 3. Initial numerical tests are presented in Section 4, while conclusions are drawn in Section 5.

## II. Model and preliminaries

**Optimization and network models**. We consider a connected network with $n$ nodes, each of which has access to a local cost function $f_i : \mathbb{R}^d \to \mathbb{R},\ i = 1, \ldots, n$. The objective for all nodes is to minimize the aggregate cost function $f : \mathbb{R}^d \to \mathbb{R}$, defined by

$$f(y) = \sum_{i=1}^{n} f_i(y). \tag{1}$$

We assume that each function $f_i,\ i = 1, \ldots, n$, is strongly convex with modulus $\mu > 0$, i.e., there holds:

$$f_i(z) \geq f_i(y) + \nabla f_i(y)^T (z-y) + \frac{\mu}{2}\|z-y\|^2,\ y, z \in \mathbb{R}^d; \tag{2}$$

Research supported by the Serbian Ministry of Education, Science, and Technological Development, Grant no. 174030. This work is also supported by the I-BiDaaS project, funded by the European Commission under Grant Agreement No. 780787. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

also, gradients of the $f_i$'s are Lipschitz continuous with constant $L$, i.e.:

$$\|\nabla f_i(y) - \nabla f_i(z)\| \leq L\|y - z\|,\ y, z \in \mathbb{R}^d,\ i = 1, \ldots, n. \tag{3}$$

Under Assumptions (2) and (3), problem (1) is solvable and has the unique solution, denoted by $y^\star$. For future reference, introduce also function $F : \mathbb{R}^{nd} \to \mathbb{R}$, defined by $F(x) = \sum_{i=1}^n f_i(x_i)$, where the argument $x \in \mathbb{R}^{nd}$ consists of $n$ blocks $x_i \in \mathbb{R}^d$, i.e., $x = (\,(x_1)^T, ..., (x_n)^T\,)^T$.

We assume that the network of nodes is an undirected, connected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges, i.e., all pairs $\{i, j\}$ of nodes which can exchange information through a communication link. We denote by $O_i$ the set of nodes that are connected with node $i$ through a direct link (neighborhood set), and let $\bar{O}_i = O_i \bigcup \{i\}$. A symmetric, doubly stochastic $n \times n$ matrix $W$ with strictly positive diagonal entries is associated with $\mathcal{G}$. Denote by $\lambda_1 \geq \ldots \geq \lambda_n$ the eigenvalues of $W$. It can be shown that $\lambda_1 = 1$, and $|\lambda_i| < 1$, $i = 2, ..., n$.

For future reference, the matrix $\mathcal{W} = W \otimes I$, where $\otimes$ denotes the Kronecker product and $I$ is the (here of size $d \times d$) identity matrix is introduced. It can be seen that matrix $\mathcal{W}$'s $d \times d$ block on the $(i, j)$-th position equals $w_{ij}\, I$. By properties of the Kronecker product, the eigenvalues of $\mathcal{W}$ take values $\lambda_1, ..., \lambda_n$, each occurring with multiplicity $d$.

**Centralized spectral gradient method**. We briefly review the spectral gradient (SG) method in centralized optimization. Consider unconstrained minimization of a generic objective function $\phi : \mathbb{R}^d \to \mathbb{R}$ which is continuously differentiable. Let the initial solution estimate be an arbitrary $x^0 \in \mathbb{R}^d$. The SG method generates the sequence of iterates $\{x^k\}$ as follows:

$$x^{k+1} = x^k - \frac{1}{\sigma_k} \nabla\phi(x^k),\ k = 0, 1, \ldots, \tag{4}$$

where the initial spectral coefficient $\sigma_0 > 0$ is arbitrary and $\sigma_k$, $k = 1, 2, ...$, is given by

$$\sigma_k = \mathcal{P}_{[\,\underline{\sigma}, \overline{\sigma}\,]}\left\{\sigma_{k,\mathrm{aux}}\right\},\quad \sigma_{k,\mathrm{aux}} = \frac{(s^{k-1})^T y^{k-1}}{(s^{k-1})^T s^{k-1}}. \tag{5}$$

Here, $0 < \underline{\sigma} < \overline{\sigma} < +\infty$ are given constants, $s^{k-1} = x^k - x^{k-1}$, $y^{k-1} = \nabla\phi(x^k) - \nabla\phi(x^{k-1})$, and $\mathcal{P}_{[a,b]}$ stands for the projection of a scalar onto the interval $[a, b]$. The projection onto the interval $[\underline{\sigma}, \overline{\sigma}]$ is the safeguarding that is necessary for convergence under a generic cost function $\phi$. The quantity $\sigma_{k,\mathrm{aux}}$ can be interpreted as follows. Assume that we seek Hessian approximation in the form $B_k = \sigma_k I$, i.e., we seek scalar $\sigma_k$, such that the secant equation approximation $B_k s^{k-1} \approx y^{k-1}$ is the best possible, in the least squares sense. It is easy to show that this requirement yields (5). For future reference, we briefly review a result on the evolution of error with the SG method. Consider the special case of a strongly convex quadratic function $\phi(x) = \frac{1}{2} x^T A x + b^T x$ for a symmetric positive definite matrix $A$, and denote by $e^k := x^\star - x^k$ the error at iteration $k$, where $x^\star$ is the minimizer of $\phi$. Then, it can be shown that the error evolution can be expressed as [4]:

$$e^{k+1} = (I - \sigma_k^{-1} A) e^k. \tag{6}$$

The above relation will play a key role in the intuitive explanation of the distributed spectral gradient method that we propose.

## III. Distributed spectral-like method

**The algorithm**. We now present the novel distributed spectral gradient method. The algorithm is based on the exact distributed first order method in [11] and incorporates into this method a spectral-like step size policy; the utilized step-sizes vary both across nodes and across iterations. The algorithm maintains over iterations $k = 0, 1, ...$, at each node $i$, solution estimate $x_i^k \in \mathbb{R}^d$ and an auxiliary variable $g_i^k \in \mathbb{R}^d$. The initial solution estimate $x_i^0$ is arbitrary, while $g_i^0 = \nabla f_i(x_i^0)$, $i = 1, ..., n$. The update rule is the following

$$x_i^{k+1} = \sum_{j \in \bar{O}_i} W_{ij}\, x_j^k - \frac{1}{\sigma_i^k}\, g_i^k \tag{7}$$

$$g_i^{k+1} = \sum_{j \in \bar{O}_i} W_{ij}\, g_j^k + \left(\nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k)\right),\ k = 0, 1, ... \tag{8}$$

The inverse step-sizes $\sigma_i^k$ are given by

$$\sigma_i^k = \mathcal{P}_{[\underline{\sigma}, \overline{\sigma}]}\{\sigma_{i,\mathrm{aux}}^k\} \tag{9}$$

$$\sigma_{i,\mathrm{aux}}^k = \frac{(s_i^{k-1})^T y_i^{k-1}}{(s_i^{k-1})^T s_i^{k-1}} + \sigma_i^{k-1} \sum_{j \in \bar{O}_i} W_{ij} \left(1 - \frac{(s_j^{k-1})^T s_j^{k-1}}{(s_i^{k-1})^T s_i^{k-1}}\right)$$

$$s_i^{k-1} = x_i^k - x_i^{k-1},\ y_i^{k-1} = \nabla f_i(x^k) - \nabla f_i(x_i^{k-1}),$$

where $0 < \underline{\sigma} < \overline{\sigma} < +\infty$ are, as before, the safeguarding parameters. The initial step-size $\sigma_i^{-1}$ is an arbitrary scalar in the interval $[\underline{\sigma}, \overline{\sigma}]$, $i = 1, ..., n$.

The method in (7)–(8) and (9) at iteration $k$ is implemented as follows. First, each node $i$ transmits to all its neighbors $j \in O_i$ the vector $x_i^k$ and receives $x_j^k$, $j \in O_i$. Next, each node calculates $\sigma_i^k$ according to (9) and afterwards performs the update (7). Subsequently, each node $i$ transmits to all its neighbors $j \in O_i$ quantity $g_i^k$ and receives $g_j^k$, $j \in O_i$, and finally performs update (8). Clearly, each of the two steps (7) and (8) is fully distributed and each requires one $d$-dimensional vector exchange between the neighboring nodes. Note that step-size calculation (9) is local to each node $i$ and does not require inter-neighbor communications.[1]

We now comment on the structure of the presented method (7)–(8). The method incorporates spectral step

[1] In order to implement step (9), each node $i$ needs quantities $s_j^{k-1} = x_j^k - x_j^{k-1}$, for all $j \in O_i$; they are available to node $i$ thanks to the two most recent receptions from neighbors, i.e., thanks to the availability of quantities $x_j^k$ and $x_j^{k-1}$, $j \in O_i$.

sizes (9) into the exact distributed first order method in [11]; in other words, when one sets $(\sigma_i^k)^{-1} = \alpha$, for all $i, k$, with $\alpha > 0$ sufficiently small, algorithm (7)–(8) reduces to the method in [11]. The spectral step sizes definition in (9) arises from a non-trivial analysis and is based on an analogy with centralized spectral gradient methods, as discussed ahead in more detail. For further details see [16]. Quantity $g_i^k$ – as with the method in [11] – serves as node $i$'s estimate of the network-wide average gradient $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)$; quantity $g_i^k$ is placed in (7) as a substitute of current node $i$'s local gradient $\nabla f_i(x_i^k)$ with standard distributed gradient methods, e.g., [15], which brings significant benefits to convergence properties of the method. For details on the choice of safeguarding parameters in (9), we refer to [16].

**Step-size derivation**. We now provide a derivation and an intuitive justification of the step-size choice (9). The derivation is based on a primal-dual interpretation of the method in [11] available in [13] and on an analogy with the error evolution with the centralized spectral gradient method [4]. Recall that $y^\star \in \mathbb{R}^d$ is the solution to (1); further, denote by $x^k = (\,(x_i^k)^T, ..., (x_n^k)^T\,)^T \in \mathbb{R}^{nd}$, $g^k = (\,(g_i^k)^T, ..., (g_n^k)^T\,)^T \in \mathbb{R}^{nd}$, and let $\Sigma_k$ be the diagonal matrix whose $i$-th diagonal entry equals $\sigma_i^k$. Next, let $e_{\rm p}^k := x^k - x^\bullet$, where $x^\bullet = (\,(y^\star)^T, ..., (y^\star)^T\,)^T$ ($y^\star$ is repeated $n$ times), and $e_{\rm d}^k := g^k + \nabla F(x^\bullet) - \nabla F(x^k)$. Quantity $e_{\rm p}^k$ represents the primal error, while quantity $e_{\rm d}^k$ represents the dual error at iteration $k$.[2] For notational simplicity, assume that $d = 1$, while similar considerations hold for $d > 1$ as well. Also, denote by $J$ the $n \times n$ matrix with all entries equal to $1/n$. Let each $f_i$ be a strongly convex quadratic function, i.e., $f_i(x_i) = \frac{1}{2}h_i(x_i - b_i)^2$, and $H = diag(h_1, \ldots, h_n)$, $h_i > 0$, for all $i$. Then, the following recursion holds, see [13]:

$$\begin{bmatrix} e_{\rm p}^{k+1} \\ e_{\rm d}^{k+1} \end{bmatrix} = \begin{bmatrix} W - \Sigma_k^{-1} H & -\Sigma_k^{-1} \\ (W-I)H & W - J \end{bmatrix} \cdot \begin{bmatrix} e_{\rm p}^k \\ e_{\rm d}^k \end{bmatrix} \qquad (10)$$

We now present an analogy between the error recursion with the centralized spectral gradient method for a strongly convex quadratic cost with leading matrix $A$, see equation (6), and the error recursion of the presented method (10). With the centralized spectral gradient method, the error recursion's matrix equals $I - \sigma_k^{-1}A$, and the (new) spectral coefficient $\sigma_{k+1}$ is set such that the secant equation

$$\sigma_{k+1}(x^{k+1} - x^k) = A(x^{k+1} - x^k) \qquad (11)$$

is fitted with least mean square deviation. In other words, the error recursion matrix $I - \sigma_{k+1}^{-1}A$ is set in such a way that $\sigma_{k+1}\, I$ is a scalar matrix approximation of $A$.

We next consider the error recursion (10) of the presented distributed method, and we specifically focus on the primal error update:

$$\begin{aligned} e_{\rm p}^{k+1} &= \left(W - \Sigma_k^{-1} H\right)\, e_{\rm p}^k + \Sigma_k^{-1}\, e_{\rm d}^k \\ &= \left(\, I - \Sigma_k^{-1}\, \left[\Sigma_k\,(I - W) + H\right]\,\right)\, e_{\rm p}^k \\ &+ \Sigma_k^{-1}\, e_{\rm d}^k. \end{aligned} \qquad (12)$$

As we can see, the second equation above does not involve $\Sigma_k$. When (6) is compared with (12), we can note that both the primal error and the dual error affect (12). We may reduce the effect of the dual error $e_{\rm d}^k$ by letting $\Sigma_k^{-1}$ be small enough. This motivates safeguarding of $\Sigma_k$ from below. On the other hand, regarding the effect of $e_{\rm p}^k$, we can see that it is achieved through matrix $I - \Sigma_k^{-1}\,[\Sigma_k\,(I - W) + H]$. We make this matrix small as with the centralized spectral gradient method's case, through the following identification: $A \equiv \Sigma_k\,(I - W) + H$, and $\sigma_{k+1} \equiv \Sigma_{k+1}$. Hence, we look for the matrix $\Sigma_{k+1}$ as the least-mean-squares-error fit of the following:

$$\Sigma_{k+1}\, \left(x^{k+1} - x^k\right) = (\Sigma_k(I - W) + H)\, \left(x^{k+1} - x^k\right).$$

When cost functions are generic (non-quadratic), the equation above translates into:

$$\begin{aligned} \Sigma_{k+1}\, \left(x^{k+1} - x^k\right) &= (\,\Sigma_k(I - W)\,)\, \left(x^{k+1} - x^k\right) \\ &+ \left(\,\nabla F(x^{k+1}) - \nabla F(x^k)\,\right). \end{aligned}$$

Matrix $\Sigma_{k+1,\rm aux}$ is then obtained by minimizing:

$$\begin{aligned} \|\, \Sigma_{k+1}\, \left(x^{k+1} - x^k\right) - (\,\Sigma_k(I - W)\,)\, \left(x^{k+1} - x^k\right) \\ - \left(\,\nabla F(x^{k+1}) - \nabla F(x^k)\,\right)\, \|^2. \end{aligned}$$

The minimizer $\Sigma_{k+1,\rm aux}$ here equals $diag(\sigma_{1,\rm aux}^{k+1}, ..., \sigma_{n,\rm aux}^{k+1})$, with $\sigma_{i,\rm aux}^{k+1}$ as in (9) for $k$ replaced with $k+1$, $i = 1, ..., n$. In order to ensure both strictly positive step-sizes and a bounded effect of the dual error, $\Sigma_{k+1,\rm aux}$ is finally projected entry-wise onto the interval $[\,\underline{\sigma}, \overline{\sigma}\,]$, yielding (9).

**Convergence and convergence rate**. For a sufficiently conservative choice of the safeguarding coefficients $\underline{\sigma}$ and $\overline{\sigma}$, it can be shown that the presented distributed spectral method converges to the exact solution at an R-linear rate. This can be shown through an extension of Theorem 2 in [12]. Details are available in [16]. Namely, reference [12] analyses a variant of the method in [11] with node-varying and time-invariant step-sizes (without consideration of spectral-like step-sizes), but the result therein can be extended to time varying step-sizes like in (9). [3] More precisely, it can be shown that there exist positive constants $c'$ and $c''$ that depend on the number of nodes $n$, weight matrix $W$, and the $f_i$'s parameters $\mu$ and $L$, such that, for $\frac{1}{\underline{\sigma}} < c' \leq 1/(2\,L)$ and $1 \leq \overline{\sigma}/\underline{\sigma} < c''$. The error $\|x_i^k - y^\star\|$ converges to zero R-linearly, for each node $i = 1, ..., n$. Extensive simulations on strongly convex quadratic and logistic losses indicate that method (7)–(8) always converges for $\frac{1}{\underline{\sigma}} < c/L$ and $\frac{1}{\underline{\sigma}} > 1/\theta$, where $c$ and

[2] More precisely, $e_{\rm d}^k$ is a linear transformation of the dual error defined with respect to an augmented Lagrangian dual reformulation of (1); see [13] for details.

[3] This result can be shown through a "worst case" analysis that does not take into account the specific form of $\sigma_{j,\rm aux}^k$ in (9) but only utilizes information on the safeguarding parameters $\underline{\sigma}$ and $\overline{\sigma}$.

$\theta$ can be taken at least as large as $c = 100$, $\theta = 10^8$. Pursuing convergence analysis for a less conservative safeguarding is left for future work.

## IV. Simulations

This section provides a numerical example to illustrate the performance of the presented novel distributed spectral method. The example demonstrates a significant speedup gained through the presented spectral-like step-size policy with respect to the counterpart constant step-size method in [11].

We let $f_i : \mathbb{R}^d \to \mathbb{R}$, $f_i(x) = \frac{1}{2}(x-b_i)^T A_i (x-b_i)$, $d = 10$, where $b_i \in \mathbb{R}^d$ and $A_i \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix. The quantities $A_i, b_i$ are generated randomly, independently across nodes, where each $b_i$'s entry is generated mutually independently from the uniform distribution on $[1, 31]$. Further, each $B_i$ is generated as $B_i = Q_i\, D_i\, Q_i^T$, where $Q_i$ is the matrix of orthonormal eigenvectors of $\frac{1}{2}(\widehat{B}_i + \widehat{B}_i^T)$, and $\widehat{B}_i$ has independent, identically distributed (i.i.d.) entries from standard normal distribution. Next, $D_i$ is a diagonal matrix with the entries drawn in an i.i.d. fashion from the uniform distribution on $[1, 101]$. The underlying graph has $n = 30$-nodes and is generated as a connected graph instance of the random geometric model with radius $r = \sqrt{\frac{\ln(n)}{n}}$. Finally, matrix $W$ is as follows: for $\{i,j\} \in E$, $i \neq j$, $w_{ij} = \frac{1}{2(1+\max\{d_i,d_j\})}$, where $d_i$ is the node $i$'s degree; for $\{i,j\} \notin E$, $i \neq j$, $w_{ij} = 0$; and $w_{ii} = 1 - \sum_{j\neq i} w_{ij}$, for all $i = 1, ..., n$.

We compare the presented method and the method in [11]. The comparison allows to assess the benefits of incorporating spectral-like step-sizes into distributed first order methods. We use the relative error (averaged across nodes) as the solution estimate quality metric: $\frac{1}{n}\sum_{i=1}^n \frac{\|x_i - y^\star\|}{\|y^\star\|}$, $y^* \neq 0$. Both methods have all parameters equal except the step-sizes. With the method in [11], we set step-size as $\alpha = 1/(3L)$, where $L = \max_{i=1,...,n} \mu_i$, and $\mu_i$ is the maximal eigenvalue of $A_i$. This value is the maximal possible step-size for the method in [11], according to the empirical evaluations in [11]. In general, optimal tuning of the step-size with [11] requires beforehand tuning and is problem and network dependent. For more details, we refer to [16]. With the presented method, all nodes' step-sizes are initialized to the value $1/(3L)$. Further, we let $\overline{\sigma} = 10^8$ and $\underline{\sigma} = \frac{3L}{10}$. This means that the presented method allows the step-sizes to reach up to 10 times larger values than the maximal possible value with [11].

Figure 1 plots the relative error versus number of iterations with the two methods. We observed that [11] with step-size equal to $1/\underline{\sigma} = 10/(3L)$ diverges on this example.

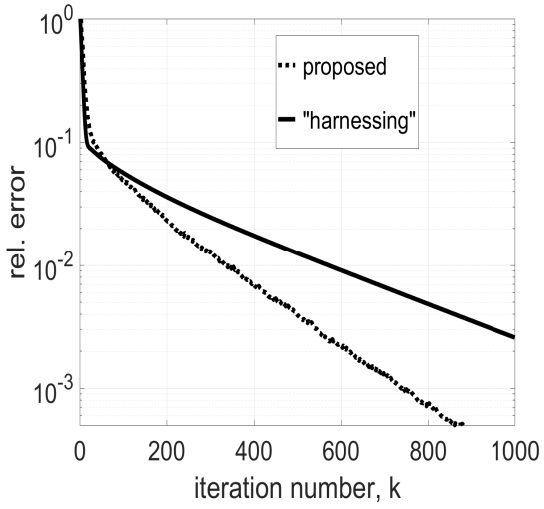


Fig. 1. Relative error versus iteration number for the method in [11] ("harnessing", solid line) and the novel presented method (dotted line).

## V. Conclusion

We presented an exact distributed spectral-like gradient method for distributed optimization. The construction of the novel node- and time-varying step-sizes is based on a primal-dual interpretation of the method and on an analogy that we draw with centralized spectral gradient algorithms. The presented method exhibits R-linear convergence to the exact solution under standard assumptions on the nodes' local costs and under appropriate safeguarding of the step-sizes. Simulation examples on strongly convex quadratic costs illustrate performance of the presented method.

## References


[1] Schizas, I. D. , Ribeiro, A., Giannakis, G. B., Consensus in ad hoc WSNs with noisy links – Part I: Distributed estimation of deterministic signals, *IEEE Transactions on Signal Processing*, vol. 56, no. 1, (2009) pp. 350–364.

[2] Cattivelli, F., Sayed, A. H., Diffusion LMS strategies for distributed estimation, *IEEE Transactions on Signal Processing*, vol. 58, no. 3, (2010) pp. 1035–1048.

[3] Mota, J., Xavier, J., Aguiar, P., Püschel, M., Distributed optimization with local domains: Applications in MPC and network flows, *to appear in IEEE Transactions on Automatic Control*, 2015.

[4] Raydan, M., On the Barzilai and Borwein Choice of Steplength for the Gradient Method, IMA Journal of Numerical Analysis, 13 (1993), 321-326.

[5] Raydan, M., Barzilai and Borwein Gradient Method for the Large Scale Unconstrained Minimization Problem, SIAM Journal on Optimization 7 (1997), 26 - 33.

[6] Barzilai J, Borwein JM, Two Point Step Size Gradient Methods, IMA Journal of Numerical Analysis, 8 (1988), 141 - 148.

[7] Birgin, E.G, Martínez, J.M, Raydan M., Nonmonotone Spectral Projected Gradient Methods on Convex Sets, SIAM Journal on Optimization, 10, (2000), 1196-1211.

[8] Birgin, E.G., Martínez, J.M., Raydan M Spectral Projected Gradient Methods: Review and Perspectives, Journal of Statistical Software 60(3), (2014), 1-21.

[9] Dai, Y.H., Liao, L.Z., R-Linear Convergence of the Barzilai and Borwein Gradient Method, IMA Journal on Numerical Analysis, 22 (2002), 1-10.

[10] Shi, W., Ling, Q., Wu, G., Yin, W., EXTRA: an Exact First-Order Algorithm for Decentralized Consensus Optimization, *SIAM Journal on Optimization*, No. 25 vol. 2, (2015) pp. 944-966.

[11] Qu, G., Li, N., Harnessing smoothness to accelerate distributed optimization, IEEE Transactions on Control of Network Systems (to appear)

[12] Nedic, A., Olshevsky, A., Shi, W., Uribe, C.A., Geometrically convergent distributed optimization with uncoordinated step-sizes, 2016

[13] Jakovetić, D., A Unification, Generalization and Acceleration of Exact Distributed First Order Methods, 2017

[14] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning, Volume 3, Issue 1, (2011) pp. 1-122.
[15] Nedić, A., Ozdaglar, A., Distributed subgradient methods for multi-agent optimization, *IEEE Transactions on Automatic Control*, vol. 54, no. 1, (2009) pp. 48–61.
[16] Jakovetić, D., Krejić, N., Krklec Jerinkić, N., Exact Spectral-Like Method for Distributed Optimization, arxiv.org/abs/1901.05682
[17] G. Calafiore, L. Carlone, M. Wei, A distributed gradient method for localization of formations using relative range measurements, IEEE International Symposium on Computer-Aided Control System Design, 2010.