Multi-step gradient methods for networked optimization

Euhanna Ghadimi, Iman Shames, and Mikael Johansson

Abstract

We develop multi-step gradient methods for network-constrained optimization of strongly convex functions with Lipschitz-continuous gradients. Given the topology of the underlying network and bounds on the Hessian of the objective function, we determine the algorithm parameters that guarantee the fastest convergence and characterize situations when significant speed-ups over the standard gradient method are obtained. Furthermore, we quantify how uncertainty in problem data at design-time affects the run-time performance of the gradient method and its multi-step counterpart, and conclude that in most cases the multi-step method outperforms gradient descent. Finally, we apply the proposed technique to three engineering problems: resource allocation under network-wide budget constraint, distributed averaging, and Internet congestion control. In all cases, our proposed algorithms converge significantly faster than the state-of-the art.

I. Introduction

Distributed optimization has recently attracted significant attention from several research communities. Examples include the work on network utility maximization for resource allocation in communication networks [1], distributed coordination of multi-agent systems [2], collaborative estimation in wireless sensor networks [3], distributed machine learning [4], and many others. The majority of these praxes apply gradient or sub-gradient methods to the dual formulation of the decision problem. Although gradient methods are easy to implement and require modest computations, they suffer from slow convergence. In some cases, such as the development of distributed power control algorithms for cellular phones [5], one can replace gradient methods by fixed-point iterations and achieve improved convergence rates. For other problems, such as average consensus [6], a number of heuristic methods have been proposed that improve the convergence time of the standard method [7], [8]. However, we are not interested in tailoring techniques to individual problems; our aim is to develop general-purpose schemes that retain the simplicity of the gradient method but improve convergence times.

Even if the optimization problem is convex and the subgradient method is guaranteed to converge to an optimal solution, the rate of convergence is very modest. The convergence rate of the gradient method is improved if the objective function is differentiable with Lipschitz-continuous gradient, and even more so if the function is also strongly convex [9]. When the objective and constraint functions are smooth, several techniques exist that allow for even shorter solution times. One such technique is higher-order methods, such as Newton's method [10], which use both the gradient and the Hessian of the objective function. Although distributed Newton methods have recently been developed for special problem classes (*e.g.*, [11], [12]), they impose large communication overhead for collecting global Hessian information. Another way to obtain faster convergence is to use

E. Ghadimi and M. Johansson are with the ACCESS Linnaeus Center, Electrical Engineering, Royal Institute of Technology, Stockholm, Sweden. {euhanna, mikaelj}@ee.kth.se. I. Shames is with the Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, Australia. iman.shames@unimelb.edu.au.

multi-step methods [13], [10]. These methods rely only on gradient information but use a history of the past iterates when computing the future ones. We explore the latter approach for distributed optimization.

This paper makes the following contributions. First, we develop a multi-step weighted gradient method that maintains a network-wide constraint on the decision variables throughout the iterations. The accelerated algorithm is based on the heavy ball method by Polyak [13] extended to the networked setting. We derive optimal algorithm parameters, show that the method has linear convergence rate and quantify the improvement in convergence factor over the gradient method. Our analysis shows that method is particularly advantageous when the eigenvalues of the Hessian of the objective function and/or the eigenvalues of the graph Laplacian of the underlying network have a large spread. Second, we investigate how similar techniques can be used to accelerate dual decomposition across a network of decision-makers. In particular, given smoothness parameters of the objective function, we present closed-form expressions for the optimal parameters of an accelerated gradient method for the dual. Third, we quantify how the convergence properties of the algorithm are affected when the algorithm is tuned using misestimated problem parameters. This robustness analysis shows that the accelerated algorithm endures parameter violations well and in most cases outperforms its non-accelerated counterpart. Finally, we apply the developed algorithms to three case studies: networked resource allocation, consensus, and network flow control. In each application we demonstrate superior performance compared to alternatives from the literature.

The paper is organized as follows. In Section II, we introduce our networked optimization problem. Section III reviews multi-step gradient techniques. Section IV proposes a multi-step weighted gradient algorithm, establishes conditions for its convergence and derives optimal step-size parameters. Section V develops a technique for accelerating the dual problem based on parameters for the (smooth) primal. Section VI presents a robustness analysis of the multi-step algorithm in the presence of uncertainty. Section VII applies the proposed techniques to three engineering problems: resource allocation, consensus and network flow control; numerical results and performance comparisons are presented for each case study. Finally, concluding remarks are given in Section VIII.

II. ASSUMPTIONS AND PROBLEM FORMULATION

This paper is concerned with collaborative optimization by a network of decision-makers. Each decision-maker v is endowed with a loss function $f_v : \mathbf{R} \mapsto \mathbf{R}$, has control of one decision-variable $x_v \in \mathbf{R}$, and collaborates with the others to solve

$$\begin{array}{ll} \text{minimize} & \sum_{v\in\mathcal{V}} f_v(x_v) \\ \text{subject to} & Ax = b, \end{array} \tag{1}$$

for given matrices $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$. We will assume that b lies in the range space of A, *i.e.* that there exists at least one decision vector x that satisfies the constraints. The physical information exchange between decision-makers is represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \{1, 2, \dots, n\}$ and edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Specifically, at each time t, we will assume that decision-maker v has access to $\nabla f_w(x_w(t))$ for all its neighbors $w \in \mathcal{N}_v \triangleq \{w \mid (v, w) \in \mathcal{E}\}$.

The formulation (1) is more general that it might first appear. For example, problems with coupled objectives can be put on the form (1) by first introducing new variables, which act as local copies of the shared variable, and then constraining these copies to be equal (see, *e.g.*, [14]). We will use this trick in Section VII-B. To use such modeling techniques in their fullest generality (*e.g.* to allow loss functions to be coupled through several distinct decision variables), one would need to allow the loss functions to be multivariable. While the results in this paper are generalizable to multivariable loss functions, we have chosen to present the scalar case for ease of notation.

Most acceleration techniques in the literature (*e.g.* [9], [15], [16]) require that the loss functions are smooth and convex. Similarly, we will make the following assumptions:

Assumption 1: Each loss function f_v is convex and twice continuously differentiable with

$$l_v \leq \nabla^2 f_v(x_v) \leq u_v, \quad \forall x_v, \tag{2}$$

for some positive real constants l_v, u_v with $0 < l_v \leq u_v$.

Some remarks are in order. Let $l = \min_{v \in \mathcal{V}} l_v$, $u = \max_{v \in \mathcal{V}} u_v$ and define $f(x) := \sum_{v \in \mathcal{V}} f_v(x_v)$. Then, Assumption 1 ensures that $f(x)$ is strongly convex with modulus l

$$f(y) \geq f(x) + (y-x)^\top \nabla f(x) + \frac{l}{2}\|y-x\|^2 \quad \forall (x,y),$$

that its gradient is Lipschitz-continuous with constant u

$$f(y) \leq f(x) + (y-x)^\top \nabla f(x) + \frac{u}{2}\|y-x\|^2 \quad \forall (x,y),$$

and that the Hessian of f satisfies

$$lI \leq \nabla^2 f(x) \leq uI \quad \forall x. \tag{3}$$

See, *e.g*, [9, Lemma 1.2.2 and Theorem 2.1.11] for details. Furthermore, Assumption 1 guarantees that (1) is a convex optimization problem whose unique optimizer $x^\star$ satisfies

$$Ax^\star = b, \qquad \nabla f(x^\star) = A^\top \mu^\star, \tag{4}$$

where $\mu^\star \in \mathbf{R}^m$ is the vector of optimal Lagrange multipliers for the linear constraints.

III. Background on multi-step methods

The basic gradient method for unconstrained minimization of a convex function $f(x)$ takes the form

$$x(k+1) = x(k) - \alpha \nabla f(x(k)), \tag{5}$$

where $\alpha > 0$ is a fixed step-size parameter. Assume that $f(x)$ is strongly convex with modulus l and has Lipschitz-continuous gradient with constant u. Then if $\alpha < 2/u$, the sequence $\{x(k)\}$ generated by (5) converges to $x^\star$ at linear rate, *i.e.* there exists a convergence factor $q \in (0,1)$ such that $\|x(k+1) - x^\star\| \leq q\|x(k) - x^\star\|$ for all k. The smallest convergence factor is $q = (u-l)/(u+l)$ obtained for $\alpha = 2/(l+u)$ (see, *e.g.*, [13]).

While the *convergence rate* cannot be improved unless higher-order information is considered [13], the *convergence factor* q can be meliorated by accounting for the history of iterates when computing the ones to come. Methods in which the next iterate depends not only on the current iterate but also on the preceding ones are called *multi-step methods*. The simplest multi-step extension of the gradient method is

$$x(k+1) = x(k) - \alpha \nabla f(x(k)) + \beta \left(x(k) - x(k-1) \right), \tag{6}$$

for fixed step-size parameters $\alpha > 0$ and $\beta > 0$. This technique, originally proposed by Polyak, is sometimes called the heavy-ball method from the physical interpretation of the added “momentum term”. For a centralized set-up, Polyak derived the optimal step-size parameters and showed that these guarantee a convergence factor of $(\sqrt{u}-\sqrt{l})/(\sqrt{u}+\sqrt{l})$, which is always smaller than the convergence factor for the gradient method and significantly so when $\sqrt{u}/\sqrt{l}$ is large.

In what follows, we will develop multi-step gradient methods for network-constrained optimization, analyze their convergence properties and derive optimal step-size parameters.

IV. A MULTI-STEP WEIGHTED GRADIENT METHOD

In the absence of constraints, (1) is trivial to solve since the objective function is separable and each decision-maker could simply minimize its loss independently of the others. Hence, it is the existence of constraints that makes (1) challenging. In the optimization literature, there are essentially two ways of dealing with constraints. One way is to project iterates onto the constraint set to maintain feasibility at all times; such a method will be developed in this section. The other way is to use dual decomposition to eliminate couplings between decision-makers and solve the associated dual problem; we will consider such techniques in Section V.

Computing the Euclidean projection onto the constraint of (1) typically requires the full decision vector x, which is not available to the decision-makers in our setting. An alternative, explored *e.g.* in [17], is to consider *weighted* gradient methods which use a linear combination of the information available to nodes to ensure that iterates remain feasible. For our problem (1) the weighted gradient method takes the form

$$x(k+1)=x(k)-\alpha W\nabla f(x(k)). \tag{7}$$

Here, $W\in\mathbf{R}^{n\times n}$ is a weight matrix that should satisfy the following three conditions: (i) the locality of information exchange between the decision makers should be preserved; (ii) provided that the initial point $x(0)$ is feasible, the iterates generated by (7) should remain feasible; and (iii), the fixed-points of (7) should satisfy the optimality conditions (4).

To ensure condition (i), W has to have the same sparsity pattern as the information graph $\mathcal{G}$, *i.e.* $W_{vw}=0$ if $v\neq w$ and $(v,w)\notin\mathcal{E}$. In this way, the iterations (7) read

$$x_v(k+1)=x_v(k)-\alpha\sum_{w\in v\cup\mathcal{N}_v}W_{vw}\nabla f_w(x_w(k)),$$

and can be executed by individual decision-makers based on the information that they have access to. Conditions (ii) and (iii) translate into the following requirements (see [17] for details)

$$AW=0, \qquad\qquad WA^\top=0. \tag{8}$$

The next example describes one particular problem instance.

Example 1: When the decision-makers are only constrained by a global resource budget, (1) reduces to

$$\begin{array}{ll}\text{minimize} & \sum_{v\in\mathcal{V}}f_v(x_v)\\ \text{subject to} & \sum_{v\in\mathcal{V}}x_v=x_{\text{tot}}.\end{array}$$

A distributed algorithm for this problem was developed in [18] and interpreted as a weighted gradient method in [17]. In our notation, $A=\mathbf{1}^T$ and $b=x_{\text{tot}}$ so in addition to the sparsity pattern, W should also satisfy $\mathbf{1}^\top W=0^\top$ and $W\mathbf{1}=0$. In [17], it was shown that the symmetric W that satisfies these constraints and guarantees the smallest convergence factor of the weighted gradient iterations can be found by solving a convex optimization problem. In addition, [17] proposed several heuristics for constructing a good W in a distributed manner.

A few comments are in order. First, not all constraint matrices A admit a weight matrix W that satisfies the above constraints, hence not all problems on the form (1) are amendable to a distributed solution using a weighted gradient method. Second, to find a feasible initial point $x(0)$, typically, one needs to find a solution for the linear system of equations $Ax = b$ in a distributed way. This generally requires a separate distributed mechanism; see, *e.g.* [18] and [19, Chapter 2].

A. A multi-step weighted gradient method and its convergence

Consider the following multi-step weighted gradient iteration

$$x(k+1) = x(k) - \alpha W \nabla f(x) + \beta \left(x(k) - x(k-1) \right). \tag{9}$$

Under the sparsity constraint on W detailed above, these iterations can be implemented by individual decision-makers. Moreover, (8) ensures that if $Ax(0) = Ax(1) = b$ then every iterate produced by (9) will also satisfy the linear constraints. The next theorem characterizes the convergence of the iterations (9) and derives optimal step-size parameters α and β.

Theorem 1: Consider the optimization problem (1) under Assumption 1, and let $x^\star$ denote its unique optimizer. Assume that W has $m < n$ eigenvalues at 0 and satisfies $AW = 0$ and $WA^\top = 0$. Let $H = \nabla^2 f(x^\star)$ and $\lambda_1(WH) \leq \lambda_2(WH) \leq \cdots \leq \lambda_n(WH)$ be the (ordered) eigenvalues of WH so that $\underline{\lambda} = \lambda_{m+1}(WH)$ is the smallest non-zero eigenvalue of WH and $\overline{\lambda} = \lambda_n(WH)$ is the largest. Then, if

$$0 \leq \beta \leq 1, \qquad\qquad 0 < \alpha < \frac{2\,(1+\beta)}{u\,\lambda_n(W)},$$

the iterates (9) converge to $x^\star$ at linear rate

$$\|x(k+1) - x^\star\| \leq q \|x(k) - x^\star\| \qquad \forall k \geq 0,$$

with $q = \max\left\{\sqrt{\beta}, |1+\beta-\alpha\underline{\lambda}| - \sqrt{\beta}, |1+\beta-\alpha\overline{\lambda}| - \sqrt{\beta}\right\}$. Moreover, the minimal value of q is

$$q^\star = \frac{\sqrt{\overline{\lambda}} - \sqrt{\underline{\lambda}}}{\sqrt{\overline{\lambda}} + \sqrt{\underline{\lambda}}},$$

obtained for step-sizes $\alpha = \alpha^\star$ and $\beta = \beta^\star$ where

$$\alpha^\star = \left(\frac{2}{\sqrt{\overline{\lambda}} + \sqrt{\underline{\lambda}}} \right)^2, \quad \beta^\star = \left(\frac{\sqrt{\overline{\lambda}} - \sqrt{\underline{\lambda}}}{\sqrt{\overline{\lambda}} + \sqrt{\underline{\lambda}}} \right)^2. \tag{10}$$

Proof: See appendix for this and all other proofs. ■

Similar to the discussion in Section III, it is interesting to investigate when (9) significantly improves over the single-step algorithm. In [17], it is shown that the best convergence factor of the weighted gradient iteration (7) is

$$q_0^\star = \frac{\overline{\lambda} - \underline{\lambda}}{\overline{\lambda} + \underline{\lambda}}.$$

One can verify that $q^\star \leq q_0^\star$, *i.e.* the optimally tuned multi-step method is never slower than the single-step method. Moreover, the improvement in convergence factor depends on the quantity $\kappa = \overline{\lambda}/\underline{\lambda}$: when κ is large, the speed-up is roughly proportional to $\sqrt{\kappa}$. In the networked setting, there are two reasons for a large value of κ. One is simply that the Hessian of the objective function is ill-conditioned, so that the ratio u/l is large. The other is that the matrix W is ill-conditioned, *i.e.* that $\lambda_n(W)/\lambda_{m+1}(W)$ is large. As we will see in the examples, the graph Laplacian is often a valid choice for W. Thus, there is a direct connection between the topology of the underlying information graph and the convergence rate (improvements) of the multi-step weighted gradient method. We will discuss this connection in detail in Section VII.

In many applications, we will not know $H = \nabla^2 f(x^\star)$, but only bounds such as (3). The next result can then be useful

Proposition 1: Let $\underline{\lambda}_W = l\lambda_{m+1}(W)$ and $\overline{\lambda}_W = u\lambda_n(W)$. Then $\underline{\lambda}_W \leq \underline{\lambda}$ and $\overline{\lambda}_W \geq \overline{\lambda}$. Moreover, the step-sizes

$$\alpha = \left(\frac{2}{\sqrt{\overline{\lambda}_W} + \sqrt{\underline{\lambda}_W}}\right)^2, \qquad \beta = \left(\frac{\sqrt{\overline{\lambda}_W} - \sqrt{\underline{\lambda}_W}}{\sqrt{\overline{\lambda}_W} + \sqrt{\underline{\lambda}_W}}\right)^2,$$

ensure the linear convergence of (9) with the convergence factor

$$\tilde{q} = \frac{\sqrt{\overline{\lambda}_W} - \sqrt{\underline{\lambda}_W}}{\sqrt{\overline{\lambda}_W} + \sqrt{\underline{\lambda}_W}}.$$

B. *Optimal weight selection for the multi-step method*

The results in the previous subsection provide optimal step-size parameters α and β for a given weight matrix W. However, the expressions for the associated convergence factors depend on the eigenvalues of WH and optimizing the entries in W jointly with the step-size parameters can yield even further speed-ups. We make the following observation.

Proposition 2: Under the hypotheses of Proposition 1,

(i) If H is known, then minimizing the convergence factor $q^\star$ is equivalent to minimizing $\overline{\lambda}/\underline{\lambda}$.

(ii) If H is not known, while l and u in (3) are, then the weight matrix that minimizes $\tilde{q}$ is the one with minimal value of $\overline{\lambda}_W/\underline{\lambda}_W$.

The next result shows how the optimal weight selection for both scenarios can be found via convex optimization.

Proposition 3: Let $\mathcal{M}_\mathcal{G}$ be the span of real symmetric matrices, $\mathcal{S}^n$, whose sparsity pattern is induced by the graph $\mathcal{G}$, *i.e.*

$$\mathcal{M}_\mathcal{G} = \{M \in \mathcal{S}^n \mid M_{vw} = 0 \text{ if } v \neq w \text{ and } (v,w) \notin \mathcal{E}\}.$$

Then the weight matrix for (9) that minimizes $\overline{\lambda}/\underline{\lambda}$ can be found by solving the convex optimization problem

$$\begin{array}{ll} \underset{W,t}{\text{minimize}} & t \\ \text{subject to} & I_{n-m} \leq P^\top H^{1/2} W H^{1/2} P \leq t I_{n-m} \\ & W \in \mathcal{M}_\mathcal{G},\ W \geq 0,\ H^{1/2} W H^{1/2} V = 0, \end{array} \tag{11}$$

where $V = H^{-1/2}A^\top$ and $P \in \mathbf{R}^{n \times n-m}$ is a matrix of orthonormal vectors spanning the null space of $V^\top$.

Remark 1: When H is not known but the bounds l and u in (3) are, Proposition 2 suggests that one should look for a weight matrix W that minimizes $\overline{\lambda}_W/\underline{\lambda}_W$. This can be done using the formulation in Proposition 3 by setting $H = I$.

Remark 2: In addition to the basic conditions that admissible weight matrices have to satisfy, Proposition 3 also requires that W be positive semi-definite. Without this additional requirement (11) is not a convex optimization problem and generically has no tractable solution (see [20], [21]). Designing distributed algorithms for constructing weight matrices that guarantee optimal or close-to-optimal convergence factors is an open question and a subject for future investigations.

V. A MULTI-STEP DUAL ASCENT METHOD

When the constraint matrix A is such that the weight optimization problem (11) does not admit a solution, then we have no structured approach to find a weight matrix W that guarantees convergence of the weighted multi-step gradient iterations. An alternative approach for solving (1) can then be to use Lagrange relaxation, *i.e.* to introduce Lagrange multipliers $\mu \in \mathbf{R}^m$ for the equality constraints and solve the dual problem. The dual function associated with (1) is

$$d(\mu) \triangleq \inf_x \sum_v \left\{ f_v(x_v) + \left(\sum_{r=1}^m \mu_r A_{rv} \right) x_v \right\} - \sum_{r=1}^m \mu_r b_r. \tag{12}$$

Since the dual function is separable in x, it can be evaluated in parallel. For a given Lagrange multiplier vector μ, each decision-maker then needs to compute

$$x_v^\star(\mu) = \arg\min_z f_v(z) + \left(\sum_{r=1}^m \mu_r A_{rv} \right) z. \tag{13}$$

The dual problem is to maximize $d(\mu)$ with respect to μ, i.e.,

$$\underset{\mu}{\text{minimize}} \quad -d(\mu) = f_\star(-A^\top \mu) + b^\top \mu,$$

where $f_\star(y) \triangleq \sup_x \; y^\top x - f(x)$ is the conjugate function. Recall that if f is strongly convex then $f_\star$ and hence $-d(\cdot)$ are convex and continuously differentiable [22]. Moreover, $\nabla d(\mu) = Ax^\star(\mu) - b$. In light of our earlier discussion, it is natural to attempt to solve the dual problem using a multi-step iteration on the form

$$\begin{aligned} \mu_r(k+1) \quad &= \quad \mu_r(k) + \alpha \left(\textstyle\sum_v A_{rv} x_v^\star(\mu(k)) - b_r \right) \\ & \qquad + \beta \left(\mu_r(k) - \mu_r(k-1) \right). \end{aligned} \tag{14}$$

To be able to execute the multi-step dual ascent iterations (13) and (14) in a distributed manner, decision-maker v needs to be able to collect the Lagrange multipliers μ_r for all r such that $A_{rv} \neq 0$, and the decision-maker in charge of updating μ_r needs to be able to collect all $x_v^\star(\mu)$ for all v with $A_{rv} \neq 0$. This is certainly not always possible, but we will give two examples that satisfy these requirements in Section VII.

To find the optimal step-sizes and estimate the convergence factors of the iterations, we need to be able to bound the strong convexity modulus of $d(\mu)$ and the Lipschitz constant of its gradient. The following observation is in order:

Lemma 1: Consider the optimization problem (1) with associated dual function (12). Let f be a continuously differentiable and closed convex function. Then,

(i) If f is strongly convex with modulus l, then $-\nabla d$ is Lipschitz continuous with constant $\lambda_n(AA^\top)/l$.

(ii) If ∇f is Lipschitz continuous with constant u, then $-d$ is strongly convex with modulus $\lambda_1(AA^\top)/u$.

These dual bounds can be used to derive the following result:

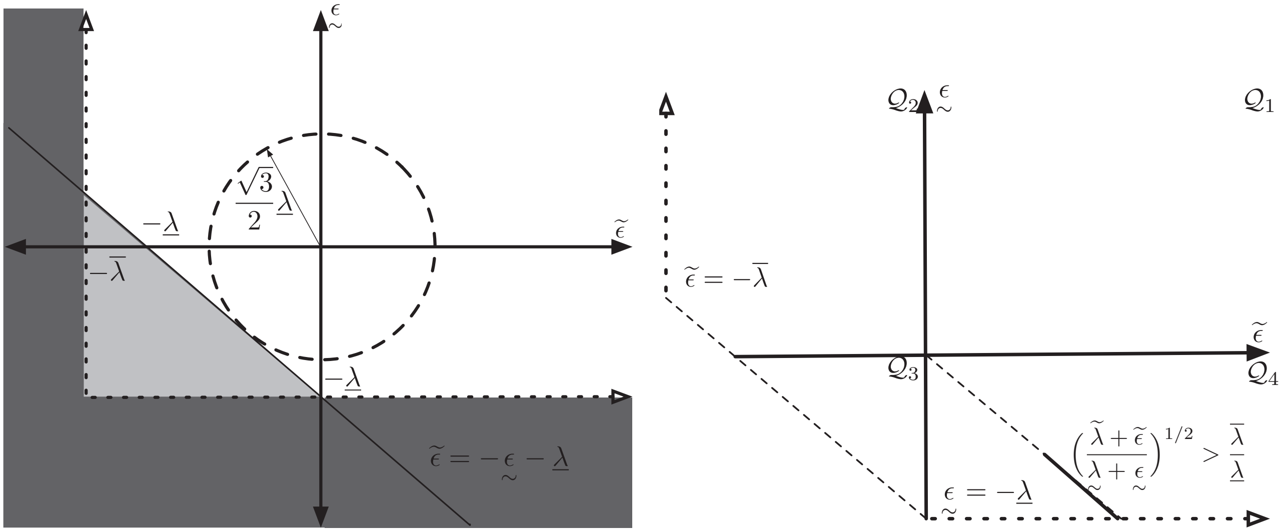
Theorem 2: For the optimization problem (1) under Assumption 1, the multi-step dual ascent iterations (14) converge to $\mu^\star$ at linear rate with the guaranteed convergence factor

$$q^\star = \frac{\sqrt{u\lambda_n(AA^\top)} - \sqrt{l\lambda_1(AA^\top)}}{\sqrt{u\lambda_n(AA^\top)} + \sqrt{l\lambda_1(AA^\top)}},$$

obtained for step-sizes:

$$\alpha^\star = \left(\frac{2}{\sqrt{u\lambda_n(AA^\top)} + \sqrt{l\lambda_1(AA^\top)}} \right)^2,$$

$$\beta^\star = \left(\frac{\sqrt{u\lambda_n(AA^\top)} - \sqrt{l\lambda_1(AA^\top)}}{\sqrt{u\lambda_n(AA^\top)} + \sqrt{l\lambda_1(AA^\top)}} \right)^2.$$


$\underset{\sim}{\epsilon}$
$\frac{\sqrt{3}}{2}\underline{\lambda}$
$-\underline{\lambda}$
$\widetilde{\epsilon}$
$-\overline{\lambda}$
$-\underline{\lambda}$
$\widetilde{\epsilon} = -\underset{\sim}{\epsilon} - \underline{\lambda}$
$\mathcal{Q}_2$
$\underset{\sim}{\epsilon}$
$\mathcal{Q}_1$
$\widetilde{\epsilon} = -\overline{\lambda}$
$\widetilde{\epsilon}$
$\mathcal{Q}_3$
$\mathcal{Q}_4$
$\left(\frac{\widetilde{\lambda}+\widetilde{\epsilon}}{\underset{\sim}{\lambda}+\underset{\sim}{\epsilon}}\right)^{1/2} > \frac{\overline{\lambda}}{\underline{\lambda}}$
$\underset{\sim}{\epsilon} = -\underline{\lambda}$

(a) Stability regions

(b) Different perturbation regions

Fig. 1. Perturbations in the white and gray area correspond to the stable and unstable regions of multi-step algorithm respectively. (b) Multi-step algorithm outperforms gradient iterations in $(\underset{\sim}{\epsilon}, \widetilde{\epsilon}) \in \mathcal{C}\backslash\mathcal{Q}_4$. For symmetric errors in $\mathcal{Q}_4$ (along the line $\widetilde{\epsilon} = -\underset{\sim}{\epsilon}$) gradient might outperform multi-step algorithm.

The advantage of Theorem 2 is that it provides step-size parameters with guaranteed convergence factor using readily available data of the primal problem. How close to optimal these results are depends on how tight the bounds in Lemma 1 are. If the bounds are tight, then the step-sizes in Theorem 2 are truly optimal. The next example shows that a certain degree of conservatism may be present, even for quadratic problems.

Example 2: Consider the quadratic minimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^\top Qx \\ \text{subject to} & Ax = b, \end{array}$$

where $Q > 0$ is positive-definite, $A \in \mathbf{R}^{n\times n}$ is nonsingular and $b \in \mathbf{R}^n$. This implies that the objective function is strongly convex with modulus $\lambda_1(Q)$ and that its gradient is Lipschitz-continuous with constant $\lambda_n(Q)$. Hence, according to Lemma 1, $-d$ is strongly convex with modulus $\lambda_1(AA^\top)/\lambda_n(Q)$ and its gradient is Lipschitz continuous with constant $\lambda_n(AA^\top)/\lambda_1(Q)$. However, direct calculations reveal that

$$d(\mu) = -\frac{1}{2}\mu^\top AQ^{-1}A^\top\mu - \mu^\top b,$$

from which we see that $-d$ has convexity modulus $\lambda_1(AQ^{-1}A^\top)$ and that its gradient is Lipschitz continuous with constant $\lambda_n(AQ^{-1}A^\top)$. By [23, p. 225], these bounds are tighter than those offered by Lemma 1. Specifically, for congruent matrices Q^{-1} and $AQ^{-1}A^\top$ there exists nonnegative real numbers θ_k such that $\lambda_1(AA^\top) \leq \theta_k \leq \lambda_n(AA^\top)$ and $\theta_k\lambda_k(Q^{-1}) = \lambda_k(AQ^{-1}A^\top)$. For $k = 1$ and n we obtain

$$\frac{\lambda_1(AA^\top)}{\lambda_n(Q)} \leq \lambda_1(AQ^{-1}A^\top), \quad \lambda_n(AQ^{-1}A^\top) \leq \frac{\lambda_n(AA^\top)}{\lambda_1(Q)}.$$

For some important classes of problems, the bounds are, however, tight. One such example is the average consensus application considered in Section VII.

VI. Robustness analysis

The proposed multi-step methods have significantly improved convergence factors compared to the gradient iterations, and particularly so when the Hessian of the loss function and/or the graph Laplacian of the network is ill-conditioned. The results of Theorem 1 and Proposition 1 specify sufficient conditions for the convergence of multi-step iterations in terms of the design parameters α, β and W. However, these parameters are determined based on upper and lower bounds on the Hessian and the largest and smallest non-zero eigenvalue of W. In many applications, W and H might not be perfectly known, and $\overline{\lambda}$ and $\underline{\lambda}$ have to be estimated based on available data. It is therefore important to analyze the sensitivity of the multi-step methods to errors in these parameters to assess if the performance benefits prevail when the step-sizes are calculated using (slightly) misestimated $\overline{\lambda}$ and $\underline{\lambda}$. Such an analysis will be performed in this section.

Let $\underset{\sim}{\lambda}$ and $\widetilde{\lambda}$ denote the estimates of $\underline{\lambda}$ and $\overline{\lambda}$ available when tuning the step-sizes. We are interested in quantifying how the convergence and the convergence factors of the gradient and the multi-step methods are affected when $\underset{\sim}{\lambda}$ and $\widetilde{\lambda}$ are used in the step-size formulas that we have derived earlier. Theorem 1 provides some useful observations for the multi-step method. The corresponding results for the weighted gradient method are summarized in the following lemma:

Lemma 2: Consider the weighted gradient iterations (7) and let $\overline{\lambda}$ and $\underline{\lambda}$ denote the largest and smallest non-zero eigenvalue of WH, respectively. Then, for fixed step-size $0 < \alpha < 2/\overline{\lambda}$ (7) converges to $x^{\star}$ at linear rate with convergence factor

$$q_G = \max\left\{|1-\alpha\underline{\lambda}|,\ |1-\alpha\overline{\lambda}|\right\}.$$

The minimal value $q_G^{\star} = (\overline{\lambda}-\underline{\lambda})/(\overline{\lambda}+\underline{\lambda})$ is obtained for the step-size $\alpha = 2/(\overline{\lambda}+\underline{\lambda})$.

Combining this lemma with our previous results from Theorem 1 yields the following observation.

Proposition 4: Let $\underset{\sim}{\lambda}$ and $\widetilde{\lambda}$ be estimates of $\underline{\lambda}$ and $\overline{\lambda}$, respectively, and assume that $0 < \underset{\sim}{\lambda} < \widetilde{\lambda}$. Then, for all values of $\underset{\sim}{\lambda}$ and $\widetilde{\lambda}$ such that $\overline{\lambda} < \widetilde{\lambda} + \underset{\sim}{\lambda}$, both the weighted gradient iteration (7) with step-size

$$\widetilde{\alpha} = 2/(\widetilde{\lambda}+\underset{\sim}{\lambda}), \tag{15}$$

and the multi-step method variant (9) with

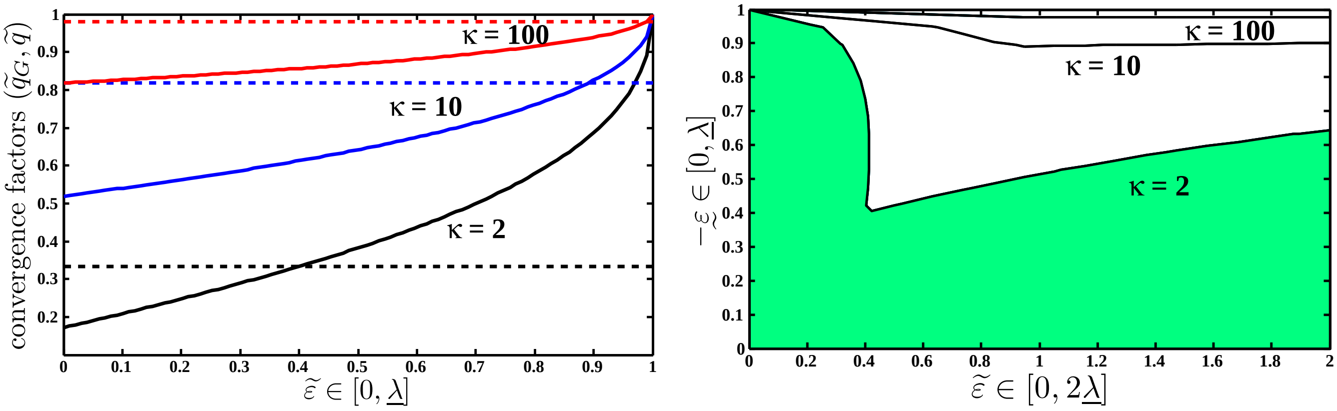
$$\widetilde{\alpha} = \left(\frac{2}{\sqrt{\widetilde{\lambda}}+\sqrt{\underset{\sim}{\lambda}}}\right)^2,\ \widetilde{\beta} = \left(\frac{\sqrt{\widetilde{\lambda}}-\sqrt{\underset{\sim}{\lambda}}}{\sqrt{\widetilde{\lambda}}+\sqrt{\underset{\sim}{\lambda}}}\right)^2, \tag{16}$$

converge to the optimizer $x^{\star}$ of (1).

In practice, one should expect that $\widetilde{\lambda}$ is overestimated, in which case both methods converge. However, convergence can be guaranteed for a much wider range of perturbations. Fig. 1(a) considers perturbations of the form $\underset{\sim}{\lambda} = \underline{\lambda} + \underset{\sim}{\varepsilon}$ and $\widetilde{\lambda} = \overline{\lambda} + \widetilde{\varepsilon}$. The white area is the locus of perturbations for which convergence is guaranteed, while the dark area represents inadmissible perturbations which render either $\underset{\sim}{\lambda}$ or $\widetilde{\lambda}$ negative. Note that both algorithms are robust to a continuous departure from the true values of $\underline{\lambda}$ and $\overline{\lambda}$, since there is a ball with radius $\sqrt{3}\underline{\lambda}/2$ around the origin for which both methods are guaranteed to converge.

Next, we compare the convergence *factors* of the two methods when the step-sizes are tuned based on inaccurate parameters. The following Lemma is then useful.

Lemma 3: Let $\underset{\sim}{\lambda}$ and $\widetilde{\lambda}$ satisfy $0 < \overline{\lambda} < \underset{\sim}{\lambda} + \widetilde{\lambda}$. The convergence factor of the weighted gradient method (7) with step-size



(a) Symmetric perturbations in $\mathcal{Q}_4$

(b) General perturbation in $\mathcal{Q}_4$

Fig. 2. (a) Convergence factor of multi-step and gradient algorithms under the condition described by (19). Solid lines belong to $\widetilde{q}$ while the dashed lines depict $\widetilde{q}_G$. (b) Level curves of $\widetilde{q} - \widetilde{q}_G$ around the origin for $(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_4$.

(15) is given by

$$\widetilde{q}_G = \begin{cases} 2\overline{\lambda}/(\underset{\sim}{\lambda} + \widetilde{\lambda}) - 1 & \text{if } \underset{\sim}{\lambda} + \widetilde{\lambda} \leq \underline{\lambda} + \overline{\lambda} \\ 1 - 2\underline{\lambda}/(\underset{\sim}{\lambda} + \widetilde{\lambda}) & \text{otherwise,} \end{cases} \tag{17}$$

while the multi-step weighted gradient method (9) with step-sizes (16) has convergence factor

$$\widetilde{q} = \max\left\{ \sqrt{\widetilde{\beta}}, |1 + \widetilde{\beta} - \widetilde{\alpha}\underline{\lambda}| - \sqrt{\widetilde{\beta}}, |1 + \widetilde{\beta} - \widetilde{\alpha}\overline{\lambda}| - \sqrt{\widetilde{\beta}} \right\}. \tag{18}$$

The convergence factor expressions derived in Lemma 3 allow us to come to the following conclusions:

Proposition 5: Let $\underset{\sim}{\lambda} = \underline{\lambda} + \underset{\sim}{\varepsilon}$, $\widetilde{\lambda} = \overline{\lambda} + \widetilde{\varepsilon}$ and define the set of perturbation under which the methods converge

$$\mathcal{C} = \{(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \mid \underset{\sim}{\varepsilon} \geq -\underline{\lambda}, \, \widetilde{\varepsilon} \geq -\overline{\lambda}, \, \underset{\sim}{\varepsilon} + \widetilde{\varepsilon} \geq -\underline{\lambda}\},$$

and the fourth quadrant in the perturbation space $\mathcal{Q}_4 = \{(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \mid \underset{\sim}{\varepsilon} < 0 \cap \widetilde{\varepsilon} > 0\}$. Then, for all $(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{C} \backslash \mathcal{Q}_4$, it holds that $\widetilde{q} \leq \widetilde{q}_G$. However, there exists $(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_4$ for which the scaled gradient has a smaller convergence factor than the multi-step variant. In particular, for

$$(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_4 \text{ and } (\overline{\lambda} + \widetilde{\varepsilon})/(\underline{\lambda} + \underset{\sim}{\varepsilon}) \geq (\overline{\lambda}/\underline{\lambda})^2, \tag{19}$$

the multi-step iterations (9) converge slower than (7).

Fig. 1(b) illustrates the different perturbations considered in Proposition 5. While the multi-step method has superior convergence factors for many perturbations, the troublesome region $\mathcal{Q}_4$ is probably common in engineering applications since it represents perturbations where the smallest eigenvalue is underestimated while the largest eigenvalue is overestimated. To shed more light on the convergence properties in this region, we perform a numerical study on a quadratic function with $\underline{\lambda} = 1$ and $\overline{\lambda}$ varying from 2 to 100. We first consider symmetric perturbations $\underset{\sim}{\varepsilon} = -\widetilde{\varepsilon}$, in which case the convergence factor of the gradient method is $\widetilde{q}_G = 1 - 2/(1 + \overline{\lambda}/\underline{\lambda})$ while the convergence factor of the multi-step method is $\widetilde{q} = 1 - 2/\sqrt{1 + \overline{\lambda}/\underline{\lambda}}$. The convergence factor of the gradient iterations is insensitive to symmetric perturbations, while the performance of the multi-step iterations degrades with the size of the perturbation and eventually becomes inferior to the gradient; see Fig. 2(a). To complement

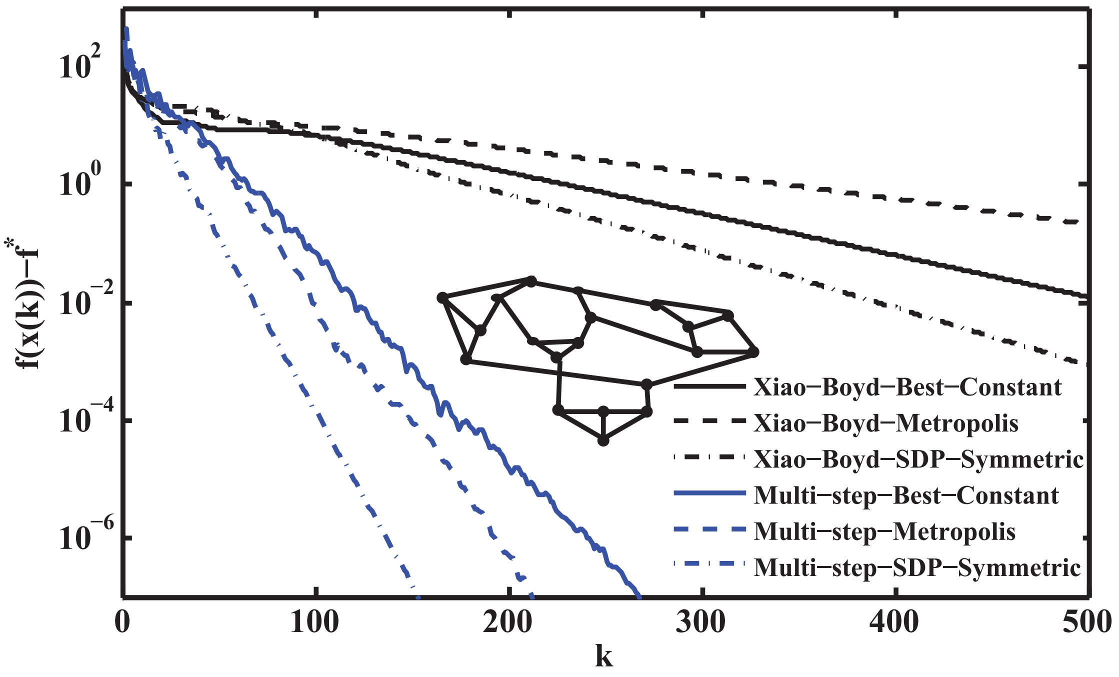


Fig. 3. Convergence behavior convergence behavior for weighted and multi-step weighted gradient iterations using randomly generated network and the heuristic weights. Plot shows $f\big(x(k)\big) - f^\star$ versus iteration number k.

this study, we also sweep over $(\underset{\sim}{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{C} \cap \mathcal{Q}_4$ and compute the convergence factors for the two methods for problems with different $\overline{\lambda}$. Fig. 2(b) indicates that when the condition number $\overline{\lambda}/\underline{\lambda}$ increases, the area where the gradient method is superior (the area above the contour line) shrinks. It also shows that when $\underset{\sim}{\lambda}$ tends to zero or $\widetilde{\lambda}$ is very large, the performance of the multi-step method is severely degraded.

VII. Applications

In this section, we apply the developed techniques to three classes of engineering problems: resource-allocation under network-wide resource constraints, distributed averaging, and Internet congestion control. In all cases, we demonstrate that direct applications of our techniques yield algorithms with significantly faster convergence than state-of-the art algorithms that have been tailor-made to the specific applications.

A. Accelerated resource allocation

Our first application is the distributed resource allocation problem under a network-wide resource constraint [18], [17] discussed in Example 1. We compare the multi-step method developed in this paper with the optimal and suboptimal tuning for the standard weighted gradient iterations proposed in [17]. Similarly to [17] we create problem instances by generating random networks and assigning loss functions on the form $f_v(x_v) = a_v(x_v - c_v)^2 + \log[1 + \exp(x_v - d_v)]$ to nodes. The parameters a_v, b_v, c_v and d_v are drawn uniformly from intervals $(0, 2]$, $[-2, 2]$, $[-10, 10]$ and $[-10, 10]$, respectively. In [17] it was shown that the second derivatives of these functions are bounded by $l_v = a_v$ and $u_v = a_v + b_v^2/4$.

Fig. 3 shows a representative problem instance along with the convergence behavior for weighted and multi-step weighted gradient iterations under several weight choices. The optimal weights for the weighted gradient method are found by solving a semi-definite program derived in [17], and by Proposition 3, setting $H = I$, for the multi-step variant. In addition, we evaluate the

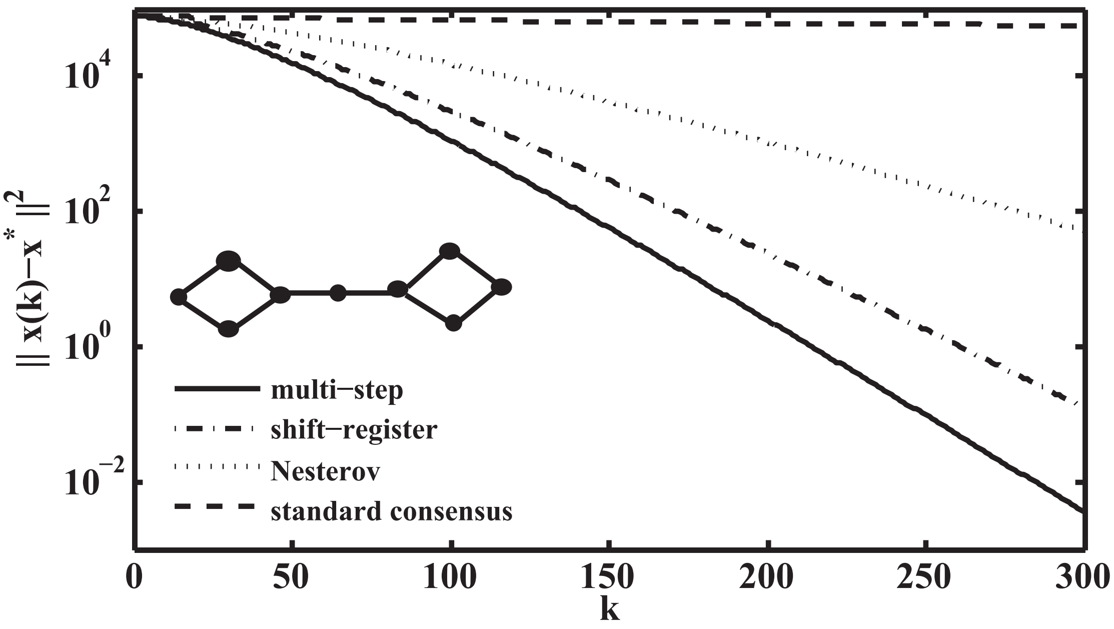


Fig. 4. Comparison of standard, multi-step, shift-register, and Nesterov consensus algorithms using metropolis wights. Simulation on a dumbbell of 100 nodes: log scale of objective function $\|x(k) - x^\star\|_2^2$ versus iteration number k. Algorithms start from common initial point $x(0)$.

heuristic weights "best constant" and "metropolis" introduced in [17]. In all cases, we observe significantly improved convergence factors for the multi-step method.

B. Distributed averaging

Our second application is devoted to the distributed averaging. Distributed algorithms for consensus seeking have been researched intensively for decades; see *e.g.* [6], [24], [25]. Here, each node v in the network initially holds a value c_v and coordinates with neighbors in the graph to find the network-wide average. Clearly, this average can be found by applying any distributed optimization technique to the problem

$$\underset{x}{\text{minimize}} \quad \textstyle\sum_{v\in\mathcal{V}} \frac{1}{2}(x - c_v)^2, \tag{20}$$

since the optimal solution to this problem is the network-wide average of the constants c_v. In particular, we will explore how the multi-step technique with our optimal parameter selection rule compares with the state-of-the art distributed averaging algorithms from the literature.

The basic consensus algorithms use iterations on the form

$$x_v(k+1) = Q_{vv}x_v(k) + \sum_{w\in\mathcal{N}_v} Q_{vw}x_w(k), \tag{21}$$

where Q_{vw} are scalar weights, and the node states are initialized with $x_v(0) = c_v$. The paper [26] provides necessary and sufficient conditions on the matrix $Q = [Q_{vw}]$ to ensure that the iterations converge to the network-wide average of the initial values. Specially, it is required that $\mathbf{1}^\top Q = \mathbf{1}^\top$, $Q\mathbf{1} = \mathbf{1}$, and $\rho(Q - (1/n)\mathbf{1}\mathbf{1}^\top) < 1$ where $\rho(\cdot)$ denotes the spectral radius of a matrix. Although the convergence conditions do not require that Q is symmetric, techniques for minimizing the convergence factor often assume Q to be symmetric [26], [8].

Following the steps given in Section V, the optimization approach to consensus would suggest the iterations (see [27] for a detailed derivation)

$$x(k+1) = x(k) - \alpha W x(k), \tag{22}$$

where $W = A^\top A$ and A is the incidence matrix of $\mathcal{G}$. These iterations are on the same form as (21) but use a particular weight matrix. The multi-step counterpart of (22) is

$$x(k+1) = \left((1+\beta)I - \alpha W\right) x(k) - \beta x(k-1). \tag{23}$$

In a fair comparison between the multi-step iterations (23) and the basic consensus iterations, the weight matrices of the two approaches should not necessarily be the same, nor necessarily equal to the graph Laplacian. Rather, the weight matrix for the consensus iterations (21) should be optimized using the results from [26] and the weight matrix for the multi-step iteration should be computed by using Proposition 3.

In addition to the basic consensus iterations with optimal weights, we will also compare our multi-step iterations with two alternative acceleration schemes from the literature. The first one comes from the literature on accelerated consensus and uses shift registers [7], [28], [29]. Similarly to the multi-step method, these techniques use a history of past iterates, stored in local registers, when computing the next. For the consensus iterations (21), the corresponding shift register iterations are

$$x(k+1) = \zeta Q x(k) + (1-\zeta) x(k-1). \tag{24}$$

The current approaches to consensus based on shift-registers assume that Q is given and design ζ to minimize the convergence factor of the iterations. The key results can be traced back to Golub and Varga [30] who determined the optimal ζ and the associated convergence factor to be

$$\zeta^\star = \frac{2}{1+\sqrt{1-\lambda_{n-1}^2(Q)}}, \quad q_{SR}^\star = \sqrt{\frac{1-\sqrt{1-\lambda_{n-1}^2(Q)}}{1+\sqrt{1-\lambda_{n-1}^2(Q)}}}. \tag{25}$$

In our comparisons, the shift-register iterations will use the Q-matrix optimized for the basic consensus iterations and the associated $\zeta^\star$ given above. The second acceleration technique that we will compare with is the order-optimal gradient methods developed by Nesterov [9]. While these techniques have optimal *convergence rate*, also in the absence of strong convexity, they are not guaranteed to obtain the best convergence factors. When the objective function is strongly convex with modulus l and its gradient is Lipschitz continuous with constant u, the following iterations are proposed in [9]:

$$\begin{aligned}
\hat{x}(k+1) &= x(k) - \nabla f(x(k))/u \\
x(k+1) &= \hat{x}(k+1) + \frac{\sqrt{u}-\sqrt{l}}{\sqrt{u}+\sqrt{l}}(\hat{x}(k+1) - \hat{x}(k)),
\end{aligned}$$

initialized with $\hat{x}(0) = x(0)$. When we apply this technique to the consensus problem, we arrive at the iterations

$$x(k+1) = (I - \alpha W)\left(x(k) + b(x(k) - x(k-1))\right), \tag{26}$$

with $W = AA^\top, a = \lambda_n^{-1}(W)$ and $b = (\sqrt{\lambda_n(W)} - \sqrt{\lambda_2(W)})/(\sqrt{\lambda_n(W)} + \sqrt{\lambda_2(W)})$, where $\lambda_2(\cdot)$ and $\lambda_n(\cdot)$ are the smallest and largest non-zero eigenvalues of their variables.

Fig. 4 compares the multi-step iterations (23) developed in this paper with (a) the basic consensus iterations (21) using a weight matrix determined using the metropolis scheme, (b) the shift-register acceleration (24) with the same weight matrix and the optimal ζ, and (c) the order-optimal method (26). The particular results shown are for a network of 100 nodes in a dumbbell topology. The simulations show that all three methods yield a significant improvement in convergence factors over the basic iterations, and that the multi-step method developed in this paper outperforms the alternatives.

Several remarks are in order. First, since the Hessian of (20) is equal to identity matrix, the speed-up of the multi-step iterations is proportional to $\sqrt{\kappa} = \sqrt{\lambda_n(W)/\lambda_2(W)}$. When W equals $\mathcal{L}$, the Laplacian of the underlying graph, we can quantify the speed-ups for certain classes of graphs using spectral graph theory [31]. For example, the complete graph has $\lambda_2(\mathcal{L}) = \lambda_n(\mathcal{L})$ so $\kappa = 1$ and there is no real advantage of the multi-step iterations. On the other hand, for a ring network the eigenvalues of $\mathcal{L}$ are given by $1 - \cos(2\pi v)/|\mathcal{V}|$, so κ grows quickly with the number of nodes, and the performance improvements of (23) over (22) could be substantial.

Our second remark pertains to the shift-register iterations. Since these iterations have the same form as (23), we can go beyond the current literature on shift-register consensus (which assumes Q to be given and optimizes ζ) and provide jointly optimal weight matrix and ζ-parameter:

Proposition 6: The weight matrix $Q^\star$ and constant $\zeta^\star$ that minimizes the convergence factor of the shift-register consensus iterations (24) are $Q^\star = I - \theta^\star W^\star$, where $W^\star$ is computed in Proposition 3 with $H = I$ and $\theta^\star = \frac{2}{\lambda_2(W^\star)+\lambda_n(W^\star)}$ while $\zeta^\star = 1 + \beta^\star$ and $\beta^\star$ is given in Theorem 1.

C. Internet congestion control

Our final application is to the area of Internet congestion control, where Network Utility Maximization (NUM) has emerged as powerful framework for studying various important resource allocation problems, see, *e.g.*, [1], [32], [33], [34]. The vast majority of the work in this area is based on the dual decomposition approach introduced in [32]. Here, the optimal bandwidth sharing among S flows in a data network is posed as the optimizer of a convex optimization problem

$$\begin{array}{ll} \underset{x}{\text{maximize}} & \sum_s u_s(x_s) \\ \text{subject to} & x_s \in [m_s, M_s] \\ & Rx \leq c. \end{array} \tag{27}$$

Here, x_s is the communication rate of flow s, and $u_s(x_s)$ is a strictly concave and increasing function that describes the utility that source s has of communicating at rate x_s. The communication rate is subject to upper and lower bounds. Finally, $R \in \{0,1\}^{L\times S}$ is a routing matrix whose entries $R_{\ell s}$ are 1 if flow s traverses link ℓ and are 0 otherwise. In this way, Rx is the total traffic on links, that cannot exceed the link capacities $c \in \mathbf{R}^n$. We make the following assumptions.

Assumption 2: For the problem (27) it holds that

(i) Each $u_s(x_s)$ is twice continuously differentiable and satisfies $0 < l < -\nabla^2 u_s(x_s) < u$ for $x_s \in [m_s, M_s]$

(ii) For every link ℓ, there exists a source s whose flow only traverses ℓ, *i.e.* $R_{\ell s} = 1$ and $R_{\ell' s} = 0$ for all $\ell' \neq \ell$.

While these assumptions appear restrictive, they are often postulated in the literature (*e.g.* [32, Assumptions C1-C4]). Note that under Assumption 2, the routing matrix has full row-rank and all the link constraints hold with equality at optimum. Hence, we can replace $Rx \leq c$ in (27) with $Rx = c$. Following the steps of the dual ascent method in Section V, we have the following primal-dual iterations

$$x_s^\star(\mu) = \underset{z\in[m_s,M_s]}{\arg\max}\, u_s(z) - z\sum_\ell R_{\ell s}\mu_\ell \tag{28}$$

$$\mu_\ell(k+1) = \mu_\ell(k) + \alpha\left(\sum_\ell R_{\ell s}x_s^\star(\mu(k)) - c_\ell\right). \tag{29}$$

Note that each source solves a localized minimization problem based on the sum of the Lagrange multipliers for the links that the flow traverses; this information can be effectively signaled back to the source explicitly or implicitly using the end-to-end acknowledgements. The Lagrange multipliers, on the other hand, are updated by individual links based on the difference between the total traffic imposed by the sources and the capacity of link. Clearly, this information is also locally available. It is possible to show that under the conditions of Assumption 2, the dual function is strongly concave, differentiable and has a Lipschitz-continuous gradient [32]. Hence, by standard arguments, the updates (28), (29) converge to a primal-dual optimal point $(x^\star, \mu^\star)$ for appropriately chosen step-size α. Our results from Section V indicate that substantially improved convergence factors could be obtained by the following class of multi-step updates of the Lagrange multipliers

$$\begin{aligned}\mu_\ell(k+1) = \mu_\ell(k) + \alpha\left(\sum_\ell R_{\ell s}x_s^\star(\mu(k)) - c_\ell\right)\\ + \beta(\mu_\ell(k) - \mu_\ell(k-1)).\end{aligned} \tag{30}$$

To tune the step-sizes in an optimal way, we bring the techniques from Section V into action. To do so, we first bound the eigenvalues of $RR^\top$ using the following result:

Lemma 4: Let $R \in \{0,1\}^{L\times S}$ satisfy Assumption 2. Then

$$1 \leq \lambda_1(RR^\top), \quad \lambda_n(RR^\top) \leq \ell_{\max}s_{\max},$$

where $\ell_{\max} = \max_s \sum_\ell R_{\ell s}$ and $s_{\max} = \max_\ell \sum_s R_{\ell s}$.

The optimal step-size parameters and corresponding convergence factor now follow from Lemma 4 and Theorem 2:

Proposition 7: Consider the network utility maximization problem (27) under Assumption 2. Then, for $0 \leq \beta < 1$ and $0 < \alpha < 2(1+\beta)/(u\ell_{\max}s_{\max})$ the iterations (28) and (30) converge linearly to a primal-dual optimal pair. The step-sizes

$$\alpha = \left(\frac{2}{\sqrt{u\ell_{\max}s_{\max}} + \sqrt{l}}\right)^2, \ \beta = \left(\frac{\sqrt{u\ell_{\max}s_{\max}} - \sqrt{l}}{\sqrt{u\ell_{\max}s_{\max}} + \sqrt{l}}\right)^2,$$

ensure that the convergence factor of the dual iterates is

$$q_{\text{NUM}} = \frac{\sqrt{u\ell_{\max}s_{\max}} - \sqrt{l}}{\sqrt{u\ell_{\max}s_{\max}} + \sqrt{l}}.$$

Note that an upper bound of the Hessian of the dual function was also derived in [32]. However, strong concavity was not explored and the associated bounds were not derived.

To compare the gradient iterations with the multi-step congestion control mechanism, we present representative results from a network with 10 links and 20 flows which satisfies Assumption 2. The utility functions are on the form $-(M_s - x_s)^2/2$ and $m_s = 0$ and $M_s = 10^5$ for all sources. As shown in Fig. 5, substantial speedups are obtained.

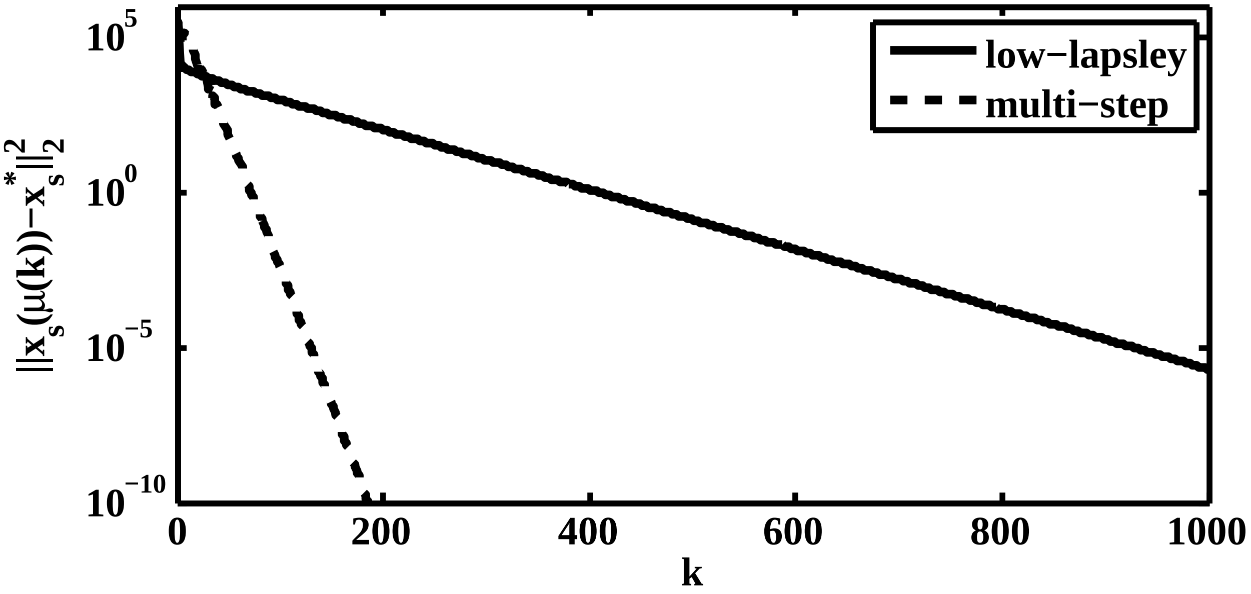


Fig. 5. Convergence of Low-Lapsley and multi-step. Plot shows log scale of the Euclidian distance from optimal source rates $\|x_s(\mu(k)) - x_s^\star\|_2^2$ vs. the iteration number k.

As a final remark, note that Lemma 4 underestimates λ_1 and overestimates λ_n, so we have no formal guarantee that the multi-step method will always outperform the gradient-based algorithm. However, in our experiments with a large number of randomly generated networks, the disadvantageous situation identified in Section VI never occurred.

VIII. Conclusions

We have studied accelerated gradient methods for network-constrained optimization problems. In particular, given the bounds of the Hessian of the objective function and the Laplacian of the underlying communication graph, we derived primal and dual multi-step techniques that allow improving the convergence factors significantly compared to the standard gradient-based techniques. We derived optimal parameters and convergence factors, and characterized the robustness of our tuning rules to errors that occur when critical problem parameters are not known but have to be estimated. Our multi-step techniques were applied to three classes of problems: distributed resource allocation under a network-wide resource constraint, distributed average consensus, and Internet congestion control. We demonstrated, both analytically and in numerical simulations, that the approaches developed in this paper often significantly outperform alternatives from the literature.

Appendix

A. *Proof of Theorem 1*

Let $x^\star$ be the optimizer of (1). The Taylor series expansion of $\nabla f\big(x(k)\big)$ around $x^\star$ yields

$$\begin{aligned} W\nabla f\big(x(k)\big) &\cong W(\nabla f(x^\star) + \nabla^2 f(x^\star)(x(k) - x^\star)) \\ &= W\nabla^2 f(x^\star)(x(k) - x^\star), \end{aligned}$$

since $W\nabla f(x^\star) = 0$ by (4) and (8). Introducing $z(k) \triangleq [x(k) - x^\star,\, x(k-1) - x^\star]^\top$, we can thus re-write (9) as

$$z(k+1) = \underbrace{\begin{bmatrix} B & -\beta I \\ I & 0 \end{bmatrix}}_{\Gamma} z(k) + o(z(k)^2), \tag{31}$$

where $B = (1+\beta)I - \alpha WH$ and $H = \nabla^2 f(x^\star)$. Now, for non-zero vectors v_1 and v_2, consider the eigenvalue equation

$$\begin{bmatrix} B & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda(\Gamma) \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

Since $v_1 = \lambda(\Gamma) v_2$, the first row can be re-written as

$$\left(-\lambda^2(\Gamma) I + \lambda(\Gamma) B - \beta I\right) v_2 = 0. \tag{32}$$

Note that (32) is a polynomial in B and B is in turn a polynomial in WH. Hence, if μ and λ denote the eigenvalues of B and WH, respectively, we have

$$\lambda^2(\Gamma) - (1+\beta-\alpha\lambda)\,\lambda(\Gamma) + \beta = 0. \tag{33}$$

The roots of (33) have the form

$$\lambda(\Gamma) = \frac{1+\beta-\alpha\lambda \pm \sqrt{\Delta}}{2}, \quad \Delta = (1+\beta-\alpha\lambda)^2 - 4\beta. \tag{34}$$

If $\Delta \geq 0$, then $|\lambda(\Gamma)| < 1$ is equivalent to

$$\begin{aligned} &(1+\beta-\alpha\lambda)^2 - 4\beta \geq 0 \\ &-2 < 1+\beta-\alpha\lambda \pm \sqrt{(1+\beta-\alpha\lambda)^2 - 4\beta} < 2, \end{aligned}$$

which, after simplifications, yield $0 < \alpha < 2(1+\beta)/\lambda$.

Furthermore, if $\Delta < 0$, then $|\lambda(\Gamma)| < 1$ is equivalent to

$$0 \leq \frac{(1+\beta-\alpha\lambda)^2 - \Delta}{4} < 1,$$

which, after similar simplifications, implies that $0 \leq \beta < 1$.

Note that the upper bound for α gives a necessary condition for λ. Here we find an upper bound for this eigenvalue. Since H is a positive diagonal matrix, under similarity equivalence we have $WH \sim H^{1/2}WHH^{-1/2} = H^{1/2}WH^{1/2}$. Without loss of generality assume $x \in \mathbf{R}^n$ and $x^\top x = 1$, Then $x^\top WHx = x^\top H^{1/2}WH^{1/2}x = y^\top Wy$, where $y = H^{1/2}x$. Clearly, for $y^\top Wy$ it holds that $\lambda_1(W)y^\top y \leq y^\top Wy \leq \lambda_n(W)y^\top y$. Now, $l \leq y^\top y = x^\top Hx \leq u$, implies $l\lambda_1(W) \leq x^\top WHx \leq u\lambda_n(W)$. and hence, a sufficient condition on α reads

$$0 < \alpha < \frac{2(1+\beta)}{u\lambda_n(W)}. \tag{35}$$

Having proven the sufficient conditions for convergence stated in the theorem, we now proceed to estimate the convergence factor. To this end, we need the following lemmas describing the eigenvalue characteristics of WH and Γ.

Lemma 5: If W has $m < n$ zero eigenvalues, then WH has exactly $n - m$ nonzero eigenvalues, i.e. $\lambda_1(WH) = \cdots = \lambda_m(WH) = 0$, $\lambda_i(WH) \neq 0 \quad i = m+1, \cdots, n$.

Proof: From [23] we know that if and only if all the principal sub-matrices of a matrix have nonnegative determinants then that matrix is positive semi-definite. Note that the i-th principal sub-matrix of WH, WH_i, is obtained by multiplication of the corresponding principal sub-matrix of W, W_i by the same principal sub-matrix of H, H_i from the right, and we have $\det(WH_i) = \det(W_i)\det(H_i)$. We know $\det(H_i) > 0$ and $\det(W_i) \geq 0$ because $W \geq 0$, thus $\det(WH_i) \geq 0$ and WH is positive semi-definite. Furthermore rank$(WH) =$ rank(W). So rank$(WH) = n - m$ and it means that WH has exactly m zero eigenvalues. ■

Lemma 6: For any WH such that $\lambda_i(WH) = 0$ for $i = 1, \cdots, m$, and $\lambda_i(WH) \neq 0$, for $i = m + 1, ..., n$., the matrix Γ has m eigenvalues equal to 1 and the absolute values of the rest of the $2n - m$ eigenvalues are strictly less than 1.

Proof: For complex $\lambda_i(\Gamma)$ we have $|\lambda_i(\Gamma)| = \beta < 1$. For real-valued $\lambda_i(\Gamma)$, on the other hand, the bound on α implies that $\alpha(\lambda(WH))$ is a decreasing function of λ. In this case, $0 < \alpha < \frac{2(1+\beta)}{\overline{\lambda}}$ guarantees that $0 < \alpha < \frac{2(1+\beta)}{\lambda_i(WH)}$ for any $0 < \lambda_i(WH) \leq \overline{\lambda}$. Note that if we set a tighter bound on α, then it does not change satisfactory condition for having $|\lambda(\Gamma)| < 1$. Only when $\lambda_i(WH) = 0$, we have $\lim_{x\to 0} \alpha = \infty$. For this case, if we substitute $\lambda_i(WH) = 0$ in (33) we obtain $\lambda_{2i-1}(\Gamma) = 1$ and $\lambda_{2i}(\Gamma) = \beta < 1$. ■

We are now ready to prove the remaining parts of Theorem 1. By the Lemmas above, Γ has $m < n$ eigenvalues equal to 1, which correspond to the m zero eigenvalues of W implied by the optimality condition (8). Hence, minimizing $m + 1$-th largest eigenvalue of (31) leads to the optimum convergence factor of the multi-step weighted gradient iterations (9). Calculating $\overline{\lambda}_\Gamma \triangleq \min_{\alpha,\beta} \max_{1\leq j\leq 2n-m} |\lambda_j(\Gamma)|$ yields the optimum $\alpha^\star$ and $\beta^\star$. Considering that (34) are the eigenvalues of Γ,

$$\overline{\lambda}_\Gamma = \frac{1}{2}\max\left\{|1+\beta-\alpha\lambda_i| + \sqrt{(1+\beta-\alpha\lambda_i)^2 - 4\beta}\right\},$$

where $\lambda_i \triangleq \lambda_i(WH)$, $\forall i = m + 1, .., n$. There are two cases:

Case 1: $(1 + \beta - \alpha\lambda_i)^2 - 4\beta \geq 0$. Then, a and b are non-negative and real with $a \geq b$. Hence, $a^2 - b^2 \geq (a - b)^2$ and consequently $a + \sqrt{a^2 - b^2} \geq 2a - b \geq b$.

Case 2: $(1+\beta-\alpha\lambda_i)^2-4\beta < 0$. In this case, $\lambda_i(\Gamma)$ is complex-valued. Consider $c, d \in \mathbf{R}^+$ with $c < d$. Then, $|c+\sqrt{c^2 - d}| = \sqrt{c^2 - c^2 + d} = \sqrt{d} \geq 2c - \sqrt{d}$.

If we substitute these results into $\overline{\lambda}_\Gamma$ with $a = 1 + \beta - \alpha\lambda_i$, $b = 2\sqrt{\beta}$, $c = |1 + \beta - \alpha\lambda_i|$ and $d = 4\beta$ we get

$$\overline{\lambda}_\Gamma \geq \max\left\{\sqrt{\beta}, \max\left\{|1+\beta-\alpha\lambda_i| - \sqrt{\beta}\right\}\right\},$$

which can be expressed in terms of $\underline{\lambda}$ and $\overline{\lambda}$:

$$\overline{\lambda}_\Gamma \geq \max\left\{\sqrt{\beta}, |1+\beta-\alpha\underline{\lambda}| - \sqrt{\beta}, |1+\beta-\alpha\overline{\lambda}| - \sqrt{\beta}\right\}. \tag{36}$$

It can be verified that

$$\begin{aligned}\max\quad &\left\{|1+\beta-\alpha\underline{\lambda}| - \sqrt{\beta}, |1+\beta-\alpha\overline{\lambda}| - \sqrt{\beta}\right\}\\ &\geq |1+\beta-\alpha'\underline{\lambda}| - \sqrt{\beta},\end{aligned} \tag{37}$$

where α' is such that $|1 + \beta - \alpha'\underline{\lambda}| = \left|1 + \beta - \alpha'\overline{\lambda}\right|$, *i.e.*

$$\alpha' = \frac{2(1+\beta)}{\underline{\lambda} + \overline{\lambda}}. \tag{38}$$

From (36), (37) and (38), we thus obtain

$$\overline{\lambda}_\Gamma \geq \max\left\{\sqrt{\beta}, (1+\beta)\frac{\overline{\lambda}-\underline{\lambda}}{\overline{\lambda}+\underline{\lambda}} - \sqrt{\beta}\right\}. \tag{39}$$

Again, the max-operator can be bounded from below by its value at the point where the arguments are equal. To this end, consider β' which satisfies $\sqrt{\beta'} = (1+\beta')\frac{\overline{\lambda}-\underline{\lambda}}{\overline{\lambda}+\underline{\lambda}} - \sqrt{\beta'}$, that is,

$$\beta' = \left(\frac{\sqrt{\overline{\lambda}}-\sqrt{\underline{\lambda}}}{\sqrt{\overline{\lambda}}+\sqrt{\underline{\lambda}}}\right)^2. \tag{40}$$

Since $\max\left\{\sqrt{\beta}, (1+\beta)\frac{\overline{\lambda}-\underline{\lambda}}{\overline{\lambda}+\underline{\lambda}} - \sqrt{\beta}\right\} \geq \sqrt{\beta'}$, we can combine this with (39) to conclude that

$$\overline{\lambda}_\Gamma \geq \sqrt{\beta'} = \frac{\sqrt{\overline{\lambda}}-\sqrt{\underline{\lambda}}}{\sqrt{\overline{\lambda}}+\sqrt{\underline{\lambda}}}. \tag{41}$$

Our proof is concluded by noting that equality in (41) is attained for the smallest non-zero eigenvalue of Γ and the optimal step-sizes $\beta^\star$ and $\alpha^\star$ stated in the body of the theorem.

B. *Proof of Proposition 1*

As shown in the proof of Theorem 1, the eigenvalues of WH are equal to those of $H^{1/2}WH^{1/2}$. According to [23, p.225] for matrices W and $H^{1/2}WH^{1/2}$, there exists a nonnegative real number θ_k such that $\lambda_1(H) \leq \theta_k \leq \lambda_n(H)$ and $\lambda_k(H^{1/2}WH^{1/2}) = \theta_k\lambda_k(W)$. Letting $k = m+1$ and $k = n$, yields $\underline{\lambda} \geq l\underline{\lambda}_W$ and $\overline{\lambda} \leq u\overline{\lambda}_W$. The rest of the proof is similar to that of Theorem 1 and is omitted for brevity.

C. *Proof of Proposition 2*

Direct calculations yield $q^\star = (\sqrt{\overline{\lambda}}-\sqrt{\underline{\lambda}})/(\sqrt{\overline{\lambda}}+\sqrt{\underline{\lambda}}) = 1-2/((\overline{\lambda}/\underline{\lambda})^{1/2}+1)$. Similarly, $\widetilde{q} = 1-2/((\overline{\lambda}_W/\underline{\lambda}_W)^{1/2}+1)$. Hence, minimizing $q^\star$ and $\widetilde{q}$ are equivalent to minimizing the condition number of WH and W, respectively.

D. *Proof of Proposition 3*

As shown in the proof of Theorem 1, the eigenvalues of WH are equal to those of $\Omega \triangleq H^{1/2}WH^{1/2}$. Thus, combined with the constraint that $W \geq 0$ the problem of minimizing $\overline{\lambda}/\underline{\lambda}$ is equivalent to minimizing u/l, where u and l are the largest and the smallest non-zero eigenvalues of Ω. Next we will construct the constraint set of this optimization problem. First, recall that W should provide the sparsity pattern induced by $\mathcal{G}$, i.e., $W \in \mathcal{M}_\mathcal{G}$. Second, W fulfills (8). For the case that W is symmetric, this constraint can be rewritten in terms of Ω in the form $\Omega V = H^{1/2}WH^{1/2}V = 0$ where $V = H^{-1/2}A^\top$. Third constraint is to bound the remaining $n-m$ eigenvalues of Ω away from zero. Let $v \in \mathbf{R}^n$ be a column of V, and let $v^\perp \in \mathbf{R}^n$ be orthogonal to v. Since $\Omega \geq 0$ and $\Omega v = 0$ then we have $\quad x^\top\Omega x > 0 \quad \forall x \in v^\perp$. This condition is equivalent to $P^\top\Omega P > 0$, where $P = [p_1, p_2, ..., p_{n-m}] \in \mathbf{R}^{n\times n-m}$ is a matrix of orthonormal vectors spanning the null space of $V^\top$. More explicitly, one can define this subspace by unit vectors satisfying $p_i^\top v_j = 0,\ \forall i = 1,\dots,n-m,\ j = 1,\dots,n, \quad p_i^\top p_k = 0,\ \forall i \neq k$. The optimization problem becomes

$$\begin{array}{ll} \text{minimize} & u/l \\ \text{subject to} & lI \leq P^\top\Omega P \leq uI, W \in \mathcal{M}_\mathcal{G}, W \geq 0, \Omega V = 0. \end{array}$$

Denoting $t = u/l$ and $\gamma = 1/l$, this problem can be recast as

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & I \leq \gamma P^\top\Omega P \leq tI, W \in \mathcal{M}_\mathcal{G}, \\ & W \geq 0, \Omega V = 0, \gamma > 0. \end{array} \tag{42}$$

Finally, the constraint on γ in (42) can be omitted. Specifically, consider $(W^\star, \alpha^\star, \beta^\star)$ to be the joint-optimal weight and stepsizes given by (42) and (10), respectively. One can replace any positive scale $\gamma W^\star$ in (10) and derive the step-sizes $\alpha = \alpha^\star/\gamma$ and $\beta = \beta^\star$. It is easy to check that the triple $(\gamma W^\star, \alpha, \beta)$ leads to an identical multi-step iterations (9) as the optimal ones.

E. Proof of Lemma 1

To prove (a) we exploit the equivalence of l-strong convexity of $f(\cdot)$ and $1/l$-Lipschitz continuity of $\nabla f_\star$. Specially according to [35, Theorem 4.2.1], for nonzero $z_1, z_2 \in \mathbf{R}^n$, Lipschitz continuity of $\nabla f_\star$ implies that

$$\langle \nabla f_\star(z_1) - \nabla f_\star(z_2), z_1 - z_2 \rangle \leq \frac{1}{l} \|z_1 - z_2\|^2.$$

Now, for $-\nabla d(z) = -A \nabla f_\star(-A^\top z) + b$, change the right hand side of the former inequality to have

$$\begin{aligned} &\langle -\nabla d(z_1) + \nabla d(z_2), z_1 - z_2 \rangle \\ &= \langle \nabla f_\star(-A^\top z_1) - \nabla f_\star(-A^\top z_2), -A^\top (z_1 - z_2) \rangle. \end{aligned}$$

In light of $1/l$-Lipschitzness of $\nabla f^\star$ we get

$$\begin{aligned} &\langle \nabla f_\star(-A^\top z_1) - \nabla f_\star(-A^\top z_2), -A^\top (z_1 - z_2) \rangle \\ &\leq \frac{1}{l} \| - A^\top (z_1 - z_2)\|^2 \leq \frac{\lambda_n(AA^\top)}{l} \|z_1 - z_2\|^2. \end{aligned}$$

(b) According to [35, Theorem 4.2.2], If $\nabla f(\cdot)$ is u-Lipschitz continuous then $f_\star$ is $1/u$-strongly convex, i.e., for non-identical $z_1, z_2 \in \mathbf{R}^n$ we have $\langle \nabla f_\star(z_1) - \nabla f_\star(z_2), z_1 - z_2 \rangle \geq \frac{1}{u} \|z_1 - z_2\|^2$. One can manipulate above inequality as

$$\begin{aligned} &\langle -\nabla d(z_1) + \nabla d(z_2), z_1 - z_2 \rangle \\ &= \langle \nabla f_\star(-A^\top z_1) - \nabla f_\star(-A^\top z_2), -A^\top (z_1 - z_2) \rangle \\ &\geq \frac{1}{u} \| - A^\top (z_1 - z_2)\|^2 \geq \frac{\lambda_1(AA^\top)}{u} \|z_1 - z_2\|^2. \end{aligned}$$

It is worth noting that here we assume that A is row full rank.

F. Proof of Theorem 2

The result follows from Lemma 1 and Theorem 1 with $W = I$ and noting that $(\lambda_1(AA^\top)/u)I \leq H \leq (\lambda_n(AA^\top)/l)I$.

G. Proof of Lemma 2

Since f is twice differentiable on $[x^\star, x]$, we have

$$\begin{aligned} \nabla f(x) &= \nabla f(x^\star) + \int_0^1 \nabla^2 f(x^\star + \tau(x - x^\star))(x - x^\star) d\tau \\ &= A^\top \mu^\star + H(x)(x - x^\star), \end{aligned}$$

where we have used the fact that $\nabla f(x^\star) = A^\top \mu^\star$ and introduced $H(x) = \int_0^1 \nabla^2 f(x^\star + \tau(x - x^\star))d\tau$. By virtue of Assumption 1, $H(x)$ is symmetric and nonnegative definite and satisfies $lI \leq H(x) \leq uI$ [13] . Hence from (7) and (8)

$$\begin{aligned}
&\|x(k+1) - x^\star\| = \|x(k) - x^\star - \alpha W \nabla f\big(x(k)\big)\| \\
&= \|x(k) - x^\star - \alpha W(A^\top \mu^\star + H(x(k))(x(k) - x^\star))\| \\
&= \|(I - \alpha W H(x(k)))(x(k) - x^\star)\| \\
&\leq \|I - \alpha W H(x(k))\| \|x(k) - x^\star\|.
\end{aligned}$$

The rest of the proof follows the same steps as [13, Theorem 3]. Essentially for fixed step-size $0 < \alpha < 2/\overline{\lambda}$, (7) converge linearly with factor $q_2 = \max\{|1 - \alpha\underline{\lambda}|, |1 - \alpha\overline{\lambda}|\}$. The minimum convergence factor $q_G^\star = \frac{\overline{\lambda}-\underline{\lambda}}{\underline{\lambda}+\overline{\lambda}}$ is obtained by minimizing q_G over α, which yields the optimal step-size $\alpha^\star = \frac{2}{\underline{\lambda}+\overline{\lambda}}$.

H. *Proof of Proposition 4*

According to Lemma 2, the weighted gradient iterations (7) with estimated step-size $\widetilde{\alpha} = 2/(\utilde{\lambda} + \widetilde{\lambda})$ will converge provided that $0 < \widetilde{\alpha} < 2/\overline{\lambda}$, *i.e.* when $\overline{\lambda} < \utilde{\lambda} + \widetilde{\lambda}$. For the multi-step algorithm (9), Theorem 1 guarantees convergence if $0 \leq \widetilde{\beta} < 1$, $0 < \widetilde{\alpha} < 2(1 + \widetilde{\beta})/\overline{\lambda}$. The assumption $0 < \utilde{\lambda} \leq \widetilde{\lambda}$ implies that the condition on $\widetilde{\beta}$ is always satisfied. Regarding $\widetilde{\alpha}$, inserting the expression for $\widetilde{\beta}$ in the upper bound for $\widetilde{\alpha}$ and simplifying yields

$$\frac{4}{\left(\sqrt{\utilde{\lambda}} + \sqrt{\widetilde{\lambda}}\right)^2} < 2\frac{2(\widetilde{\lambda} + \utilde{\lambda})}{\left(\sqrt{\widetilde{\lambda}} + \sqrt{\utilde{\lambda}}\right)^2}\frac{1}{\overline{\lambda}},$$

which is satisfied if $0 < \overline{\lambda} < \widetilde{\lambda} + \utilde{\lambda}$. The statement is proven.

I. *Proof of Lemma 3*

We consider two cases. First, when $\utilde{\lambda} + \widetilde{\lambda} < \underline{\lambda} + \overline{\lambda}$ combined with the assumption that $0 < \overline{\lambda} < \utilde{\lambda} + \widetilde{\lambda}$ yields $\widetilde{\alpha}\overline{\lambda} > 1$, which means that $|1-\widetilde{\alpha}\overline{\lambda}| = \widetilde{\alpha}\overline{\lambda}-1$. Moreover, $\widetilde{\alpha}\overline{\lambda}-1 \geq 1-\widetilde{\alpha}\underline{\lambda}$, so by Lemma 2 $\widetilde{q}_G = \max\{\widetilde{\alpha}\overline{\lambda} - 1, \max\{1 - \widetilde{\alpha}\underline{\lambda}, \widetilde{\alpha}\underline{\lambda} - 1\}\} = \widetilde{\alpha}\overline{\lambda} - 1$ $= 2\overline{\lambda}/(\utilde{\lambda} + \widetilde{\lambda}) - 1$.

The second case is when $\utilde{\lambda} + \widetilde{\lambda} > \underline{\lambda} + \overline{\lambda}$. Then, $\widetilde{\alpha}\underline{\lambda} < 1$ and hence $|1 - \widetilde{\alpha}\underline{\lambda}| = 1 - \widetilde{\alpha}\underline{\lambda}$. Moreover, $1 - \widetilde{\alpha}\underline{\lambda} \geq \widetilde{\alpha}\overline{\lambda} - 1$, so

$$\widetilde{q}_G = \max\{1 - \widetilde{\alpha}\underline{\lambda}, \max\{\widetilde{\alpha}\overline{\lambda} - 1, 1 - \widetilde{\alpha}\overline{\lambda}\}\} = 1 - \widetilde{\alpha}\underline{\lambda}.$$

The convergence factor of the multi-step iterations with inaccurate step-sizes (16) follows directly from Theorem 1.

J. *Proof of Proposition 5*

We analyze the four quadrants $\mathcal{Q}_1$ through $\mathcal{Q}_4$ in order.

$\mathcal{Q}_1$: when $(\utilde{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_1$ we have $\utilde{\lambda} > \underline{\lambda}$ and $\widetilde{\lambda} > \overline{\lambda} > \overline{\lambda}$. From the convergence factor of multi-step gradient method (18) it then follows that $\widetilde{q} = 1 + \widetilde{\beta} - \widetilde{\alpha}\underline{\lambda} - \widetilde{\beta}^{1/2}$. Moreover, since in this quadrant $\widetilde{\lambda}+\utilde{\lambda} \geq \overline{\lambda}+\underline{\lambda}$, from (17) we have $\widetilde{q}_G = 1-2\underline{\lambda}/(\utilde{\lambda}+\widetilde{\lambda})$. A direct comparison between the two expressions yields that $\widetilde{q} \leq \widetilde{q}_G$.

$\mathcal{Q}_2$: when $(\utilde{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_2$ we have $\underline{\lambda} < \utilde{\lambda}$ and $\widetilde{\lambda} < \overline{\lambda}$. Combined with the stability assumption $\utilde{\lambda} + \widetilde{\lambda} > \overline{\lambda}$, straightforward calculations show that the convergence factor of the multi-step iterations with inaccurate step-sizes (16) is

$$\widetilde{q} = \begin{cases} \widetilde{\alpha}\overline{\lambda} - \widetilde{\beta} - 1 - \sqrt{\widetilde{\beta}} & \widetilde{\lambda} + \widetilde{\lambda} \leq \underline{\lambda} + \overline{\lambda}, \\ 1 + \widetilde{\beta} - \widetilde{\alpha}\underline{\lambda} - \sqrt{\widetilde{\beta}} & \text{otherwise}, \end{cases}$$

Moreover, for this quadrant the convergence factor of weighted gradient method is given by (17). To verify that $\widetilde{q} < \widetilde{q}_G$ we perform the following comparisons:

(a) If $\utilde{\lambda} + \widetilde{\lambda} < \underline{\lambda} + \overline{\lambda}$ then we have $\widetilde{q} = \widetilde{\alpha}\overline{\lambda} - \widetilde{\beta} - 1 - \widetilde{\beta}^{1/2}$ and $\widetilde{q}_G = (2\overline{\lambda})/(\utilde{\lambda} + \widetilde{\lambda}) - 1$. To show that $\widetilde{q} < \widetilde{q}_G$ we rearrange it to obtain the following inequality

$$\Delta \triangleq (\overline{\lambda} - \widetilde{\lambda} + \widetilde{\lambda}^{1/2}\utilde{\lambda}^{1/2})(\widetilde{\lambda} + \utilde{\lambda}) - 2\overline{\lambda}\widetilde{\lambda}^{1/2}\utilde{\lambda}^{1/2} < 0.$$

After simplifications

$$\Delta = (\widetilde{\lambda}^{1/2} - \utilde{\lambda}^{1/2})\Big(-\widetilde{\lambda}^{1/2}(\widetilde{\lambda} + \utilde{\lambda} - \overline{\lambda}) - \utilde{\lambda}^{1/2}\overline{\lambda}\Big) < 0.$$

Note that the negativity of above quantity comes from the stability condition, $\widetilde{\lambda} + \utilde{\lambda} > \overline{\lambda}$.

(b) If $\utilde{\lambda} + \widetilde{\lambda} > \underline{\lambda} + \overline{\lambda}$ then we have $\widetilde{q} = 1 + \widetilde{\beta} - \widetilde{\alpha}\underline{\lambda} - (\widetilde{\beta})^{1/2}$ and $\widetilde{q}_G = 1 - (2\underline{\lambda})/(\utilde{\lambda} + \widetilde{\lambda})$. After some simplifications, we see that $\widetilde{q} < \widetilde{q}_G$ boils down to the inequality $-(\utilde{\lambda} + \widetilde{\lambda})\utilde{\lambda}^{1/2}\widetilde{\lambda}^{1/2} + 2\underline{\lambda}\utilde{\lambda}^{1/2}\widetilde{\lambda}^{1/2} - \underline{\lambda}(\utilde{\lambda} + \widetilde{\lambda}) < 0$ or equivalently $-(\utilde{\lambda} + \widetilde{\lambda} - 2\underline{\lambda})\utilde{\lambda}^{1/2}\widetilde{\lambda}^{1/2} - \underline{\lambda}(\utilde{\lambda} + \widetilde{\lambda}) < 0$ which holds by noting that $\utilde{\lambda} + \widetilde{\lambda} > \underline{\lambda} + \overline{\lambda} > 2\underline{\lambda}$.

(c) for the case $\utilde{\lambda} + \widetilde{\lambda} = \underline{\lambda} + \overline{\lambda}$, we have $\widetilde{q} = 1 + \widetilde{\beta} - \widetilde{\alpha}\underline{\lambda} - (\widetilde{\beta})^{1/2}$ and $\widetilde{q}_G = (\overline{\lambda} - \underline{\lambda})/(\underline{\lambda} + \overline{\lambda})$ which coincides with the optimal convergence factor of unperturbed gradient method. After some rearrangements we notice that $\widetilde{q} < \widetilde{q}_G$ reduces to checking that $(\widetilde{\lambda}^{1/2} - \utilde{\lambda}^{1/2})(\overline{\lambda} - \underline{\lambda}) < (\utilde{\lambda}^{1/2} + \widetilde{\lambda}^{1/2})(\utilde{\lambda} + \widetilde{\lambda})$ that holds since $\widetilde{\lambda}^{1/2} - \utilde{\lambda}^{1/2} < \utilde{\lambda}^{1/2} + \widetilde{\lambda}^{1/2}$ and $\overline{\lambda} - \underline{\lambda} < \overline{\lambda} + \underline{\lambda} = \utilde{\lambda} + \widetilde{\lambda}$.

$\mathcal{Q}_3$: if $(\utilde{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_3$ we have $0 < \utilde{\lambda} < \underline{\lambda}$ and $\widetilde{\lambda} < \overline{\lambda}$. Combined with the stability assumption $\widetilde{\lambda} + \utilde{\lambda} > \overline{\lambda}$, one can verify that the convergence factors of the two perturbed iterations are $\widetilde{q}_G = (2\overline{\lambda})/(\utilde{\lambda} + \widetilde{\lambda}) - 1$ and $\widetilde{q} = \widetilde{\alpha}\overline{\lambda} - \widetilde{\beta} - 1 - (\widetilde{\beta})^{1/2}$, respectively. The fact that $\widetilde{q} < \widetilde{q}_G$ was proven in step (a) of the analysis of $\mathcal{Q}_2$.

$\mathcal{Q}_4$: if $(\utilde{\varepsilon}, \widetilde{\varepsilon}) \in \mathcal{Q}_4$ then, (18) implies that $\widetilde{q} = \widetilde{\beta}^{1/2}$. On the other hand, for this region (17) yields $\widetilde{q}_G = (\overline{\lambda} - \underline{\lambda})/(\underline{\lambda} + \overline{\lambda})$. To conclude, we need to verify that there exists $\widetilde{\lambda}$ and $\utilde{\lambda}$ such that $\widetilde{q} > \widetilde{q}_G$, *i.e.* $(\widetilde{\lambda}^{1/2} - \utilde{\lambda}^{1/2})/(\widetilde{\lambda}^{1/2} + \utilde{\lambda}^{1/2}) > (\overline{\lambda} - \underline{\lambda})/(\underline{\lambda} + \overline{\lambda})$. We do so by multiplying both sides with $(\underline{\lambda} + \overline{\lambda})(\widetilde{\lambda}^{1/2} + \utilde{\lambda}^{1/2})$ and simplifying to find that the inequality holds if $\underline{\lambda}\widetilde{\lambda}^{1/2} > \overline{\lambda}\utilde{\lambda}^{1/2}$, or equivalently $\widetilde{\lambda}/\utilde{\lambda} > \overline{\lambda}^2/\underline{\lambda}^2$. The statement is proven.

K. *Proof of Proposition 6*

The iterations (23) and (24) are equivalent when

$$(1 - \zeta) = -\beta, \qquad (1 + \beta)I - \alpha W = \zeta Q.$$

The first condition implies that $\zeta^\star = (1 + \beta^\star)$. Combining this expression with the second condition, we find

$$Q^\star = I - \frac{\alpha^\star}{1 + \beta^\star}W^\star = I - \frac{2}{\underline{\lambda} + \overline{\lambda}}W^\star.$$

Noting that for the consensus case, $\underline{\lambda} = \lambda_2(W^\star)$ and $\overline{\lambda} = \lambda_n(W^\star)$ concludes the proof.

L. *Proof of Lemma 4*

For the upper bound on $\lambda_n(RR^\top)$, we use a similar approach as [32, Lemma 3]. Specially, from [23, p.313],

$$\lambda_n^2(RR^\top) = \|RR^\top\|_2^2 \leq \|RR^\top\|_\infty \|RR^\top\|_1 = \|RR^\top\|_\infty^2.$$

Hence,

$$\begin{aligned}\lambda_n(RR^\top) &= \max_\ell \sum_{\ell'} [RR^\top]_{\ell\ell'} = \max_\ell \sum_{\ell'} \sum_s R_{\ell s} R_{\ell' s} \\ &\le \max_\ell \sum_s R_{\ell s} \ell_{\max} \le s_{\max} \ell_{\max}.\end{aligned}$$

To find a lower bound on $\lambda_1(RR^\top)$ we consider the definition $\lambda_1(RR^\top) = \min_{\|x\|_2=1} \|R^\top x\|_2^2$. We have

$$[R^\top x]_s = \sum_{\ell=1}^{L} R^\top_{s\ell} x_\ell = \sum_{\ell=1}^{L} R_{\ell s} x_\ell.$$

According to Assumption 2, $R^\top$ has L independent rows that have only one non-zero (equal to 1) component. Hence,

$$\begin{aligned}\|R^\top x\|_2^2 &= \sum_{s=1}^{L} x_s^2 + \sum_{s=S-L+1}^{S} \left(\sum_{\ell=1}^{L} R_{\ell s} x_\ell \right)^2 \\ &= 1 + \sum_{s=S-L+1}^{n} \left(\sum_{\ell=1}^{L} R_{\ell s} x_\ell \right)^2 \ge 1,\end{aligned}$$

where the last equality is due to $\|x\|_2 = 1$.

References

[1] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[2] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95 Issue: 1, pp. 215–233, 2007.

[3] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: nonlinear observation models and imperfect communication," *IEEE Trans. on Information Theory*, 2012.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3 Issue: 1, pp. 1–122, 2011.

[5] D. Goodman and N. Mandayam, "Power control for wireless data," *Personal Communications, IEEE*, vol. 7 Issue:2, pp. 48–54, 2000.

[6] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *Automatic Control, IEEE Transactions on*, vol. 31 Issue: 9, pp. 803–812, 1986.

[7] M. Cao, D. A. Spielman, and E. M. Yeh, "Accelerated gossip algorithms for distributed computation," in *44th Annual Allerton Conference on Communication, Control, and Computation*, 2006, pp. 952–959.

[8] B. Johansson, "On distributed optimization in networked systems," Ph.D. dissertation, Royal Institute of Technology, 2008.

[9] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

[10] D. Bertsekas., *Nonlinear Programming*. Athena Scientific, 1999.

[11] F. Zanella, D. Varagnolo, A. Cenedese, G. Pillonetto, and L. Schenato, "Newton-raphson consensus for distributed convex optimization," in *IEEE Conference on Decision and Control (CDC)*, 2011.

[12] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed newton method for network utility maximization, i: Algorithm," *LIDS report 2832*, 2011.

[13] B. Polyak, *Introduction to Optimization*. ISBN 0-911575-14-6, 1987.

[14] S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley, "Notes on decomposition methods," Stanford University, Tech. Rep., 2008.

[15] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, 2005.

[16] O. Devolder, F. Glineur, and Y. Nesterov, "A double smoothing technique for constrained convex optimization problems and applications to optimal control," *submitted to SIAM Journal on Optimization*, 2011.

[17] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *J. Opt. Theory and Applications*, vol. 129 Issue:3, pp. 469–488, 2006.

[18] Y. C. Ho, L. Servi, and R. Suri, "A class of center-free resource allocation algorithms," *Large Scale Systems*, vol. 1, pp. 51–62, 1980.

[19] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation:Numerical Methods*. New York: Athena Scientific, 1997.

[20] P. Maréchal and J. Ye, "Optimizing condition numbers," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 935–947, 2009.

[21] Z. Lu and T. K. Pong, "Minimizing condition number via convex programming," *SIAM J. Matrix Anal. Appl.*, vol. 32, pp. 1193–1211, 2011.

[22] L. Vandenberghe, "Course notes for optimization methods for large-scale systems, ee236c, dual decomposition chapter," 2012.

[23] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, 1985.

[24] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *Automatic Control, IEEE Transactions on*, vol. 48, pp. 988 – 1001, 2003.

[25] B. Oreshkin, M. Coates, and M. Rabbat, "Optimization and analysis of distributed averaging with short node memory," *Signal Processing, IEEE Transactions on*, vol. 58 Issue: 5, pp. 2850 –2865, 2010.

[26] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53 Issue: 1, pp. 65–78, 2004.

[27] E. Ghadimi, M. Johansson, and I. Shames, "Accelerated gradient methods for networked optimization," in *American Control Conference (ACC), 2011*. IEEE, 2011, pp. 1668–1673.

[28] D. M. Young, "Second-degree iterative methods for the solution of large linear systems," *Journal of Approximation Theory*, 1972.

[29] J. Liu, B. D. O. Anderson, M. Cao, and A. S. Morse, "Analysis of accelerated gossip algorithms," in *48th IEEE Conference on Decision and Control (CDC)*, 2009.

[30] G. H. Golub and R. S. Varga, "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order richardson iterative methods," *Numerische Matematik*, vol. 3, pp. 147–156, 1961.

[31] F. R. K. Chung, *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, No. 92, American Mathematical Society, 1997.

[32] S. Low and D. Lapsley, "Optimization flow control - i: Basic algorithm and convergence." *IEEE/ACM Transactions on Networking*, vol. 7 Issue: 6, pp. 861–874, 1999.

[33] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Transactions on Communications*, vol. 52 Issue: 7, pp. 1136–1144, 2004.

[34] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95 Issue:1, pp. 255 – 312, 2007.

[35] J. B. H. Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II*. Springer, 1996.