Robust Convergence Analysis of Distributed Optimization Algorithms

Akhil Sundararajan[1] Bin Hu[2] Laurent Lessard[1,2]

Abstract

We present a unified framework for analyzing the convergence of distributed optimization algorithms by formulating a semidefinite program (SDP) which can be efficiently solved to bound the linear rate of convergence. Two different SDP formulations are considered. First, we formulate an SDP that depends explicitly on the gossip matrix of the network graph. This result provides bounds that depend explicitly on the graph topology, but the SDP dimension scales with the size of the graph. Second, we formulate an SDP that depends implicitly on the gossip matrix via its spectral gap. This result provides coarser bounds, but yields a small SDP that is independent of graph size. Our approach improves upon existing bounds for the algorithms we analyzed, and numerical simulations reveal that our bounds are likely tight. The efficient and automated nature of our analysis makes it a powerful tool for algorithm selection and tuning, and for the discovery of new algorithms as well.

1 Introduction

Consider n agents located at the nodes of an undirected graph. Each agent $i \in \{1,\dots,n\}$ has access to a local function $f_i : \mathbb{R}^d \to \mathbb{R}$ and local memory x_i. The objective is for each agent's local memory to eventually converge to $x^\star$, the minimizer of the average of the functions:

$$x^\star := \operatorname*{arg\,min}_{x\in\mathbb{R}^d} f(x), \quad \text{where } f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x). \tag{1}$$

In this paper, we assume each f_i is strongly convex with parameter m_i and has Lipschitz-continuous gradients with parameter L_i. Agents may perform computations involving their local function f_i and local memory x_i, and may exchange information with neighboring agents.

The abstraction above captures a variety of problems that require distributed computation, such as multi-agent coordination and distributed estimation and learning [1,2,5,9]. A simple algorithm is *distributed gradient descent* [6], in which agent i uses the update rule:

$$x_i^{k+1} = \sum_{i=1}^n w_{ij}x_j^k - \eta\nabla f_i(x_i^k). \tag{2}$$

Here, x_i^0 is arbitrary and $\{w_{ij}\}$ is a *gossip matrix*. That is, $\{w_{ij}\}$ is symmetric and doubly stochastic. Moreover, $w_{ij} > 0$ if and only if there is an edge connecting i and j or if $i = j$. The iterations (2) combine gradient descent on each f_i with diffusion (consensus) on the x_i. In general, this algorithm requires a diminishing stepsize η in order to converge to $x^\star$ and convergence happens at a sublinear rate even when the f_i are strongly convex. The intuition behind this fact is that the optimal point $x^\star$ is not necessarily a minimizer of the individual f_i so the agents find themselves taking counterproductive steps.

Since pure consensus achieves linear convergence [12] and so does centralized gradient descent for strongly convex functions [7,8], one would expect that a combination of consensus and gradient descent could achieve a linear rate as well. Recent efforts have focused on devising such linear-rate algorithms [4, 10, 11]. In the EXTRA algorithm [11], for example, the update equations are

$$x_i^{k+2} = x_i^{k+1} + \sum_{j=1}^n w_{ij}x_j^{k+1} - \frac{1}{2}x_i^k - \frac{1}{2}\sum_{j=1}^n v_{ij}x_j^k - \eta\left(\nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k)\right), \tag{3}$$

where $\{w_{ij}\}$ and $\{v_{ij}\}$ are gossip matrices, x_i^0 is arbitrary, and $x_i^1 = \sum_{j=1}^n w_{ij}x_j^0 - \eta\nabla f_i(x_i^0)$. The update equations for EXTRA (3) are considerably more complicated than those of distributed gradient descent (2). Consequently, algorithm analysis can be problematic. Indeed, the works [4, 10, 11] each propose algorithms and prove the existence of a worst-case linear rate, but the proofs are either non-constructive or yield conservative bounds.

Main contributions. In this paper, we present two analysis approaches for certifying worst-case linear rates for distributed optimization algorithms. In both cases, the analysis reduces to determining the feasibility of a semidefinite program (SDP) and can be carried out efficiently via computational means.

The paper is organized as follows. In Section 2, we cover notation and prove a key lemma that forms the basis for subsequent results. In Sections 3 and 4, we present

[1]A. Sundararajan and L. Lessard are with the Department of Electrical and Computer Engineering at the University of Wisconsin–Madison, Madison, WI 53706, USA.

[2]B. Hu and L. Lessard are with the Wisconsin Institute for Discovery, which is also at the University of Wisconsin–Madison. {asundararaja,bhu38,laurent.lessard}@wisc.edu

[3]This material is based upon work supported by the National Science Foundation under Grant No. 1656951.

our two SDPs and use the EXTRA algorithm [11] to illustrate the methodology. In Section 5, we show how our approach extends to other recently proposed distributed optimization algorithms.

2 Preliminaries

This section presents a key lemma we utilize to prove linear rates of convergence for distributed algorithms. First, we cover some notation.

Notation. The number of agents in the network is denoted by n, and 1_n is the n-dimensional all-ones vector. The $r \times r$ identity matrix is represented by I_r, with subscript omitted for the $n \times n$ identity. The $p \times q$ zeros matrix is $0_{p\times q}$. We refer to the class of m-strongly convex functions with L-Lipschitz gradients as $\mathcal{F}(m, L)$. The domain of f has dimension d. The i^{th} column of the identity matrix is e_i. The Kronecker product between two matrices A and B is denoted by $A \otimes B$. The P-norm of a vector x is $\|x\|_P := (x^\mathsf{T} P x)^{1/2}$. We now state a useful quadratic inequality for strongly convex functions with Lipschitz gradients.

Proposition 1. *Suppose $f \in \mathcal{F}(m, L)$, $u^k = \nabla f(y^k)$, and $u^\star = \nabla f(y^\star)$. Then the following inequality holds.*

$$\begin{bmatrix} y^k - y^\star \\ u^k - u^\star \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} \otimes I_d \right) \begin{bmatrix} y^k - y^\star \\ u^k - u^\star \end{bmatrix} \geq 0.$$

Proof. This follows from co-coercivity and Lipschitz property of the gradient. See for example [3, 7]. ∎

The following lemma shows that the state of a discrete-time linear dynamical system converges linearly (is exponential stable, in the language of control theory) provided a certain linear matrix inequality is feasible.

Lemma 2. *Suppose there exist sequences $\{\xi^k, u^k, y^k\}$ such that for all $k \geq 0$, we have*

$$\begin{aligned} \xi^{k+1} &= A\xi^k + Bu^k \\ y^k &= C\xi^k + Du^k \\ 0 &= F\xi^k + Gu^k. \end{aligned} \tag{4}$$

where $u^k := \begin{bmatrix} u^{1,k} \\ \vdots \\ u^{p,k} \end{bmatrix}$, $y^k := \begin{bmatrix} y^{1,k} \\ \vdots \\ y^{p,k} \end{bmatrix}$, and (A, B, C, D) is partitioned conformally as

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|ccc} A & B^1 & \cdots & B^p \\ \hline C^1 & D^{11} & \cdots & D^{1p} \\ \vdots & \vdots & \ddots & \vdots \\ C^p & D^{p1} & \cdots & D^{pp} \end{array}\right].$$

Also define the block-rows: $D^j := \begin{bmatrix} D^{j1} & \cdots & D^{jp} \end{bmatrix}$. For all $k \geq 0$ and $j = 1, \ldots, p$, further suppose the inputs $u^{j,k}$ and outputs $y^{j,k}$ satisfy the quadratic inequalities

$$\begin{bmatrix} y^{j,k} - y^{j,\star} \\ u^{j,k} - u^{j,\star} \end{bmatrix}^\mathsf{T} M^j \begin{bmatrix} y^{j,k} - y^{j,\star} \\ u^{j,k} - u^{j,\star} \end{bmatrix} \geq 0, \tag{5}$$

where $(\xi^\star, y^\star, u^\star)$ is a stationary point of (4). Let R be a matrix whose columns are a basis for null $\begin{bmatrix} F & G \end{bmatrix}$. *If there exists $\rho > 0$, $P \succ 0$, and $\lambda_j \geq 0$ such that*

$$R^\mathsf{T} \left(\begin{bmatrix} A^\mathsf{T} P A - \rho^2 P & A^\mathsf{T} P B \\ B^\mathsf{T} P A & B^\mathsf{T} P B \end{bmatrix} + \sum_{j=1}^p \lambda_j \begin{bmatrix} C^j & D^j \\ 0 & e_j^\mathsf{T} \end{bmatrix}^\mathsf{T} M^j \begin{bmatrix} C^j & D^j \\ 0 & e_j^\mathsf{T} \end{bmatrix} \right) R \preceq 0 \tag{6}$$

then $\|\xi^{k+1} - \xi^\star\|_P \leq \rho \|\xi^k - \xi^\star\|_P$ for all $k \geq 0$.

Proof. The columns of R span the nullspace of $\begin{bmatrix} F & G \end{bmatrix}$ so any vector $\begin{bmatrix} (\xi^k - \xi^\star)^\mathsf{T} & (u^k - u^\star)^\mathsf{T} \end{bmatrix}^\mathsf{T}$ is of the form Rw for some w. Multiply (6) on the left and right by w^T and w respectively and obtain, after simplification:

$$(\xi^{k+1} - \xi^\star)^\mathsf{T} P (\xi^{k+1} - \xi^\star) - \rho^2 (\xi^k - \xi^\star)^\mathsf{T} P (\xi^k - \xi^\star) + \sum_{j=1}^p \lambda_j \begin{bmatrix} y^{j,k} - y^\star \\ u^{j,k} - u^\star \end{bmatrix}^\mathsf{T} M^j \begin{bmatrix} y^{j,k} - y^\star \\ u^{j,k} - u^\star \end{bmatrix} \leq 0.$$

The sum is nonnegative by (5), so

$$(\xi^{k+1} - \xi^\star)^\mathsf{T} P (\xi^{k+1} - \xi^\star) \leq \rho^2 (\xi^k - \xi^\star)^\mathsf{T} P (\xi^k - \xi^\star)$$

Take square roots and the desired result follows. ∎

Recursing the result of Lemma 2 implies the linear rate bound: $\|\xi^k - \xi^\star\|_P \leq \rho^k \|\xi^0 - \xi^\star\|_P$. We can further bound this via the condition number of P to obtain

$$\|\xi^k - \xi^\star\| \leq \sqrt{\text{cond}(P)}\, \rho^k \, \|\xi^0 - \xi^\star\|$$

If the SDP (6) is feasible for some $\rho < 1$, then we have certified a linear convergence rate $O(\rho^k)$. Note that the original bound in Lemma 2 is a stronger result because it also provides a *Lyapunov function*, which is a monotonically decreasing function of the state.

3 Analysis via the exact gossip matrix

In this section, we present an analysis approach to prove the linear convergence of EXTRA [11] that depends explicitly on the gossip matrices $W := \{w_{ij}\}$ and $V := \{v_{ij}\}$. In Section 5, we will see that this approach can be analogously applied to analyze a variety of other algorithms.

Theorem 3 (W-SDP)**.** *Suppose $f_i \in \mathcal{F}(m_i, L_i)$ for $i \in \{1, \ldots, n\}$ and consider the EXTRA algorithm (3) with parameter η and gossip matrices W and V. Define the matrices A, B, C, D, F, G as follows.*

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] := \left[\begin{array}{ccc|c} W + I_n & -\frac{1}{2}(V + I_n) & \eta I_n & -\eta I_n \\ I_n & 0_n & 0_n & 0_n \\ 0_n & 0_n & 0_n & I_n \\ \hline I_n & 0_n & 0_n & 0_n \end{array}\right],$$

$$F := \begin{bmatrix} 1^\mathsf{T} & -1^\mathsf{T} & \eta 1^\mathsf{T} \end{bmatrix}, \qquad G := 0_{1\times n}.$$

Further define $\bar{m}$, $\bar{L}$, and M^1 as follows.

$$\bar{m} := \operatorname{diag}(m_1, \ldots, m_n), \quad \bar{L} := \operatorname{diag}(L_1, \ldots, L_n),$$
$$M^1 := \begin{bmatrix} -2\bar{m}\bar{L} & \bar{m} + \bar{L} \\ \bar{m} + \bar{L} & -2I_n \end{bmatrix}.$$

Consider the SDP (6) of Lemma 2 with $p = 1$ and the matrices A, B, C, D, F, G, M^1 defined as above. If this SDP is feasible for some $\rho > 0$, $P \succ 0$, and $\lambda = 1$, then EXTRA converges linearly with a rate of ρ. In other words, there exists some $c > 0$ such that

$$\|x_i^k - x^\star\| \le c\,\rho^k \qquad \text{for all } i, k.$$

Proof. Define $\xi^k := \begin{bmatrix}(x^{k+1})^\mathsf{T} & (x^k)^\mathsf{T} & (\nabla^k)^\mathsf{T}\end{bmatrix}^\mathsf{T}$ with

$$x^{k+1} := \begin{bmatrix} x_1^{k+1} \\ \vdots \\ x_n^{k+1} \end{bmatrix}, \quad x^k := \begin{bmatrix} x_1^k \\ \vdots \\ x_n^k \end{bmatrix}, \quad \nabla^k := \begin{bmatrix} \nabla f_1(x_1^k) \\ \vdots \\ \nabla f_n(x_n^k) \end{bmatrix},$$

and input

$$u^{1,k} := \nabla f(y^{1,k}) := \begin{bmatrix} \nabla f_1(y_1^{1,k}) \\ \vdots \\ \nabla f_n(y_n^{1,k}) \end{bmatrix}.$$

In these new coordinates, EXTRA (3) takes the form:

$$\xi^{k+1} = (A \otimes I_d)\xi^k + (B \otimes I_d)u^{1,k}$$
$$y^{1,k} = (C \otimes I_d)\xi^k + (D \otimes I_d)u^{1,k}$$

The stationary point of the dynamics is given by $y_i^{1,\star} = x_i^\star = x^\star$, and $u_i^{1,\star} = \nabla f_i(x^\star)$, where $x^\star$ is the global optimum (1). Since $f_i \in \mathcal{F}(m_i, L_i)$, the quadratic bound of Proposition 1 holds for each agent i. Aggregating the states of all agents we obtain

$$\begin{bmatrix} y^{1,k} - y^{1,\star} \\ u^{1,k} - u^{1,\star} \end{bmatrix}^\mathsf{T} M^1 \begin{bmatrix} y^{1,k} - y^{1,\star} \\ u^{1,k} - u^{1,\star} \end{bmatrix} \ge 0,$$

where M^1 is defined in the theorem statement. Finally, the special initialization condition of EXTRA can also be rewritten as $(F \otimes I_d)\xi^0 = 0$. Moreover,

$$(F \otimes I_d)\xi^{k+1} = (FA \otimes I_d)\xi^k + (FB \otimes I_d)u^k = (F \otimes I_d)\xi^k$$

and it follows that $(F \otimes I_d)\xi^k + (G \otimes I_d)u^k = 0$ for all k. Note that $x^k, \nabla^k, u^{1,k}, y^{1,k} \in \mathbb{R}^{nd}$ and $\xi^k \in \mathbb{R}^{3nd}$. In constructing the SDP (6), we may exploit the block-diagonal structure of the algorithm; there will always exist a solution of the form $P \otimes I_d$. See [3, §4.2] for an expanded explanation. Consequently, I_d factors out entirely and we are left with the SDP (6) with no dependence on d. By Lemma 2, feasibility of (6) certifies that $\|\xi^{k+1} - \xi^\star\|_P \le \rho\|\xi^k - \xi^\star\|_P$. Recursing the bound as explained in Section 2, we obtain $\|\xi^k - \xi^\star\| \le \sqrt{\operatorname{cond}(P)}\,\rho^k\,\|\xi^0 - \xi^\star\|$. Note that x_i^k is one of the components of ξ^k and $x^\star$ is the corresponding component of $\xi^\star$. So by the triangle inequality, we have $\|x_i^k - x^\star\| \le \sqrt{\operatorname{cond} P}\,\rho^k\,\|\xi^0 - \xi^\star\|$, as required. ∎

Remark 4. *When applying Lemma 2, the SDP (6) is homogeneous in $(P, \lambda_1, \ldots, \lambda_p)$. Therefore, we may set $\lambda_1 = 1$ without loss of generality. This is why $\lambda = 1$ in the statement of Theorem 3.*

For each fixed $\rho \ge 0$, the SDP (6) is a linear matrix inequality (LMI), which is convex and is solved efficiently using interior-point methods or other means. The smallest rate $\rho \ge 0$ for which there exists a feasible $P \succ 0$ may be found using a bisection search. Note that the SDP (6) is $4n \times 4n$ with $P \in \mathbb{R}^{3n \times 3n}$. Thus, the size of the SDP is proportional to the number of agents (n), but independent of the size of $x \in \mathbb{R}^d$.

Tightness of upper bound. Theorem 3 gives an upper bound on the worst case convergence rate. To see whether the bound is tight, we simulated the EXTRA algorithm with random initialization for a two-agent network where each local function is defined by $f_i(x) = \frac{1}{2}x^\mathsf{T}Q_ix - b_i^\mathsf{T}x$. The matrices $Q_i \in \mathbb{R}^{d \times d}$ are symmetric positive semidefinite matrices randomly generated such that $\lambda_{\min}(Q_i) = m$, $\lambda_{\max}(Q_i) = L$, and the rest of the eigenvalues are uniformly distributed in $[m, L]$. Finally, the b_i are random vectors with components independently and uniformly distributed on $[0, 1]$. The gossip matrices used for simulation were $W = V = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$ and both local functions f_1 and f_2 have a condition ratio of $L/m = 10$. The step size parameter used is $\eta = 1/L$. Figure 1 depicts several algorithm trajectories bounded above by the linear rate bound obtained from SDP (6), which appears tight.

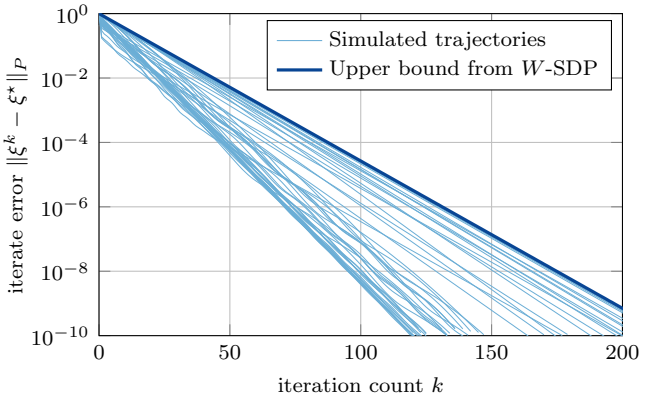

Figure 1: Numerical simulations of EXTRA for a network of $n = 2$ agents on 50 randomly generated strongly convex quadratics with $L/m = 10$. The upper bound on the iterate error is found via the W-SDP (6).

Varying the topology. We also experimented with changing the graph topology of the network. For a network with $n = 6$ agents, we consider graphs where each node has degree 5, 4, 3, and 2, respectively. The gossip matrices W and V were chosen to be symmetric and shift-invariant with a second-largest eigenvalue of $\sigma = 2/3$.

In Figure 2, we plot the worst-case convergence rate as a function of stepsize η again for the case where $L/m = 10$ for all functions. As the connectivity of the graph grows, EXTRA can tolerate larger stepsizes. The curves overlap and all start off the same, but they peel off at different values of η depending on the graph topology.

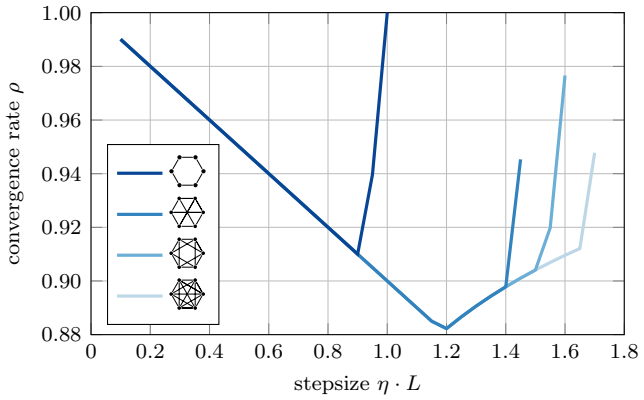

Figure 2: Linear rates obtained for EXTRA found via the W-SDP (6) as a function of stepsize η for several network topologies with $n = 6$ agents and strongly convex functions with $L/m = 10$. Results suggest that worst-case linear rates are graph-dependent.

4 Reduced SDP formulation

The approach of Section 3 and Theorem 3 provides graph-dependent bounds for worst-case performance, but involves solving a linear matrix inequality where the matrices have dimension that scales as $O(n)$. In this section, we show how to reduce the SDP (6) to one that depends only on the spectral gap of W and V and not on the number of agents n. In other words, this version of the SDP gives us a sufficient condition for linear convergence that is independent of graph size.

Our approach consists of replacing each gossip matrix by a rank-1 matrix plus a perturbation. In addition to the sector bound on the ∇f, we impose a bound on the spectral norm of the perturbation. This enables us to compute the worst case performance with respect to both the function and the graph. This formulation ultimately yields an SDP that decomposes into a pair of coupled SDPs whose sizes do not depend on the number of agents n or the domain dimension d.

Proposition 5. *Suppose Q is a matrix with spectral norm $\|Q\| \leq \sigma$. Further suppose that $u^k = Qy^k$ and $u^\star = Qy^\star$. Then, the following inequality holds.*

$$\begin{bmatrix} y^k - y^\star \\ u^k - u^\star \end{bmatrix}^\mathsf{T} \begin{bmatrix} \sigma^2 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} y^k - y^\star \\ u^k - u^\star \end{bmatrix} \geq 0.$$

Proof. By the definition of the spectral norm, we have: $\sigma \geq \|y^k - y^\star\| / \|u^k - u^\star\|$. Squaring both sides and rearranging yields the required result. ∎

The following lemma is the key result that allows us to further reduce the SDP and make it independent of n.

Lemma 6. *Suppose $Q_1, Q_2 \in R^{m\times m}$ and $J_1, J_2 \in \mathbb{R}^{n\times n}$ satisfy $J_1^2 = J_1$, $J_2^2 = J_2$, and $J_1 J_2 = J_2 J_1 = 0$. If $Q := Q_1 \otimes J_1 + Q_2 \otimes J_2$, then the following are equivalent.*

1. *$Q \succeq 0$.*
2. *$Q_1 \succeq 0$ and $Q_2 \succeq 0$.*

Proof. ($\Rightarrow$) Multiply both sides of the definition of Q by $I_m \otimes J_i$. Then $Q_i \otimes J_i \succeq 0$ and it follows that $Q_i \succeq 0$. ($\Leftarrow$) $Q_i \succeq 0$ implies $Q_i \otimes J_i \succeq 0$. Sum over i. ∎

We now present the main result: a sufficient condition for linear convergence of EXTRA.

Theorem 7 (σ-SDP)**.** *Suppose $f_i \in \mathcal{F}(m, L)$ for $i \in \{1, \ldots, n\}$ and consider the EXTRA algorithm (3) with parameter η and gossip matrices W and V such that the second-largest eigenvalue of W and V are each less than or equal to σ. Define matrices $A_i, B_i, C_i, D_i, F_i, G_i$:*

$$\left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] := \left[\begin{array}{ccc|ccc} 1 & -1/2 & \eta & -\eta & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1/2 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right] := \left[\begin{array}{ccc|ccc} 2 & -1 & \eta & -\eta & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1/2 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\left[\begin{array}{cc|ccc} F_1 & G_1 \\ \hline F_2 & G_2 \end{array}\right] := \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & -1 & \eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Further define the matrices M_i^j as follows.

$$M_1^1 = M_2^1 = \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix}$$

$$M_1^j = M_2^j = \begin{bmatrix} \sigma^2 & 0 \\ 0 & -1 \end{bmatrix} \qquad j = 2, 3.$$

Let R_i be a matrix whose columns are a basis for null $\begin{bmatrix} F_i & G_i \end{bmatrix}$ *for $i = 1, 2$. Define $J_1 := (I - \frac{1}{n} 1 1^\mathsf{T})$ and $J_2 := \frac{1}{n} 1 1^\mathsf{T}$. If there exists $\rho > 0$, $P_1, P_2 \succ 0$, and $\lambda_j \geq 0$ such that the following holds for $i = 1, 2$:*

$$R_i^\mathsf{T} \left(\begin{bmatrix} A_i^\mathsf{T} P_i A_i - \rho^2 P_i & A_i^\mathsf{T} P_i B_i \\ B_i^\mathsf{T} P_i A_i & B_i^\mathsf{T} P_i B_i \end{bmatrix} + \sum_{j=1}^{p} \lambda_j \begin{bmatrix} C_i^j & D_i^j \\ 0 & e_j^\mathsf{T} \end{bmatrix}^\mathsf{T} M_i^j \begin{bmatrix} C_i^j & D_i^j \\ 0 & e_j^\mathsf{T} \end{bmatrix} \right) R_i \preceq 0 \quad (7)$$

then EXTRA converges linearly with a rate of ρ. In other words, there exists some $c > 0$ such that

$$\|x_i^k - x^\star\| \le c\,\rho^k \qquad \textit{for all } i, k.$$

Proof. As in the proof of Theorem 3, we can factor out a $(\otimes I_d)$ term, so we omit this step and start with the dynamics in terms of n only. Write the gossip matrices as $W = \frac{1}{n}11^\mathsf{T} + \Delta W$ and $V = \frac{1}{n}11^\mathsf{T} + \Delta V$. Define two additional inputs corresponding to the uncertainties ΔW and ΔV. Then, EXTRA is given by (4) with

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{ccc|ccc} I_n+\frac{1}{n}11^\mathsf{T} & -\frac{1}{2}(I_n+\frac{1}{n}11^\mathsf{T}) & \eta I_n & -\eta I_n & I_n & I_n \\ I_n & 0_n & 0_n & 0_n & 0_n & 0_n \\ 0_n & 0_n & 0_n & I_n & 0_n & 0_n \\ \hline I_n & 0_n & 0_n & 0_n & 0_n & 0_n \\ I_n & 0_n & 0_n & 0_n & 0_n & 0_n \\ 0_n & -\frac{1}{2}I_n & 0_n & 0_n & 0_n & 0_n \end{array}\right] \tag{8}$$

and

$$\begin{aligned} u^{1,k} &= \nabla f(y^{1,k}) \\ u^{2,k} &= (\Delta W \otimes I_d) y^{2,k} \\ u^{3,k} &= (\Delta V \otimes I_d) y^{3,k}. \end{aligned}$$

Notice J_1 and J_2 satisfy $J_1^2 = J_1, J_2^2 = J_2$, and $J_1 J_2 = J_2 J_1 = 0$. Using Kronecker products, (A, B, C, D) for n agents in (8) can be split into two separate state-space representations which are independent of n.

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] \otimes J_1 + \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right] \otimes J_2$$

In this way, the dynamics of (4) with (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) correspond to the EXTRA update (3).

Since $f_i \in \mathcal{F}(m, L)$, the sector bound on $\nabla f(y^{1,k})$ applies. Since σ is the second-largest eigenvalue of W and V, $\|\Delta W\| \le \sigma$ and $\|\Delta V\| \le \sigma$. By Propositions 1 and 5, the following quadratic constraints hold for the three nonlinearities.

$$\begin{bmatrix} y^{j,k} \\ u^{j,k} \end{bmatrix}^\mathsf{T} M^j \begin{bmatrix} y^{j,k} \\ u^{j,k} \end{bmatrix} \ge 0 \qquad \text{for } j = 1, 2, 3$$

where $M^j := M_1^j \otimes J_1 + M_2^j \otimes J_2$ and the M_i^j are defined in the theorem statement.

We must also ensure that the perturbations ΔW and ΔV are such that W and V are doubly stochastic. This amounts to ensuring that $1^\mathsf{T}\Delta W = 0$ and $\Delta W 1 = 0$ and similarly for ΔV. Equivalently, we can replace C and D by $J_1 C$ and $J_1 D$ respectively and constrain the inputs for ΔW and ΔV as follows:

$$1^\mathsf{T} u^{j,k} = 0 \qquad \text{for } j = 2, 3.$$

Along with the invariant condition for EXTRA, the equality constraints on $u^{2,k}$ and $u^{3,k}$ can be expressed in the form of $0 = F\xi^k + Gu^k$ with

$$F := \begin{bmatrix} 1^\mathsf{T} & -1^\mathsf{T} & \eta 1^\mathsf{T} \\ 0_{1\times n} & 0_{1\times n} & 0_{1\times n} \\ 0_{1\times n} & 0_{1\times n} & 0_{1\times n} \end{bmatrix}, G := \begin{bmatrix} 0_{1\times n} & 0_{1\times n} & 0_{1\times n} \\ 0_{1\times n} & 1^\mathsf{T} & 0_{1\times n} \\ 0_{1\times n} & 0_{1\times n} & 1^\mathsf{T} \end{bmatrix}$$

Since $\operatorname{null}\begin{bmatrix} F & G \end{bmatrix} = \operatorname{null}\begin{bmatrix} F & G \end{bmatrix}^\mathsf{T}\begin{bmatrix} F & G \end{bmatrix}$, we observe that the nullspace only contains the J_2 component. By Lemma 2, feasibility of (7) certifies the rate bound. ∎

Varying the spectral gap. To demonstrate Theorem 7 in action, we applied the result to EXTRA with several values of σ in Figure **??** and plotted the worst-case linear rate versus step size. For each local function in the network, $L/m = 10$. Each curve represents worst case performance of EXTRA over the entire class of graphs with second-largest eigenvalue σ. In each case, there exists an optimal step size η_{opt} that achieves the smallest worst-case linear rate.

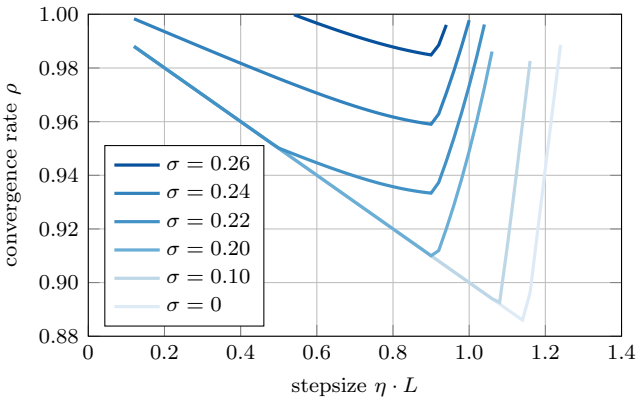

Figure 3: Linear rates obtained for EXTRA from the σ-SDP (7) as a function of stepsize η for several values of σ, which is the second-largest eigenvalue of the gossip matrices W and V.

5 Evaluating other algorithms

The methodology presented in Sections 3 and 4 can also be applied to other distributed optimization algorithms. As a proof of concept, we applied our analysis to the algorithm of Qu and Li [10] and the NIDS algorithm [4] using the reduced σ-SDP formulation of Theorem 7.

In the algorithm of Qu and Li [10], each agent performs a consensus step as well as a gradient estimation step with update equations

$$x_i^{k+1} = \sum_{j=1}^n w_{ij} x_j^k - \eta s_i^k \tag{9a}$$

$$s_i^{k+1} = \sum_{j=1}^n v_{ij} s_j^k + \nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k) \tag{9b}$$

where x_i^0 is arbitrary and $s_i(0) = \nabla f_i(x_i(0))$. The NIDS algorithm [4] update has a structure similar to EXTRA and is given by

$$x_i^{k+2} = x_i^{k+1} + \sum_{j=1}^{n} w_{ij} x_j^{k+1} - \frac{1}{2}\sum_{j=1}^{n}(1+v_{ij})(x_j^k + \eta(\nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k))) \quad (10)$$

where x_i^0 is arbitrary, and $x_i^1 = x_i^0 - \eta \nabla f_i(x_i^0)$. For both algorithms above, we defined the matrices (A, B, C, D, F, G) corresponding to the different algorithm dynamics. These results are displayed in Table 1.

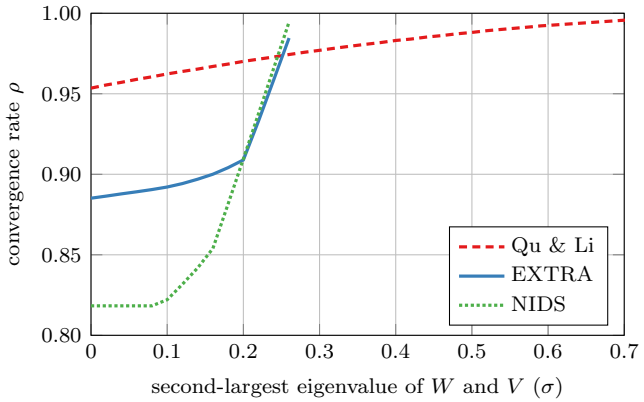

Figure 4: Worst-case linear rates obtained from the σ-SDP (7) as a function of σ using numerically determined optimal stepsizes and $L/m = 10$.

For each of the three algorithms considered in this paper, we applied Theorem 7 to obtain worst case linear rates for different choices of σ using the optimal step size η_{opt}. In Figure 4, worst-case linear rates obtained from (6) are plotted against σ and reveal that as σ increases, rate bounds worsen. All local functions in the network are assumed to be in $\mathcal{F}(m, L)$ with $L/m = 10$.

References

[1] P. A. Forero, A. Cano, and G. B. Giannakis. Consensus-based distributed support vector machines. *Journal of Machine Learning Research*, 11(May):1663–1707, 2010.

[2] B. Johansson. *On distributed optimization in networked systems*. PhD thesis, KTH, 2008.

[3] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[4] Z. Li, W. Shi, and M. Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *arXiv preprint arXiv:1704.07807*, 2017.

[5] Q. Ling and Z. Tian. Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Transactions on Signal Processing*, 58(7):3816–3827, 2010.

[6] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*.

[7] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[8] B. T. Polyak. *Introduction to optimization*. Optimization Software, Publications Division, New York, 1987.

[9] J. B. Predd, S. R. Kulkarni, and H. V. Poor. A collaborative training algorithm for distributed learning. *IEEE Transactions on Information Theory*, 55(4):1856–1871, 2009.

[10] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, (99):1–1, 2017.

[11] W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[12] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Algorithm	$\left[\begin{array}{c\|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right]$	$\left[\begin{array}{c\|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right]$	$\left[\begin{array}{c\|c} F_1 & G_1 \\ \hline F_2 & G_2 \end{array}\right]$
EXTRA [11]	$\left[\begin{array}{ccc\|ccc} 1 & -\frac{1}{2} & \eta & -\eta & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 \end{array}\right]$	$\left[\begin{array}{ccc\|ccc} 2 & -1 & \eta & -\eta & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$	$\left[\begin{array}{ccc\|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & -1 & \eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$
Qu and Li [10]	$\left[\begin{array}{ccc\|ccc} 0 & -\eta & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & -\eta & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]$	$\left[\begin{array}{ccc\|ccc} 1 & -\eta & 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & -\eta & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$	$\left[\begin{array}{ccc\|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$
NIDS [10]	$\left[\begin{array}{ccc\|ccc} 1 & -\frac{1}{2} & \frac{\eta}{2} & -\frac{\eta}{2} & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & \frac{\eta}{2} & -\frac{\eta}{2} & 0 & 0 \end{array}\right]$	$\left[\begin{array}{ccc\|ccc} 2 & -1 & \eta & -\eta & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$	$\left[\begin{array}{ccc\|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & -1 & \eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$

Table 1: Matrix parameters used in SDPs (7)for EXTRA [11], the algorithm of Qu and Li [10], and NIDS [4]. Using these definitions, Theorem 7 can be applied to obtain linear rates of convergence.