

From Mixing to Tearing: Graph Decomposition in Decentralized Optimization via Message Passing

Kuangyu Ding · Gesualdo Scutari

Abstract We study decentralized minimization of sums of smooth strongly convex functions, one per agent, with communication restricted to neighbors in a graph. Existing decentralized algorithms typically emulate a *centralized* method; *the network serves only to exchange information*, from which agents approximate the global quantities that the method needs. This paper breaks with this paradigm. We introduce a *graph decomposition framework* that uses the network to design the method, not merely execute it. It jointly determines how agreement is imposed, which variables are optimized jointly (blocks), and which connected subgraphs host the agents solving each block. The convergence analysis quantifies the trade-offs among these choices. GATE (Graph-Tearing message passing) instantiates the framework by assigning one dual variable per edge and partitioning the edges into trees. Agents interleave edge updates with local message passing recursions that summarize optimization within each tree (block). Its surrogate variant, GATE-S, uses tractable local models and compact messages to reduce computation and communication. We establish linear convergence and show that the topology-aware design acts as a *pre-conditioner*: the rate depends *jointly* on topology and function geometry through the coupling between subproblems defined on the selected subgraphs, whereas consensus-based optimization methods pay the worst case of each separately, the centralized condition number times the inverse spectral gap. On tree-dominant graphs, convergence is provably faster: the rate scales with the tree *diameter*, linearly in the number of agents on rings and barbells against quadratically for non-accelerated consensus-optimization methods. Experiments support the predicted scaling and savings in iterations and communication.

1 Introduction

The paper studies decentralized optimization over an undirected, connected, and simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (hereafter, agents), with $|\mathcal{V}| = m$, and $\mathcal{E}$ is the set of edges; neighboring agents can exchange information over single-hop links. Each agent $i \in \mathcal{V}$ knows only its own cost function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$,

K. Ding

Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN. E-mail: ding433@purdue.edu.

G. Scutari

Edwardson School of Industrial Engineering & Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. E-mail: gscutari@purdue.edu.

The work of Ding and Scutari has been supported by the ONR Grant # N00014-24-1-2751.

assumed strongly convex and smooth on $\mathbb{R}^d$. The agents seek to solve cooperatively the following optimization problem

$$\min_{x \in \mathbb{R}^d} F(x) := \sum_{i \in \mathcal{V}} f_i(x). \quad (\text{P})$$

This class of problems and setting arise in several applications of interest, including signal processing, machine learning, multi-agent systems, and communications.

(i) Inexact methods. Early methods solved (P) by decentralizing the gradient algorithm: the Distributed Gradient Descent (DGD) was introduced in [12] for least squares and later generalized to nonquadratic functions in [15], with a refined rate analysis in [29]. Each iteration combines a weighted average of the neighbors' estimates of the solution with a local gradient step; the schemes differ in how the two operations are interleaved. The stepsize sets a trade-off between speed and accuracy: a constant stepsize α yields a *linear* rate, but only to an $\mathcal{O}(\alpha)$ -neighborhood of the solution [15, 20, 29], whereas exact convergence requires a vanishing stepsize and hence a *sublinear* rate.

(ii) Exact methods. The speed-accuracy dilemma motivated a second generation of schemes aiming for *exact* solutions using a *constant* stepsize. They differ in how the local gradient direction is corrected, along three routes. **(a)** The first acts in the primal domain, either by tracking the global gradient ∇F [4, 14, 18, 19, 22, 25, 28], or through correction terms based on past iterates and gradients [1, 23]. **(b)** The second enforces consensus among the local copies x_i 's through linear constraints, whose (augmented) Lagrangian multipliers adjust the local gradient direction [2, 7, 9, 21, 24]. **(c)** The third turns the design into one of communication: dedicated routing or aggregation protocols replace gossip and approximate an *all-reduce* average of the agents' iterates or gradients over the network, which then feed the centralized update. Finite-time consensus [6, 10, 26] factorizes the averaging matrix into finitely many graph-compliant communication steps [10, 13, 16], reproducing a centralized gradient step exactly every finitely many rounds. Routing over a spanning tree achieves the same end: the agents' updates are relayed along the tree with hop-distance delays [27], or the gradients are aggregated at a root that runs a centralized accelerated method [21].

The routes are not independent: several primal exact methods admit a dual interpretation [8], and the third realizes with all-reduce averaging what the first two realize with mixing. The common approach is to rewrite (P) as a *centralized* problem with linear consensus constraints, making centralized optimization techniques readily applicable. The constraints are chosen to support implementation over the network, and the resulting formulation determines the dual variables on which the algorithm operates. Specifically, stack the copies $X = [x_1^\top; \dots; x_m^\top]$, and call $\mathbf{A} \in \mathbb{R}^{m \times q}$ *faithful* if $\ker(\mathbf{A}^\top) = \text{span}\{\mathbf{1}\}$ ($\mathbf{1}$ is the all-ones vector). Then, $\mathbf{A}^\top X = 0$ enforces exactly consensus. The lifted problem of (P) and its dual read

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times d}} \sum_{i \in \mathcal{V}} f_i(x_i) \quad \text{s.t.} \quad \mathbf{A}^\top X = 0, \\ \min_{Y \in \mathbb{R}^{q \times d}} g_{\mathbf{A}}(Y) := \sum_{i \in \mathcal{V}} f_i^*((\mathbf{A}Y)_i). \end{aligned} \quad (\text{D})$$

Here, f_i^* is the Fenchel conjugate of f_i , and each column j of $\mathbf{A}$ introduces one multiplier row $Y_j \in \mathbb{R}^d$. A dual solution Y^* recovers the primal solution through $x_i^* = \nabla f_i^*((\mathbf{A}Y^*)_i)$, $i \in \mathcal{V}$.

Faithful choices used in the decentralized methods discussed above include **(i)** $\mathbf{A} = \mathbf{I} - \mathbf{W}$, for a symmetric gossip matrix $\mathbf{W}$ with $\ker(\mathbf{I} - \mathbf{W}) = \text{span}\{\mathbf{1}\}$ [4, 8, 14, 15, 19, 22, 23, 25, 28]; **(ii)** the graph Laplacian or operators built from it [7, 9, 21]; **(iii)** edge-indexed choices—the oriented node-edge incidence matrix $\mathbf{B}$ of $\mathcal{G}$, or that of a spanning tree, as in the master-node implementation of [21]; and **(iv)** the all-reduce averaging operator $\mathbf{A} = \mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$, realized through finite-time consensus or tree routing [13, 16, 27]. These representations all enforce the *same* consensus constraint in (D), but generally induce *different* algorithms. Each column j of $\mathbf{A}$ creates a dual variable Y_j , and its support identifies the conjugates f_i^* that depend on it; grouping columns determines which dual variables are optimized jointly. This is a design layer preceding the block partition, absent from classical (centralized) block-coordinate methods (of Jacobi or Gauss–Seidel type), where the variables are given and only their grouping is a design choice. Here the representation $\mathbf{A}$ first determines the variables available for grouping and their dependencies.

The missing design step. The methods above start from a prescribed centralized method: a first- or a second-order scheme on (P), or a Lagrangian-based method on its consensus reformulation (D). A communication protocol is then added to supply the information the method needs to be implemented on the network: gossip, spanning-tree relaying, finite-time averaging, or exchanges along an edge-supported representation $\mathbf{A}$. *The network serves only as a communication medium* over which the method’s variables are exchanged, mixed, or aggregated. Topology guides communication and parameter choices, but *does not influence the decomposition of the optimization problem*. Long paths, dense regions, and sparse interfaces therefore serve communication, rather than define distinct optimization subproblems. Representative convergence rate bounds reflect this separation. Function geometry enters through the condition number of F , while communication enters through a spectral gap, an aggregation time, or a delay-dependent mixing constant. The rate combines these *separately* assessed quantities, possibly after acceleration [11, 21, 25, 27]. Lower bounds, even when matched, establish limits of *prescribed* computation and communication models, *not limits of the graph topology itself* [21]. The missing design step is thus not another improvement of the optimization and communication factors separately, but a joint design in which graph structure determines the subproblems and how agents solve them.

The ring illustrates the potential of this change. On a ring of m agents, standard nearest-neighbor gossip averaging has inverse spectral gap $\Theta(m^2)$ [17], despite a diameter of $\Theta(m)$. Non-accelerated consensus baselines inherit this quadratic network dependence (Sec. 5.4). For RelaySum on a spanning path, the analysis yields an $O(m^{3/2})$ network factor, which becomes $O(m^{5/2})$ for convex losses [27]. Yet a path can do more than transport or average iterates or gradients. With edge-indexed dual variables, *one can “open” the ring into a path*: one long path of edges is optimized jointly, with the remaining edges defining separate, individual subproblems. As shown in Sec. 3.1, the path subproblem can be solved *exactly* by eliminating one node after another in $O(m)$ steps—structural grounds for $O(m)$ overall complexity, rather than $O(m^2)$. The path determines both what is optimized jointly and how the computation is distributed. Establishing a general framework for this joint design is the goal of this paper.

1.1 From algorithm execution on a graph to graph decomposition

We rethink decomposition methods from the foundations, accounting for the graph as part of the algorithm design. The major contributions are summarized next.

(i) **Graph decomposition framework.** Starting from the dual formulation (D), we introduce a formal paradigm of algorithmic decomposition for (P) over $\mathcal{G}$. Two observations shape the construction.

- (a) **The algebraic factor graph:** The dependency of the conjugate terms $f_i^*((\mathbf{A}\mathbf{Y})_i)$ on the dual coordinates Y_j is itself a graph—the bipartite *factor graph* induced by $\mathbf{A}$:

$$\mathcal{H}_{\mathbf{A}} := (\mathcal{V} \sqcup \{1, \dots, q\}, \{(i, j) : \mathbf{A}_{ij} \neq 0\}),$$

with one factor node per conjugate f_i^* , one variable node per dual coordinate Y_j , and an edge exactly when Y_j enters f_i^* . An algebraic decomposition is then obtained by selecting p nonempty blocks of dual coordinates. A block $\mathcal{J}_r \subseteq \{1, \dots, q\}$ specifies the coordinates optimized jointly, while its neighborhood in $\mathcal{H}_{\mathbf{A}}$ identifies the conjugates entering the resulting block subproblem.

- (b) **Blocks on the physical network:** The factor graph records algebraic dependencies, not communication links: the agents involved in a block need not induce a connected subgraph of $\mathcal{G}$. Since $\mathcal{G}$ is connected, they can still cooperate through intermediate agents, possibly the whole network, at substantial communication cost. We therefore make the cooperating agents part of the design, assigning each block a *carrier* $\mathcal{C}_r \subseteq \mathcal{V}$: a set of agents containing all those whose conjugates enter the block subproblem and inducing a connected physical subgraph $\mathcal{G}[\mathcal{C}_r]$. Additional carrier agents serve as relays.

The joint choice of the three—representation, blocks, and carriers—is a *graph decomposition* of (P) over $\mathcal{G}$: a triple

$$\left(\underbrace{\mathbf{A}}_{\text{representation}}, \quad \underbrace{\{\mathcal{J}_r\}_{r=1}^p}_{\text{dual-coordinate blocks}}, \quad \underbrace{\{\mathcal{C}_r\}_{r=1}^p}_{\text{physical carriers}} \right).$$

Here, $\mathbf{A}$ is faithful; the nonempty blocks $\mathcal{J}_r$ cover $\{1, \dots, q\}$ and need not be disjoint; and each carrier $\mathcal{C}_r \subseteq \mathcal{V}$ contains every agent i such that $\mathbf{A}_{i, \mathcal{J}_r} \neq 0$ and induces a connected subgraph of $\mathcal{G}$. The three layers have distinct roles: the representation creates the variables and their dependencies; the blocks fix *what* is optimized jointly; the carriers fix *where* it is executed. Sec. 2 develops this definition and gives the triple its algorithmic meaning.

(ii) **GATE as an instance.** GATE (Graph-Tearing message passing) instantiates the framework above by tearing the graph along its edges. Agreement is imposed edge by edge, so that each edge carries one dual variable shared by its two endpoints only ($\mathbf{A} = \mathbf{B}$, the node-edge incidence matrix). The edges are partitioned into trees ($\mathcal{P} = \{\mathcal{E}_r\}_{r=1}^p$, with $\mathcal{J}_r = \mathcal{E}_r$ and $\mathcal{T}_r = (\mathcal{V}_r, \mathcal{E}_r)$ a tree). Each tree defines one block optimized cooperatively by the agents in $\mathcal{V}_r$, with carrier $\mathcal{C}_r = \mathcal{V}_r$. On a tree, fixing one edge variable splits the block subproblem into two independent ones, one on each side of the edge. With out-of-block variables held fixed, the optimal value on each side, as a function of that edge variable, defines an endpoint *cost-to-go message*, computed recursively from the leaves toward the edge. GATE interleaves one step of the message recursion with updates of the edge variables. Each edge update is based

on minimizing the sum of its two endpoint messages. The message updates use the latest out-of-block variables, accounting for the coupling between trees in a Jacobi fashion. The implementation is node-owned and uses single-hop exchanges, with no edge- or block-level coordinator (Secs. 3.1–3.2).

(iii) Delayed block-Jacobi identification and convergence analysis. We establish linear convergence of GATE through an analysis of independent interest that identifies GATE’s single-loop dynamics with an *exact delayed block-Jacobi scheme*. The identification is nontrivial: edge variables and messages evolve together, and the tree recursion is never completed while keeping the exterior coordinates fixed. Unfolding the messages through earlier rounds shows that, after a finite propagation phase, each edge response is a coordinate of an exact block minimizer. Its exterior coordinates are delayed by the distance from the edge to the boundary agents of the block. Linear convergence then follows by combining a block error bound with a perturbation analysis of these minimizers that exploits the tree structure of the delays. The resulting explicit rate is discussed next.

(iv) Block preconditioning, scalability, and partition design. The linear rate ρ reads: $\rho = 1 - \tau/(2\kappa_{\mathcal{P}})$, where τ is the relaxation stepsize and $\kappa_{\mathcal{P}}$ is a novel block error-bound constant; it measures the coupling remaining between block subproblems after exact block minimization; it depends on both the partition and the agents’ cost functions. Three further quantities govern the admissible stepsize τ , hence the final rate: block overlap, the largest number of blocks sharing an agent; propagation delays, determined by within-tree distances from boundary agents to updated edges; and “interface sensitivity”, measuring how changes in exterior coordinates affect block minimizers. Function geometry and graph topology therefore enter the rate *jointly*, unlike standard consensus bounds built from separately assessed conditioning and communication factors. The partition $\mathcal{P}$ lets us exploit this interaction: choosing which variables are minimized jointly changes the coupling between block subproblems—a *block-preconditioning effect*—whose benefits must be balanced against block overlap, propagation delays, and interface sensitivity.

These rate dependences provide concrete guidance for partition design. For instance, under the curvature conditions specified in Sec. 4.4, the bound on rings favors one long path block with $O(1)$ singleton edges over equal-length paths. The former gives $(1 - \rho)^{-1} = O(m/\log m)$ and, under mild bounds on initial transients, an ϵ -solution in $O(m + (m/\log m)\log(1/\epsilon))$ rounds, including initial propagation. For equal-length paths, the best bound we derive on $(1 - \rho)^{-1}$ is $O(m^{3/2})$, compared with $O(m^2)$ for nonaccelerated consensus baselines. The same principle gives $(1 - \rho)^{-1} = O(D/\log D)$ on tree-dominant graphs with dominant-tree diameter $\Theta(D)$, and finite termination on trees when all edges form one block and $\tau = 1$ (Sec. 4.4).

(v) Surrogate variant–GATE-S. To reduce GATE’s per-iteration computation and communication, we develop GATE-S using tractable local surrogates. Our examples include first-order quadratic models with fixed curvature and second-order models whose curvature is updated using local Hessian information at the current iterate. We provide implementations requiring only vector-sized exchanges, using scalar curvature coefficients or incremental updates of locally cached Hessians. Under regularity assumptions, we establish linear convergence with partition-dependent rates quantifying the effect of the surrogate approximation (Sec. 5).

(vi) Numerical experiments. We conduct numerical experiments to validate the theoretical results and evaluate the iteration and communication efficiency of GATE

and its surrogate variants relative to consensus-based methods. The results corroborate the theoretical scaling and demonstrate the efficiency of our algorithms in both iteration and communication.

1.2 Notation and paper organization

For a positive integer q , write $[q] := \{1, \dots, q\}$, and use $\sqcup$ for disjoint union. For functions $F(x, y)$ and $G(x, y)$, write $F \sim_x G$ when $F - G$ is independent of x . When the variable is clear from context, we write $F \sim G$.

We use row-stacked matrices for d -vectors. If the rows of a matrix A are indexed by a finite set $\mathcal{J}$, then A_α denotes the row indexed by $\alpha \in \mathcal{J}$, and A_S denotes the submatrix of rows indexed by $S \subseteq \mathcal{J}$. For matrices with both row and column indices, $A_{I,S}$ denotes the submatrix with rows I and columns S . Whenever such a row vector appears as an argument of f_i^* , in an inner product, or in a norm, it is identified with the corresponding vector in $\mathbb{R}^d$. For node variables $(x_i)_{i \in \mathcal{V}}$, X denotes the row-stacked matrix whose i -th row is $x_i^\top$. Inner products between row-stacked matrices use the Frobenius convention. For scalar functions on row-stacked matrices, $\nabla \Phi(A)^\top$ denotes the row-stacked gradient. We write $\mathbb{S}^d$ for the space of $d \times d$ symmetric matrices, and $\mathbb{S}_+^d$ and $\mathbb{S}_{++}^d$ for its positive semidefinite and positive definite cones. For a scalar differentiable function ϕ , $\nabla \phi$ is the column gradient and $\nabla^2 \phi$ is the Hessian. For a vector-valued map $F = (F_1, \dots, F_k) : \mathbb{R}^d \rightarrow \mathbb{R}^k$, we set $\nabla F(x) := [\nabla F_1(x), \dots, \nabla F_k(x)] \in \mathbb{R}^{d \times k}$; hence $\nabla F(x)^\top$ is the Jacobian and $\nabla F(x)^\top h$ is the directional variation along h . For a matrix-valued map H , $\nabla H(x)[h]$ denotes the directional derivative in direction h . For $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$, write $\text{Lip}(T) := \sup_{\eta \neq \eta'} \|T(\eta) - T(\eta')\| / \|\eta - \eta'\|$ for its global Lipschitz modulus. We use $O(\cdot)$ for asymptotic upper bounds and $\Theta(\cdot)$ for matching asymptotic upper and lower bounds, up to positive constant factors.

Organization. The remainder of the paper is organized as follows. Sec. 2 develops the graph decomposition framework. Sec. 3 specializes it to the incidence representation and tree blocks to derive GATE. Sec. 4 establishes linear convergence and partition-dependent rate bounds. Sec. 5 develops the surrogate variants of GATE, establishes their convergence, and compare their performance with consensus-based and tree-based state-of-the-art methods. Sec. 6 presents numerical results, and Sec. 7 concludes.

2 Graph Decomposition Framework

2.1 Problem and graph setting

We study (P) over the graph $\mathcal{G}$ under the following standard assumptions.

Assumption 1 (problem class) *For every $i \in \mathcal{V}$, $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ_{f_i} -strongly convex and L_{f_i} -smooth for some $\mu_{f_i}, L_{f_i} > 0$. Define*

$$\mu := \min_{i \in \mathcal{V}} \mu_{f_i}, \quad L := \max_{i \in \mathcal{V}} L_{f_i}, \quad \kappa := \frac{L}{\mu}.$$

Consequently, each f_i^* is finite-valued, differentiable, $1/L_{f_i}$ -strongly convex and $1/\mu_{f_i}$ -smooth.

Assumption 2 (Network) *The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is undirected, connected, and simple.*

Fix an arbitrary orientation of $\mathcal{E}$ and let $\mathbf{B} \in \mathbb{R}^{m \times |\mathcal{E}|}$ be the node–edge incidence matrix, with

$$\mathbf{B}_{ve} := \begin{cases} 1, & \text{if } v \text{ is the tail of } e, \\ -1, & \text{if } v \text{ is the head of } e, \\ 0, & \text{otherwise.} \end{cases}$$

For $v \in \mathcal{V}$ and $e \in \mathcal{E}$, let

$$\mathcal{A}_v := \{e \in \mathcal{E} : e \text{ is incident to } v\} \quad \text{and} \quad \mathcal{I}_e := \{v \in \mathcal{V} : v \text{ is an endpoint of } e\}.$$

The node-edge incidence graph of a graph is the bipartite graph on its nodes and edges in which a node and an edge are adjacent exactly when the node is an endpoint of the edge.

For a physical subgraph $\mathcal{G}_s \subseteq \mathcal{G}$, write $\mathcal{V}(\mathcal{G}_s)$ and $\mathcal{E}(\mathcal{G}_s)$ for its vertex and edge sets; for $S \subseteq \mathcal{V}$, $\mathcal{G}[S]$ denotes the subgraph of $\mathcal{G}$ induced by S .

2.2 Graph decomposition

This section builds the proposed decomposition framework.

2.2.1 Representations and their factor graphs

Recall from Sec. 1 that a matrix $\mathbf{A} \in \mathbb{R}^{m \times q}$ is *faithful* if $\ker(\mathbf{A}^\top) = \text{span}\{\mathbf{1}\}$, so that $\mathbf{A}^\top X = 0$ holds exactly for consensual stacked copies $X = [x_1^\top; \dots; x_m^\top]$. As derived there, the lifted consensus form of the problem has the dual formulation (D), with one multiplier row $Y_j \in \mathbb{R}^d$ per column j of $\mathbf{A}$. Every faithful $\mathbf{A}$ represents the same consensus set, but the dual (D) depends on $\mathbf{A}$ itself: the columns create the dual coordinates, and the sparsity pattern of $\mathbf{A}$ decides which coordinates enter which conjugate. This dependency structure is captured by the factor graph.

Definition 1 (Factor graph) The *factor graph induced by $\mathbf{A}$* is the bipartite graph

$$\mathcal{H}_{\mathbf{A}} := (\mathcal{V} \sqcup \{1, \dots, q\}, \{(i, j) : \mathbf{A}_{ij} \neq 0\}),$$

whose factor node i represents the term $f_i^*((\mathbf{A}Y)_i)$ and whose variable node j represents the dual coordinate Y_j .

The two sides of $\mathcal{H}_{\mathbf{A}}$ are not symmetric: each factor node is anchored at a physical agent, whereas a variable node need not have a prescribed physical location on $\mathcal{G}$. Consequently, a single variable Y_j –column $\mathbf{A}_{:,j}$ –may enter the conjugates of nonadjacent agents.

Example 1 Given the path graph $1 - 2 - 3$, with both edges oriented from left to right (Figure 1(a)), consider the following three faithful choices of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{A}_{\text{ex}} := \begin{bmatrix} 1 & 1 \\ 0 & -2 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{A} = \mathbf{L} := \frac{1}{3}\mathbf{B}\mathbf{B}^\top, \quad \text{or} \quad \mathbf{A} = \mathbf{B}. \quad (1)$$

Here, $\mathbf{B}$ is the incidence matrix of $\mathcal{G}$, and $\mathbf{L}$ is its scaled Laplacian. The three matrices enforce the same consensus constraint, yet they induce the distinct factor graphs of

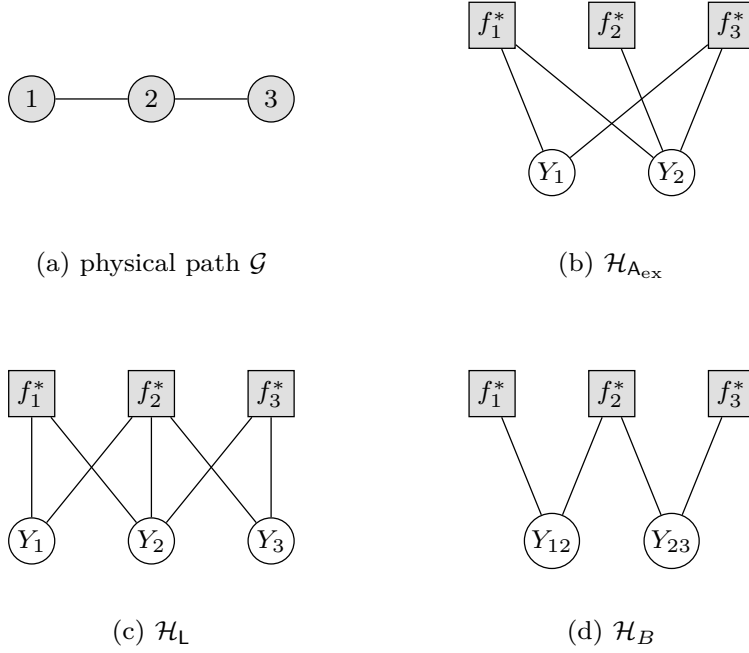


Fig. 1: Three faithful representations (see (1)) of the same path graph $1-2-3$ induce distinct factor graphs: the generic $\mathbf{A}_{\text{ex}}$ couples nonadjacent factors, the Laplacian $\mathbf{L}$ produces a cyclic factor graph, and the incidence matrix $\mathbf{B}$ reproduces the node-edge incidence structure.

Figure 1(b)-(d).¹ For $\mathbf{A}_{\text{ex}}$, the coordinate Y_1 enters the factors of the *nonadjacent* agents 1 and 3. For $\mathbf{L}$, Y_1 and Y_2 both enter f_1^* and f_2^* , so $\mathcal{H}_{\mathbf{L}}$ contains a *cycle* although $\mathcal{G}$ is a path. For $\mathbf{B}$, each coordinate corresponds to *one physical edge* and enters exactly its endpoint factors, so $\mathcal{H}_{\mathbf{B}}$ is the node-edge incidence graph of $\mathcal{G}$, that is, the bipartite graph on $\mathcal{V} \sqcup \mathcal{E}$ in which i and e are adjacent exactly when $i \in \mathcal{I}_e$. This distinction matters when the dual coordinates are grouped and optimized jointly: a group of coordinates of $\mathbf{B}$ is a set of physical edges, one per coordinate, whereas a group under $\mathbf{L}$ or $\mathbf{A}_{\text{ex}}$ may couple nonadjacent agents.

2.2.2 Block covers and conditional responses

To minimize $g_{\mathbf{A}}$ *blockwise*, we group the dual coordinates into (possibly overlapping) blocks. Each block-subproblem minimizes $g_{\mathbf{A}}$ over its block coordinates, holding the remaining coordinates fixed.

Definition 2 (Block cover and conditional responses) For a fixed $\mathbf{A}$, a *block cover* is a family $\{\mathcal{J}_r\}_{r=1}^p$ of nonempty subsets covering $[q]$ (the index set of the dual coordinates). Let $\bar{\mathcal{J}}_r := [q] \setminus \mathcal{J}_r$. For each block $\mathcal{J}_r$, define its touched factor set and block factor graph by

$$\mathcal{V}_{\mathbf{A},r} := \{i \in \mathcal{V} : \mathbf{A}_{i,\mathcal{J}_r} \neq 0\}, \quad \mathcal{F}_{\mathbf{A},r} := \mathcal{H}_{\mathbf{A}}[\mathcal{V}_{\mathbf{A},r} \sqcup \mathcal{J}_r]. \quad (2)$$

¹ In panels (b) and (c), the subscripts of Y index the columns of $\mathbf{A}_{\text{ex}}$ and $\mathbf{L}$, respectively; in panel (d), Y_{12} and Y_{23} are indexed by the physical edges $(1, 2)$ and $(2, 3)$.

Here, $\mathcal{H}_A[S]$ denotes the subgraph of $\mathcal{H}_A$ induced by the vertex set S .

The conditional response of block r is

$$\text{BR}_{A,r}(Y_{\mathcal{J}_r}) := \underset{W \in \mathbb{R}^{|\mathcal{J}_r| \times d}}{\text{argmin}} \sum_{i \in \mathcal{V}_{A,r}} f_i^*(A_{i,\mathcal{J}_r}W + A_{i,\mathcal{J}_r}Y_{\mathcal{J}_r}). \quad (3)$$

Under Assumption 1, $\text{BR}_{A,r}(Y_{\mathcal{J}_r})$ is nonempty for every $r \in [p]$ and $Y_{\mathcal{J}_r}$, but need not be a singleton. When A is fixed, we suppress it and write $\mathcal{V}_r$, $\mathcal{F}_r$, and BR_r .

Thus $\mathcal{J}_r$ identifies the coordinates optimized jointly, $\mathcal{V}_{A,r}$ the agents whose conjugates enter the block subproblem, and $\mathcal{F}_{A,r}$ records which block coordinates enter each conjugate.

Proposition 1 (Block fixed points) *Let Assumption 1 hold, let A be faithful, and let $\{\mathcal{J}_r\}_{r=1}^p$ be a block cover. Then,*

$$Y^* \in \underset{Y \in \mathbb{R}^{q \times d}}{\text{argmin}} g_A(Y) \iff Y_{\mathcal{J}_r}^* \in \text{BR}_r(Y_{\mathcal{J}_r}^*) \quad \forall r \in [p]. \quad (4)$$

Proposition 1 holds for *every* faithful A and *every* block cover: optimality restricts neither choice. Different choices produce different response maps $\text{BR}_{A,r}(Y_{\mathcal{J}_r})$ —hence, potentially, different algorithms—but the algebraic decomposition $(A, \{\mathcal{J}_r\})$ does not specify where these responses are computed on $\mathcal{G}$.

Since f_i^* is available only at agent i , computing BR_r requires the agents indexed by $\mathcal{V}_r$ to exchange information over the links of $\mathcal{G}$. These exchanges can be confined to the agents of $\mathcal{V}_r$ only if $\mathcal{G}[\mathcal{V}_r]$ is connected, which Definition 2 does not guarantee. In Example 1, the subproblem of the block $\{1\}$ (the variable Y_1) of A_{ex} contains the conjugates f_1^* and f_3^* , stored at the nonadjacent agents 1 and 3, which can communicate only through agent 2, whose conjugate does not enter the subproblem. The computation of a block response thus involves, besides the agents of $\mathcal{V}_r$, those that relay between them; the set of all agents taking part in it must induce a connected subgraph containing $\mathcal{V}_r$. This is formalized next.

2.2.3 Carriers and realizability

We associate with each block $\mathcal{J}_r$ a set of agents $\mathcal{C}_r$ that enable the joint computation of the conditional response BR_r of the block on $\mathcal{G}$. This set, called the *carrier* of $\mathcal{J}_r$, must induce a connected subgraph of $\mathcal{G}$ and contain the agents in $\mathcal{V}_r$, whose conjugates enter the block subproblem (3). We formalize the joint choice of representation, block cover, and carriers as a *graph decomposition*.

Definition 3 (Graph decomposition) A *graph decomposition over $\mathcal{G}$* is a triple

$$\mathcal{D} := (A, \{\mathcal{J}_r\}_{r=1}^p, \{\mathcal{C}_r\}_{r=1}^p), \quad (5)$$

where A is faithful, $\{\mathcal{J}_r\}_{r=1}^p$ is a block cover of g_A , and the set $\emptyset \neq \mathcal{C}_r \subseteq \mathcal{V}$ is the *carrier* of block r satisfying

$$\mathcal{V}_r \subseteq \mathcal{C}_r \quad \text{and} \quad \mathcal{G}[\mathcal{C}_r] \text{ is connected.} \quad (6)$$

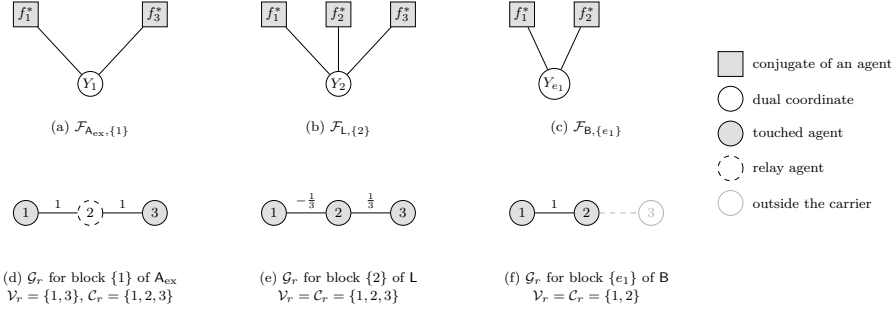


Fig. 2: Example 1 revisited. Top: block factor graphs. Bottom: carrier subgraphs $\mathcal{G}_r$, with the entries of M_r on the edges.

The three components have distinct roles: A creates the dual coordinates and their dependencies, $\mathcal{F}_r$ fixes what is optimized jointly, and $\mathcal{C}_r$ specifies the subgraph on which that computation is performed. The agents indexed by $\mathcal{C}_r \setminus \mathcal{V}_r$ contribute no term to the subproblem (3): their only role is to relay information between agents of $\mathcal{V}_r$ that are not adjacent in $\mathcal{G}$ (as agent 2 does for the block $\{1\}$ of A_{ex} in Example 1). When $\mathcal{G}[\mathcal{V}_r]$ is connected, the carrier can be taken equal to $\mathcal{V}_r$ and no additional relay agents are required.

For a subgraph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s) \subseteq \mathcal{G}$, let $B_{\mathcal{G}_s} \in \mathbb{R}^{m \times |\mathcal{E}_s|}$ be the matrix obtained from B by retaining the columns indexed by $\mathcal{E}_s$. The rows indexed by $\mathcal{V} \setminus \mathcal{V}_s$ are zero. The carrier conditions (6) admit the following equivalent matrix factorization.

Proposition 2 (realizability) Fix $r \in [p]$. Let A be faithful, let $\emptyset \neq \mathcal{J}_r \subseteq [q]$, and let $\emptyset \neq \mathcal{C}_r \subseteq \mathcal{V}$. Conditions (6) hold if and only if there exist a connected subgraph $\mathcal{G}_r \subseteq \mathcal{G}$, with $\mathcal{V}(\mathcal{G}_r) = \mathcal{C}_r$, and a matrix $M_r \in \mathbb{R}^{|\mathcal{J}_r| \times |\mathcal{E}(\mathcal{G}_r)|}$ such that

$$A_{:, \mathcal{J}_r} = B_{\mathcal{G}_r} M_r^\top. \quad (7)$$

Proposition 2 decomposes the linear couplings recorded by adjacency in the block factor graph $\mathcal{F}_r$ (see (2))—which agents' conjugate terms depend on each dual coordinate Y_j —into couplings between adjacent agents in the connected physical subgraph $\mathcal{G}_r$ (with vertex set $\mathcal{C}_r$). This opens the way to design methods that compute BR_r through neighbor-to-neighbor communication within the carrier.

To make the connection explicit, recall from (3) that each block coordinate Y_j enters the argument of f_i^* with coefficient A_{ij} . By (7), $A_{:, j}$ is a linear combination of the columns of B indexed by the edges of $\mathcal{G}_r$, with coefficients given by the corresponding row of M_r ; its nonzero entries are the agents whose factor nodes are adjacent to the variable node j in $\mathcal{F}_r$. If the row of M_r corresponding to Y_j has a single nonzero entry, Y_j appears only in the conjugates f_i^* of the two endpoint agents of one physical edge, which are neighbors in $\mathcal{G}$ and can communicate directly. If that row has several nonzero entries, Y_j may appear in the conjugates of non-adjacent agents or of more than two agents; these agents can exchange information along the edges of $\mathcal{G}_r$, with intermediate agents relaying it when needed. At agents in $\mathcal{C}_r \setminus \mathcal{V}_r$, the linear combinations in (7) vanish for every block coordinate: their conjugate terms are independent of the block variables, and these agents can serve as relays. Example 1 revisited and Figure 2 illustrate these distinctions.

Example 1 revisited. Let $e_1 = \{1, 2\}$ and $e_2 = \{2, 3\}$, with both edges oriented from left to right. For each representation in Example 1, choose the singleton block cover, with one block per dual coordinate. We examine the blocks $\{1\}$ and $\{2\}$ of $\mathbf{A}_{\text{ex}}$, the block $\{2\}$ of $\mathbf{L}$, and the block $\{e_1\}$ of $\mathbf{B}$. Their conditional responses, with terms in $g_{\mathbf{A}}$ independent of W omitted, are

$$\begin{aligned} \text{BR}_{\mathbf{A}_{\text{ex}}, \{1\}}(Y_2) &= \underset{W}{\operatorname{argmin}} \{f_1^*(W + Y_2) + f_3^*(-W + Y_2)\}, \\ \text{BR}_{\mathbf{A}_{\text{ex}}, \{2\}}(Y_1) &= \underset{W}{\operatorname{argmin}} \{f_1^*(Y_1 + W) + f_2^*(-2W) + f_3^*(-Y_1 + W)\}, \\ \text{BR}_{\mathbf{L}, \{2\}}(Y_1, Y_3) &= \underset{W}{\operatorname{argmin}} \{f_1^*\left(\frac{1}{3}(Y_1 - W)\right) + f_2^*\left(\frac{1}{3}(2W - Y_1 - Y_3)\right) + f_3^*\left(\frac{1}{3}(Y_3 - W)\right)\}, \\ \text{BR}_{\mathbf{B}, \{e_1\}}(Y_{e_2}) &= \underset{W}{\operatorname{argmin}} \{f_1^*(W) + f_2^*(Y_{e_2} - W)\}. \end{aligned}$$

The factor sets touched by the four blocks, read off from the responses, are $\mathcal{V}_r = \{1, 3\}$ for the first, as in Figure 2(a); $\{1, 2, 3\}$ for the second and third, with the Laplacian case shown in panel (b); and $\{1, 2\}$ for the last, as in panel (c). Only the first fails to induce a connected subgraph of $\mathcal{G}$. Its carrier must therefore contain agent 2, giving $\mathcal{C}_r = \{1, 2, 3\}$ (panel (d)): this agent relays between agents 1 and 3 without contributing a conjugate to the subproblem. For the other three blocks, choose $\mathcal{C}_r = \mathcal{V}_r$; panels (e),(f) illustrate the Laplacian and incidence cases. Taking $\mathcal{G}_r = \mathcal{G}[\mathcal{C}_r]$ gives the whole path for the first three blocks and the single edge e_1 for the last, so that $\mathbf{B}_{\mathcal{G}_r} = \mathbf{B}_{:,e_1}$ there. The corresponding factorizations from Proposition 2 are

$$\begin{aligned} (\mathbf{A}_{\text{ex}})_{:,1} &= \mathbf{B}_{:,e_1} + \mathbf{B}_{:,e_2}, & (\mathbf{A}_{\text{ex}})_{:,2} &= \mathbf{B}_{:,e_1} - \mathbf{B}_{:,e_2}, \\ \mathbf{L}_{:,2} &= \frac{1}{3}(-\mathbf{B}_{:,e_1} + \mathbf{B}_{:,e_2}), & \mathbf{B}_{:,e_1} &= \mathbf{B}_{\mathcal{G}_r}. \end{aligned} \quad (8)$$

The matrices M_r are respectively $(1, 1)$, $(1, -1)$, $\frac{1}{3}(-1, 1)$, and (1) ; their entries label the physical edges in Figure 2(d)–(f) for the illustrated blocks. In the first identity, the two columns of $\mathbf{B}$ have opposite entries at agent 2, which cancel. Thus $f_2^*(-2Y_2)$ is independent of W : agent 2 is absent from the block factor graph but present in the physical subgraph as a relay. The second and third combinations have nonzero entries at all three agents, so every agent contributes a conjugate to the response. For the last block, $M_r = (1)$ uses only e_1 , whose endpoint agents are neighbors and communicate directly. Here f_3^* is absent because e_1 is not incident to agent 3, rather than through cancellation.

3 Graph Tearing Along Edges: Message Passing on Tree Blocks

Proposition 1 holds for every faithful $\mathbf{A}$ and every block cover, while Proposition 2 characterizes the admissible carriers (i.e., satisfying (6)). The framework leaves these choices and the methods for computing the block responses open, with different communication and computation costs. Choices range from Laplacian-type representations in which a coordinate couples an agent with all its neighbors, to the incidence representation, in which a coordinate couples one pair of neighbors.

We now use this freedom to design a decentralized algorithm, termed GATE (Graph-Tearing message passing). We choose a graph decomposition in which *selecting dual coordinates amounts to selecting edges of $\mathcal{G}$* , so that each coordinate appears in the conjugates of two neighboring agents, and a block response can be computed by exchanges between the neighboring agents whose conjugates enter it, without relay agents. More formally, we have the following.

- (i) **Representation:** We choose $\mathbf{A}$ so that, for each block, the factorization (7) admits a matrix M_r with *one nonzero entry per row* and with *different* rows selecting *distinct* edges. Each coordinate then enters exactly the conjugates at the two endpoints of its selected edge, avoiding redundant coordinates on the same edge. Let $\mathcal{E}_r$ denote the set of these edges. Indexing the coordinates by $\mathcal{E}_r$ and absorbing each nonzero coefficient into its coordinate (a change of variables in the dual (D)) gives $\mathbf{A}_{:, \mathcal{J}_r} = \mathbf{B}_{:, \mathcal{E}_r}$.
- (ii) **Partition:** We retain one coordinate for every physical edge, leaving the whole network available for block selection. We partition the edge-set $\mathcal{E}$ into nonempty sets $\mathcal{P} := \{\mathcal{E}_r\}_{r=1}^p$, so that each variable receives one block update. Write $\mathcal{T}_r := (\mathcal{V}_r, \mathcal{E}_r)$ for the subgraph selected by block r ; by Definition 2, $\mathcal{V}_r$ is the set of endpoints of $\mathcal{E}_r$. We take each $\mathcal{T}_r$ to be connected, without loss of generality (the response (3) of a disconnected $\mathcal{T}_r$ separates into independent responses on its connected components, which can be taken as separate blocks).
- (iii) **Carriers:** We use only the agents whose conjugates enter the response: take $\mathcal{C}_r := \mathcal{V}_r$. Connectivity of $\mathcal{T}_r$ makes $\mathcal{G}[\mathcal{V}_r]$ connected, so the carrier conditions (6) hold, and no additional relay agents are needed.

The three design choices are summarized by

$$\mathbf{A}_{:, \mathcal{J}_r} = \mathbf{B}_{:, \mathcal{E}_r}, \quad \bigsqcup_{r=1}^p \mathcal{E}_r = \mathcal{E}, \quad \mathcal{C}_r = \mathcal{V}_r, \quad r \in [p]. \quad (9)$$

After ordering the columns, $\mathbf{A} = \mathbf{B}$, with $\ker(\mathbf{B}^\top) = \text{span}\{\mathbf{1}\}$ (under Assumption 2); hence $\mathbf{A}$ is faithful. Since $\{\mathcal{E}_r\}_{r=1}^p$ is a block cover and (6) holds with $\mathcal{C}_r = \mathcal{V}_r$, (9) defines a graph decomposition over $\mathcal{G}$ (Definition 3): the block factor graph $\mathcal{F}_r$ is the node-edge incidence graph of $\mathcal{T}_r$, that is, the bipartite graph on $\mathcal{V}_r \sqcup \mathcal{E}_r$ in which i and e are adjacent exactly when $i \in \mathcal{I}_e$; and (7) holds with $\mathcal{G}_r = \mathcal{T}_r$ and $M_r = \mathbf{I}$. The partition remains a design choice: finer partitions give smaller block subproblems but may need more rounds to resolve the coupling between blocks. The convergence analysis in Sec. 4 will quantify the effect of the partition on the rate and guide this computation-communication tradeoff.

The remainder of this section derives the algorithm GATE: the structure of the blocks is specified, and the block fixed-point conditions of Proposition 1 are rewritten as relations for individual edge variables and local messages. Their relaxed parallel iteration yields GATE.

3.1 Tree-block responses and message equations

Fix a block $r \in [p]$ and the exterior coordinates $Y_{\bar{\mathcal{E}}_r}$, where $\bar{\mathcal{E}}_r := \mathcal{E} \setminus \mathcal{E}_r$; under (9), the conditional response becomes

$$\text{BR}_r(Y_{\bar{\mathcal{E}}_r}) = \underset{W \in \mathbb{R}^{|\mathcal{E}_r| \times d}}{\text{argmin}} \sum_{i \in \mathcal{V}_r} f_i^*(\mathbf{B}_{i, \mathcal{E}_r} W + \mathbf{B}_{i, \bar{\mathcal{E}}_r} Y_{\bar{\mathcal{E}}_r}). \quad (10)$$

The conjugates in (10) are held by different agents. We seek subgraphs $\mathcal{T}_r$ that enable the computation of BR_r through exchanges between neighboring agents only. Fixing the coordinate Y_e separates the remaining minimization in (10) into two independent subproblems *whenever removing e disconnects $\mathcal{T}_r$* . **Trees** permit this separation at *every* edge. They also eliminate coordinate redundancy: minimizing over the $|\mathcal{V}_r| - 1$ edge variables of a spanning tree, while fixing those on the other edges of $\mathcal{T}_r$, gives the same optimal value as minimizing over the entire block. This is established in the following proposition.

Proposition 3 *Let $\mathcal{T}_r$ be connected and let $\mathcal{E}'_r$ be the edge set of any spanning tree of $\mathcal{T}_r$. Then $\text{Range}(\mathbf{B}_{:, \mathcal{E}'_r}) = \text{Range}(\mathbf{B}_{:, \mathcal{E}_r})$ and $\ker(\mathbf{B}_{:, \mathcal{E}'_r}) = \{0\}$. Furthermore, under Assumption 1, the response $\text{BR}_r(Y_{\bar{\mathcal{E}}_r})$ is a singleton for every fixed $Y_{\bar{\mathcal{E}}_r}$ if and only if $\mathcal{T}_r$ is a tree.*

We therefore choose each $\mathcal{T}_r$ to be a tree.

Assumption 3 *For every block $r \in [p]$, the associated subgraph $\mathcal{T}_r$ is a tree.*

We now express the block-optimality conditions (4) through quantities that the two endpoints of each edge can form and exchange. To distinguish edges inside and outside the block, define, for $i \in \mathcal{V}_r$,

$$\begin{aligned} \mathcal{A}_i^\circ &:= \mathcal{A}_i \cap \mathcal{E}_r, & \mathcal{A}_i^\partial &:= \mathcal{A}_i \cap \bar{\mathcal{E}}_r, \\ \mathcal{V}_r^\circ &:= \{i \in \mathcal{V}_r : \mathcal{A}_i^\partial = \emptyset\}, & \mathcal{V}_r^\partial &:= \mathcal{V}_r \setminus \mathcal{V}_r^\circ. \end{aligned}$$

Relative to $\mathcal{T}_r$, agents in $\mathcal{V}_r^\circ$ are *interior*: all their incident edges belong to $\mathcal{E}_r$. Agents in $\mathcal{V}_r^\partial$ are *boundary*: they also have incident edges outside $\mathcal{E}_r$. In (10), only the conjugate terms of boundary agents depend on the exterior coordinates; these terms may also depend on several block variables.

Splitting at the edge. For each block r , uniqueness of the conditional response (Proposition 3) gives the following equivalent conditions:

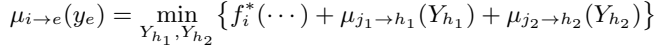
$$\begin{aligned} Y_{\mathcal{E}_r}^* &\in \text{BR}_r(Y_{\bar{\mathcal{E}}_r}^*) \\ \iff Y_e^* &\in \underset{y_e \in \mathbb{R}^{1 \times d}}{\text{argmin}} \left\{ \min_{\substack{Y_{\mathcal{E}_r} \\ Y_e = y_e}} \sum_{i \in \mathcal{V}_r} f_i^*(\mathbf{B}_{i, \mathcal{E}_r} Y_{\mathcal{E}_r} + \mathbf{B}_{i, \bar{\mathcal{E}}_r} Y_{\bar{\mathcal{E}}_r}^*) \right\} \quad \forall e \in \mathcal{E}_r. \end{aligned} \quad (11)$$

By Proposition 1, Y^* solves (D) if and only if these conditions hold for every block. To decompose the inner minimization in (11), fix $Y_e = y_e$. Removing e from $\mathcal{T}_r$ separates this minimization into two decoupled subproblems, one on each resulting connected subgraph. We solve them recursively from the leaves toward e .

For each endpoint $i \in \mathcal{I}_e$ of e , let $\mathcal{V}_e(i)$ and $\mathcal{E}_e(i)$ denote the vertex and edge sets of the resulting connected subgraph containing i . Figure 3(a) illustrates the two resulting subgraphs. Every block edge other than e lies in one of these two subgraphs. Thus, with $Y_e = y_e$, no remaining block variable enters conjugates from both subgraphs, and

$$\begin{aligned} &\min_{\substack{Y_{\mathcal{E}_r} \in \mathbb{R}^{|\mathcal{E}_r| \times d} \\ Y_e = y_e}} \sum_{k \in \mathcal{V}_r} f_k^*(\mathbf{B}_{k, \mathcal{E}_r} Y_{\mathcal{E}_r} + \mathbf{B}_{k, \bar{\mathcal{E}}_r} Y_{\bar{\mathcal{E}}_r}^*) \\ &= \sum_{i \in \mathcal{I}_e} \underbrace{\min_{Y_{\mathcal{E}_e(i)} \in \mathbb{R}^{|\mathcal{E}_e(i)| \times d}} \sum_{k \in \mathcal{V}_e(i)} f_k^*(\mathbf{B}_{k, \mathcal{E}_e(i)} Y_{\mathcal{E}_e(i)} + \mathbf{B}_{k, e} y_e + \mathbf{B}_{k, \bar{\mathcal{E}}_r} Y_{\bar{\mathcal{E}}_r}^*)}_{=: \mu_{i \rightarrow e}(y_e)}. \end{aligned} \quad (12)$$

The sum over $\mathcal{I}_e$ in (12) reads $\mu_{i \rightarrow e}(y_e) + \mu_{i' \rightarrow e}(y_e)$, where i and i' are the two endpoints of e ; $\mu_{i \rightarrow e}(y_e)$ [resp. $\mu_{i' \rightarrow e}(y_e)$] represents the cost-to-go for the subgraph $(\mathcal{V}_e(i), \mathcal{E}_e(i))$ [resp. $(\mathcal{V}_e(i'), \mathcal{E}_e(i'))$]; together, they determine Y_e^* through (11). We call them *messages*, as this information is exchanged between the endpoints of e ; the arrows in Figure 3(a) indicate the two messages. Their dependence on $Y_{\bar{\mathcal{E}}_r}^*$ is suppressed for notational convenience.



Computing $\mu_{i \rightarrow e}$ via recursive elimination along the subtrees. We evaluate the minimization defining $\mu_{i \rightarrow e}(y_e)$ by successively eliminating the edge variables $Y_{\mathcal{E}_e(i)}$ of the subgraph $(\mathcal{V}_e(i), \mathcal{E}_e(i))$, proceeding from its leaves toward e (through the endpoint i), while $Y_e = y_e$ and $Y_{\bar{\mathcal{E}}_e}^*$ remain fixed.

$$\mu_{i \rightarrow e}(y_e) = f_i^*(\mathbf{B}_{i,e} y_e + \mathbf{B}_{i,\bar{\mathcal{E}}_r} Y_{\bar{\mathcal{E}}_r}^*). \quad (13)$$

(ii) *Remaining agents.* Otherwise, $|\mathcal{V}_e(i)| > 1$. Removing i from $(\mathcal{V}_e(i), \mathcal{E}_e(i))$ leaves $|\mathcal{A}_i^\circ| - 1$ connected subgraphs, one for each edge $h \in \mathcal{A}_i^\circ \setminus \{e\}$. For each such h , let j be its endpoint other than i , so that $\{j\} = \mathcal{I}_h \setminus \{i\}$. The resulting subgraph containing j is $(\mathcal{V}_h(j), \mathcal{E}_h(j))$: its agents k are exactly those whose unique path from i to k in $\mathcal{T}_r$ begins with h . As h varies over $\mathcal{A}_i^\circ \setminus \{e\}$, the sets $\mathcal{V}_h(j)$ are pairwise disjoint and, together with $\{i\}$, form a partition of $\mathcal{V}_e(i)$. The corresponding edge sets $\{h\} \cup \mathcal{E}_h(j)$ partition $\mathcal{E}_e(i)$: any additional edge would join two distinct subgraphs

$(\mathcal{V}_h(j), \mathcal{E}_h(j))$ and create a cycle in $\mathcal{T}_r$. Hence,

$$\begin{aligned}\mathcal{V}_e(i) &= \{i\} \sqcup \bigsqcup_{\substack{h \in \mathcal{A}_i^\circ \setminus \{e\} \\ j \in \mathcal{I}_h \setminus \{i\}}} \mathcal{V}_h(j), \\ \mathcal{E}_e(i) &= \bigsqcup_{\substack{h \in \mathcal{A}_i^\circ \setminus \{e\} \\ j \in \mathcal{I}_h \setminus \{i\}}} (\{h\} \sqcup \mathcal{E}_h(j)).\end{aligned}\tag{14}$$

We proceed by induction on $|\mathcal{V}_e(i)|$, starting from the leaf case (13). For each $h \in \mathcal{A}_i^\circ \setminus \{e\}$, let $\{j\} = \mathcal{I}_h \setminus \{i\}$. Since $|\mathcal{V}_h(j)| < |\mathcal{V}_e(i)|$, the induction hypothesis ensures that $\mu_{j \rightarrow h}$ is finite-valued and strongly convex, and its defining minimum is uniquely attained for every fixed Y_h .

In the minimization defining $\mu_{i \rightarrow e}(y_e)$ in (12), first hold $Y_{\mathcal{A}_i^\circ \setminus \{e\}}$ fixed. By (14), the sum of conjugates on each $\mathcal{V}_h(j)$ can be minimized independently over $Y_{\mathcal{E}_h(j)}$, giving $\mu_{j \rightarrow h}(Y_h)$, since every block edge of an agent of $\mathcal{V}_h(j)$ lies in $\{h\} \cup \mathcal{E}_h(j)$. Minimizing the sum of these messages and agent i 's conjugate over $Y_{\mathcal{A}_i^\circ \setminus \{e\}}$, while $Y_e = y_e$ and $Y_{\bar{\mathcal{E}}_r}^*$ remain fixed, gives

$$\begin{aligned}\mu_{i \rightarrow e}(y_e) &= \min_{Y_{\mathcal{A}_i^\circ \setminus \{e\}}} \left\{ f_i^* (\mathbf{B}_{i,e} y_e + \mathbf{B}_{i, \mathcal{A}_i^\circ \setminus \{e\}} Y_{\mathcal{A}_i^\circ \setminus \{e\}} + \mathbf{B}_{i, \bar{\mathcal{E}}_r} Y_{\bar{\mathcal{E}}_r}^*) \right. \\ &\quad \left. + \sum_{h \in \mathcal{A}_i^\circ \setminus \{e\}} \sum_{j \in \mathcal{I}_h \setminus \{i\}} \mu_{j \rightarrow h}(Y_h) \right\}, \quad i \in \mathcal{V}_r, \quad e \in \mathcal{A}_i^\circ.\end{aligned}\tag{15}$$

Figure 3(b) illustrates this recursion for two edges incident to i other than e .

The objective in this recursion is finite-valued and jointly strongly convex in $(y_e, Y_{\mathcal{A}_i^\circ \setminus \{e\}})$. For each y_e , its minimum is therefore uniquely attained, and partial minimization yields a finite-valued, strongly convex $\mu_{i \rightarrow e}$. Once the minimizing values Y_h , $h \in \mathcal{A}_i^\circ \setminus \{e\}$, are fixed, the induction hypothesis uniquely determines the minimizing $Y_{\mathcal{E}_h(j)}$, where $\{j\} = \mathcal{I}_h \setminus \{i\}$. By (14), these variables constitute the unique minimizer $Y_{\mathcal{E}_e(i)}$ in (12), completing the induction. The same induction shows that any family satisfying (15) coincides with the messages defined in (12).

Proposition 4 (Message fixed points) *Under Assumptions 1, 2, and 3, consider the decomposition (9). For $Y^* \in \mathbb{R}^{|\mathcal{E}| \times d}$, let $\mu_{i \rightarrow e}$ be the messages defined in (12), with $Y_{\bar{\mathcal{E}}_r}^*$ fixed in each block. Then Y^* solves (D) if and only if*

$$\{Y_e^*\} = \operatorname{argmin}_{y_e \in \mathbb{R}^{1 \times d}} \sum_{i \in \mathcal{I}_e} \mu_{i \rightarrow e}(y_e), \quad e \in \mathcal{E}_r, \quad r \in [p].\tag{16}$$

The message recursion (15) and the edge conditions (16) form the fixed-point system whose relaxed parallel iteration defines GATE. This is formalized next.

3.2 The node-owned GATE algorithm

GATE iterates the edge conditions (16) and the recursion (15), with relaxed edge updates, as summarized in Algorithm 1. To avoid an edge-level coordinator, an edge-node assignment is introduced, $\mathbf{a} : \mathcal{E} \rightarrow \mathcal{V}$ with $\mathbf{a}(e) \in \mathcal{I}_e$. Define $\mathcal{A}_i^{\mathbf{a}} := \{e \in \mathcal{A}_i : \mathbf{a}(e) = i\}$. These sets are pairwise disjoint and cover $\mathcal{E}$. Node i updates Y_e for $e \in \mathcal{A}_i^{\mathbf{a}}$,

Algorithm 1: GATE: Graph-Tearing message passing

Input: Edge assignment $\mathbf{a}$ with ownership sets $\{\mathcal{A}_i^{\mathbf{a}}\}_{i \in \mathcal{V}}$; edge partition $\{\mathcal{E}_r\}_{r=1}^p$; initial edge variables Y^0 ; finite-valued strongly convex initial messages $\mu_{i \rightarrow e}^0$, $e \in \mathcal{E}$, $i \in \mathcal{I}_e$; relaxation parameters $\tau_r^\nu \in [0, 1]$; $N \geq 1$.

for $\nu = 0, 1, \dots, N-1$ do

Each node $i \in \mathcal{V}$ executes the following steps, in parallel with the other nodes:

(1) Assigned edge-variable update.

foreach $r \in [p]$ and $e \in \mathcal{A}_i^{\mathbf{a}} \cap \mathcal{E}_r$, compute

$$\hat{Y}_e^{\nu+1} \in \operatorname{argmin}_{y_e \in \mathbb{R}^{1 \times d}} \left\{ \sum_{j \in \mathcal{I}_e} \mu_{j \rightarrow e}^\nu(y_e) \right\}, \quad (17a)$$

$$Y_e^{\nu+1} = Y_e^\nu + \tau_r^\nu (\hat{Y}_e^{\nu+1} - Y_e^\nu). \quad (17b)$$

Send $Y_e^{\nu+1}$ to the other endpoint of e .

(2) Node-to-edge message update: foreach $r \in [p]$ and $e \in \mathcal{A}_i \cap \mathcal{E}_r$, compute

$$\begin{aligned} \mu_{i \rightarrow e}^{\nu+1}(y_e) = & \min_{Y_{\mathcal{A}_i^{\mathbf{a}} \setminus \{e\}}} \left\{ f_i^* \left(\mathbf{B}_{i,e} y_e + \mathbf{B}_{i, \mathcal{A}_i^{\mathbf{a}} \setminus \{e\}} Y_{\mathcal{A}_i^{\mathbf{a}} \setminus \{e\}} + \mathbf{B}_{i, \mathcal{A}_i^{\mathbf{a}}} Y_{\mathcal{A}_i^{\mathbf{a}}}^{\nu+1} \right) \right. \\ & \left. + \sum_{h \in \mathcal{A}_i^{\mathbf{a}} \setminus \{e\}} \sum_{j \in \mathcal{I}_h \setminus \{i\}} \mu_{j \rightarrow h}^\nu(Y_h) \right\}. \end{aligned} \quad (18)$$

Send $\mu_{i \rightarrow e}^{\nu+1}$ to the other endpoint of e .

Primal recovery. Each node $i \in \mathcal{V}$ computes $x_i^N := \nabla f_i^*((\mathbf{B}Y^N)_i)$.

while both endpoint-nodes retain a copy. For each edge e , r denotes its block index, and $\mathcal{A}_i^{\mathbf{a}}$, $\mathcal{A}_i^{\mathbf{b}}$ are understood relative to that block.

GATE is fully decentralized: each agent computes locally and communicates only with its neighbors, without a central coordinator. This is discussed next.

Initialization. For each edge e , its owner $\mathbf{a}(e)$ chooses Y_e^0 arbitrarily and sends it to the other endpoint. Each node i then selects a finite-valued strongly convex message $\mu_{i \rightarrow e}^0$ for every $e \in \mathcal{A}_i$ and sends it to the other endpoint of e . For instance, one local choice is $\mu_{i \rightarrow e}^0(y_e) = f_i^*(\mathbf{B}_{i,e} y_e + \mathbf{B}_{i, \mathcal{A}_i \setminus \{e\}} Y_{\mathcal{A}_i \setminus \{e\}}^0)$.

Local computations and communications. Step (1): Each node updates only its assigned edges, using the two endpoint messages at round ν . Relaxation damps the edge update, which is then shared with the other endpoint. **Step (2):** Each node updates a message on every incident edge, whether or not it owns that edge. The local minimization uses its conjugate and the incoming messages μ^ν , with incident exterior coordinates fixed at the newly available $Y^{\nu+1}$. Its minimizing variables are local to this computation and do not overwrite the stored edge iterates. The outgoing messages are sent to the other endpoints for round $\nu+1$.

Nodes work in parallel within each step, once the required data have arrived. Each round advances every message once and sends one d -vector and two functions per edge. Primal recovery uses only the incident edge coordinates.

Remark 1 (Relation to [5]) MP-Jacobi [5] develops message-passing algorithms for nonlinear programs with a prescribed interaction (hyper)graph. Although its hypergraph model can represent the consensus dual for a fixed $\mathbf{A}$, MP-Jacobi cannot be applied directly to obtain GATE or establish its convergence. GATE's boundary-message updates jointly minimize over other in-block variables that MP-Jacobi holds fixed in its inter-cluster factors. Moreover, the convergence theory in [5] requires strong convexity in the optimization variables, which the incidence dual lacks on

cyclic graphs. Beyond these algorithmic and analytical differences, a major departure from [5] is the development here of a general graph-decomposition framework that jointly designs the consensus representation, dual-coordinate blocks, and connected sets of agents hosting their computation; GATE is one instance.

4 Convergence Analysis

In this section, we establish convergence of Algorithm 1. We begin with the quantities the analysis is built on.

4.1 Image variables and block error bound

For the incidence representation $\mathbf{A} = \mathbf{B}$, write $g := g_{\mathbf{B}}$ and define

$$G(U) := \sum_{i \in \mathcal{V}} f_i^*(U_i), \quad U \in \mathbb{R}^{m \times d}.$$

Under Assumption 1, G is strongly convex. On loopy graphs, however, the incidence map is not injective and $g(Y) = G(\mathbf{B}Y)$ is constant along $\ker \mathbf{B}$: it is not strongly convex in the edge variables Y . The analysis therefore runs on the image variable $Z := \mathbf{B}Y$, which ranges over the subspace

$$\mathcal{S} := \{\mathbf{B}Y : Y \in \mathbb{R}^{|\mathcal{E}| \times d}\} \subseteq \mathbb{R}^{|\mathcal{V}| \times d}.$$

Changing $Y_{\mathcal{E}_r}$ alone moves Z along the affine subspace $Z + \mathcal{U}_r$, where

$$\mathcal{U}_r := \{\mathbf{B}_{\mathcal{V}, \mathcal{E}_r} H : H \in \mathbb{R}^{|\mathcal{E}_r| \times d}\} \subseteq \mathcal{S}, \quad r \in [p].$$

For $Z = \mathbf{B}Y \in \mathcal{S}$, the value of the block minimization (10) equals the minimum of G over the slice $Z + \mathcal{U}_r$:

$$\min_{W_r \in \mathbb{R}^{|\mathcal{E}_r| \times d}} g(W_r, Y_{\bar{\mathcal{E}}_r}) = \min_{U \in Z + \mathcal{U}_r} G(U) =: G(O_r(Z)), \quad (19)$$

where $O_r(Z)$ is the unique minimizer, well defined because G is strongly convex and $Z + \mathcal{U}_r$ is closed. Let

$$\Delta_r(Z) := G(Z) - G(O_r(Z)) \geq 0 \quad (20)$$

be the progress of one exact minimization of block r from Z . Let $G^* := \min_{U \in \mathcal{S}} G(U)$, attained because G is strongly convex and $\mathcal{S}$ is closed. The analysis rests on the following condition, which bounds the optimality gap $G(Z) - G^*$ by the progress of all blocks.

Assumption 4 (block error bound) *There exists a partition-dependent constant $\kappa_{\mathcal{P}} < \infty$ such that*

$$G(Z) - G^* \leq \kappa_{\mathcal{P}} \sum_{r=1}^p \Delta_r(Z), \quad Z \in \mathcal{S}. \quad (21)$$

Remark 2 The sum $\sum_{r=1}^p \Delta_r(Z)$ collects the decreases of G obtained by minimizing each block separately from the same Z , with its exterior coordinates fixed. These separate minimizations need not add up to the global gap, because the block subproblems interact through the conjugate terms of their shared agents. Condition (21) controls how much suboptimality can remain relative to the progress they offer, and its smallest constant is the largest value of this ratio over $\mathcal{S}$. Thus $\kappa_{\mathcal{P}}$ quantifies the coupling among the blocks, accounting jointly for the partition and the functions. With the single-block partition $\mathcal{P} = \{\mathcal{E}\}$ on a tree, $\mathcal{U}_1 = \mathcal{S}$, so $\Delta_1(Z) = G(Z) - G^*$ and $\kappa_{\mathcal{P}} = 1$, independently of the condition number of F and graph spectrum.

The coupling measured by $\kappa_{\mathcal{P}}$ admits a *block-preconditioning* interpretation, which becomes explicit for quadratic local functions. With the edge vectors stacked, the dual Hessian is

$$H = [H_{rs}]_{r,s \in [p]} = (\mathbf{B}^\top \otimes I_d) \text{blkdiag}\left((\nabla^2 f_1)^{-1}, \dots, (\nabla^2 f_m)^{-1}\right) (\mathbf{B} \otimes I_d) \succeq 0.$$

Set $M_{\mathcal{P}} := \text{blkdiag}(H_{11}, \dots, H_{pp})$. Each tree incidence block has full column rank, so $H_{rr} \succ 0$. Writing h for the vectorized difference between Y and a dual minimizer, exact block minimization gives $\sum_{r=1}^p \Delta_r(Z) = \frac{1}{2} h^\top H M_{\mathcal{P}}^{-1} H h$, with $Z = \mathbf{B}Y$. Thus the smallest admissible block error-bound constant is

$$\kappa_{\mathcal{P}} = \sup_{h^\top H h > 0} \frac{h^\top H h}{h^\top H M_{\mathcal{P}}^{-1} H h} = \frac{1}{\lambda_{\min}^+(M_{\mathcal{P}}^{-1/2} H M_{\mathcal{P}}^{-1/2})},$$

where $\lambda_{\min}^+$ denotes the smallest positive eigenvalue. The preconditioned Hessian has identity diagonal blocks; its off-diagonal blocks $H_{rr}^{-1/2} H_{rs} H_{ss}^{-1/2}$ capture coupling normalized by the block curvatures.

Generic bounds based only on global curvature and spectral estimates need not capture this partition dependence; exploiting it yields sharper bounds in the regimes studied in Sec. 4.4.

We prove convergence of Algorithm 1 in two steps. **Step 1** establishes an exact delayed block-Jacobi representation of its edge updates, which is a technical contribution of independent interest; **Step 2** controls the delays to prove linear convergence.

4.2 Step 1: GATE as a delayed block-Jacobi scheme

GATE interleaves message and edge updates without waiting for the message recursion to run through a whole tree (no block subproblem is ever solved to completion). Nevertheless, after a finite transient, each proposed edge update coincides with the corresponding coordinate of an exact block minimizer, evaluated at past values of the exterior coordinates, with delays that depend on the updated edge.

To establish it, we expand the messages used in **step 1** of Algorithm 1 through the recursion in **step 2**. Each substitution goes back one round. If the expansion reaches a leaf of $\mathcal{T}_r$ at distance k from the nearer endpoint of e , its message has index $\nu - k$. For $\nu - k \geq 1$, substituting the recursion replaces this message by the leaf's conjugate term; for $\nu - k = 0$, the expansion stops at the arbitrary initial function μ^0 . Since $k \leq \text{diam}(\mathcal{T}_r) - 1$, $\nu \geq \text{diam}(\mathcal{T}_r)$ ensures that no initial message remains. The sum $\sum_{j \in \mathcal{I}_e} \mu_{j \rightarrow e}^\nu(y_e)$ minimized in **step 1** then equals the minimum of the block objective over $Y_{\mathcal{E}_r \setminus \{e\}}$, with $Y_e = y_e$. In each boundary agent's conjugate term, the

exterior coordinates retain the values used when that agent formed its message. The delay associated with a boundary agent i is its distance from the nearer endpoint of the updated edge e :

$$\delta_{e,i} := \min_{j \in \mathcal{I}_e} \text{dist}_{\mathcal{T}_r}(i, j) = \frac{\text{dist}_{\mathcal{F}_r}(e, i) - 1}{2}, \quad e \in \mathcal{E}_r, \quad i \in \mathcal{V}_r^\partial. \quad (22)$$

In particular, $0 \leq \delta_{e,i} \leq \text{diam}(\mathcal{T}_r) - 1$.

The following proposition makes this representation precise.

Proposition 5 *Suppose Assumptions 1 and 3 hold. Then, for every $r \in [p]$, $e \in \mathcal{E}_r$ and $\nu \geq \text{diam}(\mathcal{T}_r)$, the updates of Algorithm 1 read*

$$Y_e^{\nu+1} = Y_e^\nu + \tau_r^\nu (\hat{Y}_e^{\nu+1} - Y_e^\nu), \quad (23a)$$

$$\hat{Y}_e^{\nu+1} = \underset{Y_e}{\text{argmin}} \min_{Y_{\mathcal{E}_r \setminus \{e\}}} \left\{ \sum_{i \in \mathcal{V}_r} f_i^* \left(\mathbf{B}_{i, \mathcal{A}_i^\circ} Y_{\mathcal{A}_i^\circ} + \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} Y_h^{\nu - \delta_{e,i}} \right) \right\}, \quad (23b)$$

where the minimizer in (23b) is unique.

Proof See Appendix A.1. $\square$

4.3 Step 2: Convergence of the delayed block-Jacobi scheme

We now establish convergence for the delayed block-Jacobi scheme in (23). The analysis runs in the image variable $Z = \mathbf{B}Y$, where G is strongly convex, and compares, for each block, the delayed edge responses that the algorithm implements with the exact block response at the current exterior coordinates. Three ingredients give the rate: **(i)** the descent of G along the iterates (Lemma 1), **(ii)** the block error bound (Lemma 2), and **(iii)** a perturbation analysis of the block minimizers, bounding the mismatch caused by the delays (Lemma 4).

Write $Z^\nu := \mathbf{B}Y^\nu$, with Z_i^ν its i -th row, and define, for $r \in [p]$,

$$\begin{aligned} Z_r^\nu &:= \mathbf{B}_{\mathcal{V}, \mathcal{E}_r} Y_{\mathcal{E}_r}^\nu, \\ \hat{Z}_r^{\nu+1} &:= \mathbf{B}_{\mathcal{V}, \mathcal{E}_r} \hat{Y}_{\mathcal{E}_r}^{\nu+1}, \\ \bar{Z}_r^{\nu+1} &:= O_r(Z^\nu) - Z^\nu + Z_r^\nu. \end{aligned} \quad (24)$$

These are the contributions of block r to the image of the current iterate, of the delayed edgewise minimizers (23b), and of the exact minimization at the current exterior coordinates. Applying $\mathbf{B}$ to the relaxed edge updates (23a) gives

$$Z^{\nu+1} = Z^\nu + \sum_{r=1}^p \tau_r^\nu (\hat{Z}_r^{\nu+1} - Z_r^\nu). \quad (25)$$

Blocks sharing an agent i all move the row Z_i , each by its own step τ_r^ν . The admissible range of τ_r^ν is therefore limited by the number of overlapping blocks at each agent, defined as

$$\omega_{\mathcal{P}} := \max_{i \in \mathcal{V}} |\{r \in [p] : i \in \mathcal{V}_r\}|, \quad (26)$$

where $1 \leq \omega_{\mathcal{P}} \leq \min\{p, \max_{i \in \mathcal{V}} |\mathcal{A}_i|\}$. This is formalized next.

(i) Descent of G along the iterates. Each round decreases G by the damped progress of the blocks, up to the error made by using delayed instead of current exterior coordinates. Along $\mathcal{U}_r$ only the conjugates of the agents in $\mathcal{V}_r$ vary, so G is $(1/L_r)$ -strongly convex and $(1/\mu_r)$ -smooth there, with

$$\mu_r := \min_{i \in \mathcal{V}_r} \mu_{f_i}, \quad L_r := \max_{i \in \mathcal{V}_r} L_{f_i}.$$

Lemma 1 *Under Assumption 1, let $0 \leq \tau_r^\nu \leq 1/\omega_{\mathcal{P}}$ for every $r \in [p]$. Then,*

$$G(Z^{\nu+1}) \leq G(Z^\nu) - \sum_{r=1}^p \tau_r^\nu \Delta_r(Z^\nu) + \sum_{r=1}^p \frac{\tau_r^\nu}{2\mu_r} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2, \quad (27)$$

with $\bar{Z}_r^{\nu+1}$ as in (24).

Proof Fix $r \in [p]$; by (24), $Z^\nu - Z_r^\nu + \bar{Z}_r^{\nu+1} = O_r(Z^\nu)$ minimizes G over $Z^\nu + \mathcal{U}_r$, so

$$\langle \nabla G(Z^\nu - Z_r^\nu + \bar{Z}_r^{\nu+1})^\top, H \rangle = 0, \quad H \in \mathcal{U}_r.$$

Since $\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1} \in \mathcal{U}_r$, the $1/\mu_r$ -smoothness of G along $\mathcal{U}_r$ and (20) give

$$G(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}) \leq G(Z^\nu - Z_r^\nu + \bar{Z}_r^{\nu+1}) + \frac{1}{2\mu_r} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2.$$

Next, $(\hat{Z}_r^{\nu+1} - Z_r^\nu)_i = 0$ for every $i \notin \mathcal{V}_r$; so (25) reads

$$Z_i^{\nu+1} = \left(1 - \sum_{r \in [p]: i \in \mathcal{V}_r} \tau_r^\nu\right) Z_i^\nu + \sum_{r \in [p]: i \in \mathcal{V}_r} \tau_r^\nu \left(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}\right)_i. \quad (28)$$

Since $\tau_r^\nu \leq 1/\omega_{\mathcal{P}}$, applying Jensen's inequality to (28) yields

$$f_i^*(Z_i^{\nu+1}) \leq \left(1 - \sum_{r \in [p]: i \in \mathcal{V}_r} \tau_r^\nu\right) f_i^*(Z_i^\nu) + \sum_{r \in [p]: i \in \mathcal{V}_r} \tau_r^\nu f_i^*\left((Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1})_i\right).$$

For $i \notin \mathcal{V}_r$, $(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1})_i = Z_i^\nu$. Hence,

$$\begin{aligned} G(Z^{\nu+1}) - G(Z^\nu) &= \sum_{i \in \mathcal{V}} \left[f_i^*(Z_i^{\nu+1}) - f_i^*(Z_i^\nu) \right] \\ &\leq \sum_{i \in \mathcal{V}} \sum_{r \in [p]: i \in \mathcal{V}_r} \tau_r^\nu \left[f_i^*\left((Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1})_i\right) - f_i^*(Z_i^\nu) \right] \\ &= \sum_{r=1}^p \tau_r^\nu \left[G(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}) - G(Z^\nu) \right]. \end{aligned}$$

Combining this inequality with the definition of $\Delta_r(Z^\nu)$ (cf. (20)) yields (27). $\square$

(ii) Block error bound. Assumption 4 converts the block progress in (27) into a contraction of the global gap. Only half of the progress is spent on it; the other half is kept as a negative term in the block increments $Z_r^{\nu+1} - Z_r^\nu$, instrumental in controlling the adversarial positive term in the mismatch $\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}$.

Lemma 2 *Under Assumptions 1 and 4, let $0 < \tau_r^\nu \leq 1/\omega_{\mathcal{P}}$ for every $r \in [p]$, and $\underline{\tau}^\nu := \min_{r \in [p]} \tau_r^\nu$. Then,*

$$\begin{aligned} G(Z^{\nu+1}) - G^* &\leq \left(1 - \frac{\underline{\tau}^\nu}{2\kappa_{\mathcal{P}}}\right) (G(Z^\nu) - G^*) - \sum_{r=1}^p \frac{1}{8L_r\tau_r^\nu} \|Z_r^{\nu+1} - Z_r^\nu\|^2 \\ &\quad + \sum_{r=1}^p \tau_r^\nu \frac{2/\mu_r + 1/L_r}{4} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2. \end{aligned} \quad (29)$$

Proof Split the block progress $\sum_{r=1}^p \tau_r^\nu \Delta_r(Z^\nu)$ in (27) into two halves. For the first, Assumption 4 and $\tau_r^\nu \geq \underline{\tau}^\nu$ give

$$\frac{1}{2} \sum_{r=1}^p \tau_r^\nu \Delta_r(Z^\nu) \geq \frac{\underline{\tau}^\nu}{2\kappa_{\mathcal{P}}} (G(Z^\nu) - G^*).$$

For the second, optimality of $O_r(Z^\nu)$ over $Z^\nu + \mathcal{U}_r$ and $1/L_r$ -strong convexity along $\mathcal{U}_r$ give

$$\Delta_r(Z^\nu) \geq \frac{1}{2L_r} \|\bar{Z}_r^{\nu+1} - Z_r^\nu\|^2 \geq \frac{1}{4L_r} \|\hat{Z}_r^{\nu+1} - Z_r^\nu\|^2 - \frac{1}{2L_r} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2.$$

Multiply the preceding lower bound by $\tau_r^\nu/2$ and use $Z_r^{\nu+1} - Z_r^\nu = \tau_r^\nu (\hat{Z}_r^{\nu+1} - Z_r^\nu)$, due to (23a). Substituting both estimates into Lemma 1 gives (29). $\square$

(iii) Perturbation analysis of the block minimizers. It remains to bound $\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2$ in Lemma 2. The contribution $\bar{Z}_r^{\nu+1}$ uses the current values Y_h^ν of the exterior coordinates. By Proposition 5, each coordinate used to form $\hat{Z}_r^{\nu+1}$ instead comes from a block minimization with edge-dependent delayed exterior coordinates $Y_h^{\nu-\delta_{e,i}}$. We first quantify the sensitivity of the block minimizer to the exterior parameters, then bound the mismatch in terms of changes in the block contributions to Z over previous iterations.

Exterior parameters and delays. For block r , let ∂_r collect the exterior edges incident to its agents, and let $\mathcal{M}_r$ collect the blocks containing these edges:

$$\partial_r := \bigcup_{i \in \mathcal{V}_r} \mathcal{A}_i^\partial, \quad \mathcal{M}_r := \{s \in [p] \setminus \{r\} : \mathcal{E}_s \cap \partial_r \neq \emptyset\}.$$

The largest delay affecting the block is

$$D_r := \begin{cases} \max_{e \in \mathcal{E}_r, i \in \mathcal{V}_r^\partial} \delta_{e,i}, & \partial_r \neq \emptyset, \\ 0, & \partial_r = \emptyset. \end{cases}$$

When $\partial_r \neq \emptyset$,

$$\frac{1}{2} \text{diam}(\mathcal{T}_r) - 1 \leq D_r \leq \text{diam}(\mathcal{T}_r) - 1.$$

An exterior edge h appears in the block subproblem inside the conjugate f_i^* of each of its endpoints $i \in \mathcal{V}_r$, through the term $\mathbf{B}_{i,h} Y_h$. Index these appearances by

$$\hat{\partial}_r := \{(i, h) : i \in \mathcal{V}_r^\partial, h \in \mathcal{A}_i^\partial\}.$$

An exterior edge incident to two agents of $\mathcal{V}_r$ therefore has two entries in $\hat{\partial}_r$, since its value can appear at different delays in their conjugate terms.

Sensitivity to the exterior parameters. For each $E \in \mathbb{R}^{|\widehat{\partial}_r| \times d}$, with rows $E_{i,h}$ indexed by the boundary incidences $(i, h) \in \widehat{\partial}_r$, define the block solution

$$Y_r(E) := \operatorname{argmin}_{Y_{\mathcal{E}_r} \in \mathbb{R}^{|\mathcal{E}_r| \times d}} \sum_{i \in \mathcal{V}_r} f_i^* \left(\mathbf{B}_{i, \mathcal{A}_i^\circ} Y_{\mathcal{A}_i^\circ} + \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} E_{i,h} \right). \quad (30)$$

The minimizer exists and is unique.

We next quantify how $Y_r(E)$ and its contribution $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_r(E)$ to Z vary with the exterior parameters E . We first express the latter in terms of E .

Let

$$\mathbb{R}^d \ni \eta \mapsto \chi_r(\eta) := \left(\sum_{i \in \mathcal{V}_r} \nabla f_i \right)^{-1}(\eta),$$

which is well defined on $\mathbb{R}^d$ because $\sum_{i \in \mathcal{V}_r} \nabla f_i$ is strongly monotone and continuous. Write $\eta(E) := \sum_{(j,h) \in \widehat{\partial}_r} \mathbf{B}_{j,h} E_{j,h}$ for the sum of the exterior contributions. At the minimizer (30) all agents of the block recover the same primal point. Lemma 9 in Appendix A.2 identifies it as $\chi_r(\eta(E))$:

$$\nabla f_i^* \left(\mathbf{B}_{i, \mathcal{E}_r} Y_r(E) + \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} E_{i,h} \right) = \chi_r(\eta(E)), \quad i \in \mathcal{V}_r,$$

which, using $\nabla f_i^* = (\nabla f_i)^{-1}$, is equivalent to

$$\mathbf{B}_{i, \mathcal{E}_r} Y_r(E) = \nabla f_i(\chi_r(\eta(E))) - \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} E_{i,h}, \quad i \in \mathcal{V}_r. \quad (31)$$

For $v \neq i$, a perturbation of $E_{i,h}$ changes $\mathbf{B}_{v, \mathcal{E}_r} Y_r(E)$ only through $\nabla f_v \circ \chi_r$.

We now use (31) to describe the dependence of the rows $(Y_r(E))_e$, $e \in \mathcal{E}_r$, on the exterior parameters. Removing $e \in \mathcal{E}_r$ from $\mathcal{T}_r$ gives two connected subgraphs. For $i \in \mathcal{V}_r$, let $\mathcal{V}_e(i)$ denote the vertex set of the subgraph containing i and set

$$\overline{\mathcal{V}}_e(i) := \mathcal{V}_r \setminus \mathcal{V}_e(i).$$

This extends the notation of (12) to agents that need not be endpoints of e . Summing (31) over $v \in \overline{\mathcal{V}}_e(i)$ gives the following identity, since every column of $\mathbf{B}_{\overline{\mathcal{V}}_e(i), \mathcal{E}_r}$ sums to zero except that of e , whose sum is ± 1 :

$$(Y_r(E))_e = \pm \sum_{v \in \overline{\mathcal{V}}_e(i)} \left[\nabla f_v(\chi_r(\eta(E))) - \sum_{h' \in \mathcal{A}_v^\partial} \mathbf{B}_{v,h'} E_{v,h'} \right], \quad (32)$$

with the sign determined by the orientation of e . A perturbation of $E_{i,h}$ leaves $\sum_{h' \in \mathcal{A}_v^\partial} \mathbf{B}_{v,h'} E_{v,h'}$ unchanged for every $v \in \overline{\mathcal{V}}_e(i)$, since $v \neq i$. It therefore affects $(Y_r(E))_e$ only through $\sum_{v \in \overline{\mathcal{V}}_e(i)} \nabla f_v(\chi_r(\eta(E)))$.

Equations (31) and (32) involve the maps $\nabla f_v \circ \chi_r$ and their sums. To treat them in a unified fashion, define, for nonempty $S \subseteq \mathcal{V}_r$,

$$\mathcal{L}_{r,S}(\eta) := \sum_{v \in S} \nabla f_v(\chi_r(\eta)), \quad \emptyset \neq S \subseteq \mathcal{V}_r.$$

The choice $S = \{v\}$ recovers the individual map $\nabla f_v \circ \chi_r$, whereas $S = \bar{\mathcal{V}}_e(i)$ gives the sum associated with $(Y_r(E))_e$. Given these two relevant subsets, we define

$$\mathcal{R}_r := \{\{v\} : (i, h) \in \hat{\partial}_r, v \in \mathcal{V}_r \setminus \{i\}\} \cup \{\bar{\mathcal{V}}_e(i) : e \in \mathcal{E}_r, (i, h) \in \hat{\partial}_r\}.$$

To express the desired sensitivity bounds using a single constant, we combine the Lipschitz moduli of $\mathcal{L}_{r,S}$, $S \in \mathcal{R}_r$, as follows: we seek a common $\kappa_{\partial,r} \geq 0$ satisfying

$$\text{Lip}(\mathcal{L}_{r,S}) \leq \kappa_{\partial,r} \frac{|S|}{|\mathcal{V}_r|}, \quad \forall S \in \mathcal{R}_r.$$

The normalization keeps the subset size explicit: singletons $S = \{v\}$ carry the factor $1/|\mathcal{V}_r|$, whereas the sets $S = \bar{\mathcal{V}}_e(i)$ carry $|\bar{\mathcal{V}}_e(i)|/|\mathcal{V}_r|$. The smallest $\kappa_{\partial,r} \geq 0$ satisfying the above inequality is

$$\kappa_{\partial,r} := \begin{cases} 0, & \partial_r = \emptyset, \\ \max_{S \in \mathcal{R}_r} \frac{|\mathcal{V}_r|}{|S|} \text{Lip}(\mathcal{L}_{r,S}), & \partial_r \neq \emptyset. \end{cases} \quad (33)$$

Combining the representations (31) and (32) with the definition (33) gives the following sensitivity bounds for the rows of $Y_r(E)$ and the rows indexed by $v \neq i$ of $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_r(E)$, with respect to perturbations of a single row $E_{i,h}$.

Lemma 3 *Suppose Assumptions 1 and 3 hold. If $E, E' \in \mathbb{R}^{|\hat{\partial}_r| \times d}$ differ only in row $(i, h) \in \hat{\partial}_r$, then*

$$\|(Y_r(E') - Y_r(E))_e\| \leq \kappa_{\partial,r} \frac{|\bar{\mathcal{V}}_e(i)|}{|\mathcal{V}_r|} \|E'_{i,h} - E_{i,h}\|, \quad e \in \mathcal{E}_r, \quad (34a)$$

$$\|\mathbf{B}_{v, \mathcal{E}_r}(Y_r(E') - Y_r(E))\| \leq \frac{\kappa_{\partial,r}}{|\mathcal{V}_r|} \|E'_{i,h} - E_{i,h}\|, \quad v \in \mathcal{V}_r \setminus \{i\}. \quad (34b)$$

Proof See Appendix A.2. $\square$

Remark 3 (On $\kappa_{\partial,r}$) Suppose $\partial_r \neq \emptyset$. By Lemma 3, $\kappa_{\partial,r}$ controls how perturbing $E_{i,h}$ affects $(Y_r(E))_e$ and $\mathbf{B}_{v, \mathcal{E}_r} Y_r(E)$ for $v \neq i$. Assumption 1 gives the crude bound $\kappa_{\partial,r} \leq L_r/\mu_r$, which uses only the block's extreme curvature bounds. When $f_i \in C^2$ for $i \in \mathcal{V}_r$, write $H_i(\eta) := \nabla^2 f_i(\chi_r(\eta))$. By the inverse function theorem, χ_r is C^1 with Jacobian $\left(\sum_{j \in \mathcal{V}_r} H_j(\eta)\right)^{-1}$, so

$$\nabla \mathcal{L}_{r,S}(\eta)^\top = \left(\sum_{i \in S} H_i(\eta)\right) \left(\sum_{j \in \mathcal{V}_r} H_j(\eta)\right)^{-1}, \quad \eta \in \mathbb{R}^d,$$

and the mean value inequality gives

$$\kappa_{\partial,r} \leq \sup_{\eta \in \mathbb{R}^d} \max_{S \in \mathcal{R}_r} \frac{|\mathcal{V}_r|}{|S|} \left\| \left(\sum_{i \in S} H_i(\eta)\right) \left(\sum_{j \in \mathcal{V}_r} H_j(\eta)\right)^{-1} \right\|. \quad (35)$$

Eq. (35) reveals an implicit preconditioning effect of exact block minimization: the Hessian sum over S is measured against the aggregate curvature $\sum_{j \in \mathcal{V}_r} H_j(\eta)$ of

the block; curvature shared by all f_i cancels in this quotient, which is why $\kappa_{\partial,r}$ can be much smaller than L_r/μ_r , as the quadratic case below shows.

Suppose $f_i(x) = \frac{1}{2}x^\top Q_i x - c_i^\top x$, $Q_i \succ 0$, $i \in \mathcal{V}_r$. Then (35) holds with equality:

$$\kappa_{\partial,r} = \max \left\{ \max_{\substack{e \in \mathcal{E}_r \\ (i,h) \in \widehat{\partial}_r}} \frac{|\mathcal{V}_r|}{|\mathcal{V}_e(i)|} \left\| \left(\sum_{j \in \mathcal{V}_e(i)} Q_j \right) \left(\sum_{j \in \mathcal{V}_r} Q_j \right)^{-1} \right\|, \max_{\substack{(i,h) \in \widehat{\partial}_r \\ v \in \mathcal{V}_r \setminus \{i\}}} |\mathcal{V}_r| \left\| Q_v \left(\sum_{j \in \mathcal{V}_r} Q_j \right)^{-1} \right\| \right\}.$$

In particular, if $Q_i = Q$ for all $i \in \mathcal{V}_r$, then $\kappa_{\partial,r} = 1$. Thus $\kappa_{\partial,r}$ can equal one while $L_r/\mu_r = \lambda_{\max}(Q)/\lambda_{\min}(Q)$ is arbitrarily large.

Bounding the mismatch $\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2$ caused by delays. We combine Lemma 3 with the tree structure of the delays to bound this mismatch. Fix $r \in [p]$ and $\nu \geq \max_{s \in [p]} \text{diam}(\mathcal{T}_s)$. The exact block response defining $\bar{Z}_r^{\nu+1}$ uses the current exterior coordinates Y_h^ν . The edge coordinates $\hat{Y}_e^{\nu+1}$ forming $\hat{Z}_r^{\nu+1}$ instead come from block minimizations with delayed exterior coordinates $Y_h^{\nu-\delta_{e,i}}$ (Proposition 5). We collect these in the matrices $E^\nu, \hat{E}_e^\nu \in \mathbb{R}^{|\widehat{\partial}_r| \times d}$, $e \in \mathcal{E}_r$, with rows

$$E_{i,h}^\nu := Y_h^\nu, \quad (\hat{E}_e^\nu)_{i,h} := Y_h^{\nu-\delta_{e,i}}, \quad (i,h) \in \widehat{\partial}_r.$$

By (19), (24) and Proposition 5,

$$\bar{Z}_r^{\nu+1} = \mathbf{B}_{\mathcal{V},\mathcal{E}_r} Y_r(E^\nu), \quad \hat{Y}_e^{\nu+1} = (Y_r(\hat{E}_e^\nu))_e.$$

The two parameter matrices E^ν and $\hat{E}_e^\nu$ differ row by row as follows:

$$E_{i,h}^\nu - (\hat{E}_e^\nu)_{i,h} = \sum_{\ell=\nu-\delta_{e,i}}^{\nu-1} (Y_h^{\ell+1} - Y_h^\ell), \quad (36)$$

at most D_r increments per row.

Appendix A.3 proves the following bound on $\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2$ by combining (34a) and (34b) with the tree structure of the delays $\delta_{e,i}$. We remark that the exploitation of the delay structure is essential: a direct Cauchy–Schwarz bound on the squared norm of (36), a sum of up to D_r increments, would deteriorate the bound by a factor D_r .

Lemma 4 *Suppose Assumptions 1 and 3 hold. For every $r \in [p]$ and $\nu \geq \max_{s \in [p]} \text{diam}(\mathcal{T}_s)$,*

$$\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 \leq 4\kappa_{\partial,r}^2 |\widehat{\partial}_r| \sum_{\ell=\nu-D_r}^{\nu-1} \sum_{(i,h) \in \widehat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2, \quad (37)$$

the sum over ℓ being zero when $D_r = 0$.

Proof See Appendix A.3. □

Lemma 4 bounds the positive mismatch term in (29) through squared changes in the dual variables Y_h , $h \in \partial_r$. The negative term in (29), however, involves squared block increments $\|Z_r^{\nu+1} - Z_r^\nu\|^2$. We therefore bound $\|Y_h^{\ell+1} - Y_h^\ell\|^2$ in terms of $\|Z_s^{\ell+1} - Z_s^\ell\|^2$, where $s \in \mathcal{M}_r$ is the block containing the exterior edge h .

The difference $Y_h^{\ell+1} - Y_h^\ell$ is the row indexed by h of $Y_{\mathcal{E}_s}^{\ell+1} - Y_{\mathcal{E}_s}^\ell$. The relation

$$Z_s^{\ell+1} - Z_s^\ell = \mathbf{B}_{\mathcal{V}, \mathcal{E}_s} (Y_{\mathcal{E}_s}^{\ell+1} - Y_{\mathcal{E}_s}^\ell)$$

allows us to recover these edge increments from $Z_s^{\ell+1} - Z_s^\ell$. Since $\mathcal{T}_s$ is a tree, $\mathbf{B}_{\mathcal{V}, \mathcal{E}_s}$ has full column rank, giving

$$Y_{\mathcal{E}_s}^{\ell+1} - Y_{\mathcal{E}_s}^\ell = (\mathbf{B}_{\mathcal{V}, \mathcal{E}_s}^\top \mathbf{B}_{\mathcal{V}, \mathcal{E}_s})^{-1} \mathbf{B}_{\mathcal{V}, \mathcal{E}_s}^\top (Z_s^{\ell+1} - Z_s^\ell). \quad (38)$$

Let $\mathcal{B}_{r,s} := \mathcal{E}_s \cap \partial_r$ collect the edges of block s touching block r ; each $h \in \mathcal{B}_{r,s}$ occurs in $|\mathcal{I}_h \cap \mathcal{V}_r|$ rows of E^ν . With

$$\Lambda_{r,s} := \text{Diag}(|\mathcal{I}_h \cap \mathcal{V}_r| : h \in \mathcal{B}_{r,s})$$

and

$$\Gamma_r := \begin{cases} \max_{s \in \mathcal{M}_r} \lambda_{\max} \left(\Lambda_{r,s}^{1/2} [(\mathbf{B}_{\mathcal{V}, \mathcal{E}_s}^\top \mathbf{B}_{\mathcal{V}, \mathcal{E}_s})^{-1}]_{\mathcal{B}_{r,s}, \mathcal{B}_{r,s}} \Lambda_{r,s}^{1/2} \right), & \mathcal{M}_r \neq \emptyset, \\ 0, & \mathcal{M}_r = \emptyset, \end{cases}$$

(38) gives

$$\sum_{\substack{(i,h) \in \widehat{\partial}_r \\ h \in \mathcal{E}_s}} \|Y_h^{\ell+1} - Y_h^\ell\|^2 = \sum_{h \in \mathcal{B}_{r,s}} |\mathcal{I}_h \cap \mathcal{V}_r| \|Y_h^{\ell+1} - Y_h^\ell\|^2 \leq \Gamma_r \|Z_s^{\ell+1} - Z_s^\ell\|^2. \quad (39)$$

If every $h \in \mathcal{B}_{r,s}$ has exactly one endpoint in $\mathcal{V}_r$, then $\Lambda_{r,s} = I$.

Lemma 5 *Suppose Assumptions 1 and 3 hold. For every $r \in [p]$ and $\nu \geq \max_{s \in [p]} \text{diam}(\mathcal{T}_s)$,*

$$\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 \leq 4\kappa_{\partial,r}^2 |\widehat{\partial}_r| \Gamma_r \sum_{\ell=\nu-D_r}^{\nu-1} \sum_{s \in \mathcal{M}_r} \|Z_s^{\ell+1} - Z_s^\ell\|^2, \quad (40)$$

the sum over ℓ being zero when $D_r = 0$.

Proof Since the sets $\mathcal{B}_{r,s}$, $s \in \mathcal{M}_r$, partition ∂_r , the inner sum in (37) splits over $s \in \mathcal{M}_r$; bounding each part by (39) gives the claim. $\square$

The mismatch is thus bounded by past increments of the neighboring blocks, in the form required by (29).

(iv) Linear convergence. Combining Lemma 5 with Lemma 2 yields the main convergence result.

Theorem 1 Suppose Assumptions 1, 3, and 4 hold, with $p > 1$. Define

$$A := \max_{s \in [p]} \sum_{r \in [p]: D_r > 0, s \in \mathcal{M}_r} \left[\left(\frac{2}{\mu_r} + \frac{1}{L_r} \right) \kappa_{\partial, r}^2 |\widehat{\partial}_r| \Gamma_r \right], \quad D := \max_{r \in [p]} D_r. \quad (41)$$

Let $X^* = \mathbf{1}(x^*)^\top$, with x^* being the minimizer of F in (P). If $\tau_r^\nu \equiv \tau$ and

$$0 < \tau \leq \min \left\{ \frac{1}{\omega_{\mathcal{P}}}, 2\kappa_{\mathcal{P}} \right\}, \quad 16LA\kappa_{\mathcal{P}}\tau \left[\left(1 - \frac{\tau}{2\kappa_{\mathcal{P}}} \right)^{-D} - 1 \right] \leq 1, \quad (42)$$

then, for some constant $C \geq 1$ and all $\nu \geq 2(D+1)$,

$$\|X^\nu - X^*\|^2 \leq C\rho^\nu, \quad \rho := 1 - \frac{\tau}{2\kappa_{\mathcal{P}}}. \quad (43)$$

Proof For each r , choose $i \in \mathcal{V}_r^\partial$. By the definition of D_r , $\text{dist}_{\mathcal{T}_r}(i, v) \leq D_r + 1$ for all $v \in \mathcal{V}_r$, so $\text{diam}(\mathcal{T}_r) \leq 2(D_r + 1) \leq 2(D + 1)$. Lemma 2 and (40), after regrouping delayed increments by block index, give

$$\begin{aligned} G(Z^{\nu+1}) - G^* &\leq \left(1 - \frac{\tau}{2\kappa_{\mathcal{P}}} \right) (G(Z^\nu) - G^*) - \frac{1}{8L\tau} \sum_{r=1}^p \|Z_r^{\nu+1} - Z_r^\nu\|^2 \\ &\quad + \tau A \sum_{\ell=\nu-D}^{\nu-1} \sum_{r=1}^p \|Z_r^{\ell+1} - Z_r^\ell\|^2. \end{aligned}$$

Set $\nu_0 := 2(D+1)$ and define

$$\mathcal{L}^\nu := G(Z^\nu) - G^* + \tau A \sum_{j=1}^D \left(\sum_{k=1}^{D-j+1} \rho^{-k} \right) \sum_{r=1}^p \|Z_r^{\nu-j+1} - Z_r^{\nu-j}\|^2. \quad (44)$$

Conditions in (42) give $8LA\tau^2 \sum_{j=1}^D \rho^{-j} \leq 1$. Then for $\nu \geq \nu_0$,

$$\mathcal{L}^{\nu+1} \leq \rho \mathcal{L}^\nu - \left(\frac{1}{8L\tau} - \tau A \sum_{j=1}^D \rho^{-j} \right) \sum_{r=1}^p \|Z_r^{\nu+1} - Z_r^\nu\|^2 \leq \rho \mathcal{L}^\nu.$$

Thus $G(Z^\nu) - G^* \leq \mathcal{L}^{\nu_0} \rho^{\nu-\nu_0}$, which yields, with $C_0 := \mathcal{L}^{\nu_0} \rho^{-\nu_0}$ and $g^* := \min_Y g(Y) = G^*$,

$$g(Y^\nu) - g^* = G(Z^\nu) - G^* \leq C_0 \rho^\nu. \quad (45)$$

Let Z^* minimize G over $\mathcal{S}$, and let X^* be given by $x_i^* = \nabla f_i^*(Z_i^*)$, $i \in \mathcal{V}$. Since G is $1/L$ -strongly convex on $\mathcal{S}$, and each ∇f_i^* is $1/\mu$ -Lipschitz,

$$\|X^\nu - X^*\|^2 \leq \frac{1}{\mu^2} \|Z^\nu - Z^*\|^2 \leq \frac{2LC_0}{\mu^2} \rho^\nu. \quad (46)$$

Absorbing constants yields (43). $\square$

The second inequality in (42) is implicit in τ . The following corollary gives an explicit sufficient condition.

Corollary 1 Suppose Assumptions 1, 3, and 4 hold, with $p > 1$. If $\tau_r^\nu \equiv \tau$ and

$$0 < \tau \leq \min \left\{ \frac{1}{\omega_{\mathcal{P}}}, \frac{2\kappa_{\mathcal{P}}}{2D+1}, \frac{1}{\sqrt{8L(2D+1)A}} \right\}, \quad (47)$$

then, for some constant $C \in (0, \infty)$ and all $\nu \geq 2(D+1)$,

$$\|X^\nu - X^\star\|^2 \leq C\rho^\nu, \quad \rho := 1 - \frac{\tau}{2\kappa_{\mathcal{P}}}.$$

Proof The first inequality in (42) follows directly from (47). For the second, when $D \geq 1$, the bound $1 - \rho \leq 1/(2D+1)$ yields

$$\sum_{j=1}^D \rho^{-j} \leq \frac{D}{\rho^D} \leq \frac{D}{1-D(1-\rho)} \leq 2D+1.$$

The third bound therefore gives $8LA\tau^2 \sum_{j=1}^D \rho^{-j} \leq 1$, which is equivalent to the second inequality in (42). When $D = 0$, we have $A = 0$, so the second condition in (42) imposes no additional restriction on τ . Theorem 1 applies. $\square$

Remark 4 (On the rate expression) Choosing τ at the upper bound in (47) gives

$$\rho = 1 - \frac{1}{2\kappa_{\mathcal{P}}} \min \left\{ \frac{1}{\omega_{\mathcal{P}}}, \frac{2\kappa_{\mathcal{P}}}{2D+1}, \frac{1}{\sqrt{8L(2D+1)A}} \right\}. \quad (48)$$

For this stepsize and $p > 1$, including the initial $2(D+1)$ rounds used in Theorem 1, the number of rounds to reach $\|X^\nu - X^\star\| \leq \epsilon$ is

$$O\left(D + \max \left\{ \kappa_{\mathcal{P}}\omega_{\mathcal{P}}, D, \kappa_{\mathcal{P}}\sqrt{LDA} \right\} \log \frac{C}{\epsilon}\right).$$

Here C is the constant in Theorem 1. The bound quantifies the interplay between function curvature and network topology through four partition/graph-dependent quantities. Two of them, $\kappa_{\mathcal{P}}$ and A , also depend on function geometry.

- (a) **Block preconditioning, $\kappa_{\mathcal{P}}$.** The parameter $\kappa_{\mathcal{P}}$ measures the coupling remaining between the block subproblems after exact block minimization over the variables grouped by $\mathcal{P}$. The partition determines which interactions are treated jointly; the agents' cost functions determine the progress these minimizations achieve. For fixed admissible τ , a smaller $\kappa_{\mathcal{P}}$ reduces ρ , improving the guaranteed convergence rate.
- (b) **Block overlap, $\omega_{\mathcal{P}}$.** This is the largest number of blocks sharing an agent. It accounts for the overlap among simultaneous block updates. Keeping each agent's incident edges in fewer blocks limits this overlap and relaxes the sufficient restriction $\tau \leq 1/\omega_{\mathcal{P}}$. The restriction depends on *local overlap, not on the total number p of blocks*.
- (c) **Message propagation, D .** The largest delay $D = \max_r D_r$ is determined by distances within each tree from boundary agents to the nearest endpoints of the updated edges. Limiting block diameters bounds these delays. A smaller D reduces the bound on the initial propagation time and relaxes the delay-dependent stepsize restrictions.

- (d) **Sensitivity to outdated information, A .** This parameter measures how sensitive block updates are to outdated information from neighboring blocks. A smaller A reduces the performance penalty of communication delays: parallel block updates can use less damping, yielding a faster convergence rate. Definition (41) combines block curvature, squared response sensitivities $\kappa_{\partial,r}^2$, boundary-incidence counts $|\hat{\partial}_r|$, and neighboring-block geometry through Γ_r . The maximum over s accounts for the largest accumulated weighted influence of one block on its neighbors' updates; the sum includes only neighbors r with $D_r > 0$. Partition design changes these factors through the location of interfaces and the conjugate functions grouped in each block. In the common-Hessian quadratic example of Remark 3, $\kappa_{\partial,r} = 1$, irrespective of Hessian conditioning.

The rate thus guides partition selection for the given functions and network. Unlike tuning gossip weights or communication rounds for prescribed optimization updates, choosing $\mathcal{P}$ provides additional design freedom to adapt the optimization itself to the topology. It changes which variables are optimized jointly and the coupling between the resulting subproblems. The convergence guarantee captures function geometry and topology *jointly*, rather than only through the separately assessed conditioning and mixing parameters of standard gossip bounds.

Remark 5 (Spanning-tree specialization) The framework includes a single-block construction on a spanning tree of any connected graph $\mathcal{G}$. For a spanning tree $\mathcal{T}$, choose

$$A = B_{\mathcal{V}, \mathcal{E}(\mathcal{T})}, \quad \mathcal{J}_1 = \mathcal{E}(\mathcal{T}), \quad \mathcal{C}_1 = \mathcal{V}.$$

This representation retains only the tree-edge variables, all in one block. There are therefore no exterior coordinates or inter-block coupling, so $A = 0$ and we may take $\kappa_{\mathcal{P}} = 1$. Under Assumption 1, Algorithm 1, run on the edges of $\mathcal{T}$ with exact message updates and $\tau_1^V \equiv 1$, returns the exact primal solution after at most $\text{diam}(\mathcal{T}) + 1$ rounds, by Proposition 5. This provides an optimization analogue of finite-time consensus. Despite this finite-round guarantee, a single spanning-tree block *need not be preferable to a multi-block decomposition*. The computational and communication considerations governing this choice are discussed in Sec. 5.5.

4.4 Rate scaling and partition design

We examine how partition selection controls network-size scaling on tree-dominant graphs and rings. Throughout this subsection, we consider connected graph sequences $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m)$, with $|\mathcal{V}_m| = m$. Problem and partition quantities may depend on m ; we suppress this dependence in the notation. All asymptotic bounds are with respect to $m \rightarrow \infty$, with implied constants independent of m .

1) Tree-dominant graphs. For $\kappa_{\mathcal{P}} = O(1)$, (48) reads $(1 - \rho)^{-1} = O(D)$ whenever $\omega_{\mathcal{P}} = O(D)$ and $LA = O(D)$. If instead $\omega_{\mathcal{P}} = O(1)$ and $LA = O(1)$, conditions in (42) admits $\tau = \Theta(\kappa_{\mathcal{P}} \log D/D)$ for sufficiently large D , yielding

$$(1 - \rho)^{-1} = O(D/\log D).$$

We next identify a graph class and partition that attain the latter regime. A graph family is *tree-dominant* if each graph contains a tree subgraph with a number of remaining edges bounded independently of network size. For each graph, we choose the edge partition consisting of the tree edges as one block and each remaining edge as a singleton block. Under the curvature assumptions below, this partition yields the required parameter bounds.

Proposition 6 *For each connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, choose a tree subgraph $\mathcal{T}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ satisfying*

$$|\mathcal{E} \setminus \mathcal{E}_1| = O(1),$$

and the edge partition

$$\mathcal{P} := \{\mathcal{E}_1\} \cup \{\{h\} : h \in \mathcal{E} \setminus \mathcal{E}_1\}.$$

Suppose Assumptions 1 and 3 hold, with $L = O(1)$ and $\mu^{-1} = O(1)$. Assume there exist $R \succ 0$ and $0 < \underline{\alpha} \leq \bar{\alpha}$, with $\bar{\alpha}/\underline{\alpha} = O(1)$, such that every agent i incident to an edge in $\mathcal{E} \setminus \mathcal{E}_1$ satisfies $f_i \in C^2$ and

$$\underline{\alpha}R \preceq \nabla^2 f_i(x) \preceq \bar{\alpha}R, \quad x \in \mathbb{R}^d.$$

For $p > 1$ and sufficiently large D , (42) admits a stepsize fixed across iterations, $\tau = \Theta(\log D/D)$, yielding

$$(1 - \rho)^{-1} = O(D/\log D).$$

Proof See Appendix A.4.

The selected tree has diameter at most $m - 1$, and possibly much smaller. The following examples have $\text{diam}(\mathcal{T}_1) = \Theta(m)$ and $D = \Theta(m)$.

- (a) **Ring.** Choose $m - q$ consecutive edges as one path block, where $1 \leq q = O(1)$, and put each remaining edge in a singleton block. Then $\text{diam}(\mathcal{T}_1) = m - q$. The common-matrix Hessian bounds in Proposition 6 are required only at the $O(q)$ agents incident to the singleton edges.
- (b) **Barbell.** Consider two cliques of bounded size, each containing at least two agents, joined by a path of $\Theta(m)$ edges. Choose this path as $\mathcal{T}_1$ and put each clique edge in a singleton block. The number of edges outside $\mathcal{T}_1$, and hence the number of agents subject to the common-matrix Hessian bounds, is $O(1)$.

Under the curvature assumptions of Proposition 6, both partitions therefore give

$$(1 - \rho)^{-1} = O(m/\log m).$$

Iteration complexity. For the ring and barbell partitions above, including the initial propagation rounds, Theorem 1 gives

$$O\left(m + \frac{m}{\log m} \log \frac{C}{\epsilon}\right)$$

rounds to reach $\|X^\nu - X^*\| \leq \epsilon$, where $0 < \epsilon < 1$ and $C \geq 1$ is an admissible prefactor in (43).

To estimate C , take $\tau = \kappa_{\mathcal{P}} \log D/D$ as in the proof of Proposition 6. With $\nu_0 := 2(D + 1)$,

$$\rho^{-\nu_0} = \left(1 - \frac{\log D}{2D}\right)^{-2(D+1)} = D(1 + o(1)).$$

Thus (46) gives $C = O(1 + D\mathcal{L}^{\nu_0})$, where $\mathcal{L}^{\nu}$ is defined in (44). For sequences of problem instances and initializations satisfying the following finite-propagation scaling conditions,

$$\frac{G(Z^{\nu_0}) - G^*}{m} = O(1), \quad \frac{1}{m} \max_{\nu_0 - D \leq \ell < \nu_0} \sum_{r=1}^p \|\hat{Z}_r^{\ell+1} - Z_r^\ell\|^2 = O(1), \quad (49)$$

with constants independent of m , we have $\mathcal{L}^{\nu_0} = O(m)$. Indeed,

$$\tau^3 A \sum_{j=1}^D \sum_{k=1}^{D-j+1} \rho^{-k} = 4A\kappa_{\mathcal{P}}^2 \tau [\rho^{-D} - 1 - D(1 - \rho)] = O\left(\frac{\log D}{\sqrt{D}}\right).$$

Thus $C = \max\{1, 2L\mathcal{L}^{\nu_0} \rho^{-\nu_0} / \mu^2\} = O(m^2)$. Since $\log C = O(\log m)$, its contribution is absorbed by the initial $O(m)$ term, giving

$$O\left(m + \frac{m}{\log m} \log \frac{1}{\epsilon}\right)$$

rounds. More generally, the same bound holds whenever an admissible C grows at most polynomially with m .

2) Rings with equal-length partitions. For $\omega_{\mathcal{P}} = O(1)$ and $LA = O(1)$, (48) gives, when $D \geq 1$,

$$(1 - \rho)^{-1} = O\left(\max\{D, \kappa_{\mathcal{P}} \sqrt{D}\}\right).$$

Shortening the blocks therefore need not improve the rate bound if the resulting increase in $\kappa_{\mathcal{P}}$ offsets the smaller delays. We examine this tradeoff by partitioning a ring into equal-length paths.

Proposition 7 *Let $\mathcal{P}$ partition the edges of a ring with m agents into $p \geq 2$ consecutive path blocks, each containing $D + 1$ edges, where $D \geq 1$ and $m = p(D + 1)$. Assume $f_i \in C^2$ for every $i \in \mathcal{V}$. Suppose there exist $R \succ 0$ and $0 < \underline{\alpha} \leq \bar{\alpha}$, with $\bar{\alpha}/\underline{\alpha} = O(1)$, such that*

$$\underline{\alpha}R \preceq \nabla^2 f_i(x) \preceq \bar{\alpha}R, \quad i \in \mathcal{V}, \quad x \in \mathbb{R}^d.$$

Then the block error bound (21) holds with

$$\kappa_{\mathcal{P}} = O(m^2/D).$$

Proof See Appendix A.5.

For this partition,

$$\omega_{\mathcal{P}} = 2, \quad D_r = D, \quad |\hat{\partial}_r| = 2, \quad \Gamma_r \leq 1, \quad r \in [p].$$

After the change of variables $u = R^{1/2}x$, the normalized Hessian bounds give $LA = O(1)$. The same contraction factor applies in the original coordinates (Appendix A.5). Consequently,

$$(1 - \rho)^{-1} = O\left(\max\left\{\frac{m^2}{D}, D, \frac{m^2}{\sqrt{D}}\right\}\right) = O(m^2/\sqrt{D}).$$

Within the equal-length family, this upper bound is minimized in order by taking $D = \Theta(m)$, equivalently a bounded number of blocks, giving $O(m^{3/2})$.

Under the curvature assumptions of Proposition 6, one long path with a bounded singleton remainder instead gives $O(m/\log m)$. Both choices have $D = \Theta(m)$, but their inter-block coupling differs. Thus partition design must account for the coupling left between blocks, rather than balance their lengths alone.

5 Surrogate enhancement

In this section, we introduce a surrogate variant of GATE to reduce per-iteration computational and communication costs. The updates in (17a) and (18) of Algorithm 1 require solving local minimization problems exactly, and each intra-cluster message $\mu_{i \rightarrow e}^\nu(\cdot)$ is a function of the receiving edge variable Y_e , which can be expensive to compute and transmit. This is particularly acute in the nonquadratic case, where messages need not admit compact finite-dimensional parametrizations; even in the quadratic case, exact messages may carry dense matrix parameters. Our approach is to replace each $g_i(Y_{\mathcal{A}_i}) = f_i^*((\mathbf{B}Y)_i)$ by a tractable surrogate: this makes the block subproblems cheaper to solve and induces messages with lightweight parametrizations.

More formally, for each node $i \in \mathcal{V}$, let

$$\mathbb{R}^{|\mathcal{A}_i| \times d} \times \mathbb{R}^{|\mathcal{A}_i| \times d} \ni (Y_{\mathcal{A}_i}, \bar{Y}_{\mathcal{A}_i}) \mapsto \tilde{g}_i(Y_{\mathcal{A}_i}; \bar{Y}_{\mathcal{A}_i})$$

be a surrogate of $g_i(Y_{\mathcal{A}_i}) = f_i^*(\mathbf{B}_{i,\mathcal{A}_i} Y_{\mathcal{A}_i})$. While we can directly work on the general $\tilde{g}_i$, we mainly focus on a structured choice obtained from a surrogate of the local conjugate f_i^* : choose

$$\mathbb{R}^d \times \mathbb{R}^d \ni (U_i, Z_i) \mapsto \tilde{f}_i^*(U_i; Z_i),$$

and set

$$\tilde{g}_i(Y_{\mathcal{A}_i}; \bar{Y}_{\mathcal{A}_i}) := \tilde{f}_i^*(\mathbf{B}_{i,\mathcal{A}_i} Y_{\mathcal{A}_i}; \mathbf{B}_{i,\mathcal{A}_i} \bar{Y}_{\mathcal{A}_i}), \quad (50)$$

where $\mathbf{B}_{i,\mathcal{A}_i}$ only gives the composition with the edge variables $Y_{\mathcal{A}_i}$ and their center $\bar{Y}_{\mathcal{A}_i}$; $\tilde{f}_i^*$ is chosen from the local conjugate f_i^* alone and is graph-independent. At iteration ν , write $Z^\nu := \mathbf{B}Y^\nu$ and $Z_i^\nu := (Z^\nu)_i$. Replacing the exact local terms in the edge-variable update (17a) and the message update (18) by these surrogates yields Algorithm 2. When $\tilde{f}_i^*(U; Z) \equiv f_i^*(U)$, Algorithm 1 is recovered.

5.1 Surrogate regularity conditions

We require the following mild regularity conditions on the local surrogates to guarantee convergence of Algorithm 2.

Assumption 5 *For each node $i \in \mathcal{V}$, the local surrogate $\tilde{f}_i^*$ is differentiable and satisfies the following conditions:*

(i) (*touching and gradient consistency*). For all $Z_i \in \mathbb{R}^d$,

$$\tilde{f}_i^*(Z_i; Z_i) = f_i^*(Z_i), \quad \nabla_U \tilde{f}_i^*(Z_i; Z_i) = \nabla f_i^*(Z_i).$$

Algorithm 2: GATE-S: Graph-Tearing message passing with surrogates

Input: Edge assignment $\mathbf{a}$ with ownership sets $\{\mathcal{A}_i^a\}_{i \in \mathcal{V}}$; edge partition $\{\mathcal{E}_r\}_{r=1}^p$; initial edge variables Y^0 ; finite-valued strongly convex initial messages $\tilde{\mu}_{i \rightarrow e}^0$, $e \in \mathcal{E}$, $i \in \mathcal{I}_e$; local surrogates $\tilde{f}_i^*(\cdot; \cdot)$; relaxation parameters $\tau_r^\nu \in [0, 1]$; $N \geq 1$.

for $\nu = 0, 1, \dots, N-1$ **do**

Each node $i \in \mathcal{V}$ **executes the following steps, in parallel with the other nodes:**

 (1) Assigned edge-variable update.

foreach $r \in [p]$ **and** $e \in \mathcal{A}_i^a \cap \mathcal{E}_r$, **compute**

$$\hat{Y}_e^{\nu+1} \in \operatorname{argmin}_{y_e \in \mathbb{R}^{1 \times d}} \left\{ \sum_{j \in \mathcal{I}_e} \tilde{\mu}_{j \rightarrow e}^\nu(y_e) \right\}, \quad (51a)$$

$$Y_e^{\nu+1} = Y_e^\nu + \tau_r^\nu (\hat{Y}_e^{\nu+1} - Y_e^\nu). \quad (51b)$$

 Send $Y_e^{\nu+1}$ to the other endpoint of e .

 (2) Node-to-edge message update.

foreach $r \in [p]$ **and** $e \in \mathcal{A}_i^a \cap \mathcal{E}_r$, **compute**

$$\begin{aligned} \tilde{\mu}_{i \rightarrow e}^{\nu+1}(y_e) = & \min_{Y_{\mathcal{A}_i^a \setminus \{e\}}} \left\{ \tilde{f}_i^* \left(\mathbf{B}_{i,e} y_e + \mathbf{B}_{i,\mathcal{A}_i^a \setminus \{e\}} Y_{\mathcal{A}_i^a \setminus \{e\}} + \mathbf{B}_{i,\mathcal{A}_i^a} Y_{\mathcal{A}_i^a}^{\nu+1}; \mathbf{B}_{i,\mathcal{A}_i} Y_{\mathcal{A}_i}^{\nu+1} \right) \right. \\ & \left. + \sum_{h \in \mathcal{A}_i^a \setminus \{e\}} \sum_{j \in \mathcal{I}_h \setminus \{i\}} \tilde{\mu}_{j \rightarrow h}^\nu(Y_h) \right\}. \end{aligned} \quad (52)$$

 Send $\tilde{\mu}_{i \rightarrow e}^{\nu+1}$ to the other endpoint of e .

Primal recovery. Each node $i \in \mathcal{V}$ computes $x_i^N := \nabla f_i^*((\mathbf{B}Y^N)_i)$.

(ii) (*majorization*). For all $U_i, Z_i \in \mathbb{R}^d$,

$$f_i^*(U_i) \leq \tilde{f}_i^*(U_i; Z_i).$$

(iii) (*strong convexity and smoothness in U_i*). There exist constants $\tilde{\mu}_i^*, \tilde{L}_i^* \in (0, +\infty)$ such that, for every center Z_i , the map $U_i \mapsto \tilde{f}_i^*(U_i; Z_i)$ is $\tilde{\mu}_i^*$ -strongly convex and $\tilde{L}_i^*$ -smooth.

(iv) (*Lipschitz continuity in Z_i*). The gradient $\nabla_U \tilde{f}_i^*(U_i; Z_i)$ is globally Lipschitz in Z_i , uniformly in U_i .

For a center $Z \in \mathcal{S}$, define

$$\tilde{G}(U; Z) := \sum_{i \in \mathcal{V}} \tilde{f}_i^*(U_i; Z_i), \quad U \in \mathbb{R}^{m \times d}.$$

For $Z = \mathbf{B}Y$, the value of the surrogate block minimization equals the minimum of $\tilde{G}(\cdot; Z)$ over the slice $Z + \mathcal{U}_r$:

$$\min_{\substack{W_r \in \mathbb{R}^{|\mathcal{E}_r| \times d} \\ i \in \mathcal{V}}} \sum_{i \in \mathcal{V}} \tilde{g}_i \left((W_r, Y_{\mathcal{E}_r})_{\mathcal{A}_i}; Y_{\mathcal{A}_i} \right) = \min_{U \in Z + \mathcal{U}_r} \tilde{G}(U; Z) =: \tilde{G}(\tilde{O}_r(Z); Z),$$

where $\tilde{O}_r(Z)$ is the unique minimizer, well defined because $\tilde{G}(\cdot; Z)$ is strongly convex and $Z + \mathcal{U}_r$ is closed. Let

$$\tilde{\Delta}_r(Z) := G(Z) - \tilde{G}(\tilde{O}_r(Z); Z) \geq 0 \quad (53)$$

be the progress of one exact minimization of the surrogate block problem from Z ; nonnegativity follows from the touching and optimality conditions.

We postulate the following surrogate block error bound condition, which is the counterpart of Assumption 4; it bounds the optimality gap by the aggregate surrogate block progress.

Assumption 6 (surrogate block error bound) *There exists a partition-dependent constant $\tilde{\kappa}_{\mathcal{P}} < \infty$ such that*

$$G(Z) - G^* \leq \tilde{\kappa}_{\mathcal{P}} \sum_{r=1}^p \tilde{\Delta}_r(Z), \quad Z \in \mathcal{S}. \quad (54)$$

As in the exact case, the smallest admissible constant $\tilde{\kappa}_{\mathcal{P}}$ in Assumption 6 is the supremum of the ratio of the global gap $G(Z) - G^*$ to the aggregate surrogate block progress $\sum_{r=1}^p \tilde{\Delta}_r(Z)$ over $Z \in \mathcal{S}$ with $G(Z) > G^*$. This constant enters the rate of Algorithm 2; its role is analogous to that of $\kappa_{\mathcal{P}}$ for Algorithm 1.

By the touching, optimality, and majorization conditions,

$$0 \leq \tilde{\Delta}_r(Z) \leq G(Z) - G(\tilde{O}_r(Z)) \leq \Delta_r(Z).$$

Consequently, the smallest admissible constants in Assumptions 4 and 6 satisfy $\kappa_{\mathcal{P}} \leq \tilde{\kappa}_{\mathcal{P}}$. If there exists $\eta \in (0, 1]$ such that

$$\tilde{\Delta}_r(Z) \geq \eta \Delta_r(Z), \quad \forall Z \in \mathcal{S}, \quad r \in [p],$$

then $\tilde{\kappa}_{\mathcal{P}} \leq \kappa_{\mathcal{P}}/\eta$. Thus, the price of using surrogate block updates is captured by this relative progress fraction: if η remains bounded away from zero, the network-size scaling of the exact block error bound is preserved up to a constant factor.

5.2 Surrogate examples

The following are two practical choices of $\tilde{f}_i^*$: a first-order surrogate with fixed metric, and a second-order surrogate that uses second-order information and allows the preconditioner to vary with the center.

Example 2 (First-order surrogate) For a fixed $M_i \in \mathbb{S}_{++}^d$, set

$$\tilde{f}_i^*(U_i; Z_i) := f_i^*(Z_i) + \langle \nabla f_i^*(Z_i), U_i - Z_i \rangle + \frac{1}{2} \langle U_i - Z_i, M_i(U_i - Z_i) \rangle. \quad (55)$$

A special case is the scalar metric: $M_i = \alpha_i^{-1} I_d$, with $\alpha_i > 0$.

Example 3 (Second-order surrogate) Let $H_i(Z_i) \in \mathbb{S}_{++}^d$ be chosen from second-order information, for instance $H_i(Z_i) = \nabla^2 f_i^*(Z_i) + \zeta_i I_d$ with $\zeta_i \geq 0$ when $f_i^* \in C^2$, and define

$$\tilde{f}_i^*(U_i; Z_i) := f_i^*(Z_i) + \langle \nabla f_i^*(Z_i), U_i - Z_i \rangle + \frac{1}{2} \langle U_i - Z_i, H_i(Z_i)(U_i - Z_i) \rangle.$$

5.2.1 Regularity and block error bound comparison

Surrogate regularity. Let $H_i(Z_i)$ denote the quadratic metric in the surrogate, with $H_i(Z_i) \equiv M_i$ in Example 2 and center-dependent H_i in Example 3. Touching and gradient consistency are immediate. If $f_i^* \in C^2$, majorization holds whenever $\nabla^2 f_i^*(U_i) \preceq H_i(Z_i)$ for all $U_i, Z_i \in \mathbb{R}^d$. Assumption 5(iii) holds with $\tilde{\mu}_i^* := \inf_{Z_i \in \mathbb{R}^d} \lambda_{\min}(H_i(Z_i))$ and $\tilde{L}_i^* := \sup_{Z_i \in \mathbb{R}^d} \lambda_{\max}(H_i(Z_i))$ whenever $0 < \tilde{\mu}_i^* \leq \tilde{L}_i^* < \infty$. For Example 2, Assumption 5(iv) follows from the global Lipschitz continuity of ∇f_i^* . For nonconstant H_i , Example 3 fails this global requirement and is retained as a practical variant. If H_i is locally Lipschitz and the edge iterates are bounded, the surrogate gradients are uniformly Lipschitz in the center on the relevant bounded sets. The subsequent center-sensitivity analysis can instead use these local Lipschitz bounds; we do not pursue this extension here.

Block error bound comparison. Given a partition $\mathcal{P}$, we next compare $\tilde{\kappa}_{\mathcal{P}}$ in Assumption 6 with the exact constant $\kappa_{\mathcal{P}}$ for the fixed-metric surrogate in Example 2. Fix r and write a block variation as $U = Z + P$ with $P \in \mathcal{U}_r$. By (53) and the change of variables $U = Z + P$, $\tilde{\Delta}_r(Z) = G(Z) - \min_{P \in \mathcal{U}_r} \tilde{G}(Z + P; Z)$. For $i \notin \mathcal{V}_r$, $B_{i, \mathcal{E}_r} = 0$, so every $P \in \mathcal{U}_r$ satisfies $P_i = 0$. Hence (55) gives

$$\tilde{G}(Z + P; Z) = G(Z) + \left\langle (\nabla G(Z)^\top)_{\mathcal{V}_r}, P_{\mathcal{V}_r} \right\rangle + \frac{1}{2} \sum_{i \in \mathcal{V}_r} \langle P_i, M_i P_i \rangle.$$

Substituting this identity into the definition gives

$$\tilde{\Delta}_r(Z) = - \min_{P \in \mathcal{U}_r} \left\{ \left\langle (\nabla G(Z)^\top)_{\mathcal{V}_r}, P_{\mathcal{V}_r} \right\rangle + \frac{1}{2} \sum_{i \in \mathcal{V}_r} \langle P_i, M_i P_i \rangle \right\}. \quad (56)$$

If the fixed metrics on $\mathcal{V}_r$ satisfy the curvature sandwich

$$\rho_i^{-1} M_i \preceq \nabla^2 f_i^*(z) \preceq M_i, \quad \forall z \in \mathbb{R}^d, \quad i \in \mathcal{V}_r, \quad (57)$$

with $\rho_i \geq 1$ and $\rho_r := \max_{i \in \mathcal{V}_r} \rho_i$, then the lower curvature bound in (57) implies

$$\begin{aligned} \Delta_r(Z) &\leq - \min_{P \in \mathcal{U}_r} \left\{ \left\langle (\nabla G(Z)^\top)_{\mathcal{V}_r}, P_{\mathcal{V}_r} \right\rangle + \frac{1}{2\rho_r} \sum_{i \in \mathcal{V}_r} \langle P_i, M_i P_i \rangle \right\} \\ &= \rho_r \left[- \min_{Q \in \mathcal{U}_r} \left\{ \left\langle (\nabla G(Z)^\top)_{\mathcal{V}_r}, Q_{\mathcal{V}_r} \right\rangle + \frac{1}{2} \sum_{i \in \mathcal{V}_r} \langle Q_i, M_i Q_i \rangle \right\} \right] \\ &= \rho_r \tilde{\Delta}_r(Z), \end{aligned}$$

where the equality uses $P = \rho_r Q$, the linearity of $\mathcal{U}_r$, and (56). Hence

$$\kappa_{\mathcal{P}} \leq \tilde{\kappa}_{\mathcal{P}} \leq \rho_{\max} \kappa_{\mathcal{P}}, \quad \rho_{\max} := \max_{r \in [p]} \rho_r.$$

Since $\rho_{\max} = \max_{i \in \mathcal{V}} \rho_i$, it is independent of the partition. If $\rho_{\max} = O(1)$, then $\tilde{\kappa}_{\mathcal{P}}$ and $\kappa_{\mathcal{P}}$ have the same network-size scaling. If the surrogate Hessian M_i matches $\nabla^2 f_i^*$ well (i.e., $\rho_i \approx 1$), the resulting degradation from $\kappa_{\mathcal{P}}$ to $\tilde{\kappa}_{\mathcal{P}}$ is only a small factor. In the scalar case, choosing $M_i = \mu_{f_i}^{-1} I_d$ ensures (57) with $\rho_i = L_{f_i}/\mu_{f_i}$, thereby retaining the network scaling of the block error bound up to a local condition number.

5.2.2 Quadratic message implementation

The surrogates in Examples 2 and 3 are quadratic in U_i . With quadratic initial messages, the updates in Algorithm 2 preserve this form, so each message is represented by a vector and a Hessian matrix. Concretely, each message can be written as

$$\tilde{\mu}_{i \rightarrow e}^\nu(Y_e) \sim \frac{1}{2} \langle Y_e, H_{i \rightarrow e}^\nu Y_e \rangle + \langle P_{i \rightarrow e}^\nu, Y_e \rangle, \quad H_{i \rightarrow e}^\nu \in \mathbb{S}_+^d, \quad P_{i \rightarrow e}^\nu \in \mathbb{R}^d. \quad (58)$$

For $h, h' \in \mathcal{A}_i$, $\nabla_{Y_h Y_{h'}}^2 \tilde{g}_i = \mathbf{B}_{i,h} \mathbf{B}_{i,h'} M_i$ in Example 2, and $\nabla_{Y_h Y_{h'}}^2 \tilde{g}_i = \mathbf{B}_{i,h} \mathbf{B}_{i,h'} H_i(Z_i^\nu)$ in Example 3. These Hessians can be dense, and so can the message Hessians $H_{i \rightarrow e}^\nu$.

Stabilization of message Hessians. In Example 2, the Hessian recursion does not depend on the surrogate centers. On the tree cluster, it is time-independent and stabilizes after finitely many sweeps. After this stabilization phase, the Hessian matrices no longer need to be transmitted; subsequent iterations update and send only the vectors $P_{i \rightarrow e}^\nu$.

Scalar-metric messages. The scalar metric even avoids matrix-valued Hessian communication entirely. Take $M_i = \alpha_i^{-1} I_d$ with $\alpha_i > 0$, and initialize the message Hessians as scalar multiples of I_d . Then $\nabla_{Y_h Y_{h'}}^2 \tilde{g}_i = \alpha_i^{-1} \mathbf{B}_{i,h} \mathbf{B}_{i,h'} I_d$, and Schur complements preserve this isotropic form. Therefore every message has the scalar-quadratic form

$$\tilde{\mu}_{i \rightarrow e}^\nu(Y_e) \sim \frac{h_{i \rightarrow e}^\nu}{2} \|Y_e\|^2 + \langle P_{i \rightarrow e}^\nu, Y_e \rangle, \quad h_{i \rightarrow e}^\nu \in \mathbb{R}_+, \quad P_{i \rightarrow e}^\nu \in \mathbb{R}^d.$$

Only one scalar Hessian coefficient and one d -vector need to be transmitted.

Slice-wise Hessian communication. The communication bottleneck is dense Hessian transmission. In both Examples 2 and 3, $H_{i \rightarrow e}^\nu$ in (58) can be a full $d \times d$ matrix. In Example 2, full-matrix transmission is confined to a finite Hessian-stabilization phase. In contrast, Example 3 has no such finite stabilization guarantee: $H_i(Z_i^\nu)$ varies with the center, so the message Hessians can remain dense and time-varying. To avoid this repeated full-matrix transmission in practice, we propose the following slice-wise Hessian communication strategy, in which only a slice of the Hessian is sent at each round. For center-dependent metrics, this is a heuristic communication strategy whose convergence is not analyzed in this paper.

Under the quadratic message representation (58), the vector $P_{i \rightarrow e}^\nu$ is sent every round, while the receiving node maintains a Hessian cache $\mathbf{H}_{i \rightarrow e}^\nu$. Decompose $\mathbb{S}^d$ as the direct sum $\mathbb{S}^d = \bigoplus_{k=1}^K \mathbb{S}_k$, where each $\mathbb{S}_k$ is chosen to admit a low-cost parametrization for communication and Π_k denotes the projection onto $\mathbb{S}_k$. Examples include diagonal bands, matrix rows or columns, and low-rank subspaces. When node i computes $\tilde{\mu}_{i \rightarrow e}^{\nu+1}$ using (52), each incoming message $\tilde{\mu}_{j \rightarrow h}^\nu$ in the local minimization is evaluated with the current vector $P_{j \rightarrow h}^\nu$ and the cached Hessian $\mathbf{H}_{j \rightarrow h}^\nu$. After this minimization, node i has computed the full Hessian $H_{i \rightarrow e}^{\nu+1}$ of $\tilde{\mu}_{i \rightarrow e}^{\nu+1}$, but sends only the slice indexed by $s(\nu) \in [K]$. The receiving node then updates only that slice of its cache:

$$\Pi_{s(\nu)} \mathbf{H}_{i \rightarrow e}^{\nu+1} = \Pi_{s(\nu)} H_{i \rightarrow e}^{\nu+1}, \quad \Pi_k \mathbf{H}_{i \rightarrow e}^{\nu+1} = \Pi_k \mathbf{H}_{i \rightarrow e}^\nu \quad (k \neq s(\nu)).$$

Thus each round sends one slice of each dense message Hessian, while the receiving node keeps the remaining slices from its Hessian cache. For the fixed-metric surrogate (55), if the updates are well defined and each slice is used at least once in every

S consecutive rounds, all Hessian caches in cluster r equal the stabilized message Hessians after at most

$$S \left(\left\lceil \frac{\text{diam}(\mathcal{F}_r)}{2} \right\rceil + 1 \right) \text{ rounds.}$$

Note that the cached Hessians may be indefinite. In practice, one can add δI_d to each cache, with $\delta > 0$ chosen to keep the local subproblem well defined.

5.3 Convergence analysis

We establish linear convergence of Algorithm 2. Similarly to the analysis of Algorithm 1 (cf. Sec. 4), we proceed in two steps. **Step 1** expresses Algorithm 2 as a delayed block-Jacobi scheme with delayed exterior coordinates and surrogate centers; **Step 2** establishes convergence of this scheme. The main difference from the exact case is that the block minimizers also depend on surrogate centers, requiring an additional perturbation analysis with respect to these centers.

5.3.1 Step 1: Equivalence to a delayed surrogate block-Jacobi scheme

Fix a block r . For $e \in \mathcal{E}_r$, extend the delays in (22) to all $i \in \mathcal{V}_r$ by

$$\delta_{e,i} := \frac{\text{dist}_{\mathcal{F}_r}(e,i) - 1}{2}, \quad e \in \mathcal{E}_r, \quad i \in \mathcal{V}_r, \quad (59)$$

$\tilde{D}_r := \max_{e \in \mathcal{E}_r, i \in \mathcal{V}_r} \delta_{e,i} = \text{diam}(\mathcal{T}_r) - 1$, and $\tilde{D} := \max_{r \in [p]} \tilde{D}_r$. Unlike D_r , which considers only boundary agents, $\tilde{D}_r$ includes delays from all agents of the block.

We expand the messages used in **step 1** of Algorithm 2 through the recursion in **step 2**, obtaining the following equivalence analogous to Proposition 5. The proof is omitted.

Lemma 6 *Under Assumptions 3 and 5, for every $r \in [p]$, $e \in \mathcal{E}_r$, and $\nu \geq \max_{s \in [p]} \text{diam}(\mathcal{T}_s)$, the updates of Algorithm 2 read*

$$\begin{aligned} Y_e^{\nu+1} &= Y_e^\nu + \tau_r^\nu (\hat{Y}_e^{\nu+1} - Y_e^\nu), \\ \hat{Y}_e^{\nu+1} &= \underset{Y_e}{\text{argmin}} \min_{Y_{\mathcal{E}_r \setminus \{e\}}} \sum_{i \in \mathcal{V}_r} \tilde{f}_i^* \left(\mathbf{B}_{i, \mathcal{A}_i^\circ} Y_{\mathcal{A}_i^\circ} + \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} Y_h^{\nu - \delta_{e,i}}; Z_i^{\nu - \delta_{e,i}} \right). \end{aligned} \quad (60)$$

5.3.2 Step 2: Linear convergence of the delayed scheme

The remaining convergence proof combines three ingredients: **(i)** a descent estimate for G along the iterates, **(ii)** the surrogate block error bound, and **(iii)** a perturbation analysis of the block minimizers, bounding the mismatch caused by delays. The arguments for **(i)** and **(ii)** parallel the exact case and are given in Appendix B.3. We focus here on **(iii)**: the block minimizers in Lemma 6 also depend on delayed surrogate centers, requiring additional control of center perturbations.

Perturbation analysis with respect to exterior parameters and surrogate centers. For exterior parameters $E \in \mathbb{R}^{|\widehat{\partial}_r| \times d}$, with rows indexed by $(i, h) \in \widehat{\partial}_r$, and surrogate centers $\vartheta \in \mathbb{R}^{|\mathcal{V}_r| \times d}$, define the block solution map

$$\tilde{Y}_r(E, \vartheta) = \operatorname{argmin}_{Y_{\mathcal{E}_r}} \sum_{i \in \mathcal{V}_r} \tilde{f}_i^* \left(\mathbf{B}_{i, \mathcal{A}_i^\vartheta} Y_{\mathcal{A}_i^\vartheta} + \sum_{h \in \mathcal{A}_i^\vartheta} \mathbf{B}_{i, h} E_{i, h}; \vartheta_i \right).$$

We now introduce the frozen-center counterparts of $\kappa_{\partial, r}$. For fixed centers ϑ , let $\tilde{\chi}_r(\cdot; \vartheta)$ be defined by

$$\sum_{i \in \mathcal{V}_r} (\nabla_u \tilde{f}_i^*(\cdot; \vartheta_i))^{-1} (\tilde{\chi}_r(\eta; \vartheta)) = \eta,$$

where $\tilde{\chi}_r(\cdot; \vartheta)$ is well defined by strong monotonicity. For nonempty $S \subseteq \mathcal{V}_r$, set

$$\tilde{\mathcal{L}}_{r, S}(\eta; \vartheta) := \sum_{i \in S} (\nabla_u \tilde{f}_i^*(\cdot; \vartheta_i))^{-1} (\tilde{\chi}_r(\eta; \vartheta)). \quad (61)$$

Using the same family of vertex subsets $\mathcal{R}_r$ as in (33), define

$$\tilde{\kappa}_{\partial, r} := \begin{cases} 0, & \partial_r = \emptyset, \\ \sup_{\vartheta \in \mathbb{R}^{|\mathcal{V}_r| \times d}} \max_{S \in \mathcal{R}_r} \frac{|\mathcal{V}_r|}{|S|} \operatorname{Lip}(\tilde{\mathcal{L}}_{r, S}(\cdot; \vartheta)), & \partial_r \neq \emptyset. \end{cases} \quad (62)$$

The supremum over ϑ makes the bounds for exterior perturbations uniform in the surrogate centers.

For a map $F(\eta; \vartheta)$, let $\operatorname{Lip}_{\vartheta_i} F(\eta; \vartheta)$ denote its Lipschitz modulus in the i -th center coordinate of ϑ , with η and the other center coordinates fixed. Define the normalized center-coordinate modulus

$$\begin{aligned} \tilde{\kappa}_{c, r} := & \sup_{\eta \in \mathbb{R}^d} \sup_{\vartheta \in \mathbb{R}^{|\mathcal{V}_r| \times d}} \max_{i \in \mathcal{V}_r} \max \left\{ \max_{e \in \mathcal{E}_r} \frac{|\mathcal{V}_r|}{|\mathcal{V}_e(i)|} \operatorname{Lip}_{\vartheta_i}(\tilde{\mathcal{L}}_{r, \overline{\mathcal{V}}_e(i)}(\eta; \vartheta)), \right. \\ & \left. \operatorname{Lip}_{\vartheta_i}(\tilde{\mathcal{L}}_{r, \{i\}}(\eta; \vartheta)), \max_{v \in \mathcal{V}_r \setminus \{i\}} |\mathcal{V}_r| \operatorname{Lip}_{\vartheta_i}(\tilde{\mathcal{L}}_{r, \{v\}}(\eta; \vartheta)) \right\}. \end{aligned} \quad (63)$$

Assumption 5(iii)–(iv) guarantees that these moduli are finite, with $\tilde{\kappa}_{c, r} = 0$ for exact surrogates; see Appendix B.1 for their interpretation. These moduli give the following sensitivity bounds.

Lemma 7 *Suppose Assumptions 3 and 5 hold. For every block r , the surrogate block solution map $\tilde{Y}_r(\cdot, \cdot)$ is single-valued.*

(I) *Sensitivity to exterior parameters. If $E, E' \in \mathbb{R}^{|\widehat{\partial}_r| \times d}$ differ only in row $(i, h) \in \widehat{\partial}_r$, then, for any center ϑ and edge $e \in \mathcal{E}_r$,*

$$\|[\tilde{Y}_r(E', \vartheta) - \tilde{Y}_r(E, \vartheta)]_e\| \leq \tilde{\kappa}_{\partial, r} \frac{|\overline{\mathcal{V}}_e(i)|}{|\mathcal{V}_r|} \|E'_{i, h} - E_{i, h}\|, \quad (64a)$$

$$\|\mathbf{B}_{v, \mathcal{E}_r}(\tilde{Y}_r(E', \vartheta) - \tilde{Y}_r(E, \vartheta))\| \leq \tilde{\kappa}_{\partial, r} \frac{1}{|\mathcal{V}_r|} \|E'_{i, h} - E_{i, h}\|, \quad v \neq i. \quad (64b)$$

(II) *Sensitivity to surrogate centers. If two centers $\vartheta, \hat{\vartheta}$ differ only in coordinate i ,*

then, for any exterior parameter E and edge $e \in \mathcal{E}_r$,

$$\|[\tilde{Y}_r(E, \hat{\vartheta}) - \tilde{Y}_r(E, \vartheta)]_e\| \leq \tilde{\kappa}_{c,r} \frac{|\bar{\mathcal{V}}_e(i)|}{|\mathcal{V}_r|} \|\hat{\vartheta}_i - \vartheta_i\|, \quad (65a)$$

$$\|\mathbf{B}_{i,\mathcal{E}_r}(\tilde{Y}_r(E, \hat{\vartheta}) - \tilde{Y}_r(E, \vartheta))\| \leq \tilde{\kappa}_{c,r} \|\hat{\vartheta}_i - \vartheta_i\|, \quad (65b)$$

$$\|\mathbf{B}_{v,\mathcal{E}_r}(\tilde{Y}_r(E, \hat{\vartheta}) - \tilde{Y}_r(E, \vartheta))\| \leq \frac{\tilde{\kappa}_{c,r}}{|\mathcal{V}_r|} \|\hat{\vartheta}_i - \vartheta_i\|, \quad v \neq i. \quad (65c)$$

Proof See Appendix B.1. $\square$

Bounding the mismatch caused by delays. As in Sec. 4.3, we bound the mismatch between the image contribution $\hat{Z}_r^{\nu+1}$ of the algorithm's delayed edge responses and the contribution $\bar{Z}_r^{\nu+1}$ obtained by surrogate block minimization at the current exterior coordinates and surrogate centers. For this comparison, write $Z^\nu := \mathbf{B}Y^\nu$, with Z_i^ν its i -th row, and define, for $r \in [p]$,

$$\begin{aligned} Z_r^\nu &:= \mathbf{B}_{\mathcal{V},\mathcal{E}_r} Y_{\mathcal{E}_r}^\nu, \\ \hat{Z}_r^{\nu+1} &:= \mathbf{B}_{\mathcal{V},\mathcal{E}_r} \hat{Y}_{\mathcal{E}_r}^{\nu+1}, \\ \bar{Z}_r^{\nu+1} &:= \tilde{O}_r(Z^\nu) - Z^\nu + Z_r^\nu. \end{aligned} \quad (66)$$

Fix $r \in [p]$ and $\nu \geq \max_{s \in [p]} \text{diam}(\mathcal{T}_s)$, and let $Z_{\mathcal{V}_r}^\nu$ denote the restriction of Z^ν to $\mathcal{V}_r$. For $e \in \mathcal{E}_r$, collect the current and delayed exterior parameters in $E^\nu, \hat{E}_e^\nu \in \mathbb{R}^{|\hat{\partial}_r| \times d}$, and the delayed centers in $Z_{r,e}^\nu$:

$$\begin{aligned} (E^\nu)_{i,h} &:= Y_h^\nu, & (\hat{E}_e^\nu)_{i,h} &:= Y_h^{\nu-\delta_{e,i}}, & (i,h) &\in \hat{\partial}_r, \\ Z_{r,e}^\nu &:= (Z_i^{\nu-\delta_{e,i}})_{i \in \mathcal{V}_r}. \end{aligned}$$

By (66), the definition of $\tilde{O}_r$, and Lemma 6,

$$\begin{aligned} \bar{Z}_r^{\nu+1} &= \mathbf{B}_{\mathcal{V},\mathcal{E}_r} \tilde{Y}_r(E^\nu, Z_{\mathcal{V}_r}^\nu), \\ \hat{Y}_e^{\nu+1} &= [\tilde{Y}_r(\hat{E}_e^\nu, Z_{r,e}^\nu)]_e. \end{aligned}$$

These sensitivity bounds yield the following image-space mismatch estimate. Compared with Lemma 5, the bound includes additional contributions involving $\tilde{\kappa}_{c,r}^2$ arising from delayed surrogate centers.

Lemma 8 *Under Assumptions 3 and 5, for each block r , define*

$$\begin{aligned} \tilde{M}_{r,\partial} &:= \Gamma_r \left(8\tilde{\kappa}_{\partial,r}^2 |\hat{\partial}_r| + 8|\partial_r| (|\mathcal{V}_r| + \tilde{D}_r) \tilde{\kappa}_{c,r}^2 \right), \\ \tilde{M}_{r,c} &:= 8(|\mathcal{V}_r| + \tilde{D}_r) \tilde{\kappa}_{c,r}^2. \end{aligned}$$

Then, for every iteration $\nu \geq \max_{r \in [p]} \text{diam}(\mathcal{T}_r)$,

$$\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 \leq \tilde{M}_{r,\partial} \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{s \in \mathcal{M}_r} \|Z_s^{\ell+1} - Z_s^\ell\|^2 + \tilde{M}_{r,c} \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \|Z_r^{\ell+1} - Z_r^\ell\|^2. \quad (67)$$

The sums over ℓ are understood to be zero when $\tilde{D}_r = 0$.

Proof See Appendix B.2. $\square$

Linear convergence. For each block r , define $\tilde{\mu}_r^* := \min_{i \in \mathcal{V}_r} \tilde{\mu}_i^*$, $\tilde{L}_r^* := \max_{i \in \mathcal{V}_r} \tilde{L}_i^*$. Using this mismatch estimate, we obtain the following convergence theorem.

Theorem 2 Suppose Assumptions 1, 3, 5, and 6 hold, with $p > 1$. Define $\underline{\tilde{\mu}}^* := \min_{r \in [p]} \tilde{\mu}_r^*$ and

$$\tilde{A} := \max_{s \in [p]} \left[\sum_{r \in [p]: \tilde{D}_r > 0, s \in \mathcal{M}_r} \frac{2\tilde{L}_r^* + \tilde{\mu}_r^*}{4} \tilde{M}_{r,\partial} + \mathbf{1}_{\{\tilde{D}_s > 0\}} \frac{2\tilde{L}_s^* + \tilde{\mu}_s^*}{4} \tilde{M}_{s,c} \right]. \quad (68)$$

Let $X^* = \mathbf{1}(x^*)^\top$ be the consensus primal optimizer. If $\tau_r^\nu \equiv \tau$ and

$$0 < \tau \leq \min \left\{ \frac{1}{\omega_{\mathcal{P}}}, 2\tilde{\kappa}_{\mathcal{P}} \right\}, \quad \frac{16\tilde{A}\tilde{\kappa}_{\mathcal{P}}\tau}{\underline{\tilde{\mu}}^*} \left[\left(1 - \frac{\tau}{2\tilde{\kappa}_{\mathcal{P}}} \right)^{-\tilde{D}} - 1 \right] \leq 1, \quad (69)$$

then, for some constant $C \in (0, \infty)$ and all $\nu \geq \tilde{D} + 1$,

$$\|X^\nu - X^*\|^2 \leq C\tilde{\rho}^\nu, \quad \tilde{\rho} := 1 - \frac{\tau}{2\tilde{\kappa}_{\mathcal{P}}}. \quad (70)$$

Proof See Appendix B.3. $\square$

Remark 6 Theorem 2 parallels Theorem 1: Algorithm 2 converges linearly, with rate $\tilde{\rho}$ now given by (70). The stepsize restriction (69) has the same form as (42). The main differences are that the exact quantities $(1/L_r, 1/\mu_r, D, A)$ are replaced by $(\tilde{\mu}_r^*, \tilde{L}_r^*, \tilde{D}, \tilde{A})$, and that $\kappa_{\mathcal{P}}$ is replaced by $\tilde{\kappa}_{\mathcal{P}}$, which may be larger when the surrogate captures less exact block progress. The resulting deviation from the exact constants is the rate price for cheaper local computation and communication. For the exact surrogate, Theorem 2 recovers the rate factor of Theorem 1 at the same admissible stepsize.

As in (47), an explicit sufficient stepsize is

$$0 < \tau \leq \min \left\{ \frac{1}{\omega_{\mathcal{P}}}, \frac{2\tilde{\kappa}_{\mathcal{P}}}{2\tilde{D} + 1}, \sqrt{\frac{\underline{\tilde{\mu}}^*}{8(2\tilde{D} + 1)\tilde{A}}} \right\}.$$

The proof is analogous to that of Corollary 1 and is omitted.

5.4 Comparison with consensus-based methods

We now compare the convergence rate of Algorithm 2 with those of consensus-based methods, such as DGD [29], EXTRA [23], and SONATA [25], in terms of network-size scaling. We show that, for a class of tree-dominant graphs, sufficiently accurate surrogates yield an acceleration in network-size scaling relative to these consensus baselines. On general graphs, our scheme achieves a rate no worse than the standard spectral-gap bound for the corresponding dual-gradient method when a suitable proximal-linear surrogate is chosen.

Acceleration in tree-dominant graphs. Consider a tree-dominant graph, and choose a partition $\mathcal{P}$ that has a dominant tree block and only $O(1)$ remaining singleton edge blocks. The analysis in Sec. 4.4 gives $\kappa_{\mathcal{P}} = O(1)$, $A = O(1)$. Theorem 1 then yields $(1 - \rho)^{-1} = O(D/\log D)$. For the ring and barbell examples, the exact inverse rate gap is $O(m/\log m)$. In contrast, consensus baselines have $O(m^2)$ rate scaling on these graphs through the inverse spectral-gap factor $1/\lambda_{\min}^+(\mathbf{B}\mathbf{B}^\top)$. The exact $O(m/\log m)$ inverse rate gap therefore gives network-size acceleration over those baselines.

The constant $\tilde{\kappa}_{\mathcal{P}}$ preserves the network-size scaling of $\kappa_{\mathcal{P}}$ whenever the surrogate block progress captures the exact block progress up to a constant fraction. Indeed, if, for some network-independent $\eta \in (0, 1]$, $\tilde{\Delta}_r(Z) \geq \eta \Delta_r(Z)$ holds for all $Z \in \mathcal{S}$ and $r \in [p]$, then $\kappa_{\mathcal{P}} \leq \tilde{\kappa}_{\mathcal{P}} \leq \kappa_{\mathcal{P}}/\eta$. For example, once a quadratic surrogate satisfies the Hessian sandwich inequality (57), it gives $\tilde{\kappa}_{\mathcal{P}} \leq \rho_{\max} \kappa_{\mathcal{P}}$ with network-independent $\rho_{\max}$. Consequently, for these fixed-metric quadratic surrogates on the tree-dominant families and partitions above, $\tilde{\kappa}_{\mathcal{P}} = O(1)$ and Lemma 8 gives $\tilde{A} = O(m)$. Theorem 2 then gives $(1 - \tilde{\rho})^{-1} = O(\sqrt{m\tilde{D}})$, for $\tilde{D} \geq 1$. For the ring and barbell examples, where $\tilde{D} = \Theta(m)$, this yields an $O(m)$ inverse rate gap, which matches the exact method's rate scaling up to a logarithmic factor.

Consensus scaling recovery. On general graphs, consider a direct proximal-linear surrogate for $g_i(Y_{\mathcal{A}_i}) = f_i^*(\mathbf{B}_{i,\mathcal{A}_i} Y_{\mathcal{A}_i})$, outside the structured subclass (50). Specifically, choose $c_{i,e} > 0$ so that the following surrogate majorizes g_i at every center:

$$\tilde{g}_i(Y_{\mathcal{A}_i}; Y_{\mathcal{A}_i}^\nu) := g_i(Y_{\mathcal{A}_i}^\nu) + \sum_{e \in \mathcal{A}_i} \langle \nabla_{Y_e} g_i(Y_{\mathcal{A}_i}^\nu), Y_e - Y_e^\nu \rangle + \frac{1}{2} \sum_{e \in \mathcal{A}_i} c_{i,e} \|Y_e - Y_e^\nu\|^2. \quad (71)$$

Since (71) is separable in $\{Y_e : e \in \mathcal{A}_i\}$, the message update satisfies

$$\tilde{\mu}_{i \rightarrow e}^{\nu+1}(Y_e) \sim \left\langle \mathbf{B}_{i,e} \nabla f_i^*((\mathbf{B}Y^{\nu+1})_i), Y_e - Y_e^{\nu+1} \right\rangle + \frac{c_{i,e}}{2} \|Y_e - Y_e^{\nu+1}\|^2.$$

For $\nu \geq 1$ and $e \in \mathcal{E}_r$ with $\mathcal{I}_e = \{i, j\}$, the edge update is the dual-gradient step

$$\hat{Y}_e^{\nu+1} = Y_e^\nu - \frac{\mathbf{B}_{i,e} \nabla f_i^*(Z_i^\nu) + \mathbf{B}_{j,e} \nabla f_j^*(Z_j^\nu)}{c_{i,e} + c_{j,e}},$$

followed by $Y_e^{\nu+1} = Y_e^\nu + \tau_r^\nu (\hat{Y}_e^{\nu+1} - Y_e^\nu)$. For this direct surrogate, the block progress depends on Y only through $Z = \mathbf{B}Y$:

$$\tilde{\Delta}_r(Z) := \sum_{e \in \mathcal{E}_r} \frac{\|\mathbf{B}_{\mathcal{V},e}^\top \nabla G(Z)^\top\|^2}{2(c_{i,e} + c_{j,e})}, \quad Z \in \mathcal{S}, \quad \mathcal{I}_e = \{i, j\}.$$

Consequently, the smallest constant in the corresponding block error bound is

$$\tilde{\kappa}_{\mathcal{P}} = \sup_{Z \in \mathcal{S}: G(Z) > G^*} \frac{G(Z) - G^*}{\frac{1}{2} \sum_{e \in \mathcal{E}} (c_{i,e} + c_{j,e})^{-1} \|\mathbf{B}_{\mathcal{V},e}^\top \nabla G(Z)^\top\|^2},$$

and is independent of the partition $\mathcal{P}$. Strong convexity of G on $\mathcal{S}$ gives

$$G(Z) - G^* \leq \frac{L}{2\lambda_{\min}^+(\mathbf{B}\mathbf{B}^\top)} \|\mathbf{B}^\top \nabla G(Z)^\top\|^2.$$

If L and the coefficients $c_{i,e} + c_{j,e}$ are bounded independently of network size, then

$$\tilde{\kappa}_{\mathcal{P}} = O\left(\frac{1}{\lambda_{\min}^+(\mathbf{B}\mathbf{B}^\top)}\right), \quad (72)$$

where $\lambda_{\min}^+$ is the smallest positive eigenvalue. Thus, this surrogate recovers the usual consensus spectral-gap rate scaling $O(1/\lambda_{\min}^+(\mathbf{B}\mathbf{B}^\top))$, matching classical nonaccelerated consensus-based methods.

5.5 Comparison with tree-based aggregation methods

Spanning-tree methods relay iterates [27] or aggregate gradients for master-node updates [21]. The tree can be constructed in a decentralized manner, by electing a root and selecting paths to it [27]. However, *using a single spanning tree need not be desirable in every application*. **(i)** It provides *no route redundancy*: a permanent failure of any tree edge or non-leaf agent disconnects it, requiring repair to restore communication. **(ii)** Distances along the tree can exceed those in the full graph, lengthening propagation delays. **(iii)** In master-node schemes, the root is additionally a single point of failure for the optimization, and each update must wait for all gradient contributions [21].

GATE and GATE-S are significantly more flexible than these schemes. **(a)** They permit both a *single* spanning-tree block and decompositions retaining *all* physical edges, without a computational master. **(b)** Their trees define optimization blocks, not merely aggregation routes: the decomposition can therefore be adapted jointly to the agents' functions and the network, exploiting the block-preconditioning effect quantified by the analysis. **(c)** Retaining edges also provides reconfiguration options. If the physical graph remains connected after a link failure, deleting the failed edge and splitting its block into the remaining nonempty tree blocks preserves a valid decomposition without changing the other blocks. The retained variables can initialize a new run under the theorem assumptions for the new partition.

Even within GATE and GATE-S, *a single spanning-tree block need not be preferable*, albeit being an option. For GATE with exact messages, it gives the finite-round guarantee of Remark 5; partitioning into smaller blocks is still convenient when some agents have high degree, since each message minimization at an agent involves all its other edges in the block. When the messages admit no closed form, GATE-S uses local surrogates to obtain tractable messages; a single block r then uses centers up to $\tilde{D}_r = \text{diam}(\mathcal{T}_r) - 1$ rounds old, a delay that affects the rate. Smaller-diameter blocks shorten these delays and can permit larger stepsizes and faster guaranteed convergence, provided the added inter-block coupling and overlap remain controlled (Theorem 2).

6 Numerical experiments

In this section, we evaluate GATE and its surrogate variants on strongly convex quadratic and regularized logistic regression problems over various graphs, with and without a dominant tree structure. The quadratic instances have the form

$$f_i(x) = \frac{1}{2}x^\top H_i x + q_i^\top x, \quad H_i \succ 0.$$

The non-quadratic instances have the form

$$f_i(x) = \frac{1}{N_i} \sum_{j=1}^{N_i} \log\left(1 + \exp\left(-b_{ij}a_{ij}^\top x\right)\right) + \frac{1}{2}x^\top R_i x - h_i^\top x, \quad R_i \succ 0.$$

The consensus baselines use the Metropolis mixing matrix W . Stepsizes for the proposed methods and the consensus baselines are tuned using an ad-hoc grid search on a logarithmic scale, yielding the optimal numerical performance. Performance is measured by the error of the iterates from the solution of the problem, $\|X^\nu - \mathbf{1}(x^*)^\top\|$.

6.1 Scaling verification and partition effects

In this part, we verify the scaling with network size and dominant-tree diameter and examine the effect of the partition.

Network-size scaling on rings. We first verify network-size scaling for quadratic problems on ring networks with $d = 20$. Each H_i is generated as a random symmetric Gaussian matrix with a positive diagonal correction chosen so that $\kappa(H_i) = 2$. The edge partition consists of one $(m - 2)$ -edge path cluster and two singleton-edge clusters. GATE-S1 uses the fixed diagonal metric with entries $(M_i)_{kk} = (H_i^{-1})_{kk} + \sum_{\ell \neq k} |(H_i^{-1})_{k\ell}|$, so $M_i \succeq H_i^{-1}$. Figure 4(a) shows that GATE and GATE-S1 are faster than EXTRA and SONATA, with the gap widening as m grows. Figure 4(b) indicates linear scaling for the GATE variants and quadratic scaling for the consensus baselines, which is consistent with the theoretical comparison in Sec. 5.4.

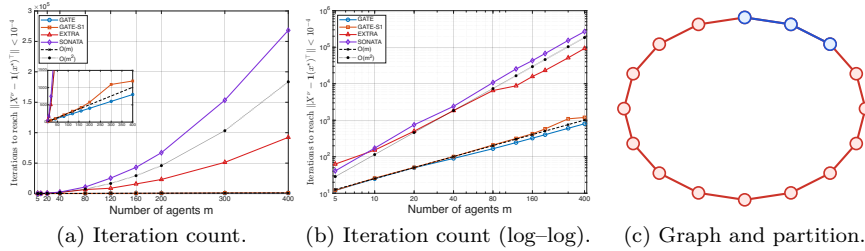


Fig. 4: Network-size scaling for quadratic problems on a ring. Panel (a) reports the iteration counts versus the number of agents m , while panel (b) presents the same data on log-log axes to make their dependence on m more apparent. Panel (c) shows the graph and edge partition. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^\top\| < 10^{-4}$.

To test whether the scaling observed above persists beyond the quadratic setting, we next consider the non-quadratic logistic regression problem, for which neither GATE surrogate is exact. We use $d = 10$, $N_i = 40$, and the same partition as in the quadratic experiment. The Hessian condition number is set to five. The quadratic-surrogate metrics are $\nabla^2 f_i^*(Z_i^\nu) + \delta I_d$ for GATE-S2 and $\text{Diag}(\nabla^2 f_i^*(Z_i^\nu)) + \delta I_d$ for GATE-S1, with δ tuned over $\{0, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$ separately for each method.

Figure 5(a) shows that GATE-S1 and GATE-S2 remain faster than EXTRA and SONATA, with the gap widening as m grows. Figure 5(b) indicates near-linear scaling for the GATE surrogates, whereas the consensus baselines approach quadratic scaling over the upper range.

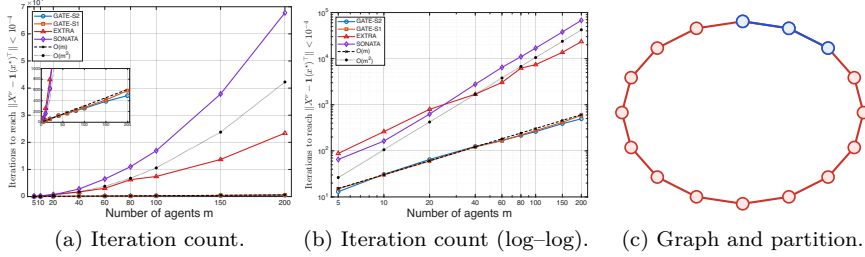


Fig. 5: Network-size scaling for logistic-regression problems on a ring. Panel (a) reports the iteration counts versus the number of agents m , while panel (b) presents the same data on log-log axes to make their dependence on m more apparent. Panel (c) shows the graph and edge partition. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^\top\| < 10^{-4}$.

Dependence on the dominant-tree diameter. We next study the dependence of the convergence rate on D for a quadratic problem over a non-ring graph. The graph consists of two three-node cliques connected by a growing non-path tree, with the partition containing one dominant tree cluster and six singleton-edge clusters. We set $H_i = H_0 \in \mathbb{S}_{++}^5$ for all i , where H_0 is dense and $\kappa(H_0) = 2$. For GATE-S1, we use a diagonal metric with entries $(M_i)_{kk} = (H_0^{-1})_{kk} + \sum_{\ell \neq k} |(H_0^{-1})_{k\ell}|$, so $M_i \succeq H_0^{-1}$.

With m fixed, Figure 6(b) indicates the $O(D)$ and $O(\sqrt{Dm})$ rate scalings for GATE and GATE-S1, respectively, consistent with Theorems 1 and 2. Figure 6(a) then considers $D = \Theta(\sqrt{m})$ as m varies and exhibits the corresponding $m^{1/2}$ and $m^{3/4}$ scalings for GATE and GATE-S1, respectively, further confirming the theoretical scaling.

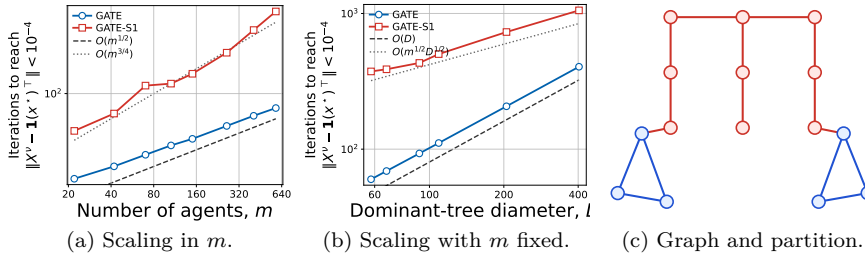


Fig. 6: Network-size and diameter scaling for quadratic problems on a non-ring graph family. Panel (a) reports the iteration counts versus the number of agents m for graphs with $D = \Theta(\sqrt{m})$. Panel (b) reports the iteration counts versus D at fixed $m = 406$. Panel (c) shows the graph and edge partition. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^\top\| < 10^{-4}$.

Effect of the partition. We next examine the effect of the partition for a quadratic problem on a ring. For each network size, we compare three partitions: P1 contains one $(m - 2)$ -edge path and two singleton-edge clusters; P2 arranges two paths of length $m/2$; and P3 contains $\sqrt{m}$ consecutive path blocks of $\sqrt{m}$ edges each.

Figure 7 shows that P1 and P2 substantially outperform P3, while P1 is modestly faster than P2. This ordering is consistent with the discussion following Proposition 7: $D = \Theta(m)$ minimizes the rate bound for equal-length partitions of the ring, whereas Proposition 6 gives the sharper $O(m)$ rate for the long-path partition.

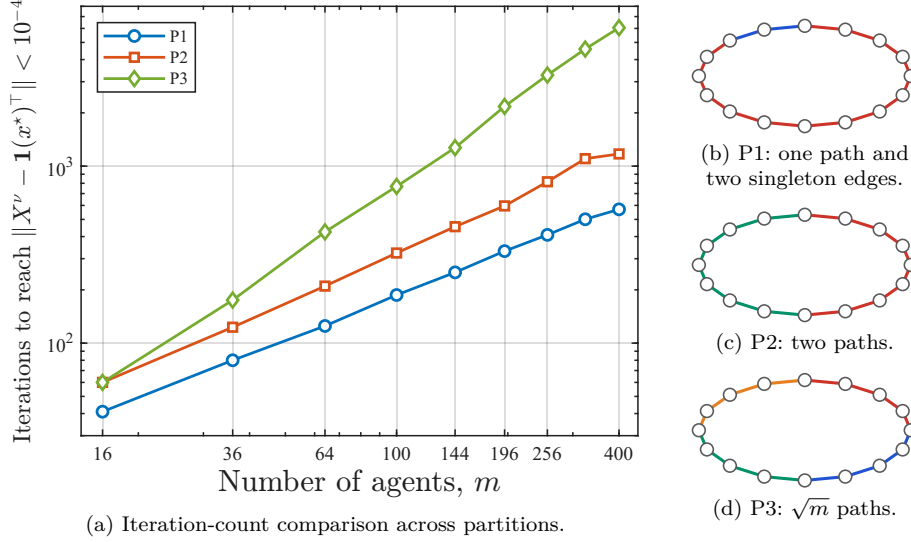


Fig. 7: Effect of the partition on GATE for quadratic problems on a ring ($d = 10$). Panel (a) reports the iteration counts versus the number of agents m for P1, P2, and P3. Panels (b)–(d) show the corresponding partitions on a representative ring. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^\top\| < 10^{-4}$.

6.2 Performance comparison

We now compare the algorithms in terms of iteration count and cumulative communication on tree-dominant barbell networks and a multi-cycle network. Communication is measured by the total number of scalars transmitted over directed links, normalized by the ambient dimension d .

Tree-dominant barbell networks. We consider barbell graphs formed by two four-node cliques connected by a path of $L \in \{10, 100\}$ edges. The partition contains one dominant tree cluster formed by the path and a three-edge star centered at the path endpoint in each clique, while each of the six remaining clique edges forms a singleton-edge cluster.

For the quadratic problems, we set $d = 100$ and use dense H_i generated from Gaussian Gram matrices with a diagonal correction, yielding $\kappa(H_i) \approx 100$. GATE uses exact quadratic messages, while GATE-S1 and GATE-S2 (cache Hessian) use a fixed diagonal metric and a slice-wise Hessian cache, respectively. Figure 8 shows that GATE is markedly faster than all other methods, reaching the target in only 31 and 114 iterations for $L = 10$ and $L = 100$, respectively. Although GATE initially transmits dense message Hessians, these matrices stabilize after finitely many rounds and only the message vectors are sent thereafter; consequently, GATE also has the lowest cumulative communication. GATE-S2 (cache Hessian) sends only message vectors and d -vector Hessian slices. Once its cache has recovered the stabilized message Hessians, its tail follows essentially the same linear rate as GATE, consistent with the finite cache-recovery property in Sec. 5.2.2. GATE-S1 is comparable to EXTRA, the fastest consensus baseline on these instances. CTA and ATC do not reach 10^{-6} within 10^5 iterations.

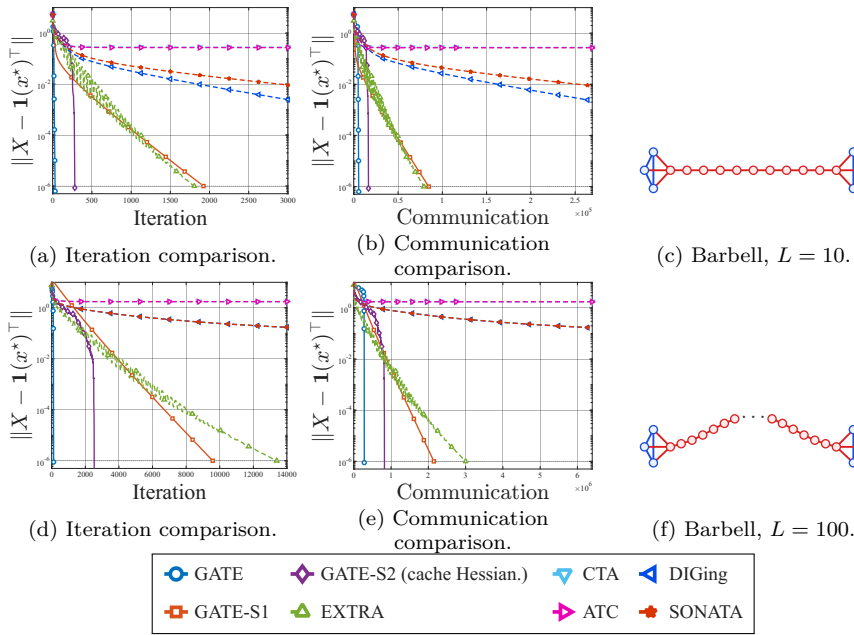


Fig. 8: Performance comparison for quadratic problems on two barbell graphs. Panels (a)–(c) use $L = 10$, while panels (d)–(f) use $L = 100$. Panels (a) and (d) report the primal error versus iteration count, while panels (b) and (e) report it versus cumulative communication. Panels (c) and (f) show the corresponding graphs and partitions; red edges form the dominant tree cluster and blue edges form singleton-edge clusters. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^\top\| \leq 10^{-6}$.

We next test logistic-regression problems on the same graphs and partitions, with $d = 100$ and $N_i = 80$. GATE-S2 (exact Hessian), GATE-S1, and GATE-S2 (cache Hessian) use the full center-dependent Hessian, its diagonal, and a slice-wise Hessian cache, respectively. The exact-Hessian variant is included only in the iteration plots

because it transmits dense Hessian matrices at every iteration. Figure 9 shows that the cache variant is second only to GATE-S2 (exact Hessian) in iteration count and requires the least communication among all other methods that reach the target. For $L = 100$, GATE-S2 (cache Hessian) requires about one third as much communication as either GATE-S1 or EXTRA.

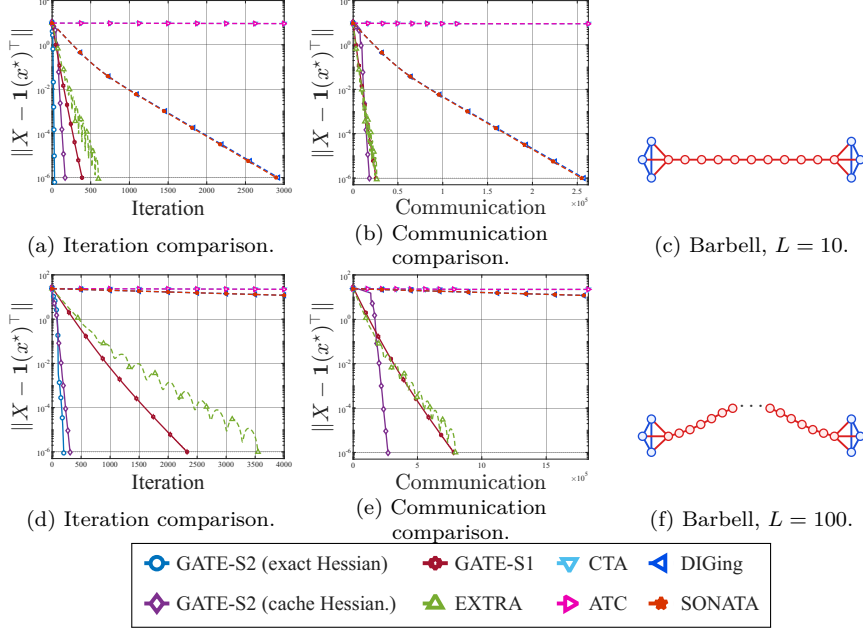


Fig. 9: Logistic-regression comparison on two barbell graphs. Panels (a)–(c) use $L = 10$, while panels (d)–(f) use $L = 100$. Panels (a) and (d) report the error versus iteration count, while panels (b) and (e) report it versus cumulative communication; GATE-S2 (exact Hessian) is omitted from the communication panels. Panels (c) and (f) show the corresponding graphs and partitions; red edges form the dominant tree cluster and blue edges form singleton-edge clusters. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^T\| \leq 10^{-6}$.

A multi-cycle network. We next test whether the gains persist on the network in Figure 10, which contains several cycles but no pronounced long-tree structure. We consider partitions beyond Assumption 3 to assess the empirical robustness of GATE to the choice of partition. The quadratic problem uses $d = 200$ and dense local Hessians with condition number eight, while the logistic-regression problem uses $d = 50$ and $N_i = 80$. For the quadratic problem, we compare exact GATE, two GATE-S1 variants, and GATE-S2 (cache Hessian) against the consensus baselines. For logistic regression, we replace exact GATE with GATE-S2 (exact Hessian) and retain the remaining methods.

Figure 10(a) shows that exact GATE converges fastest for the quadratic problem. GATE-S2 (cache Hessian) follows essentially the same tail rate, and Figure 10(b)

shows that it requires the least communication. The scalar and diagonal GATE-S1 variants also outperform all consensus baselines in both iteration count and communication. For logistic regression, Figure 10(d) shows that GATE-S2 (exact Hessian) is fastest, followed by the cache variant; Figure 10(e) shows that the cache variant again requires the least communication among the displayed methods. The diagonal GATE-S1 variant remains faster than all consensus baselines in both measures, whereas the scalar GATE-S1 variant is less competitive on this instance. EXTRA is the strongest consensus baseline in both problems; DIGing and SONATA require more communication, while CTA and ATC do not reach the target. Thus, the practical gains observed on the barbell graphs persist on this multi-cycle network.

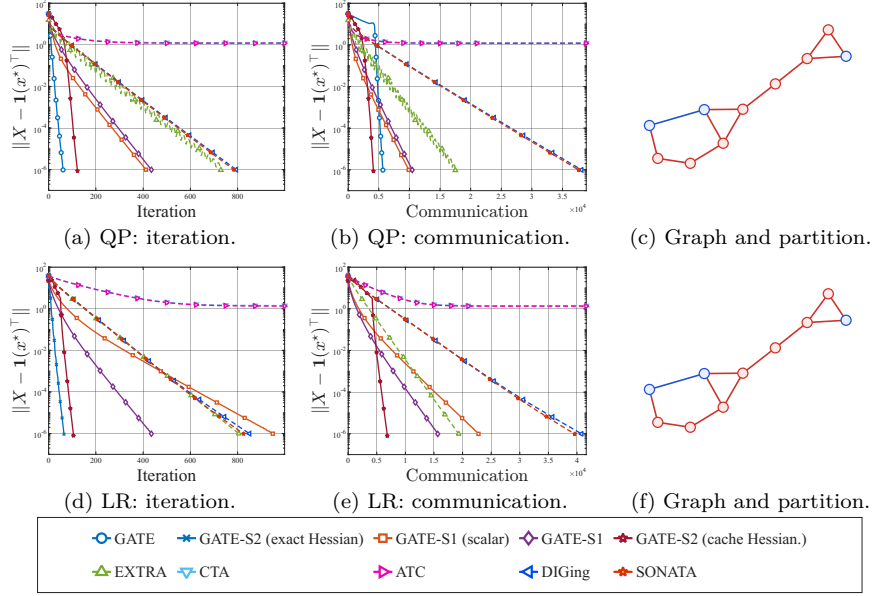


Fig. 10: Performance comparison for quadratic and logistic-regression problems on a multi-cycle network. Panels (a)–(c) show the quadratic results with $d = 200$, while panels (d)–(f) show the logistic-regression (LR) results with $d = 50$. Panels (a) and (d) report the error versus iteration count, while panels (b) and (e) report it versus cumulative communication; GATE-S2 (exact Hessian) is omitted from panel (e). Panels (c) and (f) show the common graph and partition; red edges form the message-passing cluster and the blue edge forms the singleton-edge cluster. The stopping criterion is $\|X^\nu - \mathbf{1}(x^*)^\top\| \leq 10^{-6}$.

7 Conclusion

In this work, we developed a graph-decomposition framework for jointly designing the consensus representation, dual-coordinate blocks, and connected carriers. The framework makes graph structure part of the optimization decomposition itself, beyond its

role in communicating prescribed updates. Specializing to the incidence representation and tree-structured edge blocks, we proposed **GATE**, a single-loop decentralized message-passing algorithm derived from the block fixed-point decomposition.

For smooth strongly convex local objectives, we established linear convergence and quantified a *block-preconditioning effect*: the partition changes the coupling between subproblems, so network topology and function geometry enter the rate jointly. The resulting bounds guide partition selection for the given functions and network. For the tree-dominant families and partitions considered, under the stated curvature assumptions, the guaranteed factor ρ satisfies $(1 - \rho)^{-1} = O(D/\log D)$, with D comparable to the dominant-tree diameter and $O(D)$ initial propagation rounds. GATE-S reduces per-iteration computation and communication through tractable local surrogates and compact message parametrizations, while retaining linear convergence under the regularity assumptions.

Numerical experiments on quadratic and logistic-regression problems support the theoretical scaling and demonstrate iteration and communication savings over the tested consensus-based baselines. Future directions include automated partition selection guided by the rate bounds, conjugate-free variants, and extensions to non(-strongly) convex objectives and stochastic settings.

Appendix

A Proofs of Results in Sec. 4

A.1 Proof of Proposition 5

Fix a block r , an edge $e \in \mathcal{E}_r$, and an iteration $\nu \geq \text{diam}(\mathcal{T}_r)$. Since $e \in \mathcal{E}_r$ is a vertex of the factor graph $\mathcal{F}_r$, root $\mathcal{F}_r$ at e . For every vertex $v \neq e$, let $\text{pa}(v)$ be its parent. For $i \in \mathcal{V}_r$, let $\mathcal{T}_i$ be the component of $\mathcal{F}_r \setminus \{(i, \text{pa}(i))\}$ containing i , and denote by $\mathcal{V}_i$ and $\mathcal{E}_i$ its vertices in $\mathcal{V}_r$ and $\mathcal{E}_r$, respectively. Thus $\text{pa}(i) \notin \mathcal{E}_i$, where $\mathcal{E}_i$ consists exactly of the active edge variables strictly below i . For $\ell \in \mathcal{V}_r$, define the rooted local factor

$$\Psi_{\ell}^{\nu,e}(Y_{\mathcal{A}_{\ell}^{\circ}}) := f_{\ell}^* \left(\mathbf{B}_{\ell, \mathcal{A}_{\ell}^{\circ}} Y_{\mathcal{A}_{\ell}^{\circ}} + \sum_{h \in \mathcal{A}_{\ell}^{\partial}} \mathbf{B}_{\ell,h} Y_h^{\nu - \delta_{e,\ell}} \right),$$

where the second term is absent when $\ell \in \mathcal{V}_r^{\circ}$.

Step 1: Subtree message identity. For each $i \in \mathcal{V}_r$, set $q_i := (\text{dist}_{\mathcal{F}_r}(e, i) - 1)/2$. Thus $q_i = \delta_{e,i}$ for boundary nodes. Since $q_i + 1 \leq \text{diam}(\mathcal{T}_r) \leq \nu$, every message in the expansion has index $\nu - q_i \geq 1$ and satisfies (18); no initial message function remains. We prove that

$$\mu_{i \rightarrow \text{pa}(i)}^{\nu - q_i}(Y_{\text{pa}(i)}) = \min_{Y_{\mathcal{E}_i}} \sum_{\ell \in \mathcal{V}_i} \Psi_{\ell}^{\nu,e}(Y_{\mathcal{A}_{\ell}^{\circ}}). \quad (73)$$

We prove (73) by induction on the height of $\mathcal{T}_i$.

Base case. If $\mathcal{V}_i = \{i\}$, then $\mathcal{E}_i = \emptyset$ and $\mathcal{A}_i^{\circ} = \{\text{pa}(i)\}$. Evaluating (18) with $\nu + 1$ replaced by $\nu - q_i$ gives

$$\mu_{i \rightarrow \text{pa}(i)}^{\nu - q_i}(Y_{\text{pa}(i)}) = \Psi_i^{\nu,e}(Y_{\mathcal{A}_i^{\circ}}),$$

which is (73) for the base case.

Inductive step. Assume now that (73) holds for all strict descendants of i . The same substitution in (18) gives

$$\mu_{i \rightarrow \text{pa}(i)}^{\nu - q_i}(Y_{\text{pa}(i)}) = \min_{Y_{\mathcal{A}_i^{\circ} \setminus \{\text{pa}(i)\}}} \left\{ \Psi_i^{\nu,e}(Y_{\mathcal{A}_i^{\circ}}) + \sum_{h \in \mathcal{A}_i^{\circ} \setminus \{\text{pa}(i)\}} \sum_{j \in \mathcal{I}_h \setminus \{i\}} \mu_{j \rightarrow h}^{\nu - q_i - 1}(Y_h) \right\}. \quad (74)$$

For each pair (h, j) in the double sum, $\text{pa}(h) = i$, $\text{pa}(j) = h$, and $q_j = q_i + 1$. The induction hypothesis therefore gives

$$\mu_{j \rightarrow h}^{\nu - q_i - 1}(Y_h) = \mu_{j \rightarrow \text{pa}(j)}^{\nu - q_j}(Y_{\text{pa}(j)}) = \min_{Y_{\mathcal{E}_j}} \sum_{\ell \in \mathcal{V}_j} \Psi_{\ell}^{\nu, e}(Y_{\mathcal{A}_{\ell}^{\circ}}).$$

Since $h = \text{pa}(j)$, the pairs (h, j) correspond one-to-one to $j \in \mathcal{V}_i \setminus \{i\}$ with $\text{pa}(\text{pa}(j)) = i$. The sets $\mathcal{A}_i^{\circ} \setminus \{\text{pa}(i)\}$ and these $\mathcal{E}_j$ partition $\mathcal{E}_i$; hence substituting into (74), reindexing by j , and merging the minimizations gives

$$\mu_{i \rightarrow \text{pa}(i)}^{\nu - q_i}(Y_{\text{pa}(i)}) = \min_{Y_{\mathcal{E}_i}} \left\{ \Psi_i^{\nu, e}(Y_{\mathcal{A}_i^{\circ}}) + \sum_{\substack{j \in \mathcal{V}_i \setminus \{i\} \\ \text{pa}(\text{pa}(j)) = i}} \sum_{\ell \in \mathcal{V}_j} \Psi_{\ell}^{\nu, e}(Y_{\mathcal{A}_{\ell}^{\circ}}) \right\}.$$

Using $\mathcal{V}_i = \{i\} \sqcup \bigsqcup_{\substack{j \in \mathcal{V}_i \setminus \{i\} \\ \text{pa}(\text{pa}(j)) = i}} \mathcal{V}_j$, we obtain

$$\mu_{i \rightarrow \text{pa}(i)}^{\nu - q_i}(Y_{\text{pa}(i)}) = \min_{Y_{\mathcal{E}_i}} \sum_{\ell \in \mathcal{V}_i} \Psi_{\ell}^{\nu, e}(Y_{\mathcal{A}_{\ell}^{\circ}}).$$

Step 2: Sum of messages at the root. Applying the identity to (17a), we can write

$$\hat{Y}_e^{\nu+1} \in \operatorname{argmin}_{Y_e} \left\{ \sum_{i \in \mathcal{I}_e} \mu_{i \rightarrow e}^{\nu}(Y_e) \right\}.$$

For $i \in \mathcal{I}_e$, one has $q_i = 0$, so (73) applies with $\text{pa}(i) = e$. The sets $\mathcal{V}_i$, $i \in \mathcal{I}_e$, partition $\mathcal{V}_r$, and the corresponding sets $\mathcal{E}_i$ partition $\mathcal{E}_r \setminus \{e\}$. Therefore

$$\hat{Y}_e^{\nu+1} \in \operatorname{argmin}_{Y_e} \min_{Y_{\mathcal{E}_r \setminus \{e\}}} \sum_{\ell \in \mathcal{V}_r} \Psi_{\ell}^{\nu, e}(Y_{\mathcal{A}_{\ell}^{\circ}}).$$

This is exactly (23b). $\square$

A.2 Proof of Lemma 3

The next two lemmas isolate the two ingredients for Lemma 3: for each fixed boundary vector, the minimization over $Y_{\mathcal{E}_r}$ generates a common primal variable on $\mathcal{T}_r$, and the linear system $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_{\mathcal{E}_r} = q$ admits an explicit graph-compliant solution. For each block r , write $N_r := |\mathcal{V}_r|$ and $\hat{\mathcal{J}}_r := |\hat{\partial}_r|$. For $E \in \mathbb{R}^{\hat{\mathcal{J}}_r \times d}$ with rows indexed by $(i, h) \in \hat{\partial}_r$, define

$$c_i(E) := \sum_{h \in \mathcal{A}_i^{\partial}} \mathbf{B}_{i, h} E_{i, h}, \quad \eta(E) := \sum_{i \in \mathcal{V}_r} c_i(E) = \sum_{(i, h) \in \hat{\partial}_r} \mathbf{B}_{i, h} E_{i, h}.$$

Also write, for a minimizer $Y_r(E)$ in (30),

$$\bar{x}_i(E) := \nabla f_i^*((\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_r(E))_i + c_i(E)), \quad i \in \mathcal{V}_r.$$

For $e \in \mathcal{E}_r$ and $v \in \mathcal{V}_r$, write

$$Y_{r, e}(E) := (Y_r(E))_e, \quad Z_{r, v}(E) := \mathbf{B}_{v, \mathcal{E}_r} Y_r(E).$$

Lemma 9 *Under Assumptions 1 and 3, the vectors $\bar{x}_i(E)$, $i \in \mathcal{V}_r$, have a common value, which depends on E only through $\eta(E)$; write it as $\chi_r(\eta(E))$. That is*

$$\bar{x}_i(E) = \chi_r(\eta(E)), \quad i \in \mathcal{V}_r, \quad (75a)$$

$$\sum_{i \in \mathcal{V}_r} \nabla f_i(\chi_r(\eta(E))) = \eta(E), \quad (75b)$$

$$(\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_r(E))_i = \nabla f_i(\chi_r(\eta(E))) - c_i(E), \quad i \in \mathcal{V}_r. \quad (75c)$$

Proof Fix E and the minimizer $Y_r(E)$. For $e = (u, v) \in \mathcal{E}_r$, the optimality condition gives

$$0 = \mathbf{B}_{u,e} \bar{x}_u(E) + \mathbf{B}_{v,e} \bar{x}_v(E) = \mathbf{B}_{u,e} (\bar{x}_u(E) - \bar{x}_v(E)).$$

Thus $\bar{x}_u(E) = \bar{x}_v(E)$ for every $e = (u, v) \in \mathcal{E}_r$. Since $\mathcal{T}_r$ is connected, all $\bar{x}_i(E)$ are equal. Because $\nabla f_i = (\nabla f_i^*)^{-1}$, $\nabla f_i(\bar{x}_i(E)) = (\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_r(E))_i + c_i(E)$ for $i \in \mathcal{V}_r$. Summing over $i \in \mathcal{V}_r$ and using the zero column sums of $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r}$ gives $\sum_{i \in \mathcal{V}_r} \nabla f_i(\bar{x}_i(E)) = \eta(E)$. Since all $\bar{x}_i(E)$ are equal, this equation identifies their common value as $\chi_r(\eta(E))$. Thus the common value depends on E only through $\eta(E)$, and (75a)–(75b) hold. Rearranging the first identity before summation gives (75c). $\square$

Lemma 10 *For any $q = (q_i)_{i \in \mathcal{V}_r}$ with $q_i \in \mathbb{R}^d$ and $\sum_{i \in \mathcal{V}_r} q_i = 0$, the equation $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_{\mathcal{E}_r} = q$ has a unique solution. If $e \in \mathcal{E}_r$ and $\mathcal{S}$ is one of the two connected components of $\mathcal{T}_r \setminus \{e\}$, then $Y_e = \varepsilon_{e, \mathcal{S}} \sum_{i \in \mathcal{S}} q_i$ for some sign $\varepsilon_{e, \mathcal{S}} \in \{\pm 1\}$.*

Proof Each column of $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r}$ sums to zero. Since $\mathcal{T}_r$ is a tree, $\text{rank}(\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r}) = |\mathcal{E}_r| = |\mathcal{V}_r| - 1$, and hence $\text{Range}(\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r}) = \{q \in \mathbb{R}^{|\mathcal{V}_r| \times d} : \sum_{i \in \mathcal{V}_r} q_i = 0\}$. Thus $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_{\mathcal{E}_r} = q$ has a unique solution. Fix one component $\mathcal{S}$ of $\mathcal{T}_r \setminus \{e\}$. Summing the rows indexed by $\mathcal{S}$ gives

$$\sum_{i \in \mathcal{S}} q_i = \sum_{a \in \mathcal{E}_r} \left(\sum_{i \in \mathcal{S}} \mathbf{B}_{i,a} \right) Y_a = \left(\sum_{i \in \mathcal{S}} \mathbf{B}_{i,e} \right) Y_e, \quad \sum_{i \in \mathcal{S}} \mathbf{B}_{i,e} \in \{\pm 1\},$$

where the terms with $a \neq e$ vanish because those edges do not cross from $\mathcal{S}$ to $\mathcal{V}_r \setminus \mathcal{S}$. This gives the formula. $\square$

Proof of Lemma 3. For $E \in \mathbb{R}^{\hat{\mathcal{T}}_r \times d}$ with rows indexed by $(i, h) \in \hat{\partial}_r$, set $q_i(E) := \nabla f_i(\chi_r(\eta(E))) - c_i(E)$. Lemma 9 gives $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_r(E) = q(E)$, and $\sum_{i \in \mathcal{V}_r} q_i(E) = 0$. Lemma 10 gives a unique solution of $\mathbf{B}_{\mathcal{V}_r, \mathcal{E}_r} Y_{\mathcal{E}_r} = q(E)$. Hence $Y_r(E)$ is single-valued. It also gives, for a boundary incidence $(i, h) \in \hat{\partial}_r$ and the component $\bar{\mathcal{V}}_e(i)$ not containing i ,

$$Y_{r,e}(E) = \varepsilon_{e,i} \sum_{j \in \bar{\mathcal{V}}_e(i)} (\nabla f_j(\chi_r(\eta(E))) - c_j(E)), \quad \varepsilon_{e,i} \in \{\pm 1\}. \quad (76)$$

Let E, E' differ only in coordinate (i, h) . Since $i \notin \bar{\mathcal{V}}_e(i)$, the boundary terms $c_j(\cdot)$ in (76) are unchanged for all $j \in \bar{\mathcal{V}}_e(i)$. Also $\eta(E') - \eta(E) = \mathbf{B}_{i,h}(E'_{i,h} - E_{i,h})$. Hence

$$Y_{r,e}(E') - Y_{r,e}(E) = \varepsilon_{e,i} (\mathcal{L}_{r, \bar{\mathcal{V}}_e(i)}(\eta(E')) - \mathcal{L}_{r, \bar{\mathcal{V}}_e(i)}(\eta(E))),$$

and (34a) follows from the definition of $\kappa_{\partial, r}$.

For $v \neq i$, the coordinate (i, h) does not enter c_v , so $c_v(E') = c_v(E)$. By (75c) and the definition of $Z_{r,v}$, $Z_{r,v}(E) = \nabla f_v(\chi_r(\eta(E))) - c_v(E)$. Since $\mathcal{L}_{r, \{v\}}(\eta) = \nabla f_v(\chi_r(\eta))$, we obtain

$$Z_{r,v}(E') - Z_{r,v}(E) = \mathcal{L}_{r, \{v\}}(\eta(E')) - \mathcal{L}_{r, \{v\}}(\eta(E)).$$

Then, (34b) follows from the definition of $\kappa_{\partial, r}$. $\square$

A.3 Proof of Lemma 4

The proof combines the sensitivity bounds in Lemma 3 with the following counting bounds for the rooted tree.

Lemma 11 Fix $(i, h) \in \widehat{\partial}_r$ and root the active tree $\mathcal{T}_r$ at the boundary node i . For $v \in \mathcal{V}_r$, set

$$\delta_{r,i}(v) := \max\{\text{dist}_{\mathcal{T}_r}(i, v) - 1, 0\}. \quad (77)$$

For $v \neq i$, let $\text{Ch}_{r,i}(v)$ be the set of active edges from v to its children in this rooted tree; set $\text{Ch}_{r,i}(i) := \emptyset$. Let $p(v)$ be the active edge from $v \neq i$ to its parent, and set

$$n_{r,i}(i) := N_r, \quad n_{r,i}(v) := |\overline{\mathcal{V}}_{p(v)}(i)|, \quad v \neq i.$$

Then

$$\sum_{e \in \text{Ch}_{r,i}(v)} |\overline{\mathcal{V}}_e(i)| \leq n_{r,i}(v) - 1, \quad \sum_{\text{dist}_{\mathcal{T}_r}(i,v)=t} \frac{(n_{r,i}(v) - 1)^2}{N_r^2} \leq 1, \quad t \geq 0. \quad (78)$$

Proof For the first bound in (78), take $v \neq i$ and put $p = p(v)$. The components $\overline{\mathcal{V}}_e(i)$ with $e \in \text{Ch}_{r,i}(v)$ are pairwise disjoint, and their union is $\overline{\mathcal{V}}_p(i) \setminus \{v\}$. Hence equality holds in the first bound of (78) for $v \neq i$. If $v = i$, then $\text{Ch}_{r,i}(v) = \emptyset$, and the displayed inequality is immediate.

Fix $t \geq 1$. If two distinct nodes v, w have $\text{dist}_{\mathcal{T}_r}(i, v) = \text{dist}_{\mathcal{T}_r}(i, w) = t$, then $\overline{\mathcal{V}}_{p(v)}(i) \cap \overline{\mathcal{V}}_{p(w)}(i) = \emptyset$: otherwise some node u would have both v and w on the unique path from i to u , and the two vertices at depth t on this path would have to coincide. Hence

$$\begin{aligned} \sum_{\text{dist}_{\mathcal{T}_r}(i,v)=t} n_{r,i}(v) &\leq N_r. \\ \sum_{\text{dist}_{\mathcal{T}_r}(i,v)=t} (n_{r,i}(v) - 1)^2 &\leq \left(\sum_{\text{dist}_{\mathcal{T}_r}(i,v)=t} (n_{r,i}(v) - 1) \right)^2 \leq N_r^2, \end{aligned}$$

The same conclusion is immediate for $t = 0$, since $n_{r,i}(i) = N_r$. This proves the second bound in (78). $\square$

Proof of Lemma 4. By (24), $\bar{Z}_r^{\nu+1} = \mathbf{B}_{\mathcal{V}, \mathcal{E}_r} Y_r(E^\nu)$ with $(E^\nu)_{(i,h)} = Y_h^\nu$, and $\hat{Y}_e^{\nu+1} = \mathbf{Y}_{r,e}(\hat{E}_e^\nu)$ with $(\hat{E}_e^\nu)_{(i,h)} = Y_h^{\nu-\delta_{e,i}}$ for $(i, h) \in \widehat{\partial}_r$. For $v \in \mathcal{V}_r$, define the node mismatch

$$m_v := (\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1})_v.$$

The vector $\hat{E}_e^\nu$ depends on the incident edge e , whereas m_v fixes the node v . To estimate m_v , insert the auxiliary node-aligned boundary vector

$$(E_v^\nu)_{(i,h)} := Y_h^{\nu-\delta_{r,i}(v)}, \quad (i, h) \in \widehat{\partial}_r.$$

For $e \in \mathcal{A}_v^\circ$, the definition of $\delta_{e,i}$ gives $\delta_{e,i} = \delta_{r,i}(v)$ unless $e \in \text{Ch}_{r,i}(v)$, in which case $\delta_{e,i} = \delta_{r,i}(v) + 1$. At $v = i$, both delays are zero for every incident edge. Thus E_v^ν uses, in each boundary coordinate (i, h) , the freshest value seen by any active edge incident to v . Since $(\hat{Z}_r^{\nu+1})_v = \sum_{e \in \mathcal{A}_v^\circ} \mathbf{B}_{v,e} \mathbf{Y}_{r,e}(\hat{E}_e^\nu)$ and $(\bar{Z}_r^{\nu+1})_v = \mathbf{Z}_{r,v}(E^\nu)$, inserting and subtracting $\mathbf{Z}_{r,v}(E_v^\nu)$ gives

$$m_v = \underbrace{\sum_{e \in \mathcal{A}_v^\circ} \mathbf{B}_{v,e} \left(\mathbf{Y}_{r,e}(\hat{E}_e^\nu) - \mathbf{Y}_{r,e}(E_v^\nu) \right)}_{=: p_v} + \underbrace{\mathbf{Z}_{r,v}(E_v^\nu) - \mathbf{Z}_{r,v}(E^\nu)}_{=: q_v}.$$

Write $\Delta Y_h^\ell := Y_h^{\ell+1} - Y_h^\ell$.

Estimate of p_v . Fix an ordering of the boundary coordinates $(i, h) \in \widehat{\partial}_r$. For each incident edge $e \in \mathcal{A}_v^\circ$, telescope the difference from E_v^ν to $\hat{E}_e^\nu$ by changing one boundary coordinate at a time. In the step that changes coordinate (i, h) , let $E_{e,(i,h)}^-$ and $E_{e,(i,h)}^+$ denote the two consecutive boundary vectors. Then

$$\mathbf{Y}_{r,e}(\hat{E}_e^\nu) - \mathbf{Y}_{r,e}(E_v^\nu) = \sum_{(i,h) \in \widehat{\partial}_r} \left(\mathbf{Y}_{r,e}(E_{e,(i,h)}^+) - \mathbf{Y}_{r,e}(E_{e,(i,h)}^-) \right),$$

where the (i, h) -summand is zero if $(\widehat{E}_e^\nu)_{(i, h)} = (E_v^\nu)_{(i, h)}$. The preceding observation shows that this coordinate difference is zero unless $e \in \text{Ch}_{r, i}(v)$. If $e \in \text{Ch}_{r, i}(v)$, then

$$(\widehat{E}_e^\nu)_{(i, h)} - (E_v^\nu)_{(i, h)} = -\Delta Y_h^{\nu - (\delta_{r, i}(v) + 1)}.$$

Therefore

$$p_v = \sum_{(i, h) \in \widehat{\partial}_r} \underbrace{\sum_{e \in \text{Ch}_{r, i}(v)} \text{B}_{v, e}(\Upsilon_{r, e}(E_{e, (i, h)}^+) - \Upsilon_{r, e}(E_{e, (i, h)}^-))}_{=: p_{v, (i, h)}}.$$

If $\text{Ch}_{r, i}(v) = \emptyset$, then $p_{v, (i, h)} = 0$. Otherwise, using the sensitivity bound (34a) and the first bound in (78),

$$\|p_{v, (i, h)}\| \leq \kappa_{\partial, r} \frac{n_{r, i}(v) - 1}{N_r} \|\Delta Y_h^{\nu - (\delta_{r, i}(v) + 1)}\|.$$

By Cauchy-Schwarz over the $\widehat{J}_r$ boundary coordinates,

$$\|p_v\|^2 \leq \widehat{J}_r \kappa_{\partial, r}^2 \sum_{\substack{(i, h) \in \widehat{\partial}_r \\ \text{Ch}_{r, i}(v) \neq \emptyset}} \frac{(n_{r, i}(v) - 1)^2}{N_r^2} \|\Delta Y_h^{\nu - (\delta_{r, i}(v) + 1)}\|^2. \quad (79)$$

Summing (79) over v and exchanging the v - and (i, h) -sums gives

$$\sum_{v \in \mathcal{V}_r} \|p_v\|^2 \leq \widehat{J}_r \kappa_{\partial, r}^2 \sum_{(i, h) \in \widehat{\partial}_r} \sum_{\substack{v \in \mathcal{V}_r \\ \text{Ch}_{r, i}(v) \neq \emptyset}} \frac{(n_{r, i}(v) - 1)^2}{N_r^2} \|\Delta Y_h^{\nu - (\delta_{r, i}(v) + 1)}\|^2.$$

For fixed (i, h) , group the inner sum by rooted depth $t = \text{dist}_{\mathcal{T}_r}(i, v)$. On the nodes with $\text{Ch}_{r, i}(v) \neq \emptyset$, one has $v \neq i$ and $1 \leq \delta_{r, i}(v) + 1 = t \leq D_r$. Thus all nodes in the same depth shell use the same increment $\Delta Y_h^{\nu - t}$, and the second bound in (78) gives

$$\sum_{\substack{v \in \mathcal{V}_r \\ \text{Ch}_{r, i}(v) \neq \emptyset}} \frac{(n_{r, i}(v) - 1)^2}{N_r^2} \|\Delta Y_h^{\nu - (\delta_{r, i}(v) + 1)}\|^2 \leq \sum_{t=1}^{D_r} \|\Delta Y_h^{\nu - t}\|^2.$$

Summing over (i, h) gives

$$\sum_{v \in \mathcal{V}_r} \|p_v\|^2 \leq \widehat{J}_r \kappa_{\partial, r}^2 \sum_{\ell=\nu-D_r}^{\nu-1} \sum_{(i, h) \in \widehat{\partial}_r} \|\Delta Y_h^\ell\|^2. \quad (80)$$

Estimate of q_v . Telescope the difference from E^ν to E_v^ν by changing one boundary coordinate at a time. If $v = i$, then $\delta_{r, i}(v) = 0$, so the corresponding coordinate difference is zero. Hence only coordinates (i, h) with $v \neq i$ remain, and the sensitivity bound (34b) applies:

$$\|q_v\|^2 \leq \widehat{J}_r \kappa_{\partial, r}^2 \sum_{(i, h) \in \widehat{\partial}_r: i \neq v} \frac{1}{N_r^2} \|Y_h^{\nu - \delta_{r, i}(v)} - Y_h^\nu\|^2. \quad (81)$$

For every integer $t \geq 0$,

$$\|Y_h^{\nu - t} - Y_h^\nu\|^2 \leq t \sum_{m=1}^t \|\Delta Y_h^{\nu - m}\|^2.$$

Substituting this into (81), summing over v , and using the bound $\sum_{v \in \mathcal{V}_r} \delta_{r, i}(v) \leq N_r D_r$, which follows from $0 \leq \delta_{r, i}(v) \leq D_r$, together with $D_r/N_r \leq 1$, we obtain

$$\sum_{v \in \mathcal{V}_r} \|q_v\|^2 \leq \widehat{J}_r \kappa_{\partial, r}^2 \sum_{\ell=\nu-D_r}^{\nu-1} \sum_{(i, h) \in \widehat{\partial}_r} \|\Delta Y_h^\ell\|^2. \quad (82)$$

Finally,

$$\|\widehat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 = \sum_{v \in \mathcal{V}_r} \|m_v\|^2 \leq 2 \sum_{v \in \mathcal{V}_r} \|p_v\|^2 + 2 \sum_{v \in \mathcal{V}_r} \|q_v\|^2.$$

Combining (80) and (82) proves (37). $\square$

A.4 Proof of Proposition 6

For $Z \in \mathcal{S}$, put $x_i := \nabla f_i^*(Z_i)$. We first express the block progress terms in (19). Let $h = \{i, j\} \in \bar{\mathcal{E}}_1$ be a singleton edge block, and write Δ_h for the progress of this block. Changing only Y_h by s changes the image coordinates to $Z_i + \mathbf{B}_{i,h}s$ and $Z_j + \mathbf{B}_{j,h}s$. Since $\mathbf{B}_{i,h} + \mathbf{B}_{j,h} = 0$ and $|\mathbf{B}_{i,h}| = 1$, the singleton block minimization in (19) reduces to

$$\min_{u_i + u_j = Z_i + Z_j} f_i^*(u_i) + f_j^*(u_j).$$

Let (u_i^h, u_j^h) be its minimizer. The first-order condition gives $\nabla f_i^*(u_i^h) = \nabla f_j^*(u_j^h) = p_h$. Thus $u_i^h = \nabla f_i(p_h)$ and $u_j^h = \nabla f_j(p_h)$, where p_h is the unique solution of $\nabla f_i(p_h) + \nabla f_j(p_h) = Z_i + Z_j$. By the definition of the block progress and this identity,

$$\begin{aligned} \Delta_h(Z) &= f_i^*(Z_i) + f_j^*(Z_j) - f_i^*(u_i^h) - f_j^*(u_j^h) \\ &= \sum_{k \in \{i, j\}} [f_k^*(Z_k) - f_k^*(\nabla f_k(p_h)) - \langle p_h, Z_k - \nabla f_k(p_h) \rangle]. \end{aligned}$$

By the definition of x_k , $\nabla f_k(x_k) = Z_k$. Fenchel's identity gives

$$f_k(x_k) + f_k^*(Z_k) = \langle Z_k, x_k \rangle, \quad f_k(p_h) + f_k^*(\nabla f_k(p_h)) = \langle \nabla f_k(p_h), p_h \rangle.$$

Substituting these identities yields, for $k \in \{i, j\}$,

$$f_k^*(Z_k) - f_k^*(\nabla f_k(p_h)) - \langle p_h, Z_k - \nabla f_k(p_h) \rangle = f_k(p_h) - f_k(x_k) - \langle \nabla f_k(x_k), p_h - x_k \rangle.$$

Define the Bregman divergence $D_{f_i}(u, v) := f_i(u) - f_i(v) - \langle \nabla f_i(v), u - v \rangle$. Then

$$\Delta_h(Z) = D_{f_i}(p_h, x_i) + D_{f_j}(p_h, x_j).$$

Write $\|u\|_R^2 := u^\top R u$ for $u \in \mathbb{R}^d$. The lower Hessian bound in Proposition 6 yields

$$\Delta_h(Z) \geq \frac{\alpha}{2} (\|p_h - x_i\|_R^2 + \|p_h - x_j\|_R^2) \geq \frac{\alpha}{4} \|x_i - x_j\|_R^2. \quad (83)$$

The same argument applied to (19) for the tree block yields $u_i = \nabla f_i(p_0)$, $i \in \mathcal{V}_1$, where p_0 is the unique solution of $\sum_{i \in \mathcal{V}_1} \nabla f_i(p_0) = \sum_{i \in \mathcal{V}_1} Z_i$. Let $\mathcal{I}_\partial := \bigcup_{h \in \bar{\mathcal{E}}_1} \mathcal{I}_h$ be the set of agents incident to the singleton edge blocks. The corresponding progress Δ_1 satisfies

$$\Delta_1(Z) = \sum_{i \in \mathcal{V}_1} D_{f_i}(p_0, x_i) \geq \frac{\alpha}{2} \sum_{i \in \mathcal{I}_\partial \cap \mathcal{V}_1} \|x_i - p_0\|_R^2. \quad (84)$$

We next compare the global gap with the block progress. By Fenchel's identity,

$$\sum_{i \in \mathcal{V}} D_{f_i}(p, x_i) = \sum_{i \in \mathcal{V}} (f_i(p) - f_i(x_i) - \langle Z_i, p - x_i \rangle) = G(Z) + \sum_{i \in \mathcal{V}} f_i(p) - \left\langle \sum_{i \in \mathcal{V}} Z_i, p \right\rangle.$$

Since $Z \in \mathcal{S} = \text{Range}(\mathbf{B})$, $\sum_{i \in \mathcal{V}} Z_i = 0$. Taking $p = p_0$, using $\mathcal{V} \setminus \mathcal{V}_1 \subseteq \mathcal{I}_\partial$, and applying weak duality for (P) and (D), which gives $-G^* \leq \sum_{i \in \mathcal{V}} f_i(p_0)$, we obtain

$$\begin{aligned} G(Z) - G^* &\leq G(Z) + \sum_{i \in \mathcal{V}} f_i(p_0) = \sum_{i \in \mathcal{V}} D_{f_i}(p_0, x_i) = \Delta_1(Z) + \sum_{i \in \mathcal{V} \setminus \mathcal{V}_1} D_{f_i}(p_0, x_i) \\ &\leq \Delta_1(Z) + \frac{\bar{\alpha}}{2} \sum_{i \in \mathcal{I}_\partial} \|x_i - p_0\|_R^2. \end{aligned}$$

It remains to bound the finite boundary term by block progress. Extending $\mathcal{T}_1$ to a spanning tree of $\mathcal{G}$ gives $|\mathcal{V} \setminus \mathcal{V}_1| \leq |\bar{\mathcal{E}}_1|$. Each vertex outside $\mathcal{V}_1$ has a path to $\mathcal{V}_1$ with at most $|\bar{\mathcal{E}}_1|$

edges in $\bar{\mathcal{E}}_1$, ending at a vertex of $\mathcal{I}_\partial \cap \mathcal{V}_1$. Applying Cauchy's inequality along these paths and summing gives, with $C := 1 + |\bar{\mathcal{E}}_1|(|\bar{\mathcal{E}}_1| + 1) = O(1)$,

$$\sum_{i \in \mathcal{I}_\partial} \|x_i - p_0\|_R^2 \leq C \left(\sum_{i \in \mathcal{I}_\partial \cap \mathcal{V}_1} \|x_i - p_0\|_R^2 + \sum_{\{i,j\} \in \bar{\mathcal{E}}_1} \|x_i - x_j\|_R^2 \right).$$

Combining the last two bounds with (84) and (83) gives

$$G(Z) - G^* \leq \left(1 + 2C \frac{\bar{\alpha}}{\underline{\alpha}}\right) \left(\Delta_1(Z) + \sum_{h \in \bar{\mathcal{E}}_1} \Delta_h(Z) \right),$$

which is exactly the block error bound for the partition $\mathcal{P} = \{\mathcal{E}_1\} \cup \{\{h\} : h \in \bar{\mathcal{E}}_1\}$. Since $C = O(1)$ and $\bar{\alpha}/\underline{\alpha} = O(1)$, this proves $\kappa_{\mathcal{P}} = O(1)$.

Finally we bound the constants in Theorem 1. The partition has $p = 1 + |\bar{\mathcal{E}}_1| = O(1)$. Since $\Delta_r(Z) \leq G(Z) - G^*$, the block error bound implies $\kappa_{\mathcal{P}} \geq 1/p$, and hence $\kappa_{\mathcal{P}} = \Theta(1)$. Singleton blocks have no interior propagation, so $D_r = 0$ for them; only block 1 can contribute to A . For block 1, $\hat{J}_1 \leq 2|\bar{\mathcal{E}}_1| = O(1)$. Every neighboring block is a singleton edge $\mathcal{E}_s = \{h\}$, and its incidence column has norm $\sqrt{2}$, so $\Gamma_1 \leq 1$ by its definition. Moreover, for $S \subseteq \mathcal{V}_1$, the map $\sum_{i \in S} \nabla f_i$ is $|S|L$ -Lipschitz, while $(\sum_{i \in \mathcal{V}_1} \nabla f_i)^{-1}$ is $(N_1 \mu_1)^{-1}$ -Lipschitz. Hence $\kappa_{\partial,1} \leq L/\mu_1 = O(1)$. Together with the uniform bounds on μ_1^{-1} and L , the definition of A gives $A = O(1)$. Choose $\tau = \kappa_{\mathcal{P}} \log D/D = \Theta(\log D/D)$. The first inequality in (42) clearly holds for sufficiently large D . Moreover, $\rho = 1 - \log D/(2D)$, so $\rho^{-D} = D^{1/2}(1 + o(1))$ and

$$8LA\tau^2 \sum_{j=1}^D \rho^{-j} = 16LA\kappa_{\mathcal{P}}\tau(\rho^{-D} - 1) = O\left(\frac{\log D}{\sqrt{D}}\right) \rightarrow 0.$$

Thus, (42) holds for all sufficiently large D , yielding $(1 - \rho)^{-1} = 2D/\log D$. $\square$

A.5 Proof of Proposition 7

For $Z \in \mathcal{S}$, set $G_0(Z) := \frac{1}{2} \text{tr}(ZR^{-1}Z^\top)$. Let $\kappa_{\mathcal{P}}^{(0)}$ be its smallest admissible block error-bound constant and set $W := ZR^{-1/2}$. This bijection maps $\mathcal{S}$ and each $\mathcal{U}_r$ onto themselves, so block error-bound constants are invariant under the corresponding change of objective. Moreover, $G_0(WR^{1/2}) = \frac{1}{2} \|W\|^2$. Below, we write Z for W and work with $G_0(Z) = \frac{1}{2} \|Z\|^2$. We first prove

$$\frac{D+2}{4\sin^2(\pi/p)} \leq \kappa_{\mathcal{P}}^{(0)} \leq 1 + \frac{3(D+1)}{2\sin^2(\pi/p)}, \quad \kappa_{\mathcal{P}}^{(0)} = \Theta\left(\frac{m^2}{D}\right), \quad (85)$$

For a path block $\mathcal{E}_r$, the incidence matrix has rank $|\mathcal{V}_r| - 1$, and hence

$$\mathcal{U}_r = \left\{ U \in \mathbb{R}^{m \times d} : U_i = 0 \ (i \notin \mathcal{V}_r), \sum_{i \in \mathcal{V}_r} U_i = 0 \right\}.$$

By definition, $\Delta_r^{(0)}(Z) = \frac{1}{2} \|Z\|^2 - \min_{U \in \mathcal{U}_r} \frac{1}{2} \|Z + U\|^2$. The minimum is attained when $Z_i + U_i = \bar{Z}_{\mathcal{V}_r} := |\mathcal{V}_r|^{-1} \sum_{j \in \mathcal{V}_r} Z_j$ for $i \in \mathcal{V}_r$. Hence

$$\Delta_r^{(0)}(Z) = \frac{1}{2} \sum_{i \in \mathcal{V}_r} \|Z_i - \bar{Z}_{\mathcal{V}_r}\|^2.$$

Relabel the blocks by $r = 0, \dots, p-1$ and index the ring nodes cyclically as $v_0, \dots, v_{m-1}$. Let

$$\mathcal{Q}_r := \{v_{r(D+1)+j} : 0 \leq j \leq D\}, \quad b_r := v_{(r+1)(D+1)},$$

where indices are taken modulo m . The sets Q_r form a disjoint partition of $\mathcal{V}$, with $|Q_r| = D+1$. Since b_r is the first node of Q_{r+1} , the nodes incident to block r are $\mathcal{V}_r = Q_r \cup \{b_r\}$. With $\xi_r := (D+1)^{-1} \sum_{i \in Q_r} Z_i$, we have

$$2\Delta_r^{(0)}(Z) = \sum_{i \in Q_r} \|Z_i - \xi_r\|^2 + \frac{D+1}{D+2} \|Z_{b_r} - \xi_r\|^2.$$

Thus

$$\sum_{i \in Q_r} \|Z_i - \xi_r\|^2 \leq 2\Delta_r^{(0)}(Z), \quad \|Z_{b_r} - \xi_r\|^2 \leq 4\Delta_r^{(0)}(Z). \quad (86)$$

Since $b_r \in Q_{r+1}$, (86) gives

$$\|\xi_{r+1} - \xi_r\|^2 \leq 2\|\xi_r - Z_{b_r}\|^2 + 2\|Z_{b_r} - \xi_{r+1}\|^2 \leq 8\Delta_r^{(0)}(Z) + 4\Delta_{r+1}^{(0)}(Z).$$

Summing cyclically yields

$$\sum_{r=0}^{p-1} \|\xi_{r+1} - \xi_r\|^2 \leq 12 \sum_{r=0}^{p-1} \Delta_r^{(0)}(Z). \quad (87)$$

Since $Z \in \mathcal{S}$ and the sets Q_r form a partition of $\mathcal{V}$, $\sum_{r=0}^{p-1} \xi_r = (D+1)^{-1} \sum_{i \in \mathcal{V}} Z_i = 0$. Set $\Xi := (\xi_0^\top, \dots, \xi_{p-1}^\top)^\top \in \mathbb{R}^{p \times d}$ and

$$C_p := \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ 1 & & & -1 \end{bmatrix}.$$

The cycle spectrum satisfies $\lambda_{\min}^+(C_p^\top C_p) = 4 \sin^2(\pi/p)$; see, e.g., [3]. Therefore,

$$\sum_{r=0}^{p-1} \|\xi_{r+1} - \xi_r\|^2 = \|C_p \Xi\|_F^2 \geq 4 \sin^2(\pi/p) \|\Xi\|_F^2 = 4 \sin^2(\pi/p) \sum_{r=0}^{p-1} \|\xi_r\|^2. \quad (88)$$

Combining (86), (87), and (88) yields

$$\|Z\|^2 = \sum_{r=0}^{p-1} \left[\sum_{i \in Q_r} \|Z_i - \xi_r\|^2 + (D+1) \|\xi_r\|^2 \right] \leq 2 \left(1 + \frac{3(D+1)}{2 \sin^2(\pi/p)} \right) \sum_{r=0}^{p-1} \Delta_r^{(0)}(Z).$$

This proves the upper bound in (85).

For the lower bound, choose a nonzero $\Xi \in \mathbb{R}^{p \times d}$ whose columns lie in the eigenspace of $C_p^\top C_p$ associated with $4 \sin^2(\pi/p)$, and denote its rows by ξ_r . Then

$$\sum_{r=0}^{p-1} \xi_r = 0, \quad \sum_{r=0}^{p-1} \|\xi_{r+1} - \xi_r\|^2 = 4 \sin^2(\pi/p) \sum_{r=0}^{p-1} \|\xi_r\|^2.$$

Set $Z_i = \xi_r$ for $i \in Q_r$, $r = 0, \dots, p-1$. Then $\sum_{i \in \mathcal{V}} Z_i = (D+1) \sum_{r=0}^{p-1} \xi_r = 0$ and

$$G_0(Z) = \frac{1}{2} \sum_{r=0}^{p-1} \sum_{i \in Q_r} \|Z_i\|^2 = \frac{D+1}{2} \sum_{r=0}^{p-1} \|\xi_r\|^2.$$

Moreover, $\mathcal{V}_r = Q_r \cup \{b_r\}$, with $Z_i = \xi_r$ for $i \in Q_r$ and $Z_{b_r} = \xi_{r+1}$, so $\bar{Z}_{\mathcal{V}_r} = ((D+1)\xi_r + \xi_{r+1})/(D+2)$. Substitution into the definition of $\Delta_r^{(0)}$ yields

$$\Delta_r^{(0)}(Z) = \frac{1}{2} [(D+1)\|\xi_r - \bar{Z}_{\mathcal{V}_r}\|^2 + \|\xi_{r+1} - \bar{Z}_{\mathcal{V}_r}\|^2] = \frac{D+1}{2(D+2)} \|\xi_{r+1} - \xi_r\|^2.$$

Hence $G_0(Z)/\sum_{r=0}^{p-1}\Delta_r^{(0)}(Z) = (D+2)/(4\sin^2(\pi/p))$ and (85) follows.

It remains to transfer the bound to G . Define $\widehat{G}(W) := G(WR^{1/2})$ for $W \in \mathcal{S}$. The Hessian bounds in Proposition 7 give $\bar{\alpha}^{-1}I \preceq \nabla^2[f_i^*(R^{1/2}\cdot)](W_i) \preceq \underline{\alpha}^{-1}I$. Let P_S and P_r be the Euclidean projectors onto $\mathcal{S}$ and $\mathcal{U}_r$, respectively, and define $\widehat{\Delta}_r(W) := \widehat{G}(W) - \min_{U \in W + \mathcal{U}_r} \widehat{G}(U)$. For $G_0(V) = \frac{1}{2}\|V\|^2$, the minimizer over $V + \mathcal{U}_r$ is $V - P_r V$; hence $\Delta_r^{(0)}(V) = \frac{1}{2}\|V\|^2 - \frac{1}{2}\|V - P_r V\|^2 = \frac{1}{2}\|P_r V\|^2$. The quadratic block error bound therefore gives $\sum_{r=0}^{p-1}\|P_r V\|^2 \geq (\kappa_{\mathcal{P}}^{(0)})^{-1}\|V\|^2$ for $V \in \mathcal{S}$. Since $\mathcal{U}_r \subseteq \mathcal{S}$, $P_r P_S = P_r$. The smoothness and strong convexity bounds above then yield

$$\sum_{r=0}^{p-1} \widehat{\Delta}_r(W) \geq \frac{\underline{\alpha}}{2} \sum_{r=0}^{p-1} \|P_r \nabla \widehat{G}(W)^\top\|^2 \geq \frac{\underline{\alpha}}{2\kappa_{\mathcal{P}}^{(0)}} \|P_S \nabla \widehat{G}(W)^\top\|^2 \geq \frac{\underline{\alpha}}{\bar{\alpha}\kappa_{\mathcal{P}}^{(0)}} (\widehat{G}(W) - G^*).$$

Therefore $\kappa_{\mathcal{P}} \leq (\bar{\alpha}/\underline{\alpha})\kappa_{\mathcal{P}}^{(0)}$, and (85) proves $\kappa_{\mathcal{P}} = O(m^2/D)$. $\square$

To obtain the rate bound, we apply Theorem 1 to $\widehat{f}_i(u) := f_i(R^{-1/2}u)$, which are $\underline{\alpha}$ -strongly convex and $\bar{\alpha}$ -smooth. Thus $\kappa_{\partial_r} \leq \bar{\alpha}/\underline{\alpha}$. With two boundary incidences, at most two neighbors per block, and $I_r \leq 1$, (41) gives

$$LA \leq 4 \left(2\frac{\bar{\alpha}}{\underline{\alpha}} + 1\right) \left(\frac{\bar{\alpha}}{\underline{\alpha}}\right)^2 = O(1).$$

With the same stepsize and initial data $Y^0 R^{-1/2}$ and $\mu_{i \rightarrow e}^0(\cdot R^{1/2})$, Algorithm 1 for $\widehat{f}_i$ generates edge iterates $Y^\nu R^{-1/2}$ and primal iterates $X^\nu R^{1/2}$. These fixed invertible transformations change only the constant C in the convergence bound, so the same ρ applies to the original iterates.

B Proofs of Results in Sec. 5

For the proof notation, fix a block r . Write $N_r := |\mathcal{V}_r|$, $J_r := |\partial_r|$, and $\widehat{J}_r := |\widehat{\partial}_r|$. For $E \in \mathbb{R}^{\widehat{J}_r \times d}$ with rows indexed by $(i, h) \in \widehat{\partial}_r$, use the boundary contribution $c_i(E) = \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} E_{i,h}$ from Appendix A.2, and write $c_r(E) := (c_i(E))_{i \in \mathcal{V}_r}$. For $U, \vartheta \in \mathbb{R}^{N_r \times d}$, the exact and surrogate block slices are

$$G_r(U) := \sum_{i \in \mathcal{V}_r} f_i^*(U_i), \quad \widetilde{G}_r(U; \vartheta) := \sum_{i \in \mathcal{V}_r} \widetilde{f}_i^*(U_i; \vartheta_i).$$

Only nodes in $\mathcal{V}_r$ appear in these slices; the remaining terms are constant under a block- r update. For the block solution map, write

$$\begin{aligned} \widetilde{Y}_{r,e}(E, \vartheta) &:= [\widetilde{Y}_r(E, \vartheta)]_e, & e \in \mathcal{E}_r, \\ \widetilde{Z}_{r,v}(E, \vartheta) &:= \mathbf{B}_{v,\mathcal{E}_r} \widetilde{Y}_r(E, \vartheta), & v \in \mathcal{V}_r. \end{aligned}$$

B.1 Proof of Lemma 7

Proof of (65a)–(65c). Fix a block r . For $E \in \mathbb{R}^{\widehat{J}_r \times d}$ with rows indexed by $(i, h) \in \widehat{\partial}_r$, set $\eta(E) := \sum_{i \in \mathcal{V}_r} c_i(E)$. Suppose $\vartheta, \widehat{\vartheta} \in \mathbb{R}^{N_r \times d}$ differ only in coordinate i . Define

$$u_j^0 := \widetilde{Z}_{r,j}(E, \vartheta) + c_j(E), \quad u_j^1 := \widetilde{Z}_{r,j}(E, \widehat{\vartheta}) + c_j(E).$$

Applying the same optimality condition for each active edge variable as in the proof of Lemma 9 to the two minimizers $\widetilde{Y}_r(E, \vartheta)$ and $\widetilde{Y}_r(E, \widehat{\vartheta})$, with f_j^* replaced by $\widetilde{f}_j^*(\cdot; \vartheta_j)$ and $\widetilde{f}_j^*(\cdot; \widehat{\vartheta}_j)$, respectively, gives common values $x^0, x^1 \in \mathbb{R}^d$ such that

$$\begin{aligned} u_j^0 &= (\nabla_u \widetilde{f}_j^*(\cdot; \vartheta_j))^{-1}(x^0), & u_j^1 &= (\nabla_u \widetilde{f}_j^*(\cdot; \widehat{\vartheta}_j))^{-1}(x^1), & j \in \mathcal{V}_r, \\ \sum_{j \in \mathcal{V}_r} u_j^0 &= \sum_{j \in \mathcal{V}_r} u_j^1 = \eta(E). \end{aligned}$$

By the definition of $\tilde{\mathcal{L}}_{r,S}$ in (61),

$$u_j^0 = \tilde{\mathcal{L}}_{r,\{j\}}(\eta(E); \vartheta), \quad u_j^1 = \tilde{\mathcal{L}}_{r,\{j\}}(\eta(E); \hat{\vartheta}).$$

Using (63) at the fixed input $\eta(E)$ gives

$$\|\tilde{\mathcal{Z}}_{r,i}(E, \hat{\vartheta}) - \tilde{\mathcal{Z}}_{r,i}(E, \vartheta)\| = \|\tilde{\mathcal{L}}_{r,\{i\}}(\eta(E); \hat{\vartheta}) - \tilde{\mathcal{L}}_{r,\{i\}}(\eta(E); \vartheta)\| \leq \tilde{\kappa}_{c,r} \|\hat{\vartheta}_i - \vartheta_i\|.$$

For $v \neq i$,

$$\|\tilde{\mathcal{Z}}_{r,v}(E, \hat{\vartheta}) - \tilde{\mathcal{Z}}_{r,v}(E, \vartheta)\| = \|\tilde{\mathcal{L}}_{r,\{v\}}(\eta(E); \hat{\vartheta}) - \tilde{\mathcal{L}}_{r,\{v\}}(\eta(E); \vartheta)\| \leq \frac{\tilde{\kappa}_{c,r}}{N_r} \|\hat{\vartheta}_i - \vartheta_i\|.$$

By Lemma 10, for some sign $\varepsilon_{e,i} \in \{\pm 1\}$,

$$\tilde{\mathcal{Y}}_{r,e}(E, \hat{\vartheta}) - \tilde{\mathcal{Y}}_{r,e}(E, \vartheta) = \varepsilon_{e,i} \sum_{v \in \bar{\mathcal{V}}_e(i)} (u_v^1 - u_v^0).$$

The choice $i \notin \bar{\mathcal{V}}_e(i)$ matches the edge term in (63): the center perturbation is in coordinate i , while the aggregate is over the opposite component $\bar{\mathcal{V}}_e(i)$. By (61),

$$\sum_{v \in \bar{\mathcal{V}}_e(i)} (u_v^1 - u_v^0) = \tilde{\mathcal{L}}_{r,\bar{\mathcal{V}}_e(i)}(\eta(E); \hat{\vartheta}) - \tilde{\mathcal{L}}_{r,\bar{\mathcal{V}}_e(i)}(\eta(E); \vartheta).$$

Therefore

$$\|\tilde{\mathcal{Y}}_{r,e}(E, \hat{\vartheta}) - \tilde{\mathcal{Y}}_{r,e}(E, \vartheta)\| \leq \tilde{\kappa}_{c,r} \frac{|\bar{\mathcal{V}}_e(i)|}{N_r} \|\hat{\vartheta}_i - \vartheta_i\|.$$

This proves (65a)–(65c).

Proof of (64a)–(64b). Fix a center tuple $\vartheta = (\vartheta_i)_{i \in \mathcal{V}_r}$, and suppose E, E' differ only in boundary coordinate (i, h) . With ϑ fixed, (61) is the surrogate counterpart of the aggregate map used in Lemma 3. Repeating the proof of Lemma 3 with $\mathcal{L}_{r,S}$ replaced by $\tilde{\mathcal{L}}_{r,S}(\cdot; \vartheta)$ gives

$$\|\tilde{\mathcal{Y}}_{r,e}(E', \vartheta) - \tilde{\mathcal{Y}}_{r,e}(E, \vartheta)\| \leq \tilde{\kappa}_{\partial,r} \frac{|\bar{\mathcal{V}}_e(i)|}{N_r} \|E'_{i,h} - E_{i,h}\|$$

and

$$\|\tilde{\mathcal{Z}}_{r,v}(E', \vartheta) - \tilde{\mathcal{Z}}_{r,v}(E, \vartheta)\| \leq \tilde{\kappa}_{\partial,r} \frac{1}{N_r} \|E'_{i,h} - E_{i,h}\|, \quad v \neq i.$$

This proves (64a)–(64b). $\square$

Remark on $\tilde{\kappa}_{c,r}$. The constant $\tilde{\kappa}_{c,r}$ in (63) measures the center sensitivity of the frozen aggregate maps and can reflect the associated surrogate approximation quality. It vanishes for the exact surrogate $\tilde{f}_i^*(U_i; Z_i) = f_i^*(U_i)$. Finiteness requires control of the center dependence in addition to the curvature bounds in Assumption 5(iii). When $\tilde{f}_j^* \in C^2$ in (u, θ) for $j \in \mathcal{V}_r$, fix $i \in \mathcal{V}_r$, $\eta \in \mathbb{R}^d$, and a center tuple $\vartheta \in \mathbb{R}^{|\mathcal{V}_r| \times d}$. To identify the local derivative entering Lip_{ϑ_i} in (63), vary only the i -th center, denoted by θ . By (61), along this one-coordinate variation, for any nonempty $S \subseteq \mathcal{V}_r$,

$$\tilde{\mathcal{L}}_{r,S}(\eta; \vartheta) \Big|_{\vartheta_i = \theta} = \sum_{j \in S} u_j(\theta),$$

where $u_j(\theta)$, $j \in \mathcal{V}_r$, are determined by

$$\nabla_u \tilde{f}_j^*(u_j(\theta); \vartheta_j) = x(\theta) \quad (j \neq i), \quad \nabla_u \tilde{f}_i^*(u_i(\theta); \theta) = x(\theta), \quad \sum_{j \in \mathcal{V}_r} u_j(\theta) = \eta.$$

Applying the implicit function theorem at $\theta = \vartheta_i$, write

$$H_j := \left(\nabla_{uu}^2 \tilde{f}_j^*(u_j(\vartheta_i); \vartheta_j) \right)^{-1} \quad (j \neq i),$$

$$H_i := \left(\nabla_{uu}^2 \tilde{f}_i^*(u_i(\vartheta_i); \vartheta_i) \right)^{-1}, \quad C_i := \left(\nabla_{\theta} [\nabla_u \tilde{f}_i^*(u_i(\vartheta_i); \theta)] \right)^{\top} \Big|_{\theta=\vartheta_i}, \quad H_{\Sigma} := \sum_{j \in \mathcal{V}_r} H_j.$$

Differentiating the defining equations gives

$$\nabla_{\theta} u_j(\vartheta_i)^{\top} = H_j \nabla_{\theta} x(\vartheta_i)^{\top} \quad (j \neq i), \quad \nabla_{\theta} u_i(\vartheta_i)^{\top} = H_i \nabla_{\theta} x(\vartheta_i)^{\top} - H_i C_i.$$

Differentiating $\sum_{j \in \mathcal{V}_r} u_j(\theta) = \eta$ then yields

$$0 = H_{\Sigma} \nabla_{\theta} x(\vartheta_i)^{\top} - H_i C_i, \quad \nabla_{\theta} x(\vartheta_i)^{\top} = H_{\Sigma}^{-1} H_i C_i.$$

Therefore, for every component $\bar{\mathcal{V}}_e(i)$ not containing i ,

$$\nabla_{\theta} \left(\sum_{v \in \bar{\mathcal{V}}_e(i)} u_v \right) (\vartheta_i)^{\top} = \left(\sum_{v \in \bar{\mathcal{V}}_e(i)} H_v \right) H_{\Sigma}^{-1} H_i C_i,$$

$$\nabla_{\theta} u_i(\vartheta_i)^{\top} = H_i (H_{\Sigma}^{-1} H_i - I) C_i, \quad \nabla_{\theta} u_v(\vartheta_i)^{\top} = H_v H_{\Sigma}^{-1} H_i C_i \quad (v \neq i).$$

Thus $\tilde{\kappa}_{c,r}$ is certified by any upper bound on

$$\sup_{\eta, \vartheta} \max_{i \in \mathcal{V}_r} \left\{ \max_{e \in \mathcal{E}_r} \frac{|\mathcal{V}_r|}{|\bar{\mathcal{V}}_e(i)|} \left\| \left(\sum_{v \in \bar{\mathcal{V}}_e(i)} H_v \right) H_{\Sigma}^{-1} H_i C_i \right\|, \right. \\ \left. \left\| H_i (H_{\Sigma}^{-1} H_i - I) C_i \right\|, \quad \max_{v \neq i} |\mathcal{V}_r| \left\| H_v H_{\Sigma}^{-1} H_i C_i \right\| \right\}.$$

Thus $\tilde{\kappa}_{c,r}$ exhibits a preconditioning effect analogous to that in Remark 3. The additional factor C_i captures the center sensitivity of the local surrogate gradient. Assumption 5(iv) ensures that the matrices C_i are uniformly bounded. Together with the uniform curvature bounds in Assumption 5(iii), this ensures that $\tilde{\kappa}_{c,r} < \infty$.

B.2 Proof of Lemma 8

Fix a block r and an iteration $\nu \geq \max_{s \in [p]} \text{diam}(\mathcal{T}_s)$. For $i, v \in \mathcal{V}_r$, set $d_i(v) := 0$ if $i = v$, and $d_i(v) := \min_{e \in \mathcal{A}_v^{\circ}} \delta_{e,i}$ otherwise. Use the node-aligned boundary vector E_v^{ν} from the proof of Lemma 4, and set $Z_{r,v}^{\nu} := (Z_i^{\nu - d_i(v)})_{i \in \mathcal{V}_r}$ for the center coordinates. In particular, $[Z_{r,v}^{\nu}]_v = Z_v^{\nu}$.

Mismatch decomposition. For $v \in \mathcal{V}_r$, define the node mismatch

$$m_v := (\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1})_v.$$

The pair $(\hat{E}_e^{\nu}, Z_{r,e}^{\nu})$ depends on the incident edge e , whereas m_v fixes the node v . To estimate m_v , insert the node-aligned pair $(E_v^{\nu}, Z_{r,v}^{\nu})$. Since

$$(\hat{Z}_r^{\nu+1})_v = \sum_{e \in \mathcal{A}_v^{\circ}} \mathbf{B}_{v,e} \tilde{\mathbf{Y}}_{r,e}(\hat{E}_e^{\nu}, Z_{r,e}^{\nu})$$

and $(\bar{Z}_r^{\nu+1})_v = \tilde{\mathbf{Z}}_{r,v}(E_v^{\nu}, Z_{r,v}^{\nu})$, adding and subtracting

$$\sum_{e \in \mathcal{A}_v^{\circ}} \mathbf{B}_{v,e} \tilde{\mathbf{Y}}_{r,e}(E_v^{\nu}, Z_{r,v}^{\nu}) = \tilde{\mathbf{Z}}_{r,v}(E_v^{\nu}, Z_{r,v}^{\nu})$$

gives the exact decomposition

$$m_v = p_v + q_v,$$

where

$$p_v := \sum_{e \in \mathcal{A}_v^\circ} \mathbf{B}_{v,e} \left[\tilde{\mathbf{Y}}_{r,e}(\hat{E}_e^\nu, Z_{r,e}^\nu) - \tilde{\mathbf{Y}}_{r,e}(E_v^\nu, Z_{r,v}^\nu) \right], \quad (89)$$

$$q_v := \tilde{\mathbf{Z}}_{r,v}(E_v^\nu, Z_{r,v}^\nu) - \tilde{\mathbf{Z}}_{r,v}(E^\nu, Z_{\mathcal{V}_r}^\nu). \quad (90)$$

Estimate of p_v . Insert $\tilde{\mathbf{Y}}_{r,e}(E_v^\nu, Z_{r,e}^\nu)$ in each summand of (89). This gives

$$p_v = p_v^\partial + p_v^c,$$

where

$$p_v^\partial := \sum_{e \in \mathcal{A}_v^\circ} \mathbf{B}_{v,e} \left[\tilde{\mathbf{Y}}_{r,e}(\hat{E}_e^\nu, Z_{r,e}^\nu) - \tilde{\mathbf{Y}}_{r,e}(E_v^\nu, Z_{r,e}^\nu) \right],$$

$$p_v^c := \sum_{e \in \mathcal{A}_v^\circ} \mathbf{B}_{v,e} \left[\tilde{\mathbf{Y}}_{r,e}(E_v^\nu, Z_{r,e}^\nu) - \tilde{\mathbf{Y}}_{r,e}(E_v^\nu, Z_{r,v}^\nu) \right].$$

The term p_v^∂ changes only the boundary vector. For $e \in \mathcal{A}_v^\circ$ and $(i, h) \in \hat{\partial}_r$, the delay definition and (77) give $\delta_{e,i} \in \{\delta_{r,i}(v), \delta_{r,i}(v) + 1\}$. Hence the same argument in (80) gives

$$\sum_{v \in \mathcal{V}_r} \|p_v^\partial\|^2 \leq \hat{J}_r \tilde{\kappa}_{\partial,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2. \quad (91)$$

For the center part, write $\Delta Z_i^\ell := Z_i^{\ell+1} - Z_i^\ell$. For each $i \in \mathcal{V}_r$, root the tree at i , and let $\text{Ch}_{r,i}(v) := \{e \in \mathcal{A}_v^\circ : \delta_{e,i} = d_i(v) + 1\}$, whose elements are among the child edges of v in this rooted tree. For $v \neq i$, let $p_i(v)$ be the active edge from v to its parent, and set

$$n_{r,i}(i) := N_r, \quad n_{r,i}(v) := |\bar{\mathbf{V}}_{p_i(v)}(i)|, \quad v \neq i.$$

Since $\delta_{e,i} = 0$ for every $e \in \mathcal{A}_i^\circ$, $\text{Ch}_{r,i}(i) = \emptyset$. For each $e \in \mathcal{A}_v^\circ$, telescope differences from $Z_{r,v}^\nu$ to $Z_{r,e}^\nu$ by changing center coordinates one at a time. In the coordinate- i step, let $\vartheta_{e,i}^-$ and $\vartheta_{e,i}^+$ denote the consecutive center tuples, and define

$$p_{v,i}^c := \sum_{e \in \mathcal{A}_v^\circ} \mathbf{B}_{v,e} \left[\tilde{\mathbf{Y}}_{r,e}(E_v^\nu, \vartheta_{e,i}^+) - \tilde{\mathbf{Y}}_{r,e}(E_v^\nu, \vartheta_{e,i}^-) \right].$$

Then $p_v^c = \sum_{i \in \mathcal{V}_r} p_{v,i}^c$. Since $\delta_{e,i} \in \{d_i(v), d_i(v) + 1\}$ for every $e \in \mathcal{A}_v^\circ$, the i -coordinate difference is zero unless $e \in \text{Ch}_{r,i}(v)$. If $e \in \text{Ch}_{r,i}(v)$, then

$$(\vartheta_{e,i}^+ - \vartheta_{e,i}^-)_i = [Z_{r,e}^\nu]_i - [Z_{r,v}^\nu]_i = Z_i^{\nu-\delta_{e,i}} - Z_i^{\nu-d_i(v)} = -\Delta Z_i^{\nu-d_i(v)-1}.$$

If $\text{Ch}_{r,i}(v) = \emptyset$, then $p_{v,i}^c = 0$. Otherwise $d_i(v) + 1 \leq \tilde{D}_r$, and the sensitivity bound (65a) together with the counting argument for $v \neq i$ in (78) gives

$$\begin{aligned} \|p_{v,i}^c\|^2 &\leq \tilde{\kappa}_{c,r}^2 \frac{\left(\sum_{e \in \text{Ch}_{r,i}(v)} |\bar{\mathbf{V}}_e(i)| \right)^2}{N_r^2} \|\Delta Z_i^{\nu-d_i(v)-1}\|^2 \\ &\leq \tilde{\kappa}_{c,r}^2 \frac{(n_{r,i}(v) - 1)^2}{N_r^2} \|\Delta Z_i^{\nu-d_i(v)-1}\|^2. \end{aligned}$$

By Cauchy-Schwarz,

$$\|p_v^c\|^2 \leq N_r \tilde{\kappa}_{c,r}^2 \sum_{i \in \mathcal{V}_r, \text{Ch}_{r,i}(v) \neq \emptyset} \frac{(n_{r,i}(v) - 1)^2}{N_r^2} \|\Delta Z_i^{\nu-d_i(v)-1}\|^2.$$

Summing over v and rearranging terms gives

$$\begin{aligned}
\sum_{v \in \mathcal{V}_r} \|p_v^c\|^2 &\leq N_r \tilde{\kappa}_{c,r}^2 \sum_{i \in \mathcal{V}_r} \sum_{v \in \mathcal{V}_r: \text{Ch}_{r,i}(v) \neq \emptyset} \frac{(n_{r,i}(v) - 1)^2}{N_r^2} \|\Delta Z_i^{\nu-d_i(v)-1}\|^2 \\
&= N_r \tilde{\kappa}_{c,r}^2 \sum_{i \in \mathcal{V}_r} \sum_{t=1}^{\tilde{D}_r} \left(\sum_{v \in \mathcal{V}_r: \text{Ch}_{r,i}(v) \neq \emptyset, d_i(v)+1=t} \frac{(n_{r,i}(v) - 1)^2}{N_r^2} \right) \|\Delta Z_i^{\nu-t}\|^2 \\
&\leq N_r \tilde{\kappa}_{c,r}^2 \sum_{i \in \mathcal{V}_r} \sum_{t=1}^{\tilde{D}_r} \|\Delta Z_i^{\nu-t}\|^2 = N_r \tilde{\kappa}_{c,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2, \tag{92}
\end{aligned}$$

Here $d_i(v) = \max\{\text{dist}_{\mathcal{T}_r}(i, v) - 1, 0\}$, so each inner sum is over vertices at depth t and is at most one by (78). Combining (91) and (92) yields

$$\sum_{v \in \mathcal{V}_r} \|p_v\|^2 \leq 2\hat{J}_r \tilde{\kappa}_{\partial,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2 + 2N_r \tilde{\kappa}_{c,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2. \tag{93}$$

Estimate of q_v . Split $q_v = q_v^\partial + q_v^c$, where

$$q_v^\partial := \tilde{Z}_{r,v}(E_v^\nu, Z_{r,v}^\nu) - \tilde{Z}_{r,v}(E^\nu, Z_{r,v}^\nu), \quad q_v^c := \tilde{Z}_{r,v}(E^\nu, Z_{r,v}^\nu) - \tilde{Z}_{r,v}(E^\nu, Z_{\mathcal{V}_r}^\nu).$$

With $Z_{r,v}^\nu$ fixed, the proof of Lemma 4 gives the same q_v -estimate with $\tilde{\kappa}_{\partial,r}$ in place of $\kappa_{\partial,r}$:

$$\sum_{v \in \mathcal{V}_r} \|q_v^\partial\|^2 \leq \hat{J}_r \tilde{\kappa}_{\partial,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2. \tag{94}$$

For the center part, keep the boundary vector fixed at E^ν and change $Z_{\mathcal{V}_r}^\nu$ to $Z_{r,v}^\nu$ one center coordinate at a time. Let $q_{v,i}^c$ be the change in $\tilde{Z}_{r,v}$ during the i -coordinate replacement. Since $d_v(v) = 0$, we have $q_{v,v}^c = 0$. Thus $q_v^c = \sum_{i \in \mathcal{V}_r \setminus \{v\}} q_{v,i}^c$, and (65c) gives $\|q_{v,i}^c\| \leq \frac{\tilde{\kappa}_{c,r}}{N_r} \|Z_i^{\nu-d_i(v)} - Z_i^\nu\|$ for $i \neq v$. Hence, using Cauchy-Schwarz and $(N_r - 1)/N_r^2 \leq 1/N_r$,

$$\|q_v^c\|^2 \leq \frac{\tilde{\kappa}_{c,r}^2}{N_r} \sum_{i \in \mathcal{V}_r} \|Z_i^{\nu-d_i(v)} - Z_i^\nu\|^2. \tag{95}$$

For every i , telescoping gives

$$\|Z_i^{\nu-d_i(v)} - Z_i^\nu\|^2 \leq d_i(v) \sum_{m=1}^{d_i(v)} \|Z_i^{\nu-m+1} - Z_i^{\nu-m}\|^2, \quad 0 \leq d_i(v) \leq \tilde{D}_r.$$

Therefore

$$\frac{1}{N_r} \sum_{v \in \mathcal{V}_r} \sum_{i \in \mathcal{V}_r} \|Z_i^{\nu-d_i(v)} - Z_i^\nu\|^2 \leq \tilde{D}_r \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2, \tag{96}$$

where we used $\sum_{v \in \mathcal{V}_r} d_i(v) \leq N_r \tilde{D}_r$. Combining (95) and (96) yields

$$\sum_{v \in \mathcal{V}_r} \|q_v^c\|^2 \leq \tilde{D}_r \tilde{\kappa}_{c,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2.$$

Together with (94),

$$\sum_{v \in \mathcal{V}_r} \|q_v\|^2 \leq 2\hat{J}_r \tilde{\kappa}_{\partial,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2 + 2\tilde{D}_r \tilde{\kappa}_{c,r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2. \tag{97}$$

Since

$$\|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 = \sum_{v \in \mathcal{V}_r} \|m_v\|^2 \leq 2 \sum_{v \in \mathcal{V}_r} \|p_v\|^2 + 2 \sum_{v \in \mathcal{V}_r} \|q_v\|^2,$$

combining (93) and (97) gives

$$\begin{aligned} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 &\leq 8\hat{J}_r \tilde{\kappa}_{\partial, r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2 \\ &\quad + 4(N_r + \tilde{D}_r) \tilde{\kappa}_{c, r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2. \end{aligned} \quad (98)$$

Reduction to block increments. It remains to replace the coordinate increments by block and boundary increments. Since $Z^\ell = \mathbf{B}Y^\ell$, $Z_r^\ell = \mathbf{B}_{\mathcal{V}, \mathcal{E}_r} Y_{\mathcal{E}_r}^\ell$, and $\mathcal{A}_i = \mathcal{A}_i^\circ \sqcup \mathcal{A}_i^\partial$, for $i \in \mathcal{V}_r$,

$$Z_i^{\ell+1} - Z_i^\ell = \sum_{h \in \mathcal{A}_i} \mathbf{B}_{i,h} (Y_h^{\ell+1} - Y_h^\ell) = (Z_r^{\ell+1} - Z_r^\ell)_i + \sum_{h \in \mathcal{A}_i^\partial} \mathbf{B}_{i,h} (Y_h^{\ell+1} - Y_h^\ell).$$

Cauchy's inequality, $J_{r,i} := |\mathcal{A}_i^\partial| \leq J_r$, and $\hat{\partial}_r = \bigsqcup_{i \in \mathcal{V}_r} (\{i\} \times \mathcal{A}_i^\partial)$ give

$$\sum_{i \in \mathcal{V}_r} \|Z_i^{\ell+1} - Z_i^\ell\|^2 \leq 2\|Z_r^{\ell+1} - Z_r^\ell\|^2 + 2J_r \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2.$$

Summing over $\ell = \nu - \tilde{D}_r, \dots, \nu - 1$ and substituting into (98) gives

$$\begin{aligned} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2 &\leq \left(8\tilde{\kappa}_{\partial, r}^2 \hat{J}_r + 8J_r(N_r + \tilde{D}_r) \tilde{\kappa}_{c, r}^2 \right) \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2 \\ &\quad + 8(N_r + \tilde{D}_r) \tilde{\kappa}_{c, r}^2 \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \|Z_r^{\ell+1} - Z_r^\ell\|^2. \end{aligned} \quad (99)$$

By (39),

$$\sum_{\substack{(i,h) \in \hat{\partial}_r \\ h \in \mathcal{E}_s}} \|Y_h^{\ell+1} - Y_h^\ell\|^2 \leq \Gamma_r \|Z_s^{\ell+1} - Z_s^\ell\|^2, \quad s \in \mathcal{M}_r.$$

Since $\hat{\partial}_r = \bigsqcup_{s \in \mathcal{M}_r} \{(i,h) \in \hat{\partial}_r : h \in \mathcal{E}_s\}$, summing this estimate over s and $\ell = \nu - \tilde{D}_r, \dots, \nu - 1$ yields

$$\sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{(i,h) \in \hat{\partial}_r} \|Y_h^{\ell+1} - Y_h^\ell\|^2 \leq \Gamma_r \sum_{\ell=\nu-\tilde{D}_r}^{\nu-1} \sum_{s \in \mathcal{M}_r} \|Z_s^{\ell+1} - Z_s^\ell\|^2.$$

Substitution into (99), together with the definitions of $\widetilde{M}_{r,\partial}$ and $\widetilde{M}_{r,c}$, proves (67). $\square$

B.3 Proof of Theorem 2

As in the proof of Theorem 1, we first establish a descent estimate and apply the surrogate block error bound. Combining the resulting recursion with the mismatch bound in Lemma 8 then yields linear convergence.

Lemma 12 (Descent) *Under Assumption 5, if $0 \leq \tau_r^\nu \leq 1/\omega_{\mathcal{P}}$ for every $r \in [p]$, then*

$$G(Z^{\nu+1}) \leq G(Z^\nu) - \sum_{r=1}^p \tau_r^\nu \tilde{\Delta}_r(Z^\nu) + \sum_{r=1}^p \tau_r^\nu \frac{\tilde{L}_r^*}{2} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2. \quad (100)$$

Proof Since $\tau_r^\nu \leq 1/\omega_{\mathcal{P}}$, applying Jensen's inequality to each f_i^* and summing over the nodes yields

$$G(Z^{\nu+1}) - G(Z^\nu) \leq \sum_{r=1}^p \tau_r^\nu \left[G(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}) - G(Z^\nu) \right].$$

For fixed r , all outside coordinates are unchanged. Majorization and touching give

$$G(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}) - G(Z^\nu) \leq \tilde{G}(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}; Z^\nu) - \tilde{G}(Z^\nu; Z^\nu).$$

Since $Z^\nu - Z_r^\nu + \bar{Z}_r^{\nu+1}$ minimizes $\tilde{G}(\cdot; Z^\nu)$ over $Z^\nu + \mathcal{U}_r$ and $\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1} \in \mathcal{U}_r$, first-order optimality gives

$$\left\langle \nabla_U \tilde{G}(Z^\nu - Z_r^\nu + \bar{Z}_r^{\nu+1}; Z^\nu)^\top, \hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1} \right\rangle = 0.$$

The $\tilde{L}_r^*$ -smoothness on this slice then yields

$$\tilde{G}(Z^\nu - Z_r^\nu + \hat{Z}_r^{\nu+1}; Z^\nu) \leq \tilde{G}(Z^\nu - Z_r^\nu + \bar{Z}_r^{\nu+1}; Z^\nu) + \frac{\tilde{L}_r^*}{2} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2.$$

Combining these bounds with the definition of $\tilde{\Delta}_r(Z^\nu)$ proves (100). $\square$

Lemma 13 (Surrogate block error bound) *Under Assumptions 5 and 6, assume that $0 < \tau_r^\nu \leq 1/\omega_{\mathcal{P}}$ for every $r \in [p]$, and let $\tau^\nu := \min_{r \in [p]} \tau_r^\nu$. Then*

$$\begin{aligned} G(Z^{\nu+1}) - G^* &\leq \left(1 - \frac{\tau^\nu}{2\tilde{\kappa}_{\mathcal{P}}}\right) (G(Z^\nu) - G^*) - \sum_{r=1}^p \frac{\tilde{\mu}_r^*}{8\tau_r^\nu} \|Z_r^{\nu+1} - Z_r^\nu\|^2 \\ &\quad + \sum_{r=1}^p \tau_r^\nu \frac{2\tilde{L}_r^* + \tilde{\mu}_r^*}{4} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2. \end{aligned} \quad (101)$$

Proof Split $\sum_{r=1}^p \tau_r^\nu \tilde{\Delta}_r(Z^\nu)$ in (100) into two halves. The surrogate block error bound gives

$$\frac{1}{2} \sum_{r=1}^p \tau_r^\nu \tilde{\Delta}_r(Z^\nu) \geq \frac{\tau^\nu}{2\tilde{\kappa}_{\mathcal{P}}} (G(Z^\nu) - G^*).$$

For the second half, touching and strong convexity of the nondelayed surrogate slice give

$$\tilde{\Delta}_r(Z^\nu) \geq \frac{\tilde{\mu}_r^*}{2} \|\bar{Z}_r^{\nu+1} - Z_r^\nu\|^2.$$

Using

$$\|\bar{Z}_r^{\nu+1} - Z_r^\nu\|^2 \geq \frac{1}{2} \|\hat{Z}_r^{\nu+1} - Z_r^\nu\|^2 - \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2$$

and $Z_r^{\nu+1} - Z_r^\nu = \tau_r^\nu (\hat{Z}_r^{\nu+1} - Z_r^\nu)$, the second half of the progress satisfies

$$\frac{1}{2} \tau_r^\nu \tilde{\Delta}_r(Z^\nu) \geq \frac{\tilde{\mu}_r^*}{8\tau_r^\nu} \|Z_r^{\nu+1} - Z_r^\nu\|^2 - \tau_r^\nu \frac{\tilde{\mu}_r^*}{4} \|\hat{Z}_r^{\nu+1} - \bar{Z}_r^{\nu+1}\|^2.$$

Substituting these bounds into Lemma 12 gives (101). $\square$

Proof of Theorem 2. Set $\nu_0 := \tilde{D} + 1 = \max_{r \in [p]} \text{diam}(\mathcal{T}_r)$. Substituting (67) into (101) and using (68) gives for all $\nu \geq \nu_0$

$$\begin{aligned} G(Z^{\nu+1}) - G^* &\leq \left(1 - \frac{\tau}{2\tilde{\kappa}_{\mathcal{P}}}\right) (G(Z^\nu) - G^*) - \frac{\tilde{\mu}^*}{8\tau} \sum_{r=1}^p \|Z_r^{\nu+1} - Z_r^\nu\|^2 \\ &\quad + \tau \tilde{A} \sum_{\ell=\nu-\tilde{D}}^{\nu-1} \sum_{r=1}^p \|Z_r^{\ell+1} - Z_r^\ell\|^2. \end{aligned}$$

Define the Lyapunov function

$$\tilde{\mathcal{L}}^\nu := G(Z^\nu) - G^* + \tau \tilde{A} \sum_{j=1}^{\tilde{D}} \left(\sum_{k=1}^{\tilde{D}-j+1} \tilde{\rho}^{-k} \right) \sum_{r=1}^p \|Z_r^{\nu-j+1} - Z_r^{\nu-j}\|^2.$$

Condition (69) gives $\frac{8\tilde{A}\tau^2}{\underline{\mu}^*} \sum_{j=1}^{\tilde{D}} \tilde{\rho}^{-j} \leq 1$. Then for $\nu \geq \nu_0$,

$$\tilde{\mathcal{L}}^{\nu+1} \leq \tilde{\rho} \tilde{\mathcal{L}}^\nu - \left(\frac{\tilde{\mu}^*}{8\tau} - \tau \tilde{A} \sum_{j=1}^{\tilde{D}} \tilde{\rho}^{-j} \right) \sum_{r=1}^p \|Z_r^{\nu+1} - Z_r^\nu\|^2 \leq \tilde{\rho} \tilde{\mathcal{L}}^\nu.$$

Thus $G(Z^\nu) - G^* \leq \tilde{\mathcal{L}}^{\nu_0} \tilde{\rho}^{\nu-\nu_0}$, and

$$\|X^\nu - X^*\|^2 \leq \frac{1}{\mu^2} \|Z^\nu - Z^*\|^2 \leq \frac{2L\tilde{\mathcal{L}}^{\nu_0}}{\mu^2} \tilde{\rho}^{\nu-\nu_0}, \quad \nu \geq \nu_0.$$

Absorbing constants gives (70). $\square$

References

1. A. Berahas, R. Bollapragada, N. Keshkar, and E. Wei. Balancing communication and computation in distributed optimization. *IEEE Trans. Autom. Control*, 64(8):3141–3155, 2018.
2. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.
3. F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997.
4. P. Di Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Trans. Signal Inf. Process. Netw.*, 2(2):120–136, 2016.
5. K. Ding, M. Maros, and G. Scutari. A new decomposition paradigm for graph-structured nonlinear programs via message passing. *arXiv:2512.24676*, 2025.
6. J. M. Hendrickx, R. M. Jungers, A. Olshevsky, and G. Vankeerberghen. Graph diameter, eigenvalues, and minimum-time consensus. *Automatica*, 50(2):635–640, 2014.
7. H. Hendrikx, F. Bach, and L. Massoulié. An optimal algorithm for decentralized finite-sum optimization. *SIAM J. Optim.*, 31(4):2753–2783, 2021.
8. D. Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Trans. Signal Inf. Process. Netw.*, 5(1):31–46, 2018.
9. D. Jakovetić, J. Moura, and J. Xavier. Linear convergence rate of a class of distributed augmented Lagrangian algorithms. *IEEE Trans. Autom. Control*, 60(4):922–936, 2015.
10. X. Jiang, E. D. H. Nguyen, C. A. Uribe, and B. Ying. Sparse factorization of the square all-ones matrix of arbitrary order. *SIAM J. Matrix Anal. Appl.*, 46(1):466–486, 2025.
11. H. Li and Z. Lin. Revisiting extra for smooth distributed optimization. *SIAM J. Optim.*, 30(3):1795–1821, 2020.
12. C. Lopes and A. Sayed. Diffusion Least-Mean Squares Over Adaptive Networks: Formulation and Performance Analysis. *IEEE Trans. Signal Process.*, 56(7):3122–3136, 2008.
13. Y. Lu and C. De Sa. Optimal complexity in decentralized training. In M. Meila and T. Zhang, editors, *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 139 of *Proc. of Machine Learning Research*, pages 7111–7123. PMLR, 2021.
14. A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM J. Optim.*, 27(4):2597–2633, 2017.
15. A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Autom. Control*, 54(1):48–61, 2009.
16. E. D. H. Nguyen, X. Jiang, B. Ying, and C. A. Uribe. On graphs with finite-time consensus and their use in gradient tracking. *SIAM J. Optim.*, 35(2):872–898, 2025.
17. A. Olshevsky and J. Tsitsiklis. Convergence speed in distributed consensus and averaging. *SIAM J. Control Optim.*, 48(1):33–55, 2009.

18. S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Math. Program.*, 187(1):409–457, 2021.
19. G. Qu and N. Li. Harnessing Smoothness to Accelerate Distributed Optimization. *IEEE Trans. Control Netw. Syst.*, 5(3):1245–1260, 2018.
20. A. Sayed et al. Adaptation, learning, and optimization over networks. *Found. Trends Mach. Learn.*, 7(4-5):311–801, 2014.
21. K. Scaman, F. Bach, S. Bubeck, Y. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In D. Precup and Y. Teh, editors, *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 70 of *Proc. of Machine Learning Research*, pages 3027–3036. PMLR, 2017.
22. G. Scutari and Y. Sun. Distributed nonconvex constrained optimization over time-varying digraphs. *Math. Program.*, 176(1):497–544, 2019.
23. W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM J. Optim.*, 25(2):944–966, 2015.
24. W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Trans. Signal Process.*, 62(7):1750–1761, 2014.
25. Y. Sun, G. Scutari, and A. Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM J. Optim.*, 32(2):354–385, 2022.
26. S. Sundaram and C. N. Hadjicostis. Distributed function calculation and consensus using linear iterative strategies. *IEEE J. Sel. Areas Commun.*, 26(4):650–660, 2008.
27. T. Vogels, L. He, A. Koloskova, S. P. Karimireddy, T. Lin, S. U. Stich, and M. Jaggi. RelaySum for decentralized deep learning on heterogeneous data. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pages 28004–28015, 2021.
28. J. Xu, S. Zhu, Y. C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *Proc. IEEE Conf. Decis. Control (CDC)*, pages 2055–2060, 2015.
29. K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM J. Optim.*, 26(3):1835–1854, 2016.