AN INEXACT SEMI-PROXIMAL ALM WITH ADAPTIVE COMMUNICATION-FREE LOCAL SOLVES FOR DECENTRALIZED COMPOSITE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized optimization enables agents to collaboratively solve learning and decision-making problems using only local computation and neighbor communication. In many practical networks, communication is substantially more expensive than local computation, motivating algorithms that perform multiple local updates between successive exchanges. Augmented Lagrangian methods provide an effective framework for composite optimization, but their decentralized subproblems are typically coupled through the network and may require repeated communication during the inner iterations. To alleviate this issue, we propose CFI-spALM, a Communication-Free Inexact Semi-Proximal Augmented Lagrangian Method that uses a structured semi-proximal term to decompose the coupled subproblem into independent nodewise problems. Each agent adaptively performs local proximal-gradient updates under an absolute or relative stopping rule, requiring only one model-vector exchange per outer iteration. For convex composite objectives, we establish global convergence, sublinear communication complexity, and linear convergence under metric subregularity. Experiments on logistic regression and rank-deficient LASSO demonstrate that CFI-spALM can achieve the prescribed optimization target with fewer model-vector exchanges than the comparison methods.

1 INTRODUCTION

Decentralized optimization has attracted increasing attention with the growing demand for collaborative learning and decision-making over networked systems (Tsitsiklis et al., 1986; Nedić & Ozdaglar, 2009; Duchi et al., 2012; Sayed, 2014). Without relying on a central coordinator, it enables agents to optimize a common objective through neighbor communication while keeping their raw data and objective functions local, thereby reducing the direct exposure of private information. This framework has found broad applications in distributed estimation, networked learning, resource allocation, and statistical inference (Lopes & Sayed, 2008; Predd et al., 2006; Xiao & Boyd, 2006; Jordan et al., 2019). In this paper, we consider the decentralized composite convex problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \big(f_i(x) + g_i(x)\big), \tag{1}$$

where f_i is smooth convex and g_i is proper closed convex with an efficiently computable proximal mapping. This general formulation covers a wide range of applications, such as regularized empirical-risk minimization (Xiao & Zhang, 2014), distributed sparse regression (Mateos et al., 2010), constrained optimization through indicator functions (Parikh & Boyd, 2014), and imaging and inverse problems (Beck & Teboulle, 2009; Chambolle & Pock, 2011).

Related decentralized algorithms can be broadly divided into single-loop and double-loop methods. Single-loop methods typically exchange model vectors after each gradient or proximal-gradient update (Nedić & Ozdaglar, 2009; Shi et al., 2015a;b; Li et al., 2019), and may require many communication rounds when the problem is poorly conditioned or high accuracy is needed. Double-loop multiplier methods instead approximately solve a primal subproblem before each dual update (Arjevani et al., 2020; Yang et al., 2026). However, their subproblems often contain graph-dependent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

quadratic terms evaluated at each inner iterate, requiring neighbor exchanges throughout the local solve. Since communication is often substantially more expensive than local computation under limited bandwidth, high latency, or weak network connectivity (Nedić et al., 2018; Scaman et al., 2017; Alistarh et al., 2017; McMahan et al., 2017; Chen et al., 2025; Hu et al., 2026), these inner exchanges can diminish the communication benefit of performing multiple updates per outer iteration. This raises the following question: *Can a decentralized multiplier method solve its subproblems through communication-free local updates while retaining locally checkable stopping rules and rigorous primal–dual convergence guarantees?*

We answer this question affirmatively by proposing CFI-spALM, an inexact semi-proximal augmented Lagrangian method for decentralized optimization with communication-free local solves. A structured semi-proximal term replaces the candidate-dependent graph coupling with cached network information and decomposes the multiplier subproblem into independent nodewise problems. Each agent applies proximal-gradient or accelerated proximal-gradient updates from its current outer iterate until a locally computable residual satisfies an absolute or relative criterion. The amount of local work therefore adapts across agents and outer iterations. All local updates and residual tests use only local objective information and cached graph data. After the local solves terminate, the agents exchange their new model vectors and update the multipliers. Thus, each outer iteration requires only one exchange of a d-dimensional model vector, regardless of the number of local updates.

The main contributions are as follows:

- We propose CFI-spALM with a structured semi-proximal term that makes the multiplier subproblem agent-separable, enabling adaptive residual-driven local updates with one model-vector exchange per outer iteration.
- We provide convergence guarantees for both stopping rules. Summable absolute residuals ensure convergence of the primal–dual sequence, while a nodewise relative rule guarantees KKT-set convergence without a prescribed error schedule. Both rules attain an $O(K^{-1/2})$ best-iterate KKT residual and $O(K^{-1})$ ergodic primal errors. Under metric subregularity, geometrically decreasing absolute errors and sufficiently small relative thresholds yield linear convergence.
- Experiments on smooth strongly convex logistic regression and nonsmooth rank-deficient LASSO show that both variants reliably reach the prescribed optimization targets with fewer model-vector exchanges than the graph-coupled multiplier methods in the reported settings.

2 Related Work

Related decentralized algorithms can be grouped by how they organize computation between communication rounds: single-loop methods exchange information after each first-order correction, whereas double-loop methods approximately solve an inner subproblem before completing the outer update.

Single-loop decentralized methods. Early decentralized gradient methods alternate neighbor averaging with local gradient steps (Tsitsiklis et al., 1986; Nedić & Ozdaglar, 2009). However, with a fixed stepsize, ordinary decentralized gradient descent generally converges only to a neighborhood of the solution (Yuan et al., 2016). To recover exact convergence, gradient-tracking methods recursively estimate the global gradient (Nedić et al., 2017). Another line of work introduces corrections based on successive mixing steps: EXTRA (Shi et al., 2015a) handles smooth objectives, while PG-EXTRA (Shi et al., 2015b) extends this mechanism to composite problems. Building on related correction structures, NIDS (Li et al., 2019) separates the objective and network effects to reduce the network dependence of primal stepsizes, whereas D-iPGM (Guo et al., 2023) adopts a combine-then-adapt proximal-gradient scheme for decentralized composite optimization. Despite their different update mechanisms, these methods generally perform one explicit first-order correction per model-vector exchange.

Double-loop decentralized methods. A different line of work uses double-loop schemes, in which an augmented-Lagrangian subproblem is solved through multiple inner iterations before the outer variables are updated. IDEAL (Arjevani et al., 2020) approximately solves graph-coupled subproblems using prescribed absolute tolerances or fixed inner budgets and accelerates the multiplier

sequence. To reduce the associated tuning burden, D-ripALM (Zhu et al., 2026) adopts a tuning-friendly relative error criterion inspired by (Solodov & Svaiter, 1999; Eckstein & Silva, 2013) for general convex objectives. However, the standard inner solvers of both methods require neighbor communication at every inner iteration to evaluate graph-dependent terms, which may still result in considerable communication. CFI-spALM preserves the subproblem-based double-loop structure while making its inner iterations communication-free.

3 Problem Setup and Preliminaries

3.1 Equivalent consensus formulations

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the connected undirected communication graph, where $\mathcal{V} = \{1, \dots, n\}$. Introducing a local copy $x_i \in \mathbb{R}^d$ at each agent yields the following equivalent problem

$$\min_{\{x_i\}_{i=1}^n} \quad \sum_{i=1}^n \big(f_i(x_i) + g_i(x_i)\big) \qquad \text{s.t.} \quad x_i = x_j, \qquad \forall\, \{i,j\} \in \mathcal{E}. \tag{2}$$

Because $\mathcal{G}$ is connected, the edge constraints in (2) are equivalent to the global consensus condition $x_1 = \cdots = x_n$. Stack the local variables as $X = (x_1^\mathsf{T}, \dots, x_n^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{nd}$. Using a communication matrix $W = [w_{ij}] \in \mathbb{R}^{n\times n}$, the edge-consensus formulation (2) can be expressed as

$$\min_{X\in\mathbb{R}^{nd}} \quad \Phi(X) := F(X) + G(X) \qquad \text{s.t.} \quad X = (W \otimes \mathrm{I}_d)X. \tag{3}$$

Here, $F(X) := \sum_{i=1}^n f_i(x_i)$, $G(X) := \sum_{i=1}^n g_i(x_i)$, and $\varphi_i := f_i + g_i$ is the local composite objective. The matrix W describes information exchange over the network. We adopt the following standard conditions on W, under which $X = (W \otimes \mathrm{I}_d)X$ is equivalent to $x_1 = \cdots = x_n$ (Nedić & Ozdaglar, 2009; Shi et al., 2015a; Li et al., 2019).

Assumption 1 (Communication matrix)**.** *The matrix W is symmetric, satisfies $W\mathbf{1} = \mathbf{1}$, $\mathrm{I}_n - W \succeq 0$, $\ker(\mathrm{I}_n - W) = \mathrm{span}\{\mathbf{1}\}$, and $w_{ij} = 0$ whenever $i \neq j$ and $\{i,j\} \notin \mathcal{E}$.*

For a linearly constrained formulation, define $Z := \mathrm{I}_n - W$ and $\mathcal{Z} := Z \otimes \mathrm{I}_d$. Choose a full-row-rank matrix $\mathcal{A} \in \mathbb{R}^{(n-1)d\times nd}$ such that $\mathcal{A}^\mathsf{T}\mathcal{A} = \mathcal{Z}$. The conditions $X = (W \otimes \mathrm{I}_d)X$, $\mathcal{Z}X = 0$, and $\mathcal{A}X = 0$ define the same consensus subspace. Therefore, problem (3) can be written as

$$\min_{X\in\mathbb{R}^{nd}} \quad \Phi(X) \qquad \text{s.t.} \quad \mathcal{A}X = 0. \tag{4}$$

3.2 Inexact augmented-Lagrangian methods: IDEAL and D-ripALM

Problem (4) is a linearly constrained convex problem. Its augmented Lagrangian is

$$\mathcal{L}_\rho(X, y) := \Phi(X) + \left\langle \mathcal{A}^\mathsf{T} y, X \right\rangle + \frac{\rho}{2} \left\| \mathcal{A}X \right\|^2,$$

where $\rho > 0$ is the penalty parameter and y is the multiplier associated with $\mathcal{A}X = 0$. A standard *inexact augmented-Lagrangian iteration* takes the form

$$X^{t+1} \approx \arg\min_X \ \mathcal{L}_{\rho_t}(X, y^t), \qquad y^{t+1} = y^t + \rho_t \mathcal{A}X^{t+1}.$$

Introducing the transformed multiplier $\Omega^t := \mathcal{A}^\mathsf{T} y^t$ gives the equivalent multiplier update

$$\Omega^{t+1} = \Omega^t + \rho_t \mathcal{Z}X^{t+1}.$$

IDEAL and D-ripALM are two representative double-loop implementations of this framework. Both approximately solve an augmented-Lagrangian subproblem before updating the multiplier, but they use different subproblems and inexactness rules. For smooth strongly convex objectives with $g_i \equiv 0$ for all i, the nonsmooth term $G(X)$ vanishes and $\Phi(X) = F(X)$. In this setting, **IDEAL** (Arjevani et al., 2020) uses the iteration

$$\begin{aligned} X^{t+1} &\approx \arg\min_X \left\{ F(X) + \left\langle \Omega^t, X \right\rangle + \frac{\rho}{2} \left\langle X, \mathcal{Z}X \right\rangle \right\}, \\ \Lambda^{t+1} &= \Omega^t + \eta \mathcal{Z}X^{t+1}, \\ \Omega^{t+1} &= \Lambda^{t+1} + \beta_{t+1}\big(\Lambda^{t+1} - \Lambda^t\big). \end{aligned}$$

The primal subproblem is solved approximately, either to a prescribed accuracy or with a prescribed number of inner iterations. The last two lines accelerate the multiplier sequence. By comparison, **D-RIPALM** (Zhu et al., 2026) handles general closed convex objectives through a proximal augmented-Lagrangian subproblem. At outer iteration t, it approximately minimizes

$$\Psi_t(X) := \Phi(X) + \left\langle \Omega^t, X \right\rangle + \frac{\rho_t}{2} \left\langle X, \mathcal{Z}X \right\rangle + \frac{\tau_t}{2\rho_t} \left\| X - X^t \right\|^2.$$

More precisely, it computes X^{t+1} and an error vector $\Delta^{t+1} \in \partial\Psi_t(X^{t+1})$ satisfying $2\left|\left\langle w^t - X^{t+1}, \rho_t\Delta^{t+1}\right\rangle\right| + \left\|\rho_t\Delta^{t+1}\right\|^2 \le \theta\left(\rho_t^2\left\|\mathcal{A}X^{t+1}\right\|^2 + \tau_t\left\|X^{t+1} - X^t\right\|^2\right)$ with $\theta \in [0,1)$. The multiplier and auxiliary variable are then updated by

$$\Omega^{t+1} = \Omega^t + \rho_t \mathcal{Z}X^{t+1}, \qquad w^{t+1} = w^t - \rho_t\Delta^{t+1}.$$

The subproblems of IDEAL and D-ripALM contain graph quadratics that require a new evaluation of $\mathcal{Z}U$, and hence another vector exchange, whenever the inner candidate U changes. The relative stopping test of D-ripALM also requires scalar aggregation across agents. Consequently, these methods may still require considerable communication. *Our goal is to avoid communication while solving subproblems within the ALM framework, which may help reduce total communication cost.*

4 The Proposed CFI-spALM Method

This section introduces CFI-spALM, which uses a semi-proximal term to obtain separable local subproblems and residual-based rules to adapt local computation without additional communication.

4.1 Separable Multiplier Subproblems

Motivated by the preconditioned proximal ALM framework in (Yang et al., 2026), we introduce a semi-proximal term designed to eliminate the candidate-dependent graph products discussed in the preceding section. For a penalty parameter $\rho > 0$, we first consider the following exact update

$$X^{t+1} \in \arg\min_X \left\{ \Phi(X) + \left\langle \mathcal{A}^{\mathsf{T}} y^t, X \right\rangle + \frac{\rho}{2}\left\|\mathcal{A}X\right\|^2 + \frac{1}{2}\left\|X - X^t\right\|_S^2 \right\}, \quad y^{t+1} = y^t + \rho\mathcal{A}X^{t+1}.$$

Here $S \succeq 0$, and $\|u\|_H^2 := \langle u, Hu\rangle$ for a symmetric matrix $H \succeq 0$. Choose a block-diagonal matrix $M = \mathrm{diag}(M_1, \ldots, M_n) \succ 0$ such that $S = M - \rho\mathcal{Z} \succeq 0$. Since $\mathcal{A}^{\mathsf{T}}\mathcal{A} = \mathcal{Z}$, the quadratic terms satisfy $\frac{\rho}{2}\|\mathcal{A}X\|^2 + \frac{1}{2}\|X - X^t\|_S^2 = \frac{1}{2}\|X - X^t\|_M^2 + \rho\langle \mathcal{Z}X^t, X\rangle + C_t$, where C_t is independent of X. Thus, the graph product acts only on the fixed point X^t, and the primal subproblem becomes

$$\min_X \left\{ \Phi(X) + \left\langle \mathcal{A}^{\mathsf{T}} y^t + \rho\mathcal{Z}X^t, X \right\rangle + \frac{1}{2}\left\|X - X^t\right\|_M^2 \right\}. \tag{5}$$

For a distributed implementation, define $\Omega^t := \mathcal{A}^{\mathsf{T}} y^t \in \mathrm{range}(\mathcal{Z})$ and $q^t := \mathcal{Z}X^t$. The transformed multiplier update is

$$\Omega^{t+1} = \Omega^t + \rho q^{t+1}, \qquad q^{t+1} = \mathcal{Z}X^{t+1}.$$

Under the consensus initialization $X^0 = \mathbf{1} \otimes x^0$, we have $q^0 = 0$. In each subsequent round, q^t comes from the preceding neighbor exchange and remains fixed during the local solve, so the implementation need not form $\mathcal{A}$. Moreover, the separability of Φ and block-diagonal structure of M reduce the primal update to independent local subproblems. Specifically, node i computes

$$\bar{x}_i^{t+1} = \arg\min_x \left\{ f_i(x) + g_i(x) + \left\langle \omega_i^t + \rho q_i^t, x \right\rangle + \frac{1}{2}\left\|x - x_i^t\right\|_{M_i}^2 \right\}. \tag{6}$$

where ω_i^t and q_i^t are the ith blocks of Ω^t and q^t. Both vectors remain unchanged during the local solve, and hence repeated local updates require no additional neighbor exchange. Since $M_i \succ 0$, the local objective is strongly convex and has a unique minimizer.

Remark 1 (Choice and role of M)**.** *The role of M can be seen by writing the primal subproblems of CFI-spALM, IDEAL, and D-ripALM in the common form*

$$\min_X \left\{ \Phi(X) + \langle \Omega^t + \rho_t \mathcal{Z} X^t, X \rangle + \frac{1}{2} \left\| X - X^t \right\|_{M_t}^2 \right\},$$

up to a term independent of X. IDEAL corresponds to $\rho_t = \rho$ and $M_t^{\mathrm{I}} = \rho\mathcal{Z}$, while D-ripALM uses $M_t^{\mathrm{D}} = (\tau_t/\rho_t)\mathrm{I}_{nd} + \rho_t\mathcal{Z}$. In both cases, M_t contains $\mathcal{Z}$, so evaluating the inner gradient at each new candidate requires a graph product and hence neighbor communication. CFI-spALM instead uses a block-diagonal majorizer. When W has nonnegative edge weights, $Z \preceq 2D_Z$ with $D_Z := \mathrm{diag}(Z_{11}, \ldots, Z_{nn})$, so one may choose $M = 2\rho(D_Z \otimes \mathrm{I}_d) + \gamma_M \mathrm{I}_{nd}$ with $\gamma_M \geq 0$. This choice is agent-separable and satisfies $M - \rho\mathcal{Z} \succeq 0$. Consequently, the only graph product in the local subproblem is $\mathcal{Z}X^t$, which remains fixed and can be cached before the local iterations.

4.2 Local Stopping Criteria

The local subproblems generally do not have closed-form solutions, and computing highly accurate solutions at every outer iteration may require many local updates. In this paper, we use the first-order optimality residual to determine when each local solve can terminate.

For a candidate x_i^{t+1}, choose $\zeta_i^{t+1} \in \partial g_i(x_i^{t+1})$ and define $v_i^{t+1} := \nabla f_i(x_i^{t+1}) + \zeta_i^{t+1} \in \partial\varphi_i(x_i^{t+1})$. The corresponding residual of the local subproblem is

$$\varepsilon_i^{t+1} := v_i^{t+1} + \omega_i^t + \rho q_i^t + M_i(x_i^{t+1} - x_i^t). \tag{7}$$

Every term in Eq. (7) can be computed at node i, and $\varepsilon_i^{t+1} = 0$ is precisely the first-order optimality condition for (6). Let $\varepsilon^{t+1} := (\varepsilon_1^{t+1}; \ldots; \varepsilon_n^{t+1})$ and $\bar{X}^{t+1} := (\bar{x}_1^{t+1}; \ldots; \bar{x}_n^{t+1})$. The following lemma relates the residual to the local solution error, with its proof provided in Appendix B.1.

Lemma 1. *For any local output X^{t+1} and the corresponding residual ε^{t+1} defined by (7), $\left\| X^{t+1} - \bar{X}^{t+1} \right\|_M \leq \left\| \varepsilon^{t+1} \right\|_{M^{-1}}$.*

The lemma shows that the locally computed residual controls the error in the local solution. Motivated by this bound, we consider the following two nodewise stopping criteria:

- *Absolute stopping rule.* Given a tolerance $\eta_{i,t+1} > 0$, node i stops when

$$\left\| \varepsilon_i^{t+1} \right\|_{M_i^{-1}} \leq \eta_{i,t+1}. \tag{8}$$

  The tolerances may vary across nodes and rounds. Suitable schedules are specified in the convergence analysis.

- *Relative stopping rule.* To adapt the local accuracy to the progress of each node, we use a criterion inspired by relative error frameworks for proximal methods (Solodov & Svaiter, 1999; Eckstein & Silva, 2013; Zhu et al., 2026). Specifically, let $B = \mathrm{diag}(B_1, \ldots, B_n) \succ 0$ satisfy $B \preceq S$, initialize $\nu_i^0 = x_i^0$, and choose $0 < \theta_i < 1$. Node i stops when

$$2\left|\langle \nu_i^t - x_i^{t+1}, \varepsilon_i^{t+1} \rangle\right| + \left\| \varepsilon_i^{t+1} \right\|_{B_i^{-1}}^2 \leq \theta_i^2 \left\| x_i^{t+1} - x_i^t \right\|_{B_i}^2. \tag{9}$$

  Once the criterion is satisfied, node i returns x_i^{t+1} and updates the auxiliary vector as

$$\nu_i^{t+1} = \nu_i^t - B_i^{-1}\varepsilon_i^{t+1}. \tag{10}$$

Both criteria are evaluated independently at each node, allowing agents to perform different numbers of local updates in the same round. We denote the two modes by ABS and REL. The resulting method is given in Algorithm 1, and its proximal-gradient implementation is provided in Appendix A.1.

The following lemma ensures finite termination of both stopping criteria when the proximal-gradient solver is initialized at the current outer iterate and its proof is provided in Appendix B.1.

Lemma 2 (Finite termination)**.** *Fix a round t and a node i. Let $\{u_i^{t,e}\}$ be generated by (13) from $u_i^{t,0} = x_i^t$, with $0 < \gamma_i \leq (L_i + \lambda_{\max}(M_i))^{-1}$, and let $\varepsilon_i^{t,e}$ be defined by (14) for $e \geq 1$. Then $u_i^{t,e} \to \bar{x}_i^{t+1}$ and $\varepsilon_i^{t,e} \to 0$. Consequently, the local iterations terminate after finitely many updates under either (8) with $\eta_{i,t+1} > 0$ or (9) with $0 < \theta_i < 1$.*

Algorithm 1 CFI-spALM with residual-based local solves.

Require: $x_i^0 = x^0$, $\omega_i^0 = q_i^0 = 0$, $\rho > 0$, and $M = \mathrm{diag}(M_1, \ldots, M_n) \succ 0$ with $S = M - \rho\mathcal{Z} \succeq 0$; the parameters of ABS or REL; for REL, $\nu_i^0 = x_i^0$

for $t = 0, 1, 2, \ldots$ **do** ▷ outer rounds
  for all agents $i \in \mathcal{V}$ in parallel **do**
    Approximately solve (6) and compute (7) until the selected rule, (8) or (9), is satisfied
    if REL is used **then**
      Update ν_i^{t+1} by (10)
    end if
  end for
  Exchange x_i^{t+1} with the neighbors ▷ one model-vector round
  for all agents $i \in \mathcal{V}$ in parallel **do**
    $q_i^{t+1} \leftarrow \sum_{\ell \in \mathcal{N}_i \cup \{i\}} Z_{i\ell} x_\ell^{t+1}$
    $\omega_i^{t+1} \leftarrow \omega_i^t + \rho q_i^{t+1}$
  end for
end for

5 SUBLINEAR CONVERGENCE UNDER CONVEXITY

This section establishes global convergence and sublinear rates for both stopping rules under convexity. For problem (4), define the primal–dual KKT mapping and its solution set by

$$\mathcal{T}(X, y) := \begin{bmatrix} \nabla F(X) + \partial G(X) + \mathcal{A}^{\mathsf{T}} y \\ -\mathcal{A}X \end{bmatrix}, \qquad \mathcal{K} := \mathcal{T}^{-1}(0).$$

Thus, $z^\star := (X^\star, y^\star) \in \mathcal{K}$ if and only if it satisfies the stationarity and feasibility conditions of (4). The analysis uses the following assumption.

Assumption 2. *Each f_i is convex and continuously differentiable with an L_i-Lipschitz gradient, and each g_i is proper, closed, and convex. The function $\Phi = F + G$ is proper, the KKT set $\mathcal{K}$ is nonempty, $\rho > 0$, and the block-diagonal matrix $M \succ 0$ satisfies $S = M - \rho\mathcal{A}^{\mathsf{T}}\mathcal{A} \succeq 0$.*

5.1 CONVERGENCE FOR ABSOLUTE STOPPING RULE

The nodewise absolute rule in (8) controls each local optimality residual through a prescribed tolerance. We first show that summability of the resulting aggregate residual ensures convergence of the primal–dual sequence and gives a best-iterate KKT residual bound.

Theorem 1 (Absolute rule)**.** *Suppose Assumption 2 holds and $\sum_{t=0}^{\infty} \delta_{t+1} < \infty$ with $\delta_{t+1} := \|\varepsilon^{t+1}\|_{M^{-1}}$. Then the following statements hold:*

(i) The primal–dual sequence converges to a KKT point: $z^t := (X^t, y^t) \to \bar{z} = (\bar{X}, \bar{y}) \in \mathcal{K}$.

(ii) The local residuals and the KKT residual components satisfy

$$\varepsilon^t \to 0, \quad \mathcal{A}X^t \to 0, \quad \mathrm{dist}\big(0, \nabla F(X^t) + \partial G(X^t) + \mathcal{A}^{\mathsf{T}} y^t\big) \to 0.$$

(iii) The objective values converge to the optimal value $\Phi^\star$ of (4), that is, $\Phi(X^t) \to \Phi(\bar{X}) = \Phi^\star$.

(iv) There exists $C_{\mathrm{abs}} > 0$, independent of K, such that $\min_{0 \le t < K} \mathrm{dist}\big(0, \mathcal{T}(z^{t+1})\big) \le C_{\mathrm{abs}}/\sqrt{K}$.

The summability condition in Theorem 1 can be enforced independently at each node. If $\sum_{t=0}^{\infty} \eta_{i,t+1} < \infty$ for every agent, then $\delta_{t+1} \le \big(\sum_{i=1}^n \eta_{i,t+1}^2\big)^{1/2} \le \sum_{i=1}^n \eta_{i,t+1}$, and hence $\sum_{t=0}^{\infty} \delta_{t+1} < \infty$. Valid choices include $\eta_{i,t+1} = a_i \beta^{t+1}$ with $a_i > 0$ and $0 < \beta < 1$, and $\eta_{i,t+1} = a_i/(t+1)^{1+\gamma}$ with $\gamma > 0$.

5.2 CONVERGENCE FOR RELATIVE STOPPING RULE

The relative rule in (9) determines the required local accuracy from the current local step, without prescribing a tolerance sequence across rounds. We next show that this nodewise condition yields convergence to the KKT set and a best-iterate KKT residual bound.

Theorem 2 (Relative rule). *Suppose Assumption 2 holds and $S \succeq B = \operatorname{diag}(B_1, \dots, B_n) \succ 0$. Set $z^t := (X^t, y^t)$, $\nu^t := (\nu_1^t; \dots; \nu_n^t)$, $P := \operatorname{diag}(S, \rho^{-1}\mathrm{I})$, and $\operatorname{dist}_P(z, \mathcal{K}) := \inf_{z^\star \in \mathcal{K}} \|z - z^\star\|_P$. Let every local output satisfy (9) with $0 < \theta_i < 1$, and update ν_i^t according to (10). Then:*

(i) The sequences $\{z^t\}$ and $\{\nu^t\}$ are bounded.

(ii) The residuals satisfy

$$\varepsilon^t \to 0, \qquad \mathcal{A}X^t \to 0, \qquad \operatorname{dist}\big(0, \nabla F(X^t) + \partial G(X^t) + \mathcal{A}^{\mathsf{T}} y^t\big) \to 0, \qquad \operatorname{dist}(0, \mathcal{T}(z^t)) \to 0.$$

(iii) Every accumulation point of $\{z^t\}$ belongs to $\mathcal{K}$, while $\operatorname{dist}_P(z^t, \mathcal{K}) \to 0$ and $\Phi(X^t) \to \Phi^\star$.

(iv) For every $z^\star \in \mathcal{K}$ and $K \geq 1$, $\min_{0 \leq t < K} \operatorname{dist}(0, \mathcal{T}(z^{t+1})) \leq c_{\mathrm{rel}} \sqrt{\mathcal{E}_0(z^\star)} / (\sqrt{(1-\theta^2)K})$, where $\theta := \max_i \theta_i$, $\mathcal{E}_t(z^\star) := \|z^t - z^\star\|_P^2 + \|\nu^t - X^\star\|_B^2$, $c_{\mathrm{rel}} := \sqrt{\|P\|_2} + \theta\sqrt{\|B\|_2}$.

In addition to the best-iterate KKT residual bounds, the averaged primal iterates satisfy $O(K^{-1})$ feasibility and objective-error bounds.

Corollary 1 (Ergodic primal rates). *Suppose the assumptions of either Theorem 1 or Theorem 2 hold. For $K \geq 1$, let $\widehat{X}^K := K^{-1} \sum_{t=0}^{K-1} X^{t+1}$. Then there exist constants $C_{\mathrm{feas}} > 0$ and $C_{\mathrm{obj}} > 0$, independent of K, such that $\|\mathcal{A}\widehat{X}^K\| \leq C_{\mathrm{feas}}/K$ and $|\Phi(\widehat{X}^K) - \Phi^\star| \leq C_{\mathrm{obj}}/K$.*

Theorems 1 and 2 show that both stopping rules give an $O(K^{-1/2})$ best-iterate KKT residual bound, while Corollary 1 gives $O(K^{-1})$ ergodic primal rates. Since each outer iteration requires one model-vector exchange, these bounds correspond to $O(\epsilon^{-2})$ communication rounds for an ϵ-accurate KKT residual and $O(\epsilon^{-1})$ communication rounds for an ϵ-accurate ergodic feasibility violation and objective error. The proofs are provided in Appendix B.

6 Linear Convergence under Error Bounds

This section establishes linear convergence for both stopping rules under a local error bound that relates the KKT residual to the distance from the KKT set.

Assumption 3 (KKT error bound). *For a given KKT point $\bar{z} \in \mathcal{K}$, there exist a neighborhood $\mathcal{U}$ of $\bar{z}$ and a constant $\kappa > 0$ such that*

$$\operatorname{dist}(z, \mathcal{K}) \leq \kappa \operatorname{dist}(0, \mathcal{T}(z)), \qquad z \in \mathcal{U}. \tag{11}$$

This condition is metric subregularity of $\mathcal{T}$ at $(\bar{z}, 0)$ and ensures that, within $\mathcal{U}$, a small KKT residual implies proximity to the solution set. Detailed proofs are provided in Appendix B.

6.1 Convergence for Absolute Stopping Rule

Geometrically decreasing absolute tolerances satisfy the summability condition in Theorem 1, which guarantees convergence of the full primal–dual sequence to a KKT point. The following theorem further establishes a linear convergence rate under a local error bound at this limit point.

Theorem 3 (Absolute rule). *Use the notation z^t, δ_{t+1}, P, and dist_P from Theorems 1 and 2. Suppose Assumption 2 holds and $\delta_{t+1} \leq C\beta^{t+1}$ for some $C > 0$ and $0 < \beta < 1$. Let $\bar{z} \in \mathcal{K}$ be the limit given by Theorem 1, and suppose Assumption 3 holds at $\bar{z}$. Then there exists $t_0 \geq 0$ such that*

$$d_{t+1} \leq q_a d_t + C\beta^{t+1}, \qquad t \geq t_0,$$

where $d_t := \operatorname{dist}_P(z^t, \mathcal{K})$, $a := \kappa \|P\|_2$, and $q_a := a/\sqrt{1+a^2} < 1$. Let $r := \max\{q_a, \beta\}$, and define b_t as r^t when $q_a \neq \beta$ or tr^t otherwise. Then, for any fixed $\widehat{q} \in (r, 1)$,

$$d_t + \|z^t - \bar{z}\| + \|\mathcal{A}X^t\| + \|R^t\| = O(b_t) = O(\widehat{q}^t).$$

Here, $\xi^{t+1} := (\varepsilon^{t+1}, 0)$ and $R^{t+1} := P(z^t - z^{t+1}) + \xi^{t+1} \in \mathcal{T}(z^{t+1})$. When P is singular, $\|\cdot\|_P$ and dist_P are understood as a seminorm and semidistance, respectively. Thus, z^t converges R-linearly to $\bar{z}$, while $\mathcal{A}X^t$ and R^t converge R-linearly to zero.

The residual bound can be satisfied independently at each node by choosing $\eta_{i,t+1} \leq C_i \beta^{t+1}$ in (8), which yields $\delta_{t+1} \leq C\beta^{t+1}$ with $C = (\sum_i C_i^2)^{1/2}$. The preceding theorem shows that the asymptotic rate depends on the error-bound constant through q_a and the tolerance decay factor β.

6.2 Convergence for Relative Stopping Rule

Theorem 2 guarantees that the relative stopping rule satisfies $\mathrm{dist}_P(z^t, \mathcal{K}) \to 0$ without requiring a prescribed tolerance sequence. To establish a linear rate, we use a KKT error bound in the P-weighted norm. Since the relative rule requires $S \succ 0$, we have $P \succ 0$, and the weighted error bound is given by

$$\mathrm{dist}_P(z, \mathcal{K}) \le \kappa_P \, \mathrm{dist}_{P^{-1}}(0, \mathcal{T}(z)), \tag{12}$$

where $\mathrm{dist}_{P^{-1}}(0, \mathcal{T}(z)) := \inf_{v \in \mathcal{T}(z)} \|v\|_{P^{-1}}$. By norm equivalence, Assumption 3 implies (12) locally. The result below assumes that (12) holds with a common constant $\kappa_{P,r}$ throughout $\mathcal{U}_r^P := \{z : \mathrm{dist}_P(z, \mathcal{K}) \le r\}$ for some $r > 0$. Together with the global convergence in Theorem 2, this uniform error bound gives a linear contraction once the iterates enter $\mathcal{U}_r^P$.

Theorem 4 (Relative rule)**.** *Using the notation introduced above, suppose the assumptions of Theorem 2 hold and (12) holds throughout $\mathcal{U}_r^P$ for some $r > 0$ with constant $\kappa_{P,r}$. If $\theta < (1 + \kappa_{P,r} + \sqrt{1 + \kappa_{P,r}^2})^{-1}$, then $\chi_{\mathrm{rel}} := (\kappa_{P,r} + \frac{\theta}{1-\theta})/\sqrt{1 + \kappa_{P,r}^2} < 1$, and there exists $t_0 \ge 0$ such that*

$$d_{t+1} \le \chi_{\mathrm{rel}} d_t, \qquad t \ge t_0.$$

Consequently, d_t converges Q-linearly for all $t \ge t_0$. Moreover, there exists $\bar{z} \in \mathcal{K}$ such that

$$d_t \le \chi_{\mathrm{rel}}^{t-t_0} d_{t_0}, \qquad \|z^t - \bar{z}\| + \|\mathcal{A}X^t\| + \|R^t\| = O(\chi_{\mathrm{rel}}^{t-t_0}).$$

Thus, the full primal–dual sequence, the consensus violation, and the residual R^t converge R-linearly.

7 Numerical Experiments

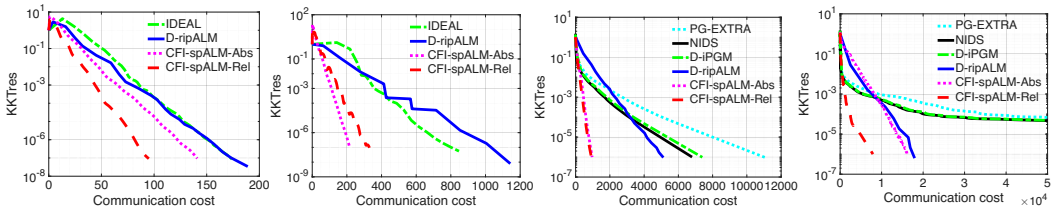
To demonstrate the effectiveness and communication advantages of CFI-spALM, we compare it with representative decentralized methods on smooth strongly convex logistic regression and nonsmooth rank-deficient LASSO. The experiments examine convergence reliability and the number of model-vector exchanges required to reach a prescribed optimization target.

Experimental setting. For both problems, the communication network is generated as a connected Erdős–Rényi graph with independent edge probability 0.2, and the mixing matrix is constructed using symmetric Metropolis weights (Xiao & Boyd, 2004; Nedić & Ozdaglar, 2009; Zhu et al., 2026). For evaluation, define the aggregate functions $f(x) := \sum_{i=1}^n f_i(x)$ and $g(x) := \sum_{i=1}^n g_i(x)$, the average iterate $\bar{x}^t := n^{-1} \sum_{i=1}^n x_i^t$, and the consensus projector $\Pi := \mathrm{I}_n - n^{-1}\mathbf{1}\mathbf{1}^\mathsf{T}$. For a fixed evaluation stepsize α_{ev}, let $\mathcal{G}_{\alpha_{\mathrm{ev}}}(x) := \alpha_{\mathrm{ev}}^{-1}[x - \mathrm{prox}_{\alpha_{\mathrm{ev}} g}(x - \alpha_{\mathrm{ev}} \nabla f(x))]$. Every trajectory is evaluated by

$$\texttt{KKTres}_t := \max\left\{ \left(\frac{1}{n} \sum_{i=1}^n \|x_i^t - \bar{x}^t\|^2 \right)^{1/2}, \frac{\|\mathcal{G}_{\alpha_{\mathrm{ev}}}(\bar{x}^t)\|}{\max\{1, \|\mathcal{G}_{\alpha_{\mathrm{ev}}}(\bar{x}^0)\|\}} \right\}.$$

Here, $\alpha_{\mathrm{ev}} = (\frac{1}{4}\|A\|_2^2 + n\lambda)^{-1}$ for logistic regression and $\alpha_{\mathrm{ev}} = \|A\|_2^{-2}$ for LASSO with $A := [A_1^\mathsf{T}, \ldots, A_n^\mathsf{T}]^\mathsf{T}$ being the vertically stacked data matrix. For CFI-spALM, we set $M_i = (2\rho Z_{ii} + \gamma_M)\mathrm{I}_d$, $B_i = \gamma_M \mathrm{I}_d$, and $\rho = s_\rho \max_i L_i / \lambda_{\max}(Z)$. The dimensionless quantities s_ρ and γ_M/ρ are treated as hyperparameters and tuned separately for each problem setting. The local subproblems are solved by accelerated proximal gradient with stepsize $0.99/(L_i + \lambda_{\max}(M_i))$ and the detailed procedure is provided in Appendix A. We denote the variants using the absolute and relative stopping rules by CFI-spALM-Abs and CFI-spALM-Rel, respectively. Further experimental details, including the hyperparameter settings of all comparison methods, are provided in Appendix C. All numerical experiments are performed using MATLAB R2026a on a MacBook Pro equipped with an Apple M5 Pro processor (18 CPU cores) and 48 GB memory.

Smooth strongly convex logistic regression. To compare CFI-spALM with the graph-coupled multiplier methods IDEAL (Arjevani et al., 2020) and D-ripALM (Zhu et al., 2026), we consider



(a) LR ($\lambda = 10^{-1}$). (b) LR ($\lambda = 10^{-3}$). (c) LASSO ($\lambda_c = 10^{-1}$). (d) LASSO ($\lambda_c = 10^{-2}$).

Figure 1: Communication performance on logistic regression (LR) and rank-deficient LASSO.

the ℓ_2-regularized binary classification problem used in Zhu et al. (2026):

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \left[\sum_{j=1}^{m_i} \log\bigl(1+\exp(-y_{ij}a_{ij}^{\mathsf{T}}x)\bigr) + \frac{\lambda}{2}\|x\|^2 \right].$$

We use $n = 10$ agents, $d = 1000$, and $m_i = 40$ samples per agent. We generate $x_{\text{true}} \sim \mathcal{N}(0, \mathrm{I}_d)$, $a_{ij} \sim \mathcal{N}(0, \mathrm{I}_d/d)$, and $\epsilon_{ij} \sim \mathcal{N}(0, 0.1^2)$, with labels $y_{ij} = \operatorname{sign}(a_{ij}^{\mathsf{T}}x_{\text{true}} + \epsilon_{ij})$. The results for $\lambda = 10^{-1}$ and $\lambda = 10^{-3}$ are reported in Figs. 1(a)–(b), respectively. All four methods reach the prescribed target in both settings. For $\lambda = 10^{-1}$, CFI-spALM-Rel requires the fewest communicated d-vectors, followed by CFI-spALM-Abs, and both outperform IDEAL and D-ripALM. For $\lambda = 10^{-3}$, the two CFI-spALM variants continue to reach the target with fewer communicated d-vectors than the graph-coupled methods, demonstrating their communication efficiency in both logistic-regression settings.

Nonsmooth LASSO. To further evaluate CFI-spALM on general composite objectives, we compare it with the single-loop methods PG-EXTRA (Shi et al., 2015b), NIDS (Li et al., 2019), and D-iPGM (Guo et al., 2023), as well as the double-loop multiplier method D-ripALM (Zhu et al., 2026), on the following LASSO problem:

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \left[\frac{1}{2}\|A_i x - b_i\|^2 + \frac{\lambda}{n}\|x\|_1 \right], \qquad \lambda = \lambda_c \left\|A^{\mathsf{T}}b\right\|_\infty.$$

We use $n = 20$, $d = 1000$, and $m_i = 10$, and test $\lambda_c = 10^{-1}$ and $\lambda_c = 10^{-2}$. The entries of $A_i \in \mathbb{R}^{m_i\times d}$ are drawn independently from $\mathcal{N}(0, 1/m_i)$. The vector x_{true} has 10% nonzero entries drawn independently from $\mathcal{N}(0, 1)$, and $b_i = A_i x_{\text{true}} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.1^2\mathrm{I}_{m_i})$. Here, $A = [A_1; \ldots; A_n]$ and $b = [b_1; \ldots; b_n]$. Since $m_i < d$, every local least-squares loss is rank deficient, while the ℓ_1 term makes the objective nonsmooth. The results for $\lambda_c = 10^{-1}$ and $\lambda_c = 10^{-2}$ are reported in Figs. 1(c)–(d), respectively. For $\lambda_c = 10^{-1}$, all compared methods reach the prescribed target within the given communication budget, while the two CFI-spALM variants require substantially fewer communicated d-vectors. For the more difficult setting $\lambda_c = 10^{-2}$, the single-loop methods do not reach the target within the communication budget, whereas the ALM-type methods do. The two CFI-spALM variants retain a large communication advantage over D-ripALM, further supporting the communication benefit observed in the logistic-regression experiments.

8 Conclusion

This paper proposed CFI-spALM for decentralized composite optimization. A structured semi-proximal term decouples the graph-dependent multiplier subproblem, enabling adaptive local solves without inner communication and requiring only one model-vector exchange per outer iteration. We established global convergence and sublinear rates under absolute and relative stopping rules, together with linear convergence under metric subregularity. Numerical results demonstrate its communication efficiency on smooth and nonsmooth problems. This communication advantage may come at the cost of additional local computation. Future work will investigate how different semi-proximal matrices affect communication complexity and develop algorithms with a better balance between communication and computation.

AI USE STATEMENT

We used generative AI tools solely to improve the language and readability of this manuscript. We carefully reviewed and revised the AI-assisted text and take full responsibility for the manuscript.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1709–1720, 2017.

Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gurbuzbalaban, Stefanie Jegelka, and Hongzhou Lin. IDEAL: Inexact decentralized accelerated augmented lagrangian method. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20648–20659, 2020.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

Chuan Chen, Tianchi Liao, Xiaojun Deng, Zihou Wu, Sheng Huang, and Zibin Zheng. Advances in robust federated learning: A survey with heterogeneity considerations. *IEEE Transactions on Big Data*, 11(3):1548–1567, 2025.

Liang Chen, Xudong Li, Defeng Sun, and Kim-Chuan Toh. On the equivalence of inexact proximal ALM and ADMM for a class of convex composite programming. *Mathematical Programming*, 185:111–161, 2021.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57 (3):592–606, 2012.

Jonathan Eckstein and Paulo J. S. Silva. A practical relative error criterion for augmented Lagrangians. *Mathematical Programming*, 141:319–348, 2013.

Luyao Guo, Xinli Shi, Jinde Cao, and Zihao Wang. Decentralized inexact proximal gradient method with network-independent stepsizes for convex composite optimization. *IEEE Transactions on Signal Processing*, 71:786–801, 2023.

Qing Hu, Tianchi Liao, Shuyi Wu, Lei Yang, and Chuan Chen. FedLAGC: Towards high performance system-heterogeneous federated learning via layer-adaptive submodel extraction and gradient correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21858–21866, 2026.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Dennis Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360(2):681–688, 2009.

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67 (17):4494–4506, 2019.

Cassio G. Lopes and Ali H. Sayed. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7):3122–3136, 2008.

Gonzalo Mateos, Juan Andrés Bazerque, and Georgios B. Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276, 2010.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1273–1282, 2017.

Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Angelia Nedić, Alex Olshevsky, and Michael G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.

Joel B. Predd, Sanjeev R. Kulkarni, and H. Vincent Poor. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):56–69, 2006.

R. Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976a.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976b.

Ali H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7(4-5):311–801, 07 2014.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3027–3036, 2017.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015a.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, 2015b.

Mikhail V. Solodov and Benar F. Svaiter. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7:323–345, 1999.

John N. Tsitsiklis, Dimitri P. Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9): 803–812, 1986.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Lin Xiao and Stephen Boyd. Optimal scaling of a gradient method for distributed resource allocation. *Journal of Optimization Theory and Applications*, 129(3):469–488, 2006.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Lei Yang, Jiayi Zhu, Ling Liang, and Kim-Chuan Toh. Convergence analysis of a relative-type inexact preconditioned proximal ALM for convex nonlinear programming. *arXiv preprint arXiv:2510.25261*, 2026.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Jiayi Zhu, Hong Wang, Ling Liang, and Lei Yang. D-ripALM: A tuning-friendly decentralized relative-type inexact proximal augmented lagrangian method. *arXiv preprint arXiv:2602.06398*, 2026.

Algorithm 2 Distributed implementation of CFI-spALM with proximal-gradient local solves.

Require: $x_i^0 = x^0$, $\omega_i^0 = q_i^0 = 0$, and $\rho > 0$; for REL, $\nu_i^0 = x_i^0$; $M = \mathrm{diag}(M_1, \dots, M_n) \succ 0$ with $S = M - \rho\mathcal{Z} \succeq 0$; for REL, $S \succeq B = \mathrm{diag}(B_i) \succ 0$; $0 < \gamma_i \le (L_i + \lambda_{\max}(M_i))^{-1}$; the parameters and mode of the selected stopping rule

for $t = 0, 1, 2, \dots$ **do** ▷ outer rounds
  for all agents $i \in \mathcal{V}$ in parallel **do**
    $c_i^t \leftarrow \omega_i^t + \rho q_i^t - M_i x_i^t$
    $u_i^{t,0} \leftarrow x_i^t$
    $p_i^{t,0} \leftarrow \nabla f_i(u_i^{t,0}) + M_i u_i^{t,0} + c_i^t$
    for $e = 0, 1, 2, \dots$ **do**
      $u_i^{t,e+1} \leftarrow \mathrm{prox}_{\gamma_i g_i}(u_i^{t,e} - \gamma_i p_i^{t,e})$
      $d_i^{t,e+1} \leftarrow u_i^{t,e} - u_i^{t,e+1}$
      $\zeta_i^{t,e+1} \leftarrow \gamma_i^{-1} d_i^{t,e+1} - p_i^{t,e}$
      $p_i^{t,e+1} \leftarrow \nabla f_i(u_i^{t,e+1}) + M_i u_i^{t,e+1} + c_i^t$
      $\varepsilon_i^{t,e+1} \leftarrow p_i^{t,e+1} + \zeta_i^{t,e+1}$
      if $\mathsf{Stop}_{i,t}(\varepsilon_i^{t,e+1}, u_i^{t,e+1})$ **then**
        $x_i^{t+1} \leftarrow u_i^{t,e+1}$
        if REL is used **then**
          $\nu_i^{t+1} \leftarrow \nu_i^t - B_i^{-1} \varepsilon_i^{t,e+1}$
        end if
        break
      end if
    end for
  end for
  Wait until all agents have stopped
  Exchange x_i^{t+1} with the neighbors ▷ one model-vector round
  for all agents $i \in \mathcal{V}$ in parallel **do**
    $q_i^{t+1} \leftarrow \sum_{\ell \in \mathcal{N}_i \cup \{i\}} Z_{i\ell} x_\ell^{t+1}$
    $\omega_i^{t+1} \leftarrow \omega_i^t + \rho q_i^{t+1}$
  end for
end for

APPENDIX

A ALGORITHM

A.1 DISTRIBUTED IMPLEMENTATION

Algorithm 2 implements CFI-spALM by alternating local solves with a model-vector exchange. At the start of round t, node i uses its local variables and cached graph information to form $c_i^t := \omega_i^t + \rho q_i^t - M_i x_i^t$, which remains fixed throughout the local solve. The gradient of the smooth part of (6) can therefore be written as

$$p_i^t(u) := \nabla f_i(u) + M_i u + c_i^t.$$

Starting from $u_i^{t,0} = x_i^t$, node i uses a stepsize $0 < \gamma_i \le (L_i + \lambda_{\max}(M_i))^{-1}$ and performs

$$u_i^{t,e+1} = \mathrm{prox}_{\gamma_i g_i}\left(u_i^{t,e} - \gamma_i p_i^t(u_i^{t,e})\right), \tag{13}$$

where $\mathrm{prox}_{\gamma g_i}(u) := \arg\min_x \{g_i(x) + (2\gamma)^{-1}\|x - u\|^2\}$. To evaluate the stopping criterion at the new candidate, let $d_i^{t,e+1} := u_i^{t,e} - u_i^{t,e+1}$ and use the proximal optimality condition to obtain

$$\zeta_i^{t,e+1} := \gamma_i^{-1} d_i^{t,e+1} - p_i^t(u_i^{t,e}) \in \partial g_i(u_i^{t,e+1}).$$

The corresponding residual (7) is

$$\varepsilon_i^{t,e+1} = p_i^t(u_i^{t,e+1}) + \zeta_i^{t,e+1},$$

or, equivalently,

$$\begin{aligned}\varepsilon_i^{t,e+1} = {} & \nabla f_i(u_i^{t,e+1}) - \nabla f_i(u_i^{t,e}) \\ & + (\gamma_i^{-1}\mathrm{I} - M_i)(u_i^{t,e} - u_i^{t,e+1}).\end{aligned} \tag{14}$$

Node i repeats these updates until $\mathsf{Stop}_{i,t}(\varepsilon_i^{t,e+1}, u_i^{t,e+1})$ holds, and accepts the candidate and residual as x_i^{t+1} and ε_i^{t+1}. Under REL, it also updates ν_i^{t+1} according to (10). These computations correspond to lines 7–17 of Algorithm 2 and require no neighbor exchange.

Once all agents have completed their local solves, they exchange the accepted model vectors in line 22. Each node then computes

$$q_i^{t+1} = \sum_{\ell \in \mathcal{N}_i \cup \{i\}} Z_{i\ell} x_\ell^{t+1}, \qquad \omega_i^{t+1} = \omega_i^t + \rho q_i^{t+1},$$

and retains q_i^{t+1} for the next local solve. This completes the outer round with one neighbor exchange of a d-dimensional model vector, regardless of how many local updates each agent performs.

A.2 Accelerated Local Solves

The same stopping criteria can also be used with accelerated proximal-gradient updates, as in the experiments. Set $u_i^{t,-1} = u_i^{t,0} = x_i^t$, let $\mu_i^{\mathrm{loc}} := \lambda_{\min}(M_i)$, and use the same step-size condition for γ_i. Define

$$\beta_i^{\mathrm{loc}} := \frac{1 - \sqrt{\gamma_i \mu_i^{\mathrm{loc}}}}{1 + \sqrt{\gamma_i \mu_i^{\mathrm{loc}}}}.$$

Each local iteration computes an extrapolated point and a new candidate:

$$\begin{aligned} \widetilde{u}_i^{t,e} &= u_i^{t,e} + \beta_i^{\mathrm{loc}}(u_i^{t,e} - u_i^{t,e-1}), \\ u_i^{t,e+1} &= \mathrm{prox}_{\gamma_i g_i}\left(\widetilde{u}_i^{t,e} - \gamma_i p_i^t(\widetilde{u}_i^{t,e})\right). \end{aligned} \tag{15}$$

The proximal optimality condition gives

$$\zeta_i^{t,e+1} := \gamma_i^{-1}(\widetilde{u}_i^{t,e} - u_i^{t,e+1}) - p_i^t(\widetilde{u}_i^{t,e}) \in \partial g_i(u_i^{t,e+1}),$$

so the residual at the new candidate is

$$\varepsilon_i^{t,e+1} = p_i^t(u_i^{t,e+1}) + \zeta_i^{t,e+1}. \tag{16}$$

This is precisely (7) evaluated at the accelerated candidate. The agent applies the selected ABS or REL stopping criterion to this residual and accepts the candidate only when that criterion is satisfied. These accelerated updates use the same cached network information and preserve the communication pattern of Algorithm 1.

B Proofs of the Theoretical Results

B.1 Local-Solve Properties

We first show that the local optimality residual controls the distance between an inexact local output and the exact minimizer of the corresponding subproblem.

Lemma 1 (Local solution error)**.** *Let $\bar{X}^{t+1}$ be the exact minimizer of (5). For any local output X^{t+1} and residual ε^{t+1} defined by (7),*

$$\left\|X^{t+1} - \bar{X}^{t+1}\right\|_M \le \left\|\varepsilon^{t+1}\right\|_{M^{-1}}.$$

Proof. Let $v^{t+1} := (v_1^{t+1}; \ldots; v_n^{t+1}) \in \partial\Phi(X^{t+1})$. Since $\bar{X}^{t+1}$ exactly minimizes (5), there exists $\bar{v}^{t+1} \in \partial\Phi(\bar{X}^{t+1})$ such that

$$0 = \bar{v}^{t+1} + \Omega^t + \rho q^t + M(\bar{X}^{t+1} - X^t).$$

Stacking the local residual relations in (7) gives

$$\varepsilon^{t+1} = v^{t+1} + \Omega^t + \rho q^t + M(X^{t+1} - X^t).$$

Subtracting the exact optimality condition from this identity yields

$$\varepsilon^{t+1} = v^{t+1} - \bar{v}^{t+1} + M(X^{t+1} - \bar{X}^{t+1}).$$

Because $\partial\Phi$ is monotone, we have $\langle v^{t+1} - \bar{v}^{t+1}, X^{t+1} - \bar{X}^{t+1} \rangle \geq 0$. Taking the inner product of the residual identity with $X^{t+1} - \bar{X}^{t+1}$ therefore gives

$$\left\|X^{t+1} - \bar{X}^{t+1}\right\|_M^2 \leq \langle \varepsilon^{t+1}, X^{t+1} - \bar{X}^{t+1} \rangle \leq \left\|\varepsilon^{t+1}\right\|_{M^{-1}} \left\|X^{t+1} - \bar{X}^{t+1}\right\|_M,$$

where the second inequality follows from the weighted Cauchy–Schwarz inequality. If $X^{t+1} \neq \bar{X}^{t+1}$, division by $\left\|X^{t+1} - \bar{X}^{t+1}\right\|_M$ proves the result. The case $X^{t+1} = \bar{X}^{t+1}$ is immediate. □

The next lemma shows that the two stopping rules can be satisfied after finitely many proximal-gradient updates.

Lemma 2 (Finite termination of the local solver)**.** *Fix an outer iteration t and an agent i. Let $\{u_i^{t,e}\}$ be generated by (13) from $u_i^{t,0} = x_i^t$ with $0 < \gamma_i \leq \big(L_i + \lambda_{\max}(M_i)\big)^{-1}$, and let $\varepsilon_i^{t,e}$ be defined by (14) for $e \geq 1$. Then*

$$u_i^{t,e} \to \bar{x}_i^{t+1}, \qquad \varepsilon_i^{t,e} \to 0.$$

Consequently, the local iterations terminate after finitely many updates under either the absolute rule (8) with $\eta_{i,t+1} > 0$ or the relative rule (9) with $0 < \theta_i < 1$.

Proof. Define the smooth part of the local objective in (6) by

$$h_i^t(u) := f_i(u) + \langle \omega_i^t + \rho q_i^t, u \rangle + \frac{1}{2} \left\|u - x_i^t\right\|_{M_i}^2.$$

Its gradient is μ_i-strongly monotone and $\widehat{L}_i$-Lipschitz continuous, where $\mu_i := \lambda_{\min}(M_i) > 0$ and $\widehat{L}_i := L_i + \lambda_{\max}(M_i)$. Since g_i is convex, the complete local objective $h_i^t + g_i$ is μ_i-strongly convex and therefore has the unique minimizer $\bar{x}_i^{t+1}$.

Define the proximal-gradient mapping

$$T_i^t(u) := \operatorname{prox}_{\gamma_i g_i}\big(u - \gamma_i \nabla h_i^t(u)\big).$$

The optimality condition of (6) is equivalent to $\bar{x}_i^{t+1} = T_i^t(\bar{x}_i^{t+1})$. For any u and v, the nonexpansiveness of the proximal mapping gives

$$\left\|T_i^t(u) - T_i^t(v)\right\|^2 \leq \left\|u - v - \gamma_i\big(\nabla h_i^t(u) - \nabla h_i^t(v)\big)\right\|^2.$$

Because h_i^t is μ_i-strongly convex and has an $\widehat{L}_i$-Lipschitz gradient,

$$\left\langle \nabla h_i^t(u) - \nabla h_i^t(v), u - v \right\rangle \geq \frac{\mu_i \widehat{L}_i}{\mu_i + \widehat{L}_i} \left\|u - v\right\|^2 + \frac{1}{\mu_i + \widehat{L}_i} \left\|\nabla h_i^t(u) - \nabla h_i^t(v)\right\|^2.$$

Expanding the preceding bound and using $0 < \gamma_i \leq \widehat{L}_i^{-1} \leq \frac{2}{\mu_i + \widehat{L}_i}$ yield

$$\left\|T_i^t(u) - T_i^t(v)\right\|^2 \leq \left(1 - \frac{2\gamma_i \mu_i \widehat{L}_i}{\mu_i + \widehat{L}_i}\right) \left\|u - v\right\|^2.$$

Thus, T_i^t is contractive. Since $u_i^{t,e+1} = T_i^t(u_i^{t,e})$, we have

$$\left\|u_i^{t,e+1} - \bar{x}_i^{t+1}\right\| \leq q_i \left\|u_i^{t,e} - \bar{x}_i^{t+1}\right\|, \qquad q_i := \sqrt{1 - \frac{2\gamma_i \mu_i \widehat{L}_i}{\mu_i + \widehat{L}_i}} < 1.$$

Consequently, $u_i^{t,e} \to \bar{x}_i^{t+1}$. By (14),

$$\varepsilon_i^{t,e+1} = \nabla f_i(u_i^{t,e+1}) - \nabla f_i(u_i^{t,e}) + (\gamma_i^{-1}\mathrm{I} - M_i)(u_i^{t,e} - u_i^{t,e+1}).$$

Both $u_i^{t,e}$ and $u_i^{t,e+1}$ converge to $\bar{x}_i^{t+1}$. The Lipschitz continuity of ∇f_i therefore implies $\varepsilon_i^{t,e} \to 0$.

For the absolute rule, $\eta_{i,t+1} > 0$. Hence, $\left\|\varepsilon_i^{t,e}\right\|_{M_i^{-1}} \leq \eta_{i,t+1}$ holds for all sufficiently large e, proving finite termination.

For the relative rule, ν_i^t remains fixed throughout the local solve. If $\bar{x}_i^{t+1} \neq x_i^t$, then

$$2\left|\left\langle \nu_i^t - u_i^{t,e}, \varepsilon_i^{t,e}\right\rangle\right| + \left\|\varepsilon_i^{t,e}\right\|_{B_i^{-1}}^2 \to 0,$$
$$\theta_i^2 \left\|u_i^{t,e} - x_i^t\right\|_{B_i}^2 \to \theta_i^2 \left\|\bar{x}_i^{t+1} - x_i^t\right\|_{B_i}^2 > 0.$$

Therefore, the relative stopping condition holds after finitely many updates.

If $\bar{x}_i^{t+1} = x_i^t$, then the initial point is the unique minimizer and hence a fixed point of T_i^t. Thus, $u_i^{t,1} = T_i^t(x_i^t) = x_i^t$ and $\varepsilon_i^{t,1} = 0$, so the relative stopping condition is satisfied after the first update. □

B.2 Absolute Rule under Convexity

For the remaining convergence proofs, let $z^t := (X^t, y^t)$, $P := \mathrm{diag}(S, \rho^{-1}\mathrm{I})$, $\xi^{t+1} := (\varepsilon^{t+1}, 0)$, $\delta_{t+1} := \left\|\varepsilon^{t+1}\right\|_{M^{-1}}$, and $D_t := \left\|z^{t+1} - z^t\right\|_P$. Let $v^{t+1} := (v_1^{t+1}; \ldots; v_n^{t+1}) \in \partial\Phi(X^{t+1})$. Stacking the local residual relations and using $S = M - \rho\mathcal{A}^\mathsf{T}\mathcal{A}$ give

$$\begin{aligned} v^{t+1} + \mathcal{A}^\mathsf{T} y^{t+1} + S(X^{t+1} - X^t) &= \varepsilon^{t+1}, \\ y^{t+1} - y^t &= \rho\mathcal{A}X^{t+1}. \end{aligned} \tag{17}$$

Equivalently,

$$P(z^t - z^{t+1}) + \xi^{t+1} \in \mathcal{T}(z^{t+1}),$$

which is the inexact proximal-point form of the CFI-spALM update for $\mathcal{T}$ (Rockafellar, 1976b;a; Chen et al., 2021). These relations will be used throughout the remaining proofs.

Theorem 1 (Absolute rule)**.** *Suppose Assumption 2 holds and $\sum_{t=0}^{\infty} \delta_{t+1} < \infty$. Then the following statements hold:*

(i) The full primal–dual sequence converges to a KKT point, namely, $z^t \to \bar{z} = (\bar{X}, \bar{y}) \in \mathcal{K}$.

(ii) The local residuals and the KKT residual components satisfy

$$\begin{aligned} \varepsilon^t &\to 0, \\ \mathcal{A}X^t &\to 0, \\ \mathrm{dist}\big(0, \nabla F(X^t) + \partial G(X^t) + \mathcal{A}^\mathsf{T} y^t\big) &\to 0. \end{aligned}$$

(iii) The objective values converge to the optimal value $\Phi^\star$ of (4), that is, $\Phi(X^t) \to \Phi(\bar{X}) = \Phi^\star$.

(iv) There exists $C_{\mathrm{abs}} > 0$, independent of K, such that

$$\min_{0 \le t < K} \mathrm{dist}\big(0, \mathcal{T}(z^{t+1})\big) \le \frac{C_{\mathrm{abs}}}{\sqrt{K}}, \qquad K \ge 1.$$

Proof. We first establish an estimate used throughout the proof. Fix any $z^\star = (X^\star, y^\star) \in \mathcal{K}$ and define $\mathcal{H}_t(z^\star) := \|z^t - z^\star\|_P^2$. Choose $\zeta^{t+1} \in \partial G(X^{t+1})$, $p^{t+1} := \nabla F(X^{t+1}) + \zeta^{t+1} \in \partial\Phi(X^{t+1})$, and $p^\star = -\mathcal{A}^\mathsf{T} y^\star \in \partial\Phi(X^\star)$. Combining (17) with the multiplier update and expanding the primal and dual energies gives

$$\mathcal{H}_t(z^\star) - \mathcal{H}_{t+1}(z^\star) = D_t^2 + 2\left\langle p^{t+1} - p^\star, X^{t+1} - X^\star\right\rangle - 2\left\langle \varepsilon^{t+1}, X^{t+1} - X^\star\right\rangle. \tag{18}$$

Monotonicity of $\partial\Phi$ makes the middle term nonnegative. Moreover,

$$\left\|X^{t+1} - X^\star\right\|_M^2 = \left\|X^{t+1} - X^\star\right\|_S^2 + \rho\left\|\mathcal{A}X^{t+1}\right\|^2 \le \mathcal{H}_{t+1}(z^\star) + D_t^2.$$

Therefore,

$$\left|\left\langle \varepsilon^{t+1}, X^{t+1} - X^\star\right\rangle\right| \le \delta_{t+1}\sqrt{\mathcal{H}_{t+1}(z^\star) + D_t^2}.$$

Writing $Q_t := \mathcal{H}_{t+1}(z^\star) + D_t^2$, we obtain

$$\mathcal{H}_t(z^\star) \ge Q_t - 2\delta_{t+1}\sqrt{Q_t},$$

and solving this quadratic inequality for $\sqrt{Q_t}$ gives

$$\sqrt{\mathcal{H}_{t+1}(z^\star) + D_t^2} \le \delta_{t+1} + \sqrt{\mathcal{H}_t(z^\star) + \delta_{t+1}^2} \le \sqrt{\mathcal{H}_t(z^\star)} + 2\delta_{t+1}.$$

Let

$$\Delta_1 := \sum_{t=0}^{\infty} \delta_{t+1}, \qquad \Delta_2 := \left(\sum_{t=0}^{\infty} \delta_{t+1}^2\right)^{1/2},$$
$$B := \sqrt{\mathcal{H}_0(z^\star)} + 2\Delta_1, \qquad A := \mathcal{H}_0(z^\star) + 2B\Delta_1.$$

This estimate gives $\sqrt{\mathcal{H}_{t+1}(z^\star) + D_t^2} \le B$. Summing (18) and using the preceding error bound yields

$$\sum_{t=0}^{\infty} D_t^2 \le A < \infty.$$

Step 1: Convergence of the full sequence. It follows that $y^{t+1} - y^t \to 0$ and $\mathcal{A}X^{t+1} = \rho^{-1}(y^{t+1} - y^t) \to 0$. Since $\left\|X^{t+1} - X^t\right\|_M^2 = \left\|X^{t+1} - X^t\right\|_S^2 + \rho \left\|\mathcal{A}(X^{t+1} - X^t)\right\|^2$, the relations $D_t \to 0$ and $\mathcal{A}X^t \to 0$ also give $X^{t+1} - X^t \to 0$. Since $\delta_t \to 0$ and $\|\varepsilon^t\| \le \sqrt{\|M\|_2}\,\delta_t$, the local residuals satisfy $\varepsilon^t \to 0$. The bound $\left\|X^{t+1} - X^\star\right\|_M^2 \le \mathcal{H}_{t+1}(z^\star) + D_t^2$ shows that the primal sequence is bounded; the dual sequence is bounded directly by $\mathcal{H}_t(z^\star)$. Hence, $\{z^t\}$ has a cluster point. Because $\delta_t \to 0$ and $p^t + \mathcal{A}^\mathsf{T} y^t = \varepsilon^t - S(X^t - X^{t-1}) \to 0$, every cluster point belongs to $\mathcal{K}$ by closedness of $\mathrm{gph}(\partial\Phi)$.

It remains to justify convergence of the full sequence when P is singular. For any fixed $z^\star \in \mathcal{K}$, the quasi-Fejér estimate implies that $\sqrt{\mathcal{H}_t(z^\star)} + 2\sum_{s=t}^{\infty} \delta_{s+1}$ is nonincreasing. Hence, $\mathcal{H}_t(z^\star)$ has a limit. Let $\bar{z} = (\bar{X}, \bar{y})$ be any cluster point. Along a subsequence converging to $\bar{z}$, one has $\mathcal{H}_t(\bar{z}) \to 0$, so the limit of the full sequence $\mathcal{H}_t(\bar{z})$ is zero. Thus, $\left\|X^t - \bar{X}\right\|_S \to 0$ and $y^t \to \bar{y}$. Both X^t and $\bar{X}$ are asymptotically feasible, and therefore

$$\left\|X^t - \bar{X}\right\|_M^2 = \left\|X^t - \bar{X}\right\|_S^2 + \rho \left\|\mathcal{A}(X^t - \bar{X})\right\|^2 \to 0.$$

This proves $z^t \to \bar{z} \in \mathcal{K}$ and hence establishes (i).

Step 2: Convergence of the residuals. The estimates established while proving (i) give $\varepsilon^t \to 0$, $\mathcal{A}X^t \to 0$, and $p^t + \mathcal{A}^\mathsf{T} y^t \to 0$. Because $p^t \in \nabla F(X^t) + \partial G(X^t)$, the stationarity residual converges to zero. This proves (ii).

Step 3: Convergence of the objective values. Lower semicontinuity gives $\Phi(\bar{X}) \le \liminf_t \Phi(X^t)$. Conversely, the subgradient inequality at X^t gives $\Phi(X^t) - \Phi(\bar{X}) \le \langle p^t, X^t - \bar{X}\rangle$. Since $X^t \to \bar{X}$ and $p^t \to -\mathcal{A}^\mathsf{T}\bar{y}$, the right-hand side converges to zero. Hence, $\limsup_t \Phi(X^t) \le \Phi(\bar{X})$, and therefore $\Phi(X^t) \to \Phi(\bar{X}) = \Phi^\star$. This proves (iii).

Step 4: Best-iterate KKT residual. Define $R^{t+1} := P(z^t - z^{t+1}) + \xi^{t+1} \in \mathcal{T}(z^{t+1})$, and let $p_{\max} := \|P\|_2$ and $m_{\max} := \|M\|_2$. Since $P^2 \preceq p_{\max} P$, we have

$$\left\|P(z^t - z^{t+1})\right\|^2 = \left\langle z^t - z^{t+1}, P^2(z^t - z^{t+1})\right\rangle \le p_{\max} \left\|z^t - z^{t+1}\right\|_P^2 = p_{\max} D_t^2.$$

Moreover, since $\xi^{t+1} = (\varepsilon^{t+1}, 0)$ and $M \preceq m_{\max}\mathrm{I}$, we obtain

$$\left\|\xi^{t+1}\right\|^2 = \left\|\varepsilon^{t+1}\right\|^2 \le m_{\max} \left\|\varepsilon^{t+1}\right\|_{M^{-1}}^2 = m_{\max}\delta_{t+1}^2.$$

Applying Minkowski's inequality to the product norm over $t = 0, \ldots, K-1$ gives

$$\begin{aligned}
\left(\sum_{t=0}^{K-1} \left\|R^{t+1}\right\|^2\right)^{1/2} &\le \left(\sum_{t=0}^{K-1} \left\|P(z^t - z^{t+1})\right\|^2\right)^{1/2} + \left(\sum_{t=0}^{K-1} \left\|\xi^{t+1}\right\|^2\right)^{1/2} \\
&\le \sqrt{p_{\max} \sum_{t=0}^{K-1} D_t^2} + \sqrt{m_{\max} \sum_{t=0}^{K-1} \delta_{t+1}^2} \\
&\le \sqrt{p_{\max} A} + \sqrt{m_{\max}}\Delta_2 =: C_{\mathrm{abs}}.
\end{aligned}$$

Because $R^{t+1} \in \mathcal{T}(z^{t+1})$, it follows that

$$\begin{aligned}\min_{0\le t<K} \operatorname{dist}\big(0, \mathcal{T}(z^{t+1})\big) &\le \min_{0\le t<K} \left\|R^{t+1}\right\| \\ &\le \frac{1}{\sqrt{K}} \left(\sum_{t=0}^{K-1} \left\|R^{t+1}\right\|^2\right)^{1/2} \\ &\le \frac{C_{\mathrm{abs}}}{\sqrt{K}}.\end{aligned}$$

This proves (iv) and completes the proof. □

B.3 Relative Rule under Convexity

Theorem 2 (Relative rule)**.** *Suppose Assumption 2 holds and $S \succeq B = \operatorname{diag}(B_1, \dots, B_n) \succ 0$. Let every local output satisfy (9) with $0 < \theta_i < 1$, and update each ν_i^t according to (10). Denote*

$$\theta := \max_i \theta_i, \qquad \nu^t := (\nu_1^t; \dots; \nu_n^t).$$

For any $z^\star = (X^\star, y^\star) \in \mathcal{K}$, consider

$$\mathcal{E}_t(z^\star) := \left\|z^t - z^\star\right\|_P^2 + \left\|\nu^t - X^\star\right\|_B^2.$$

Then

$$\mathcal{E}_{t+1}(z^\star) + (1-\theta^2) D_t^2 \le \mathcal{E}_t(z^\star).$$

Consequently, the following statements hold:

(i) The sequences $\{z^t\}$ and $\{\nu^t\}$ are bounded.

(ii) The local residuals and the KKT residual components satisfy

$$\begin{aligned}\varepsilon^t \to 0, \qquad \mathcal{A}X^t \to 0,\\ \operatorname{dist}\big(0, \nabla F(X^t) + \partial G(X^t) + \mathcal{A}^{\mathsf{T}} y^t\big) \to 0,\\ \operatorname{dist}(0, \mathcal{T}(z^t)) \to 0.\end{aligned}$$

(iii) Every accumulation point of $\{z^t\}$ belongs to $\mathcal{K}$, $\operatorname{dist}_P(z^t, \mathcal{K}) \to 0$, and $\Phi(X^t) \to \Phi^\star$.

(iv) With $c_{\mathrm{rel}} := \sqrt{\|P\|_2} + \theta\sqrt{\|B\|_2}$, for every $K \ge 1$,

$$\min_{0\le t<K} \operatorname{dist}(0, \mathcal{T}(z^{t+1})) \le \frac{c_{\mathrm{rel}}\sqrt{\mathcal{E}_0(z^\star)}}{\sqrt{(1-\theta^2)K}}.$$

Proof. We first derive the energy estimate used for all four conclusions. Summing the nodewise inequalities (9) and using $|\sum_i a_i| \le \sum_i |a_i|$ give

$$\begin{aligned}2\left|\left\langle \nu^t - X^{t+1}, \varepsilon^{t+1}\right\rangle\right| + \left\|\varepsilon^{t+1}\right\|_{B^{-1}}^2 &\le \sum_{i=1}^n \left(2\left|\left\langle \nu_i^t - x_i^{t+1}, \varepsilon_i^{t+1}\right\rangle\right| + \left\|\varepsilon_i^{t+1}\right\|_{B_i^{-1}}^2\right) \\ &\le \sum_{i=1}^n \theta_i^2 \left\|x_i^{t+1} - x_i^t\right\|_{B_i}^2 \le \theta^2 \left\|X^{t+1} - X^t\right\|_B^2 \le \theta^2 D_t^2,\end{aligned}$$

where the last inequality follows from $S \succeq B$ and $D_t^2 = \left\|X^{t+1} - X^t\right\|_S^2 + \rho^{-1}\left\|y^{t+1} - y^t\right\|^2$. The tracker update (10) gives

$$\left\|\nu^t - X^\star\right\|_B^2 - \left\|\nu^{t+1} - X^\star\right\|_B^2 = 2\left\langle \varepsilon^{t+1}, \nu^t - X^\star\right\rangle - \left\|\varepsilon^{t+1}\right\|_{B^{-1}}^2.$$

Adding this identity to (18), with $p^{t+1} = v^{t+1}$ and $p^\star = -\mathcal{A}^{\mathsf{T}} y^\star$, yields

$$\mathcal{E}_t(z^\star) - \mathcal{E}_{t+1}(z^\star) = D_t^2 + 2\left\langle v^{t+1} - p^\star, X^{t+1} - X^\star\right\rangle + 2\left\langle \varepsilon^{t+1}, \nu^t - X^{t+1}\right\rangle - \left\|\varepsilon^{t+1}\right\|_{B^{-1}}^2.$$

Monotonicity of $\partial\Phi$ makes the second term nonnegative. Applying the aggregate relative-error bound to the last two terms proves the energy descent inequality in Theorem 2.

Step 1: Boundedness and square-summability. Since $S \succeq B \succ 0$, both P and B are positive definite. Hence, this descent inequality shows that $\{z^t\}$ and $\{\nu^t\}$ are bounded and

$$\sum_{t=0}^{\infty} D_t^2 \leq \frac{\mathcal{E}_0(z^\star)}{1-\theta^2} < \infty.$$

This proves (i). The square-summability estimate will also be used below.

Step 2: Convergence of the residuals. The square-summability of D_t gives $z^{t+1}-z^t \to 0$. Moreover, the aggregate relative-error bound gives $\left\|\varepsilon^{t+1}\right\|_{B^{-1}} \leq \theta D_t \to 0$. Since $B \succ 0$, this also implies $\varepsilon^{t+1} \to 0$. The multiplier update and (17) now imply

$$\mathcal{A}X^{t+1} = \rho^{-1}(y^{t+1}-y^t) \to 0,$$
$$v^{t+1} + \mathcal{A}^{\mathsf{T}} y^{t+1} = \varepsilon^{t+1} - S(X^{t+1}-X^t) \to 0.$$

Finally, $\left\|R^{t+1}\right\| \leq \left(\sqrt{\|P\|_2} + \theta\sqrt{\|B\|_2}\right) D_t \to 0$, and hence $\operatorname{dist}(0, \mathcal{T}(z^{t+1})) \to 0$. These estimates prove (ii).

Step 3: KKT cluster points and objective convergence. The estimates in (ii) and the closedness of $\operatorname{gph}(\partial\Phi)$ show that every accumulation point of $\{z^t\}$ belongs to $\mathcal{K}$. Since $\{z^t\}$ is bounded and $P \succ 0$, it follows that $\operatorname{dist}_P(z^t, \mathcal{K}) \to 0$.

It remains to prove convergence of the objective values. Fix $z^\star = (X^\star, y^\star) \in \mathcal{K}$. The subgradient inequality at $X^\star$ gives $\Phi(X^t) \geq \Phi^\star - \langle y^\star, \mathcal{A}X^t\rangle$, and hence $\liminf_{t\to\infty} \Phi(X^t) \geq \Phi^\star$. On the other hand, the subgradient inequality at X^t gives

$$\Phi(X^t) - \Phi^\star \leq \left\langle v^t, X^t - X^\star \right\rangle = -\left\langle y^t, \mathcal{A}X^t \right\rangle + \left\langle v^t + \mathcal{A}^{\mathsf{T}} y^t, X^t - X^\star \right\rangle.$$

The right-hand side converges to zero because $\{X^t\}$ and $\{y^t\}$ are bounded, $\mathcal{A}X^t \to 0$, and $v^t + \mathcal{A}^{\mathsf{T}} y^t \to 0$. Thus, $\limsup_{t\to\infty} \Phi(X^t) \leq \Phi^\star$, and consequently $\Phi(X^t) \to \Phi^\star$. This proves (iii).

Step 4: Best-iterate KKT residual. Let $p_{\max} := \|P\|_2$ and $b_{\max} := \|B\|_2$. Using the definition of $R^{t+1} \in \mathcal{T}(z^{t+1})$,

$$\begin{aligned}\operatorname{dist}(0, \mathcal{T}(z^{t+1})) &\leq \left\|P(z^t - z^{t+1}) + \xi^{t+1}\right\| \\ &\leq \sqrt{p_{\max}} D_t + \sqrt{b_{\max}} \left\|\varepsilon^{t+1}\right\|_{B^{-1}} \\ &\leq c_{\mathrm{rel}} D_t.\end{aligned}$$

Combining this estimate with the bound on $\sum_t D_t^2$ and taking the smallest of the first K terms gives

$$\min_{0\leq t<K} \operatorname{dist}(0, \mathcal{T}(z^{t+1})) \leq c_{\mathrm{rel}} \left(\frac{1}{K}\sum_{t=0}^{K-1} D_t^2\right)^{1/2} \leq \frac{c_{\mathrm{rel}}\sqrt{\mathcal{E}_0(z^\star)}}{\sqrt{(1-\theta^2)K}}.$$

This proves (iv) and completes the proof. □

B.4 Ergodic Primal Rates

Corollary 1 (Ergodic primal rates)**.** *Suppose that the assumptions of either Theorem 1 or Theorem 2 hold. Then there exist constants $C_{\mathrm{feas}} > 0$ and $C_{\mathrm{obj}} > 0$, independent of K, such that*

$$\left\|\mathcal{A}\widehat{X}^K\right\| \leq \frac{C_{\mathrm{feas}}}{K}, \qquad \left|\Phi(\widehat{X}^K) - \Phi^\star\right| \leq \frac{C_{\mathrm{obj}}}{K}, \qquad K \geq 1.$$

Proof. For $K \geq 1$, define $\widehat{X}^K := K^{-1}\sum_{t=0}^{K-1} X^{t+1}$. We first derive a one-step primal–dual gap estimate and then treat the two stopping rules separately.

Step 1: One-step primal–dual gap estimate. Fix any $z^\star = (X^\star, y^\star) \in \mathcal{K}$ and an arbitrary vector y. Let $v^{t+1} \in \partial\Phi(X^{t+1})$ be the subgradient appearing in (17). Convexity gives $\Phi(X^{t+1}) - \Phi^\star \le \langle v^{t+1}, X^{t+1} - X^\star \rangle$, while (17) gives $v^{t+1} = \varepsilon^{t+1} - S(X^{t+1} - X^t) - \mathcal{A}^\mathsf{T} y^{t+1}$. Because $\mathcal{A}X^\star = 0$, it follows that

$$\begin{aligned}&\Phi(X^{t+1}) - \Phi^\star + \langle y, \mathcal{A}X^{t+1} \rangle \\ &\le \langle \varepsilon^{t+1}, X^{t+1} - X^\star \rangle - \langle S(X^{t+1} - X^t), X^{t+1} - X^\star \rangle + \langle y - y^{t+1}, \mathcal{A}X^{t+1} \rangle .\end{aligned}$$

The three-point identity gives

$$\begin{aligned}-2\langle S(X^{t+1} - X^t), X^{t+1} - X^\star \rangle = \|X^t - X^\star\|_S^2 - \|X^{t+1} - X^\star\|_S^2 \\ - \|X^{t+1} - X^t\|_S^2 .\end{aligned}$$

Moreover, the multiplier update $y^{t+1} - y^t = \rho\mathcal{A}X^{t+1}$ implies

$$\begin{aligned}2\langle y - y^{t+1}, \mathcal{A}X^{t+1} \rangle = \rho^{-1}\|y^t - y\|^2 - \rho^{-1}\|y^{t+1} - y\|^2 \\ - \rho^{-1}\|y^{t+1} - y^t\|^2 .\end{aligned}$$

Combining these relations and recalling that $D_t^2 = \|X^{t+1} - X^t\|_S^2 + \rho^{-1}\|y^{t+1} - y^t\|^2$ give

$$\begin{aligned}&2\left(\Phi(X^{t+1}) - \Phi^\star + \langle y, \mathcal{A}X^{t+1} \rangle\right) \\ &\le \|X^t - X^\star\|_S^2 - \|X^{t+1} - X^\star\|_S^2 + \rho^{-1}\|y^t - y\|^2 - \rho^{-1}\|y^{t+1} - y\|^2 \\ &\quad - D_t^2 + 2\langle \varepsilon^{t+1}, X^{t+1} - X^\star \rangle .\end{aligned}$$

Step 2: Absolute stopping rule. Under the assumptions of Theorem 1, the sequence $\{X^t\}$ is bounded. Hence, $B_X := \sup_{t\ge 0}\|X^{t+1} - X^\star\|_M < \infty$. By the weighted Cauchy–Schwarz inequality,

$$\begin{aligned}|\langle \varepsilon^{t+1}, X^{t+1} - X^\star \rangle| &\le \|\varepsilon^{t+1}\|_{M^{-1}} \|X^{t+1} - X^\star\|_M \\ &\le B_X \delta_{t+1}.\end{aligned}$$

Summing the one-step estimate from $t = 0$ to $K - 1$ gives

$$\begin{aligned}&2\sum_{t=0}^{K-1}\left(\Phi(X^{t+1}) - \Phi^\star + \langle y, \mathcal{A}X^{t+1} \rangle\right) \\ &\le \|X^0 - X^\star\|_S^2 - \|X^K - X^\star\|_S^2 + \rho^{-1}\|y^0 - y\|^2 - \rho^{-1}\|y^K - y\|^2 \\ &\quad - \sum_{t=0}^{K-1} D_t^2 + 2B_X \sum_{t=0}^{K-1} \delta_{t+1}.\end{aligned}$$

Dropping the nonpositive terms and using $\sum_{t=0}^{\infty} \delta_{t+1} < \infty$, we obtain

$$2\sum_{t=0}^{K-1}\left(\Phi(X^{t+1}) - \Phi^\star + \langle y, \mathcal{A}X^{t+1} \rangle\right) \le \|X^0 - X^\star\|_S^2 + \rho^{-1}\|y^0 - y\|^2 + 2B_X \sum_{t=0}^{\infty} \delta_{t+1}.$$

Convexity of Φ and linearity of $\mathcal{A}$ imply

$$\Phi(\widehat{X}^K) \le \frac{1}{K}\sum_{t=0}^{K-1} \Phi(X^{t+1}), \qquad \mathcal{A}\widehat{X}^K = \frac{1}{K}\sum_{t=0}^{K-1} \mathcal{A}X^{t+1}.$$

Dividing the preceding estimate by $2K$ therefore yields

$$\Phi(\widehat{X}^K) - \Phi^\star + \left\langle y, \mathcal{A}\widehat{X}^K \right\rangle \le \frac{\|X^0 - X^\star\|_S^2 + \rho^{-1}\|y^0 - y\|^2 + 2B_X \sum_{t=0}^{\infty} \delta_{t+1}}{2K}.$$

Step 3: Relative stopping rule. Under the assumptions of Theorem 2, the tracker update gives

$$\|\nu^t - X^\star\|_B^2 - \|\nu^{t+1} - X^\star\|_B^2 = 2\langle \varepsilon^{t+1}, \nu^t - X^\star \rangle - \|\varepsilon^{t+1}\|_{B^{-1}}^2 .$$

Decomposing the error term as

$$2\left\langle \varepsilon^{t+1}, X^{t+1}-X^\star\right\rangle = 2\left\langle \varepsilon^{t+1}, \nu^t - X^\star\right\rangle + 2\left\langle \varepsilon^{t+1}, X^{t+1}-\nu^t\right\rangle,$$

and using the preceding identity, we obtain

$$2\left\langle \varepsilon^{t+1}, X^{t+1}-X^\star\right\rangle \le \left\|\nu^t - X^\star\right\|_B^2 - \left\|\nu^{t+1}-X^\star\right\|_B^2 + \left\|\varepsilon^{t+1}\right\|_{B^{-1}}^2 + 2\left|\left\langle \nu^t - X^{t+1}, \varepsilon^{t+1}\right\rangle\right|.$$

The aggregate relative-error criterion and $B \preceq S$ give

$$\begin{aligned} 2\left\langle \varepsilon^{t+1}, X^{t+1}-X^\star\right\rangle &\le \left\|\nu^t - X^\star\right\|_B^2 - \left\|\nu^{t+1}-X^\star\right\|_B^2 + \theta^2\left\|X^{t+1}-X^t\right\|_B^2 \\ &\le \left\|\nu^t - X^\star\right\|_B^2 - \left\|\nu^{t+1}-X^\star\right\|_B^2 + \theta^2 D_t^2. \end{aligned}$$

Substituting this estimate into the one-step primal–dual gap inequality gives

$$\begin{aligned} &2\left(\Phi(X^{t+1}) - \Phi^\star + \left\langle y, \mathcal{A}X^{t+1}\right\rangle\right) \\ &\le \left\|X^t - X^\star\right\|_S^2 - \left\|X^{t+1}-X^\star\right\|_S^2 + \rho^{-1}\left\|y^t - y\right\|^2 - \rho^{-1}\left\|y^{t+1}-y\right\|^2 \\ &\quad + \left\|\nu^t - X^\star\right\|_B^2 - \left\|\nu^{t+1}-X^\star\right\|_B^2 - (1-\theta^2)D_t^2. \end{aligned}$$

Summing from $t=0$ to $K-1$, dropping the nonpositive terminal and increment terms, and dividing by $2K$ yield

$$\Phi(\widehat{X}^K) - \Phi^\star + \left\langle y, \mathcal{A}\widehat{X}^K\right\rangle \le \frac{\left\|X^0 - X^\star\right\|_S^2 + \rho^{-1}\left\|y^0 - y\right\|^2 + \left\|\nu^0 - X^\star\right\|_B^2}{2K}.$$

Step 4: Feasibility and objective-error bounds. For either stopping rule, summing the multiplier updates gives $y^K - y^0 = \rho\sum_{t=0}^{K-1}\mathcal{A}X^{t+1} = \rho K\mathcal{A}\widehat{X}^K$, and hence $\mathcal{A}\widehat{X}^K = (y^K - y^0)/(\rho K)$. The boundedness of $\{y^t\}$ therefore implies that there exists a constant $C_{\mathrm{feas}} > 0$, independent of K, such that

$$\left\|\mathcal{A}\widehat{X}^K\right\| \le \frac{C_{\mathrm{feas}}}{K}.$$

Taking $y = 0$ in the corresponding primal–dual gap estimate gives a constant $C_+ > 0$, independent of K, such that $\Phi(\widehat{X}^K) - \Phi^\star \le C_+/K$. On the other hand, the KKT condition $-\mathcal{A}^{\mathsf{T}}y^\star \in \partial\Phi(X^\star)$ and convexity of Φ give

$$\Phi(\widehat{X}^K) - \Phi^\star \ge -\left\langle y^\star, \mathcal{A}\widehat{X}^K\right\rangle \ge -\left\|y^\star\right\|\left\|\mathcal{A}\widehat{X}^K\right\| \ge -\frac{\left\|y^\star\right\| C_{\mathrm{feas}}}{K}.$$

Let $C_{\mathrm{obj}} := \max\left\{C_+,\ \left\|y^\star\right\| C_{\mathrm{feas}}\right\}$. Then, we have

$$\left|\Phi(\widehat{X}^K) - \Phi^\star\right| \le \frac{C_{\mathrm{obj}}}{K}.$$

This completes the proof. □

B.5 A Common Exact-Step Estimate

The linear convergence proofs compare each inexact update with the exact multiplier step starting from the same iterate. The following lemma collects the estimates used for both stopping rules.

Lemma 3 (Exact-step estimate)**.** *Let $\bar{X}^{t+1}$ be the exact minimizer of (5), and define $\bar{y}^{t+1} := y^t + \rho\mathcal{A}\bar{X}^{t+1}$ and $\bar{z}^{t+1} := (\bar{X}^{t+1}, \bar{y}^{t+1})$. Then*

(i) The exact step satisfies $P(z^t - \bar{z}^{t+1}) \in \mathcal{T}(\bar{z}^{t+1})$, and the inexact and exact updates obey

$$\left\|z^{t+1} - \bar{z}^{t+1}\right\|_P = \left\|X^{t+1} - \bar{X}^{t+1}\right\|_M \le \delta_{t+1}. \tag{19}$$

(ii) Let $d_t := \mathrm{dist}_P(z^t, \mathcal{K})$, $\bar{d}_{t+1} := \mathrm{dist}_P(\bar{z}^{t+1}, \mathcal{K})$, and $s_t := \left\|z^t - \bar{z}^{t+1}\right\|_P$. Then $\bar{d}_{t+1}^2 + s_t^2 \le d_t^2$.

(iii) If $P \succ 0$ and (12) holds at $\bar{z}^{t+1}$ with constant κ_P, then

$$\bar{d}_{t+1} \le q_P d_t, \qquad q_P := \frac{\kappa_P}{\sqrt{1+\kappa_P^2}} < 1.$$

Proof. We prove the three statements in order.

Step 1: Exact-step inclusion and perturbation bound. The exact optimality condition of (5), together with the definition of $\bar{y}^{t+1}$, gives $P(z^t - \bar{z}^{t+1}) \in \mathcal{T}(\bar{z}^{t+1})$. Moreover, $y^{t+1} - \bar{y}^{t+1} = \rho\mathcal{A}(X^{t+1} - \bar{X}^{t+1})$. Since $M = S + \rho\mathcal{A}^{\mathsf{T}}\mathcal{A}$, Lemma 1 gives

$$\left\|z^{t+1} - \bar{z}^{t+1}\right\|_P^2 = \left\|X^{t+1} - \bar{X}^{t+1}\right\|_S^2 + \rho\left\|\mathcal{A}(X^{t+1} - \bar{X}^{t+1})\right\|^2 = \left\|X^{t+1} - \bar{X}^{t+1}\right\|_M^2 \le \delta_{t+1}^2.$$

This proves (i), including the case in which S is singular.

Step 2: Fejér estimate for the exact step. For any $z^\star \in \mathcal{K}$, monotonicity of $\mathcal{T}$, $0 \in \mathcal{T}(z^\star)$, and the exact-step inclusion imply

$$\left\langle \bar{z}^{t+1} - z^\star, P(z^t - \bar{z}^{t+1})\right\rangle \ge 0.$$

Expanding $\|z^t - z^\star\|_P^2$ around $\bar{z}^{t+1}$ therefore gives

$$\left\|z^t - z^\star\right\|_P^2 \ge \left\|\bar{z}^{t+1} - z^\star\right\|_P^2 + \left\|z^t - \bar{z}^{t+1}\right\|_P^2.$$

Taking the infimum over $z^\star \in \mathcal{K}$ proves $\bar{d}_{t+1}^2 + s_t^2 \le d_t^2$ and hence (ii).

Step 3: Contraction under the error bound. If $P \succ 0$ and (12) holds at $\bar{z}^{t+1}$, then the exact-step inclusion gives

$$\bar{d}_{t+1} \le \kappa_P \operatorname{dist}_{P^{-1}}(0, \mathcal{T}(\bar{z}^{t+1})) \le \kappa_P \left\|P(z^t - \bar{z}^{t+1})\right\|_{P^{-1}} = \kappa_P s_t.$$

Combining this bound with (ii) yields $(1+\kappa_P^{-2})\bar{d}_{t+1}^2 \le d_t^2$, which is equivalent to the contraction in (iii). This is the standard proximal-point contraction under metric subregularity (Leventhal, 2009). □

B.6 Absolute Rule under the KKT Error Bound

Theorem 3 (Absolute rule)**.** *Suppose Assumption 2 holds and the aggregate residual satisfies $\delta_{t+1} \le C\beta^{t+1}$ for some $C > 0$ and $0 < \beta < 1$. Let $\bar{z} \in \mathcal{K}$ be the limit given by Theorem 1, and suppose Assumption 3 holds at $\bar{z}$. Then there exists $t_0 \ge 0$ such that*

$$d_{t+1} \le q_a d_t + C\beta^{t+1}, \qquad t \ge t_0,$$

where $d_t := \operatorname{dist}_P(z^t, \mathcal{K})$, $a := \kappa\|P\|_2$, and $q_a := a/\sqrt{1+a^2} < 1$. Let $r := \max\{q_a, \beta\}$, and define b_t as r^t when $q_a \ne \beta$ or tr^t otherwise. Then, for any fixed $\widehat{q} \in (r, 1)$,

$$d_t + \left\|z^t - \bar{z}\right\| + \left\|\mathcal{A}X^t\right\| + \left\|R^t\right\| = O(b_t) = O(\widehat{q}^t).$$

Thus, z^t converges R-linearly to $\bar{z}$, while $\mathcal{A}X^t$ and R^t converge R-linearly to zero.

Proof. We prove the result in three steps.

Step 1: Distance recursion. The geometric residual bound is summable, so Theorem 1 gives $z^t \to \bar{z}$. Moreover, Lemma 3(i) gives

$$\left\|X^{t+1} - \bar{X}^{t+1}\right\|_M \le \delta_{t+1} \to 0.$$

Since $y^{t+1} - \bar{y}^{t+1} = \rho\mathcal{A}(X^{t+1} - \bar{X}^{t+1})$, we also have $\bar{z}^{t+1} \to \bar{z}$. Hence, for all sufficiently large t, $\bar{z}^{t+1}$ lies in the neighborhood where (11) holds.

Define $\bar{d}_{t+1} := \operatorname{dist}_P(\bar{z}^{t+1}, \mathcal{K})$, $s_t := \left\|z^t - \bar{z}^{t+1}\right\|_P$, and $p := \|P\|_2$. Lemma 3(ii) gives

$$\bar{d}_{t+1}^2 + s_t^2 \le d_t^2.$$

Since $P^2 \preceq pP$, the metric-subregularity condition and the exact inclusion $P(z^t - \bar{z}^{t+1}) \in \mathcal{T}(\bar{z}^{t+1})$ yield

$$\begin{aligned}\bar{d}_{t+1} &\le \sqrt{p}\ \mathrm{dist}(\bar{z}^{t+1}, \mathcal{K}) \\ &\le \kappa\sqrt{p}\ \mathrm{dist}(0, \mathcal{T}(\bar{z}^{t+1})) \\ &\le \kappa\sqrt{p}\ \left\|P(z^t - \bar{z}^{t+1})\right\| \\ &\le \kappa p\, s_t = a s_t.\end{aligned}$$

Combining the last two inequalities gives

$$\bar{d}_{t+1} \le \frac{a}{\sqrt{1+a^2}} d_t = q_a d_t.$$

The triangle inequality and Lemma 3(i) therefore imply

$$d_{t+1} \le \bar{d}_{t+1} + \left\|z^{t+1} - \bar{z}^{t+1}\right\|_P \le q_a d_t + \delta_{t+1} \le q_a d_t + C\beta^{t+1},$$

which proves the distance recursion in Theorem 3.

Step 2: Rate of the distance to the KKT set. Let t_0 be sufficiently large that this recursion holds for every $t \ge t_0$. Iterating this recursion gives

$$d_t \le q_a^{t-t_0} d_{t_0} + C \sum_{j=t_0}^{t-1} q_a^{t-1-j} \beta^{j+1}.$$

If $q_a \ne \beta$, the geometric convolution is $O(r^t)$, where $r = \max\{q_a, \beta\}$. If $q_a = \beta = r$, it is $O(tr^t)$. Thus, in both cases, $d_t = O(b_t)$.

Step 3: Rates of the full sequence and residuals. Define $\upsilon_t := \left\|z^{t+1} - z^t\right\|_P$. Taking the infimum over $z^\star \in \mathcal{K}$ in the quasi-Fejér estimate gives

$$\sqrt{d_{t+1}^2 + \upsilon_t^2} \le d_t + 2\delta_{t+1}.$$

Since $d_t = O(b_t)$ and $\delta_{t+1} = O(b_t)$, it follows that $\upsilon_t = O(b_t)$. The multiplier update gives $\left\|\mathcal{A}X^{t+1}\right\| = \rho^{-1}\left\|y^{t+1} - y^t\right\| \le \rho^{-1/2}\upsilon_t = O(b_t)$. Moreover, using $M = S + \rho\mathcal{A}^{\mathsf{T}}\mathcal{A}$ and the multiplier updates at two consecutive iterations,

$$\begin{aligned}\left\|X^{t+1} - X^t\right\|_M &= \left(\left\|X^{t+1} - X^t\right\|_S^2 + \rho^{-1}\left\|(y^{t+1} - y^t) - (y^t - y^{t-1})\right\|^2\right)^{1/2} \\ &\le \upsilon_t + \upsilon_{t-1} = O(b_t).\end{aligned}$$

Since $M \succ 0$ and $\left\|y^{t+1} - y^t\right\| \le \sqrt{\rho}\,\upsilon_t$, we obtain $\left\|z^{t+1} - z^t\right\| = O(b_t)$. The sequence $\{b_t\}$ is summable, and Theorem 1 already gives $z^t \to \bar{z}$. Therefore,

$$\left\|z^t - \bar{z}\right\| \le \sum_{j=t}^{\infty} \left\|z^{j+1} - z^j\right\| = O(b_t).$$

Finally, letting $m := \|M\|_2$ and using $P^2 \preceq pP$,

$$\begin{aligned}\left\|R^{t+1}\right\| &\le \left\|P(z^t - z^{t+1})\right\| + \left\|\xi^{t+1}\right\| \\ &\le \sqrt{p}\,\upsilon_t + \sqrt{m}\,\delta_{t+1} = O(b_t).\end{aligned}$$

Together with $\left\|\mathcal{A}X^{t+1}\right\| = O(b_t)$, this proves the rate estimates in Theorem 3. Since $b_t = O(\widehat{q}^t)$ for every $r < \widehat{q} < 1$, all the stated quantities converge R-linearly. □

B.7 Relative Rule under the KKT Error Bound

Theorem 4 (Relative rule)**.** *Suppose the assumptions of Theorem 2 hold and (12) holds throughout $\mathcal{U}_r^P$ for some $r > 0$ with constant $\kappa_{P,r}$. If $\theta := \max_{1\le i\le n}\theta_i < (1 + \kappa_{P,r} + \sqrt{1+\kappa_{P,r}^2})^{-1}$, then $\chi_{\mathrm{rel}} := (\kappa_{P,r} + \frac{\theta}{1-\theta})/\sqrt{1+\kappa_{P,r}^2} < 1$, and there exists $t_0 \ge 0$ such that*

$$d_{t+1} \le \chi_{\mathrm{rel}} d_t, \qquad t \ge t_0.$$

Consequently, d_t converges eventually Q-linearly. Moreover, there exists $\bar{z} \in \mathcal{K}$ such that

$$d_t \le \chi_{\mathrm{rel}}^{t-t_0} d_{t_0},$$
$$\left\|z^t - \bar{z}\right\| + \left\|\mathcal{A}X^t\right\| + \left\|R^t\right\| = O(\chi_{\mathrm{rel}}^{t-t_0}).$$

Thus, the full primal–dual sequence, the consensus violation, and the residual R^t converge R-linearly.

Proof. We prove the result in three steps.

Step 1: Difference between the inexact and exact steps. Theorem 2 gives $\mathrm{dist}_P(z^t, \mathcal{K}) \to 0$. Hence, there exists t_0 such that $z^t \in \mathcal{U}_r^P$ for every $t \ge t_0$. The nodewise test and $S \succeq B \succ 0$ give

$$\begin{aligned}\left\|\xi^{t+1}\right\|_{P^{-1}} = \left\|\varepsilon^{t+1}\right\|_{S^{-1}} &\le \left\|\varepsilon^{t+1}\right\|_{B^{-1}} \\ &\le \theta \left\|X^{t+1} - X^t\right\|_B \le \theta \left\|z^{t+1} - z^t\right\|_P.\end{aligned}$$

Let $h_t := \left\|z^{t+1} - \bar{z}^{t+1}\right\|_P$ and $s_t := \left\|z^t - \bar{z}^{t+1}\right\|_P$. Monotonicity of $\mathcal{T}$ applied to the exact and inexact inclusions gives

$$h_t^2 \le \left\langle z^{t+1} - \bar{z}^{t+1}, \xi^{t+1} \right\rangle \le h_t \left\|\xi^{t+1}\right\|_{P^{-1}}.$$

Consequently, $h_t \le \left\|\xi^{t+1}\right\|_{P^{-1}} \le \theta \left\|z^{t+1} - z^t\right\|_P \le \theta(s_t + h_t)$, and hence $h_t \le c_\theta s_t$, where $c_\theta := \theta/(1-\theta)$.

Step 2: Contraction of the distance to the KKT set. If $s_t = 0$, then the exact inclusion gives $z^t = \bar{z}^{t+1} \in \mathcal{K}$, while $h_t = 0$ gives $z^{t+1} = z^t$; the contraction therefore holds trivially. Suppose now that $s_t > 0$ and define $u_t := \mathrm{dist}_P(\bar{z}^{t+1}, \mathcal{K})/s_t$. Lemma 3(ii) and the triangle inequality imply

$$\frac{\mathrm{dist}_P(z^{t+1}, \mathcal{K})}{\mathrm{dist}_P(z^t, \mathcal{K})} \le \frac{u_t + c_\theta}{\sqrt{1 + u_t^2}}.$$

For $t \ge t_0$, Lemma 3(ii) gives $\mathrm{dist}_P(\bar{z}^{t+1}, \mathcal{K}) \le \mathrm{dist}_P(z^t, \mathcal{K}) \le r$, so $\bar{z}^{t+1} \in \mathcal{U}_r^P$. The matched error bound therefore gives $0 \le u_t \le \kappa_{P,r}$. Let $f(u) := (u + c_\theta)/\sqrt{1+u^2}$. Since $f'(u) = (1 - c_\theta u)/(1+u^2)^{3/2}$, the threshold condition implies $c_\theta \kappa_{P,r} < 1$, so f is increasing on $[0, \kappa_{P,r}]$. It also gives $f(\kappa_{P,r}) = (\kappa_{P,r} + c_\theta)/\sqrt{1 + \kappa_{P,r}^2} = \chi_{\mathrm{rel}} < 1$. Therefore,

$$\frac{\mathrm{dist}_P(z^{t+1}, \mathcal{K})}{\mathrm{dist}_P(z^t, \mathcal{K})} \le f(u_t) \le \chi_{\mathrm{rel}},$$

which proves the contraction asserted in Theorem 4.

Step 3: Rates of the full sequence and residuals. Since $s_t \le \mathrm{dist}_P(z^t, \mathcal{K})$,

$$\left\|z^{t+1} - z^t\right\|_P \le s_t + h_t \le (1 + c_\theta)\,\mathrm{dist}_P(z^t, \mathcal{K}),$$

so the primal–dual increments are geometrically summable. Hence, z^t converges R-linearly to some $\bar{z} \in \mathcal{K}$. Finally,

$$\mathcal{A}X^{t+1} = \rho^{-1}(y^{t+1} - y^t), \qquad R^{t+1} = P(z^t - z^{t+1}) + \xi^{t+1},$$

and the relative residual bound gives the remaining R-linear estimates. □

C Additional Experimental Details

C.1 Hyperparameter Settings

We first specify the parameterizations used for the three single-loop baselines. The selected hyperparameters for all compared methods under the four problem settings are summarized in Table 1. Recall from (3) that $F(X) = \sum_{i=1}^n f_i(x_i)$ and $G(X) = \sum_{i=1}^n g_i(x_i)$. Let W be the mixing matrix, $Z = \mathrm{I}_n - W$, $\mathcal{W} = W \otimes \mathrm{I}_d$, and $\mathcal{Z} = Z \otimes \mathrm{I}_d$. We denote by L_i the Lipschitz constant of ∇f_i, set $L_{\max} = \max_i L_i$, and define $\Lambda = \mathrm{diag}(\alpha_1, \ldots, \alpha_n) \otimes \mathrm{I}_d$. The metric proximal mapping is

$$\mathrm{prox}_G^\Lambda(V) := \arg\min_X \left\{ G(X) + \frac{1}{2}\|X - V\|_{\Lambda^{-1}}^2 \right\}.$$

Table 1: Selected hyperparameters for the four reported problem settings.

Method	Selected hyperparameters
	Logistic regression: $\lambda = 10^{-1}$
IDEAL	$s_\rho = 2,\ \eta_0 = 1,\ \beta = 0.5$
D-ripALM	$s_\rho = 4,\ \theta = 0.99,\ \tau = 10^{-2}$
CFI-spALM-Abs	$s_\rho = 0.5,\ \gamma_M/\rho = 10^{-2},\ \eta_0 = 1,\ \beta = 0.9$
CFI-spALM-Rel	$s_\rho = 1,\ \gamma_M/\rho = 0.1,\ \theta = 0.9$
	Logistic regression: $\lambda = 10^{-3}$
IDEAL	$s_\rho = 2,\ \eta_0 = 100,\ \beta = 0.3$
D-ripALM	$s_\rho = 4,\ \theta = 0.3,\ \tau = 10^{-4}$
CFI-spALM-Abs	$s_\rho = 0.02,\ \gamma_M/\rho = 10^{-2},\ \eta_0 = 0.1,\ \beta = 0.9$
CFI-spALM-Rel	$s_\rho = 0.08,\ \gamma_M/\rho = 0.1,\ \theta = 0.9$
	Decentralized LASSO: $\lambda_c = 10^{-1}$
PG-EXTRA	$s_\alpha = 1.5$
NIDS	$s_\alpha = 2.3$
D-iPGM	$s_\alpha = 2.1,\ s_\beta = 0.1$
D-ripALM	$s_\rho = 4,\ \theta = 0.9,\ \tau = 0.1$
CFI-spALM-Abs	$s_\rho = 0.05,\ \gamma_M/\rho = 0.1,\ \eta_0 = 1,\ \beta = 0.98$
CFI-spALM-Rel	$s_\rho = 0.05,\ \gamma_M/\rho = 0.3,\ \theta = 0.99$
	Decentralized LASSO: $\lambda_c = 10^{-2}$
PG-EXTRA	$s_\alpha = 1.3$
NIDS	$s_\alpha = 2.1$
D-iPGM	$s_\alpha = 2,\ s_\beta = 0.1$
D-ripALM	$s_\rho = 4,\ \theta = 0.3,\ \tau = 0.1$
CFI-spALM-Abs	$s_\rho = 10^{-3},\ \gamma_M/\rho = 10^{-2},\ \eta_0 = 1,\ \beta = 0.999$
CFI-spALM-Rel	$s_\rho = 5 \times 10^{-3},\ \gamma_M/\rho = 10^{-2},\ \theta = 0.99$

PG-EXTRA (Shi et al., 2015b). For PG-EXTRA, let $\widetilde{\mathcal{W}} = (\mathrm{I}_{nd} + \mathcal{W})/2$ and parameterize its scalar stepsize as

$$\alpha = \frac{s_\alpha}{L_{\max}},$$

where s_α is the dimensionless stepsize scale used in our experiments. Starting from

$$U^0 = \mathcal{W}X^0 - \alpha\nabla F(X^0), \qquad X^1 = \mathrm{prox}_{\alpha G}(U^0),$$

PG-EXTRA performs

$$U^{k+1} = U^k - X^{k+1} + \widetilde{\mathcal{W}}(2X^{k+1} - X^k) - \alpha\big(\nabla F(X^{k+1}) - \nabla F(X^k)\big),$$
$$X^{k+2} = \mathrm{prox}_{\alpha G}(U^{k+1}).$$

NIDS (Li et al., 2019). NIDS uses node-dependent stepsizes

$$\alpha_i = \frac{s_\alpha}{L_i}, \qquad \Lambda = \mathrm{diag}(\alpha_1, \ldots, \alpha_n) \otimes \mathrm{I}_d,$$

where s_α is again a dimensionless stepsize scale. In our implementation,

$$c = \frac{1}{(1 - \lambda_{\min}(W)) \max_i \alpha_i}, \qquad \widetilde{\mathcal{W}}_{\mathrm{NIDS}} = \mathrm{I}_{nd} - c\Lambda\mathcal{Z}.$$

Starting from

$$U^0 = X^0 - \Lambda\nabla F(X^0), \qquad X^1 = \mathrm{prox}_G^{\Lambda}(U^0),$$

the NIDS iteration is

$$\overline{X}^{k+1} = 2X^{k+1} - X^k - \Lambda\big(\nabla F(X^{k+1}) - \nabla F(X^k)\big),$$

$$U^{k+1} = U^k - X^{k+1} + \widetilde{\mathcal{W}}_{\mathrm{NIDS}}\overline{X}^{k+1}, \qquad X^{k+2} = \mathrm{prox}_G^{\Lambda}(U^{k+1}).$$

Although PG-EXTRA and NIDS can have closely related algebraic forms under particular parameter choices, they communicate different quantities: PG-EXTRA mixes an extrapolated primal variable, whereas NIDS mixes the gradient-corrected variable $\overline{X}^{k+1}$.

D-iPGM (Guo et al., 2023). D-iPGM uses node-dependent primal stepsizes

$$\alpha_i = \frac{s_\alpha}{L_i}, \qquad \Lambda = \mathrm{diag}(\alpha_1, \ldots, \alpha_n) \otimes \mathrm{I}_d.$$

Its dual stepsize is normalized using the same problem-dependent reference scale as the penalty parameters of the ALM-type methods:

$$\beta = s_\beta \beta_{\mathrm{ref}}, \qquad \beta_{\mathrm{ref}} = \frac{L_{\max}}{\lambda_{\max}(Z)},$$

where s_α and s_β are dimensionless scale parameters. The D-iPGM iteration is

$$\widetilde{X}^{k+1} = \mathrm{prox}_G^{\Lambda}\big(X^k - \Lambda\big(\nabla F(X^k) - Y^k\big)\big),$$

$$Y^{k+1} = Y^k - \beta \mathcal{Z} \widetilde{X}^{k+1}, \qquad X^{k+1} = \widetilde{X}^{k+1} - \Lambda(Y^k - Y^{k+1}).$$

Thus, s_α controls the node-dependent primal stepsizes, whereas s_β controls the dual stepsize relative to the common spectral reference scale $L_{\max}/\lambda_{\max}(Z)$.

Based on the parameterizations above, Table 1 lists the final hyperparameter values used to produce the reported results.