Stable Outlier-Robust Signal Recovery over Networks: A Convex Analytic Approach Using Minimax Concave Loss

Maximilian H. V. Tillmann, *Graduate Student Member, IEEE*, Masahiro Yukawa, *Senior Member, IEEE*

***Abstract*—This paper presents a mathematically rigorous framework of remarkably-robust signal recovery over networks. The proposed framework is based on the *minimax concave (MC)* loss, which is a weakly convex function so that it attains (i) remarkable outlier-robustness and (ii) guarantee of convergence to a solution of the posed problem. We present a novel problem formulation which involves an auxiliary vector so that the formulation accommodates statistical properties of signal, noise, and outliers. We show the conditions to guarantee convexity of the local and global objectives. Via reformulation, the distributed triangularly preconditioned primal-dual algorithm is applied to the posed problem. The numerical examples show that our proposed formulation exhibits remarkable robustness under devastating outliers as well as outperforming the existing methods. Comparisons between the local and global convexity conditions are also presented.**

***Index Terms*—distributed optimization, outlier robustness, minimax concave penalty, proximity operator**

I. Introduction

ROBUST methods in the presence of outliers (or impulsive noise) have been studied in a variety of fields including statistics [2]–[5], control [6], optimization [7], machine learning, as well as signal processing [8]–[10]. Outliers happen frequently in wireless communication channels, radar/sonar systems, biomedical sensors, load prediction/monitoring systems, image/video sensors, and many others. Robust methods have been studied in distributed settings as well [11]–[15], where data are scattered over a network. The distributed optimization framework is useful particularly in solving large-scale problems where the data volume is too large to store at a single computer so that the data need to be stored and processed at each local node. There are two key aspects in the distributed signal recovery task: (i) the problem formulation to characterize the target signal as a minimizer of cost functions, and (ii) the algorithm to solve the formulated problem in a distributed fashion [16]–[22] (see also [23] for an extensive survey of distributed optimization algorithms). The major contributions of this work concern the former aspect basically.

Decentralized systems (having no central node) are considered in the present study, possessing advantages in many aspects: (i) no single point of failure, (ii) no need to collect data at a single node, (iii) no need for infrastructures, and (iv) suitability for edge computing. The significance of those advantages can be seen by considering the application of wireless sensor networks, for instance, where the whole system may break down due to potential node/link failures. In some applications, moreover, data are not allowed to be collected at a single point for privacy reasons.

As outliers are typically *sparse*, we consider a linear model where the observation vector at each node $i \in \mathcal{V} := \{1, 2, \ldots, N\}$ is given by

$$\boldsymbol{y}_i = \boldsymbol{A}_i \boldsymbol{x}_\star + \boldsymbol{\varepsilon}_{i\star} + \boldsymbol{o}_{i\diamond} \in \mathbb{R}^{m_i}, \tag{1}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{m_i \times n}$ is the system matrix, $\boldsymbol{x}_\star \in \mathbb{R}^n$ is the signal to be recovered obeying the i.i.d. zero-mean Gaussian distribution with variance $\sigma_{x\star}^2 > 0$, $\boldsymbol{\varepsilon}_{i\star} \in \mathbb{R}^{m_i}$ is the i.i.d. zero-mean Gaussian noise vector with variance $\sigma_{\varepsilon\star}^2 > 0$, and $\boldsymbol{o}_{i\diamond} \in \mathbb{R}^{m_i}$ is the sparse outlier vector. Here, the subscripts $\star$ and $\diamond$ symbolize Gaussian and sparse vectors, respectively. For simplicity, the Gaussian noise vectors are assumed to share the same distribution over the nodes. The same applies to the outlier vectors. The model in (1) has previously been studied in centralized (non-distributed) settings [24], [25], but it has not been studied well in the distributed settings. We also mention that, although distributed robust optimization under model uncertainty has been studied in the literature [26], [27], distributed robust signal recovery in the presence of outliers has not been investigated well so far.

With the variable vectors $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}_i$ to model $\boldsymbol{x}_\star$ and $\boldsymbol{\varepsilon}_{i,\star}$, respectively, our primal focus in the present study is on the following problem formulation:

$$\min_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{\varepsilon}_i\in\mathbb{R}^{m_i} \\ (i\in\mathcal{V})}} \sum_{i\in\mathcal{V}} \left(\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i) + \frac{\sigma_x^{-2}}{2\mu_i N}\left\|\boldsymbol{x}\right\|_2^2 + \frac{\sigma_\varepsilon^{-2}}{2\mu_i}\left\|\boldsymbol{\varepsilon}_i\right\|_2^2 \right), \tag{2}$$

where $\mu_i > 0$ is the regularization parameter, $\sigma_x^2 > 0$ is the signal power estimate, $\sigma_\varepsilon^2 > 0$ is the noise power estimate, and

$$\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{x}) := \sum_{i=1}^m \phi_\gamma^{\mathrm{MC}}(x_i) = \left\|\boldsymbol{x}\right\|_1 - {}^\gamma\|\cdot\|_1(\boldsymbol{x}), \tag{3}$$

is the MC penalty [28], [29], defined with $\phi_\gamma^{\mathrm{MC}}(x) := \begin{cases} |x| - x^2/2\gamma, & \text{if } |x| \le \gamma, \\ \gamma/2, & \text{if } |x| > \gamma. \end{cases}$ Here, $\gamma > 0$ is *the "saturation" factor* to control the saturation points from which $\Phi_\gamma^{\mathrm{MC}}$ becomes constant on each side of the real line. See Section II-B for the definition of the Moreau envelope ${}^\gamma\|\cdot\|_1$ of the ℓ_1 norm.

Corresponding author: Masahiro Yukawa

This work was supported by JSPS KAKENHI Grant Number (22H01492). A preliminary version of this work has been published in a conference [1].

Each term of the summands in (2) accommodates the prior information about the random vectors. Specifically, the first term $\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i)$ reflects the sparseness of the outlier vector $\boldsymbol{o}_{i\diamond} = \boldsymbol{y}_i - (\boldsymbol{A}_i\boldsymbol{x}_\star+\boldsymbol{\varepsilon}_{i\star})$, and the second and third terms, $\frac{\sigma_x^{-2}}{2\mu_i N}\|\boldsymbol{x}\|_2^2$ and $\frac{\sigma_\varepsilon^{-2}}{2\mu_i}\|\boldsymbol{\varepsilon}_i\|_2^2$, reflect the Gaussianity of the signal vector $\boldsymbol{x}_\star$ and the noise vector $\boldsymbol{\varepsilon}_{i\star}$, respectively. We mention here that the MC function $\phi_\gamma^{\mathrm{MC}}$ is known to bridge the ℓ_0 norm (the direct discrete measure of sparseness) and the ℓ_1 norm (its convex relaxation) by the single parameter γ [30].

The introduction of $\boldsymbol{\varepsilon}_i$ in our formulation together with the MC loss function $\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i)$ is the key to realize both outstanding robustness and "stability" (see the beginning of Section IV). Intuitively, a small σ_ε^{-2} (a large noise power estimate) allows the term $\|\boldsymbol{\varepsilon}_i\|_2^2$ to be large, modeling large Gaussian noise appropriately. We call the formulation in (2) *distributed stable outlier-robust signal recovery (D-SORR)*, because it gives "stable" estimates in the sense of [31]. The MC loss at each local node is nonconvex, as it is only weakly convex.

Our research questions concerned in this paper are presented below.

i) When is the problem (2) solvable by an iterative algorithm efficiently?
ii) How robust is the D-SORR estimator against outliers compared to existing methods?

To answer the first question, we study the convexity condition for the cost function in (2) using the framework called linearly-involved Moreau-enhance-over-subspace (LiMES) model developed in [32], [33].[1] The second question will be addressed experimentally via computer simulations. The major contributions of this paper are summarized below.

a) We start by considering a simplified version of (2) which does not involve the variable vectors $\boldsymbol{\varepsilon}_i$ to display the technical developments of the present work in a simpler manner. We show that each local objective is ensured to be convex under a certain condition on the regularization parameter (Proposition 1) based on the LiMES framework. We also show that the proposed formulation is solvable by the TriPD-Dist algorithm [21] via reformulation using Moreau's decomposition. A condition to ensure convexity of the global objective is also derived (Proposition 2); the condition is weaker than the local convexity condition.
b) We then study the formulation (2) which accommodates statistical properties of the noise and outliers, and which is thus more robust against perturbations caused by Gaussian noise as well as large outliers. See Section IV for the motivation of this specific formulation. In analogy with the first formulation, the local and global convexity conditions are presented (Propositions 4 and 5). We mention here that both local and global convexity conditions in the distributed setting differ from the centralized one. More precisely, the distributed local-convexity condition is stronger compared to the centralized case, while the distributed global-convexity condition involves the auxiliary vectors $\boldsymbol{\varepsilon}_i$ to model the noise vectors at each individual node.
c) The numerical examples show that our proposed methods lead to remarkable robustness even in catastrophic situations where data are contaminated by many and/or huge outliers, outperforming the existing methods in a variety of situations. Comparisons between the local and global convexity conditions are also presented.

We emphasize that the proposed methods have the following two properties simultaneously: (i) convergence guarantee to a solution based on the convexity condition (to be presented in Propositions 1 and 4), and (ii) remarkable robustness against outliers (to be shown by simulations) owing to the use of the nonconvex loss. This is in sharp contrast to the prior works which use either a convex loss with limited robustness or a nonconvex loss with, at most, a convergence guarantee to a stationary point. As a by-product, the convexity condition reduces the number of tuning parameters. We finally note that our approach is deterministic, and a study of the formulation (2) from the statistical perspective is beyond the scope of the present study.

A. Why to use nonconvex loss in the presence of outliers?

The least absolute deviation (LAD) loss $\psi(e) := |e|$ is less sensitive to outliers than the least square loss $\phi(e) := \frac{1}{2}e^2$, because it only grows linearly instead of quadratically.[2] Here, $e \in \mathbb{R}$ is the estimation residual. Residing between those two loss functions, Huber's loss function [3] is quadratic when $|e|$ is small, while it is linear when $|e|$ is large (see Section VI-A for the precise definition of Huber's loss). Because of this, Huber's loss is insensitive to small perturbations while possessing the same level of robustness as LAD.

The LAD and Huber's losses are convex and thus mathematically tractable, but their robustness is limited from the aspect of the so-called *influence function* [3], for which the *M-estimator* is proportional to the derivative $\psi(e) := \phi'(e)$ (defined at those points where the loss function is differentiable). Ideally, the influence of outliers, and thus the derivative, is desired to vanish for sufficiently large $|e|$. The derivative of Huber's loss, indeed, does not vanish and stays constant away from zero when $|e|$ exceeds a threshold. In fact, no convex function has a vanishing gradient, as long as such a class of loss functions are considered that satisfy the following conditions: (i) $\phi(0) = 0$, (ii) $\phi(e) > 0$ for all $e \neq 0$, and (iii) ϕ is differentiable everywhere but the origin [33].

This simple observation motivates us to explore nonconvex loss functions. Among many others [3]–[5], Tukey's biweight loss is a popular nonconvex loss function possessing the so-called *redescending property*; i.e., the derivative increases as the magnitude $|e|$ of the residual increases from zero, and it then *redescends* and keeps decreasing until it vanishes

[1] The LiMES model has been developed inspired by the linearly-involved generalized Moreau enhanced (LiGME) model [30], which includes the MC penalty [31], the generalized MC penalty [29], and many others, as its special cases. The LiMES model envisions an application to robust regression/classification as a particular example.

[2] We consider the one dimensional case for simplicity, but the arguments here can be extended to the multi-dimensional case straightforwardly.

completely at some point (see Section VI-A for the definition of Tukey's loss). Thanks to this property, Tukey's biweight loss could be fairly insensitive to large outliers, but convergence to a solution[3] is not guaranteed in general. The simulation results to be presented in Section VI will show that the formulation in (2) based on the MC loss performs better than Tukey's loss in a variety of situations.

B. Related works

In the previous work [34], the sparse signal recovery problem has been studied in a distributed setting under the use of the MC penalty to promote sparsity of estimates, where the proximal gradient EXTRA (PG-EXTRA) algorithm [19], an extension of EXTRA [18], was exploited. The algorithm, however, cannot be used in the present case because the proximity operator of the MC loss is hardly available due to the involvement of the linear composition. The recently-developed convex solver called distributed triangularly preconditioned primal-dual (TriPD-Dist) algorithm [21] can be applied to the present case as it encompasses the case when the objective function involves a composition of a "proximable" function and a linear operator based on *operator splitting* [35]–[41]. Here, the term "proximable" is used when the proximity operator can be computed efficiently.

Recall that the MC loss is a nonconvex (weakly convex) function. Nonconvex methods for distributed optimization have been studied actively both in signal processing and machine learning communities [42]. Recent developments include the heuristic approach based on the notion of graduated nonconvexity [43]. This approach starts with a relaxed (convex) version of the nonconvex loss, which gradually shifts to the original nonconvex loss for reducing the chance to be trapped by a local minimum. The graduated nonconvexity method has been used recently for outlier-robust distributed optimization [44] but with no guarantee of convergence to a solution (i.e., a global minimizer of the cost). More importantly, this approach cannot deal with a composition of a proximable function and a linear operator in analogy with most of the other existing methods. We also mention that some related problems have been studied in [45], [46] in the framework of the saddle-point problems.

Finally, when dealing with the nonconvex MC penalty term, an important question is whether convergence to a global minimizer can be guaranteed if the whole cost is convex, but the local cost at each node is not necessarily convex. Several works exist that establish the convergence to a global minimizer for distributed algorithms for nonconvex local loss functions, when the global loss function is convex [47]. For example, the in-network successive convex approximation (NEXT) algorithm demonstrates asymptotic convergence towards a local solution for the sum of a smooth (possibly nonconvex) loss function and a convex (possibly nonsmooth) regularization term at each node [48]. Another example is the distributed gradient descent algorithm, which is proven to converge towards a critical point for smooth and possibly nonconvex loss functions at each node [49]. However, these algorithms cannot be applied to our proposed formulation (2) involving the nonsmooth nonconvex MC loss function. This is because the local nonsmooth functions are different from each other, and also because the MC loss function is composed with the affine operator. In Section III-D, the convergence properties for the cases of local and global convexity are discussed for the TriPD-Dist algorithm [21], which is applicable to our proposed formulation. To the best of the authors' knowledge, no previous work guarantees convergence to a global minimizer of the whole cost when the composition of the MC loss function with a linear operator is involved in the distributed setting.

[3]The term "solution" always means a "global" minimizer of the objective function in the present work.

II. Preliminaries

We first present the notation and the assumptions used throughout the paper. We then present the convex analytic tools and briefly introduce the TriPD-Dist algorithm. We finally present the linear model in a distributed setting as well as the MC penalty.

A. Notation and assumptions

Vectors are written in boldfaced lowercase letters, and matrices are written in boldfaced uppercase letters. The transpose of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\top$. Let $\mathbb{R}^n$ denote the n dimensional Euclidean space. We define the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y}$ between $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$. The $n \times n$ identity matrix is denoted by $\boldsymbol{I}_n$. The $m \times n$ zero matrix is denoted by $\boldsymbol{O}_{m\times n}$; the $n \times n$ square zero-matrix and the length-n zero vector are particularly denoted by $\boldsymbol{O}_n$ and $\boldsymbol{0}_n$, respectively. The largest and smallest eigenvalues of a symmetric matrix $\boldsymbol{A}$ are denoted by $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$, respectively. Given any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A} \succeq \boldsymbol{O}_n$ means that $\boldsymbol{A}$ is a positive semidefinite matrix. The ℓ_1 norm of a Euclidean vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined by $\|\boldsymbol{x}\|_1 := \sum_{k=1}^n |x_k|$, and the ℓ_2 norm is defined by $\|\boldsymbol{x}\|_2 := \left(\sum_{k=1}^n x_k^2\right)^{1/2}$.

We consider *decentralized systems* equipped with a network of N nodes represented by the undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} := \{1, 2, .., N\}$ is the set of nodes (vertices), and $\mathcal{E}$ is the set of edges. Here, $(i, j) \in \mathcal{E}$ if node $i \in \mathcal{V}$ is connected to node $j \in \mathcal{V}$. The set of neighbors of node j is denoted by $\mathcal{N}_j$, where $j \in \mathcal{N}_j$ for every $j \in \mathcal{V}$. Throughout, we solely consider *connected* graphs; i.e., there exists a path among every pair of nodes when one follows the edges.

B. Convex analytic tools

A function $f : \mathbb{R}^n \to (-\infty, +\infty] := \mathbb{R} \cup \{+\infty\}$ is convex if $f(t\boldsymbol{x}+(1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x})+(1-t)f(\boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f := \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \in \mathbb{R}\}$ and every $t \in [0, 1]$. A function f is a proper convex function, if additionally $\mathrm{dom} f \neq \emptyset$. A convex function f is lower-semicontinuous (or closed) on $\mathbb{R}^n$, if the level set $\mathrm{lev}_{\leq a} f := \left\{\boldsymbol{x} \in \mathbb{R}^N : f(\boldsymbol{x}) \leq a\right\}$ is closed for any $a \in \mathbb{R}$. All continuous functions are lower-semicontinuous. A function f is μ-weakly convex, if $f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ is convex for some $\mu > 0$.

Suppose that $f : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper lower-semicontinuous convex function. Then, its Fenchel conjugate

is defined by $f^*(\boldsymbol{x}) := \sup_{\boldsymbol{y}\in\mathbb{R}^n}(\langle \boldsymbol{x},\boldsymbol{y}\rangle - f(\boldsymbol{y}))$ [50], which is again a proper lower-semicontinuous convex function. The proximity operator of f of index $\gamma > 0$ is defined by $\operatorname{prox}_{\gamma f}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathbb{R}^n}\left(f(\boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right)$ [50], [51], and the minimum value ${}^{\gamma}f(\boldsymbol{x}) := \min_{\boldsymbol{y}\in\mathbb{R}^n}\left(f(\boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right) = f(\operatorname{prox}_{\gamma f}(\boldsymbol{x})) + \frac{1}{2\gamma}\left\|\boldsymbol{x}-\operatorname{prox}_{\gamma f}(\boldsymbol{x})\right\|_2^2$ achieved by the proximity operator is called the Moreau envelope of f [50], [51].

C. TriPD-Dist: Distributed convex optimization algorithm

The TriPD-Dist algorithm is a convex analytic solver for distributed optimization problems in the following form [21]:

$$\begin{aligned}\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\in\mathbb{R}^n} \quad & \sum_{i\in\mathcal{V}} F_i(\boldsymbol{x}_i) + G_i(\boldsymbol{x}_i) + H_i(\boldsymbol{A}_i\boldsymbol{x}_i) \\ \text{s.t.} \quad & \boldsymbol{B}_{ij}\boldsymbol{x}_i + \boldsymbol{B}_{ji}\boldsymbol{x}_j = \boldsymbol{d}_{ij}, \quad (i,j)\in\mathcal{E},\end{aligned} \tag{4}$$

where each variable vector $\boldsymbol{x}_i$ is updated at each node with information exchanges allowed over the given network represented by the graph $\mathcal{G}(\mathcal{V},\mathcal{E})$. Here, $F_i : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function with a Lipschitz continuous gradient, $G_i : \mathbb{R}^n \to (-\infty,+\infty]$ and $H_i : \mathbb{R}^{m_i} \to (-\infty,+\infty]$ are (possibly nonsmooth) convex functions, and $\boldsymbol{A}_i \in \mathbb{R}^{m_i\times n}$. The consensus constraint $\boldsymbol{x}_i = \boldsymbol{x}_j, \forall i,j\in\mathcal{V}$, can be expressed by letting $\boldsymbol{B}_{ij} := \boldsymbol{I}_n$, $\boldsymbol{B}_{ji} := -\boldsymbol{I}_n$, and $\boldsymbol{d}_{ij} := \boldsymbol{0}_n$ for all $(i,j)\in\mathcal{E}$.

The TriPD-Dist algorithm [21] is given in Algorithm 1. In this study, the agents in the distributed network are assumed to be time synchronized, and therefore we adopt the synchronous version of the distributed algorithm. The UNLocBoX toolbox was used to evaluate the proximity operator in the implementation [52].

Algorithm 1: Distributed Triangular Preconditioned Primal-Dual (TriPD-Dist) algorithm from [21]

Inputs: primal variable $\boldsymbol{x}_i(0)\in\mathbb{R}^n$ for $i\in\mathcal{V}$, dual variable $\boldsymbol{z}_i(0)\in\mathbb{R}^{m_i}$ for $i\in\mathcal{V}$, edge variable $\boldsymbol{w}_{ij}(0)\in\mathbb{R}^n$ for $(i,j)\in\mathcal{E}$, step size $\tau_i > 0$, dual step size $\varsigma_i > 0$, and link weights $\kappa_{ij} > 0$

for $k = 0, 1, \ldots$ **do**

local updates

for *all neighbors* j *of agent* i **do**

$\overline{\boldsymbol{w}}_{ij}(k) = \frac{1}{2}\left[\boldsymbol{w}_{ij}(k) + \boldsymbol{w}_{ji}(k)\right] + \frac{\kappa_{ij}}{2}\left[\boldsymbol{B}_{ij}\boldsymbol{x}_i(k) + \boldsymbol{B}_{ji}\boldsymbol{x}_j(k) - \boldsymbol{d}_{ij}\right]$

end

$\overline{\boldsymbol{z}}_i(k) = \operatorname{prox}_{\varsigma_i H_i^*}\left[\boldsymbol{z}_i(k) + \varsigma_i\boldsymbol{A}_i\boldsymbol{x}_i(k)\right]$

$\boldsymbol{x}_i(k+1) = \operatorname{prox}_{\tau_i G_i}[\boldsymbol{x}_i(k) - \tau_i\boldsymbol{A}_i^\top\overline{\boldsymbol{z}}_i(k) - \tau_i\sum_{j\in\mathcal{N}_i}\boldsymbol{B}_{ij}\overline{\boldsymbol{w}}_{ij}(k) - \tau_i\nabla F_i(\boldsymbol{x}_i(k))]$

$\boldsymbol{z}_i(k+1) = \overline{\boldsymbol{z}}_i(k) + \varsigma_i\boldsymbol{A}_i\left[\boldsymbol{x}_i(k+1) - \boldsymbol{x}_i(k)\right]$

for *all neighbors* j *of agent* i **do**

$\boldsymbol{w}_{ij}(k+1) = \overline{\boldsymbol{w}}_{ij}(k) + \kappa_{ij}\boldsymbol{B}_{ij}\left[\boldsymbol{x}_i(k+1) - \boldsymbol{x}_i(k)\right]$

end

transmission of information

send $\boldsymbol{x}_i(k+1)$ and each estimate $\boldsymbol{w}_{ij}(k+1)$ to each neighbor j

end

III. Convex Analytic Framework for Distributed Outlier-Robust Signal Recovery Using MC Loss

We present our first problem formulation for distributed outlier-robust signal recovery using the MC loss function. We then analyze the convexity of the local objective, and present the optimization algorithm. We finally present additional discussions about convexity of the global objective with a remark on parameter design.

A. Distributed outlier-robust signal recovery — formulation

We start by considering the global objective function $\frac{1}{2\mu}\|\boldsymbol{x}\|_2^2 + \sum_{i\in\mathcal{V}}\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i)$ for some constant $\mu > 0$. The "global" regularizer $\frac{1}{2\mu}\|\boldsymbol{x}\|_2^2$ can then be shared by N nodes equally so that the objective becomes $\sum_{i\in\mathcal{V}}\left(\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i) + \frac{1}{2\mu N}\|\boldsymbol{x}\|_2^2\right)$. Allowing μ to be designed independently at each node, our first formulation, which we call *distributed outlier-robust signal recovery (D-ORR)*, is given as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\sum_{i\in\mathcal{V}}\left(\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i) + \frac{1}{2\mu_i N}\|\boldsymbol{x}\|_2^2\right), \tag{5}$$

where $\mu_i > 0$ is the regularization parameter.

B. Convexity condition for local objective of D-ORR

The objective function of the D-ORR formulation in (5) can be split into smooth and nonsmooth terms as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\sum_{i\in\mathcal{V}}\underbrace{\left(\frac{1}{2\mu_i N}\|\boldsymbol{x}\|_2^2 - {}^{\gamma}\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i)\right)}_{F_i^{\text{D-ORR}}(\boldsymbol{x})} + \underbrace{\|\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i\|_1}_{H_i^{\text{D-ORR}}(\boldsymbol{A}_i\boldsymbol{x})}, \tag{6}$$

where

$$F_i^{\text{D-ORR}}(\boldsymbol{x}) := \frac{1}{2\mu_i N}\|\boldsymbol{x}\|_2^2 - {}^{\gamma}\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i), \tag{7}$$

$$H_i^{\text{D-ORR}}(\boldsymbol{v}) := \|\boldsymbol{v}-\boldsymbol{y}_i\|_1, \quad \boldsymbol{v}\in\mathbb{R}^{m_i}. \tag{8}$$

The local functions $F_i^{\text{D-ORR}}(\boldsymbol{x})$ and $H_i^{\text{D-ORR}}(\boldsymbol{A}_i\boldsymbol{x})$ at each node i need to be convex to guarantee convergence of the TriPD-Dist algorithm to a solution of the problem [21]. The function $H_i^{\text{D-ORR}}(\boldsymbol{v})$ is a convex function, because it is the ℓ_1 norm with a translation by $\boldsymbol{y}_i$. Hence, its composition (the nonsmooth term) $H_i^{\text{D-ORR}}(\boldsymbol{A}_i\boldsymbol{x})$ with the linear operator $\boldsymbol{A}_i$ is also convex without any condition. We therefore present a necessary and sufficient condition below for convexity of the smooth term $F_i^{\text{D-ORR}}$.

Proposition 1. *[Convexity condition of local objective $F_i^{\text{D-ORR}}(\boldsymbol{x})$]*

(a) For each $i\in\mathcal{V}$, the local objective function $F_i^{\text{D-ORR}}(\boldsymbol{x})$ is convex if

$$\mu_i N\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i) \le \gamma. \tag{9}$$

(b) Assume in particular that $\boldsymbol{A}_i$ has full column rank or that $\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{y}_i = \boldsymbol{0}_{m_i}$ for some $\boldsymbol{x}\in\mathbb{R}^n$. Then, $F_i^{\text{D-ORR}}(\boldsymbol{x})$ is convex if and only if (9) *is satisfied.*

Proof: See Appendix A. ■

We briefly mention that Proposition 1 allows to design μ_i systematically, leaving only γ as a tuning parameter (see Remark 1 in Section III-D for details).

C. Distributed optimization algorithm for D-ORR

Using the local variable vector $\boldsymbol{x}_i \in \mathbb{R}^n$ at each node of the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the D-ORR problem to be solved by the TriPD-Dist algorithm (over the network) can be written as follows:

$$\begin{aligned} \min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \mathbb{R}^n} \sum_{i\in\mathcal{V}} \underbrace{F_i^{\text{D-ORR}}(\boldsymbol{x}_i)}_{\text{smooth}} + \underbrace{H_i^{\text{D-ORR}}(\boldsymbol{A}_i\boldsymbol{x}_i)}_{\text{nonsmooth}} \\ \text{subject to} \quad \boldsymbol{x}_i = \boldsymbol{x}_j, \quad (i,j) \in \mathcal{E}. \end{aligned} \tag{10}$$

The problem in (10) is exactly in the form of (4) with $G_i := 0$. Recall here that the consensus constraint in (10) is a linear constraint with $\boldsymbol{B}_{ij} := \boldsymbol{I}_n$, $\boldsymbol{B}_{ji} := -\boldsymbol{I}_n$, and $\boldsymbol{d}_{ij} := \boldsymbol{0}_n$ for all $(i,j) \in \mathcal{E}$. Note that $\boldsymbol{x}_i = \boldsymbol{x}_j,\ \forall (i,j) \in \mathcal{E} \Leftrightarrow \boldsymbol{x}_i = \boldsymbol{x}_j,\ \forall i,j = 1,2,\ldots,N$, i.e., consensus between every pair $(i,j) \in \mathcal{V}$ implies consensus among all nodes, because the graph is assumed to be connected.

Invoking Moreau's identity [35], the proximity operator $\text{prox}_{\varsigma_i H_i^*}$ can be computed by $\text{prox}_{\varsigma_i H_i^*}(\boldsymbol{y}) = \boldsymbol{y} - \varsigma_i \text{prox}_{\varsigma_i^{-1} H_i}(\varsigma_i^{-1}\boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{R}^{m_i}$, where $\text{prox}_{\varsigma_i^{-1} H_i^{\text{D-ORR}}}(\boldsymbol{v}) = \boldsymbol{y}_i + \text{prox}_{\varsigma_i^{-1}\|\cdot\|_1}(\boldsymbol{v} - \boldsymbol{y}_i)$. On the other hand, the proximity operator for $G_i = 0$ is given simply by $\text{prox}_{\tau_i G_i}(\boldsymbol{x}) = \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^n$. We mention that the primal-dual formulation decouples the linear operator $\boldsymbol{A}_i$ from the function $H_i^{\text{D-ORR}}$, because the proximity operator of the composition $H_i^{\text{D-ORR}} \circ \boldsymbol{A}_i$ has no closed-form expression. The gradient of the Moreau envelope term is given by

$$\nabla\, {}^{\gamma}\|\cdot\|_1 (\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i) = \boldsymbol{A}_i^\top \frac{\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i - \text{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i)}{\gamma}.$$

The following result gives a Lipschitz constant for the gradient of $F_i^{\text{D-ORR}}$.

Lemma 1. *The function $F_i^{\text{D-ORR}}$ has a Lipschitz continuous gradient operator $\nabla F_i^{\text{D-ORR}}$ with constant*

$$\beta_i^{\text{D-ORR}} := 1/(\mu_i N) + \frac{1}{2\gamma}(\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i) - \lambda_{\min}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)). \tag{11}$$

Proof: See Appendix B. ■

Note here that the bound $\beta_i^{\text{D-ORR}}$ given in (11) is tighter than the bound $1/(\mu_i N) + \frac{\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)}{\gamma}$, which is immediately obtained from (7) using the fact that the gradient of the Moreau envelope of index γ is $1/\gamma$-Lipschitz continuous [50], [51]. In typical situations, we have $m_i < n$ so that $\lambda_{\min}(\boldsymbol{A}_i^\top \boldsymbol{A}_i) = 0$, which reduces (11) to $\beta_i^{\text{D-ORR}} := 1/(\mu_i N) + \frac{1}{2\gamma}\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)$. Under the convexity condition (9) and the step size condition

$$\tau_i < \frac{1}{\beta_i^{\text{D-ORR}}/2 + \varsigma_i \left\|\boldsymbol{A}_i^\top \boldsymbol{A}_i\right\| + \sum_{j\in\mathcal{N}_i} \kappa_{ij}}, \tag{12}$$

where $\left\|\boldsymbol{A}_i^\top \boldsymbol{A}_i\right\|$ is the spectral norm of $\boldsymbol{A}_i^\top \boldsymbol{A}_i$, the sequence $(\boldsymbol{x}_i(k))_{k\in\mathbb{N}}$ generated by the TriPD-Dist algorithm at every node i converges to a common solution of (5), provided that the graph is connected (see Appendix C for details about the convergence). Note that the problem in (5) always has a solution because the objective function is clearly coercive, i.e., it tends to infinity as $\|\boldsymbol{x}\|_2 \to +\infty$.

The condition in (12) can be used to design the primal step-size τ_i. In the present study, the dual step sizes of the algorithm are set to $\varsigma_i = 0.065$, and the link weights are set to $\kappa_{ij} := 1$, if $(i,j) \in \mathcal{E}$, and $\kappa_{ij} := 0$, otherwise.

D. Convexity condition for global objective of D-ORR

The TriPD-Dist algorithm provably converges to a solution under convexity of "the local objective at each node", which has been analyzed in Proposition 1. Nevertheless, there may be a possibility for the algorithm to converge to a solution under a weaker condition. We present below a condition for convexity of "the global objective over the entire network", which is weaker than that of the local objective, because the sum of convex functions is again a convex function.

Proposition 2 (Convexity condition of global objective $F^{\text{D-ORR}}(\boldsymbol{x}) = \sum_{i\in\mathcal{V}} F_i^{\text{D-ORR}}(\boldsymbol{x})$)**.** *Let $\mu_1 = \cdots = \mu_N =: \mu > 0$. Assume that the matrix $\boldsymbol{A}_i$ has full column rank for every $i \in \mathcal{V}$. Then, the smooth part*

$$F^{\text{D-ORR}}(\boldsymbol{x}) = \sum_{i\in\mathcal{V}} \left(\frac{1}{2\mu N} \|\boldsymbol{x}\|_2^2 - {}^{\gamma}\|\cdot\|_1 (\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i) \right) \tag{13}$$

of the global objective function in (6) *is convex if and only if*

$$\mu\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A}) \le \gamma, \tag{14}$$

where $\boldsymbol{A} := [\boldsymbol{A}_1^\top\ \boldsymbol{A}_2^\top\ \cdots\ \boldsymbol{A}_N^\top]^\top$.

Proof: The proof is omitted as it is obtained in an analogous way to the proof of Proposition 5 presented in Section IV. ■

Remark 1 (Parameter design for D-ORR)**.** *The D-ORR formulation involves two kinds of parameters; i.e., the saturation factor γ and the regularization parameters μ_i. Our recommended way of choosing those parameters is the following: tune γ by grid search with μ_i (or μ) set to its upper bound based on* (9) *(or on* (14)*) for each given γ.*

We now compare the convexity condition of the global objective in (14) with that of the local one in (9). To do so, we rewrite the D-ORR formulation in (5) as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left(\frac{1}{N}\sum_{i\in\mathcal{V}} \mu_i^{-1} \right) \frac{1}{2}\|\boldsymbol{x}\|_2^2 + \sum_{i\in\mathcal{V}} \Phi_\gamma^{\text{MC}}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i), \tag{15}$$

where $\frac{1}{N}\sum_{i\in\mathcal{V}} \mu_i^{-1}$ can be regarded as the global regularization parameter, which scales the impact of the Tikhonov regularization $\frac{1}{2}\|\boldsymbol{x}\|_2^2$ over the entire network. Simple inspections of (9) and (14) suggest that

$$\begin{aligned} (9),\ \forall i \in \mathcal{V} &\Leftrightarrow \gamma\mu_i^{-1} \ge N\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i),\ \forall i \in \mathcal{V} \\ &\Rightarrow \frac{\gamma}{N}\sum_{i\in\mathcal{V}} \mu_i^{-1} \ge \sum_{i\in\mathcal{V}} \lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i) =: \alpha_{\text{local}}, \end{aligned} \tag{16}$$

$$(14) \Leftrightarrow \frac{\gamma}{N}\sum_{i\in\mathcal{V}} \mu^{-1} = \gamma\mu^{-1} \ge \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A}) =: \alpha_{\text{global}}. \tag{17}$$

Inequalities (16) and (17) imply that $\frac{\gamma}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1}$ needs to be sufficiently large to ensure the convexity of each objective. At the same time, however, each of the quantities γ and $\frac{1}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1}$ is also desired to be reasonably "small" from different aspects. Specifically, γ is desired to be small for robustness against outliers, and $\frac{1}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1}$ is desired to be small to reduce extra estimation biases caused by the regularization.

Because the convexity of the local objective is stronger than that of the global objective as mentioned previously, the condition in (16) should be stricter than that in (17). This is stated formally as follows.

Proposition 3. *The lower bounds of* $\frac{\gamma}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1}$ *given in* (16) *and* (17) *satisfy the following inequality:*

$$\alpha_{\text{local}} \geq \alpha_{\text{global}}. \tag{18}$$

Proof: This is a direct consequence of the eigenvalue inequality [53, Lemma 4.2]: $\lambda_{\max}\left(\sum_{i\in\mathcal{V}}\boldsymbol{A}_i^\top\boldsymbol{A}_i\right) \leq \sum_{i\in\mathcal{V}}\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)$. ∎

Simulation studies will be given in Section V-B to compare the convexity conditions for the global and local objectives, as well as the performances when the parameters γ and μ_i are designed based on each of those conditions. A theoretical convergence analysis under the global convexity condition is left as a future work.

IV. Distributed Stable Outlier-Robust Signal Recovery Using MC Loss and Auxiliary Vector

The derivative $\psi_\gamma^{\mathrm{MC}} := \frac{d}{dx}\phi_\gamma^{\mathrm{MC}} : \mathbb{R}\setminus\{0\} \to [-1,1]$ of $\phi_\gamma^{\mathrm{MC}}$ is given by $\psi_\gamma^{\mathrm{MC}} : x \mapsto \begin{cases}\operatorname{sign}(x) - x/\gamma, & \text{if } |x| \in (0,\gamma),\\ 0, & \text{if } |x| \geq \gamma.\end{cases}$ We inspect the behavior of the derivative $\psi_\gamma^{\mathrm{MC}}$ on the positive side of the real line. (We can also make essentially the same arguments for the negative side of the real line.) Start at $x = \gamma/2$ and increase x gradually. Then, $\psi_\gamma^{\mathrm{MC}}(x)$ decreases linearly, and it vanishes at $x = \gamma$. This property makes the MC loss remarkably robust against huge outliers in analogy with Tukey's biweight loss. Let us go back to the point $x = \gamma/2$ and now "decrease" x gradually. One can easily see that $\lim_{x\downarrow 0}\psi_\gamma^{\mathrm{MC}}(x) = 1$, meaning that the derivative does not vanish at the origin. This is an intrinsic difference from Tukey's biweight loss in addition to the weak convexity of the MC loss.

The non-vanishment of the derivative $\psi_\gamma^{\mathrm{MC}}$ mentioned above implies that the MC loss sharply increases by small deviations from zero (as the LAD loss does), and thus it would not allow small errors originated by Gaussian noise. As such, the MC loss would make the error $\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i$ in (5) be a sparse vector which does not model the "nonsparse" vector $\boldsymbol{A}_i\boldsymbol{x}_\star - \boldsymbol{y}_i$ $(= -\boldsymbol{\varepsilon}_{i\star} - \boldsymbol{o}_{i\diamond})$ well. This model mismatch may cause sensitivity to Gaussian noise. Our proposed solution for this issue is the introduction of the auxiliary vectors $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{m_i}$ involved in the D-SORR formulation (2), which is studied in this section. The employment of $\boldsymbol{\varepsilon}_i$ brings significant improvements of the performance in the case when the Gaussian noise is dominant compared to the outliers, as will be shown by simulations in Section VI.

A. Convexity condition for local objective of D-SORR

For convenience, we define $\boldsymbol{\xi}_i := [\boldsymbol{x}_i^\top\ \boldsymbol{\varepsilon}_i^\top]^\top \in \mathbb{R}^{n+m_i}$. The objective function in (2) can be split into smooth and nonsmooth terms as follows:[4]

$$\min_{\substack{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{\varepsilon}_i\in\mathbb{R}^{m_i}\\ (i\in\mathcal{V})}} \sum_{i\in\mathcal{V}}\Bigg(\underbrace{\|\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i\|_1}_{H_i^{\text{D-SORR}}(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i)} + \underbrace{\frac{\sigma_x^{-2}}{2\mu_i N}\|\boldsymbol{x}\|_2^2+\frac{\sigma_\varepsilon^{-2}}{2\mu_i}\|\boldsymbol{\varepsilon}_i\|_2^2-\gamma\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i)}_{F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i)}\Bigg). \tag{19}$$

Here, the nonsmooth term $H_i^{\text{D-SORR}}(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i)$, defined with

$$H_i^{\text{D-SORR}}(\boldsymbol{v}) := \|\boldsymbol{v}-\boldsymbol{y}_i\|_1, \tag{20}$$

is a convex function in the space $\mathbb{R}^n\times\mathbb{R}^{m_i}$ of the pair $(\boldsymbol{x},\boldsymbol{\varepsilon}_i)$ of variable vectors in analogy with the case of D-ORR by considering the linear operator $(\boldsymbol{x},\boldsymbol{\varepsilon}_i) \mapsto \boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i$. The convexity condition for the smooth term

$$F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i) := \frac{\sigma_x^{-2}}{2\mu_i N}\|\boldsymbol{x}\|_2^2 + \frac{\sigma_\varepsilon^{-2}}{2\mu_i}\|\boldsymbol{\varepsilon}_i\|_2^2 - \gamma\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i) \tag{21}$$

is analyzed below.

Proposition 4 (Convexity condition of local objective $F_i^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_i)$)**.** *For each* $i \in \mathcal{V}$*, the local function* $F_i^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_i)$ *is convex in* $(\boldsymbol{x},\boldsymbol{\varepsilon}_i)\in\mathbb{R}^n\times\mathbb{R}^{m_i}$ *if and only if*

$$\mu_i(\sigma_\varepsilon^2 + N\sigma_x^2\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)) \leq \gamma. \tag{22}$$

Proof: One can replace σ_x^2 by $N\sigma_x^2$ and let $\mu := \mu_i$, $\varepsilon := \varepsilon_i$, $\boldsymbol{A} := \boldsymbol{A}_i$, $\boldsymbol{y} := \boldsymbol{y}_i$ in [33, Proposition 3] to obtain the result. ∎

The local convexity condition for the distributed setting is from a technical standpoint similar to the centralized setting, because every local node can be regarded as an individual centralized setting for the derivation of the convexity condition. Note however that the set of local convexity conditions differs from the overall convexity condition in the centralized case, as the set of local convexity conditions is stronger. Our simulations in Section V, moreover, will give important insights into the distributed setting by investigating how the proposed methods and its parameters behave when the data are distributed among different numbers of nodes.

B. Distributed optimization algorithm for D-SORR

We reformulate the D-SORR problem with the local variables $\boldsymbol{x}_i \in \mathbb{R}^n$ into a suitable form to the TriPD-Dist algorithm. Note that the unknown vector $\boldsymbol{x}_\star$ is common to all nodes in (1), while the noise vectors $\boldsymbol{\varepsilon}_{i\star}$ are different among the nodes. This means that the consensus constraint is required among the $\boldsymbol{x}_i$'s, but it is *not* required for the $\boldsymbol{\varepsilon}_i$'s. The "partial" consensus constraint $\boldsymbol{x}_i = \boldsymbol{x}_j$ can be expressed by $\tilde{\boldsymbol{I}}_i\boldsymbol{\xi}_i = \tilde{\boldsymbol{I}}_j\boldsymbol{\xi}_j$ with $\tilde{\boldsymbol{I}}_i := [\boldsymbol{I}_n\ \boldsymbol{O}_{n\times m_i}] \in$

[4]The existing methods, such as NEXT [48] and the distributed gradient descent algorithm [49], cannot be applied to the reformulated problem (19) as well.

$\mathbb{R}^{n\times(n+m_i)}$. Let $\tilde{\boldsymbol{A}}_i = \begin{bmatrix}\boldsymbol{A}_i & \boldsymbol{I}_{m_i}\end{bmatrix} \in \mathbb{R}^{m_i\times(n+m_i)}$, and $\boldsymbol{\Lambda}_i := \begin{bmatrix}(\sigma_x^{-1}/\sqrt{\mu_i N})\boldsymbol{I}_n & \boldsymbol{O}_{n\times m_i} \\ \boldsymbol{O}_{m_i\times n} & (\sigma_\varepsilon^{-1}/\sqrt{\mu_i})\boldsymbol{I}_{m_i}\end{bmatrix} \in \mathbb{R}^{(n+m_i)\times(n+m_i)}$. Then, (19) can be reformulated as follows:

$$\min_{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_N} \sum_{i\in\mathcal{V}} \left(H_i^{\text{D-SORR}}(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i) + F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i)\right) \quad \text{s.t. } \tilde{\boldsymbol{I}}_i\boldsymbol{\xi}_i = \tilde{\boldsymbol{I}}_j\boldsymbol{\xi}_j,\ \forall i,j=1,2,\ldots,N, \tag{23}$$

where $F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i)$ can be rewritten as

$$F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i) = \frac{1}{2}\left\|\boldsymbol{\Lambda}_i\boldsymbol{\xi}_i\right\|_2^2 - {}^{\gamma}\|\cdot\|_1\left(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i - \boldsymbol{y}_i\right). \tag{24}$$

Since (23) shares the same form as (10), the TriPD-Dist algorithm can be applied with $\boldsymbol{B}_{ij} := \begin{cases}\tilde{\boldsymbol{I}}_i, & \text{if } i<j,\\ -\tilde{\boldsymbol{I}}_i, & \text{otherwise,}\end{cases}$ and $\boldsymbol{d}_{ij} := \boldsymbol{0}_n$. Analogously to the case of D-ORR (see Section III-C), the dual step sizes and the link weights are set respectively to $\varsigma_i := 0.065$ and $\kappa_{ij} := 1$, if $(i,j)\in\mathcal{E}$, and $\kappa_{ij} := 0$, otherwise.

A tight Lipschitz constant for the gradient of $F_i^{\text{D-SORR}}$ is given below.

Lemma 2. *The function $F_i^{\text{D-SORR}}$ has a Lipschitz continuous gradient operator $\nabla F_i^{\text{D-SORR}}$ with constant*

$$\beta_i^{\text{D-SORR}} := \lambda_{\max}(\boldsymbol{\Lambda}_i^2 - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i) + \frac{1}{2\gamma}\lambda_{\max}(\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i). \tag{25}$$

Proof: See Appendix D. ■

It should be mentioned that the matrix $\boldsymbol{\Lambda}_i^2 - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i$ in (25) is positive definite under the convexity condition in (22). The largest eigenvalues in (25) can be found efficiently by the power method. Note here that $\lambda_{\max}(\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i) = \lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i) + 1$, and the particular structure of $\boldsymbol{\Lambda}_i^2 - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i$ can be exploited for efficiency of the power iteration.

In analogy with D-ORR, convergence of the TriPD-Dist algorithm for the D-SORR formulation is guaranteed under the step size condition

$$\tau_i < \frac{1}{\beta_i^{\text{D-SORR}}/2 + \varsigma_i\left\|\boldsymbol{A}_i^\top\boldsymbol{A}_i\right\| + \sum_{j\in\mathcal{N}_i}\kappa_{ij}}, \tag{26}$$

the convexity condition (22), and the graph connectivity (see Appendix E for details about the convergence). Note that the problem in (2) always has a solution because the objective function is clearly coercive here again.

C. Convexity condition for global objective of D-SORR

We analyze and discuss the condition for convexity of the global objective of the D-SORR formulation, and then compare it to that of the local objective.

Proposition 5 (Convexity condition of global objective $F^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_N) = \sum_{i\in\mathcal{V}} F_i^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_i)$)**.** *Let $\mu_1 = \cdots = \mu_N =: \mu > 0$. Then, the smooth part*

$$F^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_N) = \sum_{i\in\mathcal{V}}\left(\frac{\sigma_x^{-2}}{2\mu N}\left\|\boldsymbol{x}\right\|_2^2 + \frac{\sigma_\varepsilon^{-2}}{2\mu}\left\|\boldsymbol{\varepsilon}_i\right\|_2^2 - {}^{\gamma}\|\cdot\|_1\left(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i\right)\right) \tag{27}$$

of the global objective in (19) *is convex in $(\boldsymbol{x},\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_N)\in\mathbb{R}^n\times\mathbb{R}^{m_1}\times\cdots\times\mathbb{R}^{m_N}$ if and only if*

$$\mu(\sigma_\varepsilon^2 + \sigma_x^2\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})) \le \gamma. \tag{28}$$

Proof: See Appendix F. ■

Remark 2 (On auxiliary vectors $\boldsymbol{\varepsilon}_i$ of D-SORR)**.** *As every node is corrupted by random noise statistically independent of those of the other nodes, each node has an associated auxiliary vector $\boldsymbol{\varepsilon}_i$ that is also independent of those of the other nodes. This is a significant difference from the centralized setting, where only one auxiliary vector is involved. It also explains why a consensus does not need to be imposed on the $\boldsymbol{\varepsilon}_i$'s, as doing so would potentially limit the performance of the algorithm and increase computational loads. This difference between the distributed and centralized settings necessitates a detailed proof of the global convexity condition of D-SORR (Proposition 5), which is one of our original technical contributions.*

Remark 3 (Parameter design for D-SORR)**.** *The D-SORR formulation involves two extra parameters σ_x^2 and σ_ε^2 in addition to γ and the μ_i's. Our recommended way of choosing those parameters is similar to the one given in Remark 1 for D-ORR. Especially, when estimates of $\sigma_{x\star}^2$ and $\sigma_{\varepsilon\star}^2$ are available, one may exactly follow the way in Remark 1.*

Suppose that such estimates are unavailable. Then, by letting $\bar{\mu}_i := \mu_i\sigma_x^2$ ($\bar{\mu} := \mu\sigma_x^2$) and $\varrho := \sigma_x^2/\sigma_\varepsilon^2$, the last two terms of (2) *are reduced to $\frac{1}{2\bar{\mu}_i N}\left\|\boldsymbol{x}\right\|_2^2 + \frac{\varrho}{2\bar{\mu}_i}\left\|\boldsymbol{\varepsilon}_i\right\|_2^2$. Furthermore, the convexity conditions in* (22) *and* (28) *are reduced to $\bar{\mu}_i(\varrho + N\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)) \le \gamma$ and $\bar{\mu}(\varrho + \lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})) \le \gamma$, respectively. As such, only the power ratio $\varrho_\star := \sigma_{x\star}^2/\sigma_{\varepsilon\star}^2$ needs to be estimated rather than each of $\sigma_{x\star}^2$ and $\sigma_{\varepsilon\star}^2$. When even the ratio estimate is unavailable, ϱ as well as γ can be considered as a tuning parameter. More specifically, ϱ and γ can be tuned by grid search with μ_i (or μ) set to its upper bound based on the convexity condition for each pair of (ϱ,γ). D-SORR is fairly insensitive to the choice of ϱ as will be shown by simulations in Section V-B.*[5]

Based on similar arguments to those given in Section III-D, we obtain

$$\begin{aligned}&(22),\ \forall i\in\mathcal{V} \Leftrightarrow \gamma\mu_i^{-1} \ge \sigma_\varepsilon^2 + N\sigma_x^2\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i),\ \forall i\in\mathcal{V}\\ &\Rightarrow \frac{\gamma}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1} \ge \sigma_\varepsilon^2 + \sigma_x^2\sum_{i\in\mathcal{V}}\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i) =: \hat{\alpha}_{\text{local}}, \quad (29)\\ &(28) \Leftrightarrow \frac{\gamma}{N}\sum_{i\in\mathcal{V}}\mu^{-1} = \gamma\mu^{-1}\\ &\qquad\qquad \ge \sigma_\varepsilon^2 + \sigma_x^2\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A}) =: \hat{\alpha}_{\text{global}}. \quad (30)\end{aligned}$$

As in the case of D-ORR, the quantity $\frac{1}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1}$ governs the strength of $\frac{1}{2}\left\|\boldsymbol{x}\right\|_2^2$ in the entire optimization over the network, which affects the performance of D-SORR significantly.

Proposition 6. *The lower bounds of $\frac{\gamma}{N}\sum_{i\in\mathcal{V}}\mu_i^{-1}$ given in* (29) *and* (30) *satisfy the following inequality:*

$$\hat{\alpha}_{\text{local}} \ge \hat{\alpha}_{\text{global}}. \tag{31}$$

[5]A similar tendency has already been witnessed in a centralized case [32].

Proof: The claim can be verified in an analogous way to the proof of Proposition 3. ■

Hereafter, we let $m_1 = m_2 = \cdots = m_N =: m$.

Remark 4 (On the behaviour of the lower bounds α_{local}, α_{global}, $\hat{\alpha}_{\text{local}}$, and $\hat{\alpha}_{\text{global}}$ for different network sizes N)**.** *Let us consider the situation where the total amount of measurements over the network is fixed to some constant, say $mN = 100$. For instance, the network size $N = 100$ gives $m = 1$ (meaning that each node is given one equation), while $N = 2$ gives $m = 50$ so that the system is closer to the centralized case. We assume that the components of the matrix $\boldsymbol{A}$ are generated from an i.i.d. zero-mean Gaussian distribution. A simple inspection of* (17) *and* (30) *suggests that, given a set of measurements, α_{global} and $\hat{\alpha}_{\text{global}}$ are independent of N, because $\boldsymbol{A}^\top \boldsymbol{A}$ can be expressed as a sum of $mN = 100$ rank-one matrices. In contrast, α_{local} and $\hat{\alpha}_{\text{local}}$ does depend on N, as seen from* (16) *and* (29)*. In particular, in the regime of large network sizes N (implying small m) α_{local} and $\hat{\alpha}_{\text{local}}$ grow almost linearly in N (for mN constant), as each summand $\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)$ is nearly constant in m in this regime. This is because, due to the statistical assumption on $\boldsymbol{A}$, the m rows of each $\boldsymbol{A}_i$ tend to be nearly orthogonal to each other, implying that $\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)$ is roughly given by the largest norm of the row vectors, which is nearly constant on average. Those arguments will be justified by simulations in Section V-B.*

How does the lower bounds α_{local} and $\hat{\alpha}_{\text{local}}$ affect the performance of D-ORR and D-SORR? As mentioned above, α_{local} and $\hat{\alpha}_{\text{local}}$ tend to increase in N, which makes $\frac{\gamma}{N}\sum_{i\in\mathcal{V}} \mu_i^{-1}$ be larger. The saturation factor γ needs to be reasonably small for outlier robustness, implying that the global regularization parameter $\frac{1}{N}\sum_{i\in\mathcal{V}} \mu_i^{-1}$ needs to be large. This causes larger estimation biases as will be shown in Section V-B.

Remark 5 (On the consistency of the D-SORR estimator)**.** *As we are interested in the finite-sample regime, we leave the issue of statistical consistency of the D-SORR estimator as a future work (i.e., we do not touch the issue whether the estimator converges to $\boldsymbol{x}^\star$ in probability as the sample size mN tends to infinity). We mention that the performance of ridge regression (which is well known to be a consistent estimator) is highly sensitive to outliers [3]. In stark contrast, D-SORR exhibits remarkably robust performance against huge and relatively-dense outliers, as shown by simulations later on (see Section VI).*

In the large sample-size regime, there are two options for preserving convexity according to (22)*: (i) increase the saturation factor γ in proportion to the increase of $N\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)$ with the regularization parameter μ_i fixed, or (ii) increase μ_i^{-1} in proportion to the increase of $N\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)$ with γ fixed. Note here that $N\lambda_{\max}(\boldsymbol{A}_i^\top \boldsymbol{A}_i)$ is roughly proportional to the sample size mN (for large enough m). To see how each option works in the asymptotic regime, we divide the objective function in* (2) *by mN. Then, in option (i), the effect of $1/(mN)\sum_{i\in\mathcal{V}} \sigma_x^{-2}\|\boldsymbol{x}\|_2^2/(\mu_i N) = (\sigma_x^{-2}/\mu_i)\|\boldsymbol{x}\|_2^2/(mN)$ diminishes as $mN \to \infty$, while $1/(mN)\sum_{i\in\mathcal{V}} \Phi_\gamma^{\text{MC}}(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{\varepsilon}_i - \boldsymbol{y}_i)$ approximates $E(\|\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i\|_1)$ by the law of large numbers under proper assumptions. We mention that the noise term does not vanish because the number of summands increases in mN. The MC function tends to the ℓ_1 norm as $\gamma \to +\infty$ so that the overall objective is convex without the strongly convex term $(\sigma_x^{-2}/\mu_i)\|\boldsymbol{x}\|_2^2$ in the limit. In option (ii), in contrast, the increase of μ_i^{-1} (e.g., set $\mu_i := c/(mN)$ for some constant $c > 0$) preserves the effect of $(\sigma_x^{-2}/\mu_i)\|\boldsymbol{x}\|_2^2$ even in the asymptotic regime. In addition, the MC function remains its shape in the limit, as opposed to the case of option (i). It is nontrivial to say which option works better, and further investigations will be required for that issue.*

V. Simulation Studies I — Basic Performance

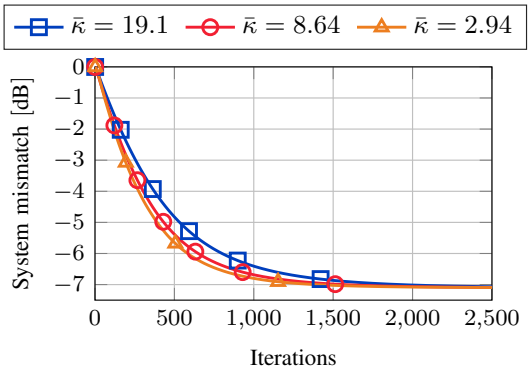
Each local matrix $\boldsymbol{A}_i \in \mathbb{R}^{m\times n}$ and the unknown vector $\boldsymbol{x}_\star \in \mathbb{R}^n$ follow the i.i.d. standard Gaussian distribution. The noise vectors $\boldsymbol{\varepsilon}_{i\star}$ are generated by scaling those temporary vectors according to SNR $:= \sum_{i\in\mathcal{V}}\|\boldsymbol{A}_i\boldsymbol{x}_\star\|_2^2/(\sum_{i\in\mathcal{V}}\|\boldsymbol{\varepsilon}_{i\star}\|_2^2)$, where the temporary vectors are generated from the i.i.d. standard Gaussian distribution. The positions of the nonzero elements of $\boldsymbol{o}_{i\diamond}$ are chosen randomly, and the nonzero values follow an i.i.d. scaled and shifted uniform distribution. Here, for all simulations, given the specified value $\overline{M}_{o_\diamond} > 0$, the interval of the uniform distribution is set to $d_{\text{uniform}} := 2\overline{M}_{o_\diamond}/9$ with its center $M_{o_\diamond}$ chosen randomly again from another uniform distribution with center and interval given by $\overline{M}_{o_\diamond}$ and d_{uniform}, respectively. In most simulations, we set $\overline{M}_{o_\diamond} := 90$, meaning that the outliers come from the interval of width $d_{\text{uniform}} = 20$ with its center chosen randomly between 80 and 100 at each independent run.

We use the system mismatch $\frac{1}{N}\sum_{i\in\mathcal{V}}\|\boldsymbol{x}_i - \boldsymbol{x}_\star\|_2^2 / \|\boldsymbol{x}_\star\|_2^2$ as our primary performance measure. Unless stated explicitly, all plots in the figures presented in this section show the averages over 250 independent runs. For D-SORR, we set $\sigma_x^2 := \sigma_{x\star}^2 := 1$ and $\sigma_\varepsilon^2 := \sigma_{\varepsilon\star}^2 := \frac{1}{mN}\sum_{i\in\mathcal{V}}\|\boldsymbol{\varepsilon}_{i\star}\|_2^2$. For the design of the saturation factor γ and the regularization parameters μ_i for D-ORR and D-SORR, see Remarks 1 and 3.

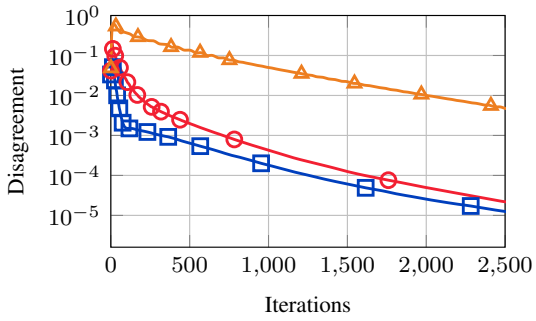
A. Impacts of connectivity per node

What is the impact of the network connectivity per node on the performance of the distributed optimization algorithm? This is the question addressed in this part. We let $n := 30$, $N := 100$, $m := 1$, SNR $:= 10$ dB, $\overline{M}_{o_\diamond} := 90$, and the outlier density is 0.3. We test D-SORR with τ_i set to the upper bound given in (26); see Section IV-B for the design of the other parameters.

Fig. 1 shows the learning curves for different degrees of connectivity for (a) system mismatch and (b) disagreement $\frac{1}{N}\sum_{i\in\mathcal{V}}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2$, where $\bar{\boldsymbol{x}}$ is the arithmetic mean of the $\boldsymbol{x}_i$'s over all nodes. The degree of connectivity is measured by the average number of connections per node $\bar{\kappa}$ in the network. The errors are averaged over all nodes at each time instance. It can be seen that an increase in the connectivity enhances the convergence speed in terms of disagreement, as expected, while it slightly slows the convergence speed in system mismatch. We mention that it would be possible to tune the link weights κ_{ij} so that the convergence speeds become the same among different degrees of connectivity.



(a) System mismatch averaged over all nodes



(b) Disagreement among nodes

Fig. 1: Convergence speed of the TriPD-Dist algorithm in terms of (a) system mismatch and (b) consensus disagreement among nodes.

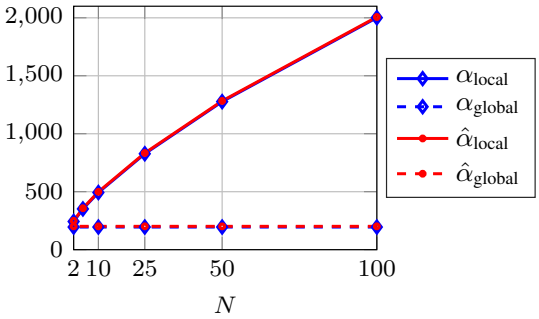


Fig. 2: The lower bounds α_{local}, α_{global}, $\hat{\alpha}_{\text{local}}$, and $\hat{\alpha}_{\text{global}}$ across the network size N with $m = 100/N$.

B. Comparisons between convexity conditions for local and global objectives

We justify the arguments in Remark 4 by simulations. We let $n := 30$ and SNR $:= 5$ dB with the total number of equations fixed to $mN = 100$ over the entire network, where N changes from 2 to 100. We first examine how the lower bounds α_{local}, α_{global}, $\hat{\alpha}_{\text{local}}$, and $\hat{\alpha}_{\text{global}}$ behave as the network size N increases (or, equivalently, the number m of equations per node decreases).

Fig. 2 depicts the results, where each plot is computed by averaging the results over 1000 independent runs. As argued in Remark 4, α_{local} increases almost linearly in the network size N, while α_{global} remains constant. Focusing on the extreme case of $N = 100$, in particular, one can see

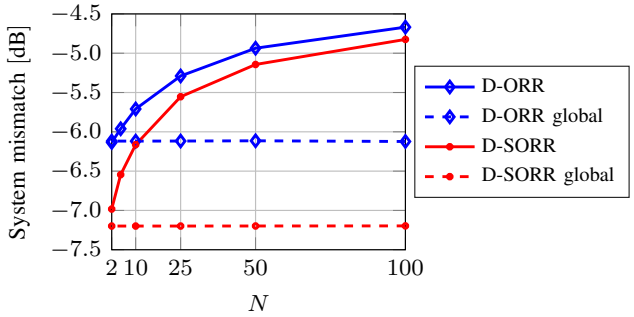


Fig. 3: System mismatch across the network size N with $m = 100/N$. SNR $= 5$ dB, $\overline{M}_{o_\diamond} = 90$, outlier density 0.3, and $n = 30$.

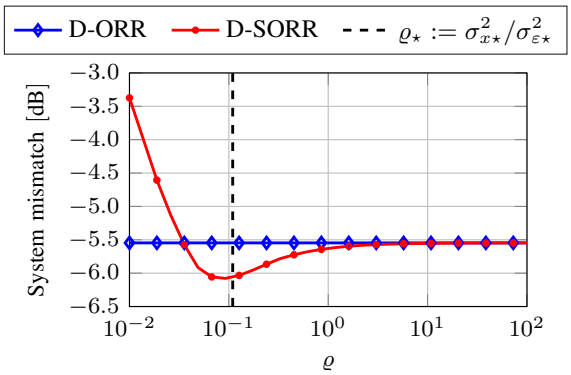


Fig. 4: Performance of D-SORR for different ϱ in comparison with D-ORR for $N = 10$, $m = 10$, SNR $= 5$ dB, $\overline{M}_{o_\diamond} = 90$, outlier density 0.3, and $n = 30$.

the notable difference between α_{local} and α_{global}. Recall that the bounds α_{local}, α_{global}, $\hat{\alpha}_{\text{local}}$, and $\hat{\alpha}_{\text{global}}$ of the quantity $\frac{\gamma}{N}\left(\sum_{i\in\mathcal{V}} \mu_i^{-1}\right)$ are desired to be small for reduced bias and/or for outlier robustness. In the above extreme case, there would be a considerable difference between the performances corresponding to the local and global convexity conditions.

To verify this, we study the performance of D-ORR and D-SORR under the two different convexity conditions, respectively. Fig. 3 shows that the system mismatch increases monotonically in the network size N for D-ORR and D-SORR with the local convexity condition. The system mismatch is low when N is small (i.e., the algorithm is more centralized) because μ_i^{-1} is allowed to be small (see Fig. 2), while it becomes higher when N becomes larger because μ_i^{-1} needs to be larger. We clarify here that the strength of the regularization changes in N when the local convexity condition is used to compute the regularization parameters μ_i, although the global loss is independent of N given the fixed total number mN of equations over the entire network. Viewing the performance corresponding to the global convexity condition (labeled as "D-ORR global" and "D-SORR global") in Fig. 3, one can see that the system mismatch stays constant, because the strength of the regularization remains the same owing to the constancy of α_{global} and $\hat{\alpha}_{\text{global}}$.

Despite the nice property of the global convexity condition shown above, at least one node (a central node) needs access to

all data in the network to use it for designing the parameters, which is prohibited in some applications for privacy reasons. (One may encode and send such information, but this would require extra computational/communication costs.) Since, in addition, convergence to a solution is not guaranteed under global convexity, the local convexity condition will be used in the following simulations.

Fig. 4 shows the system mismatch of D-SORR for different values of the tuning parameter ϱ. The dashed black line indicates the value $\varrho_\star (:= \sigma_{x\star}^2/\sigma_{\varepsilon\star}^2)$, which is used in all other simulations. For comparison, the system mismatch of D-ORR is shown, which is independent of ϱ. The other tuning parameter γ is optimized by grid search for both D-ORR and D-SORR. It is observed that D-SORR is rather robust to the choice of ϱ in the direction of $\varrho > \varrho_\star$, as the performance of D-SORR is always equal or better compared to D-ORR. However, for values of $\varrho < \varrho_\star$, the performance of D-SORR quickly becomes worse than the performance of D-ORR.

VI. Simulation Studies II — Comparisons to Existing Methods

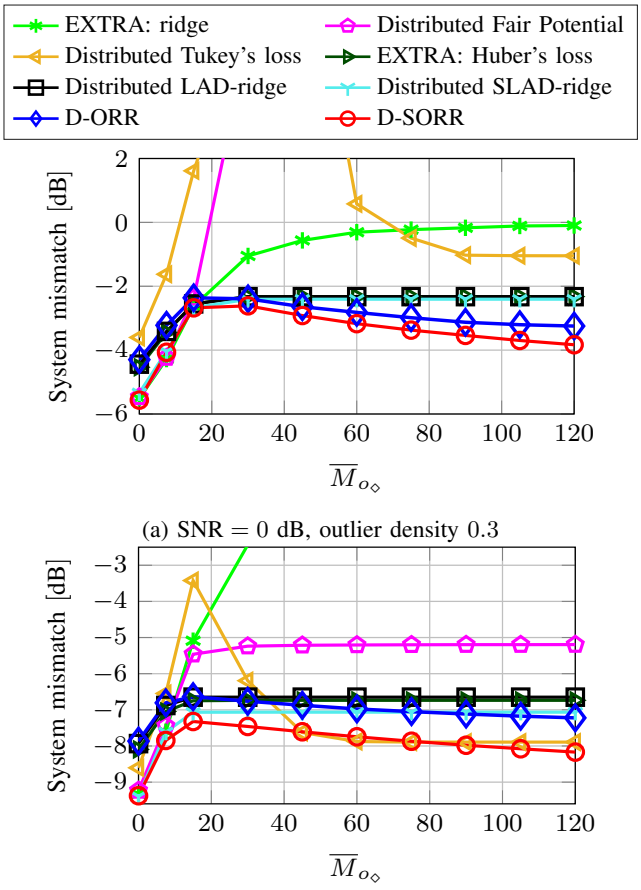
We show the advantages of the proposed methods over the existing methods in terms of outlier robustness. After showing the simulation results using toy data under various scenarios, we present the results for real and synthetic data to show that the proposed method will be useful potentially in real-world applications.

A. Toy data

The signals are generated in the same way as described in Section V. The proposed D-ORR and D-SORR methods are compared to the following robust loss functions for positive constants $\delta_{\mathrm{L}}, \delta_{\mathrm{S}_x}, \delta_{\mathrm{S}_\varepsilon} \delta_{\mathrm{H}}, \delta_{\mathrm{T}}, \delta_{\mathrm{P}} > 0$: LAD-ridge $\sum_{i\in\mathcal{V}} \|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i\|_1 + \delta_{\mathrm{L}} \|\boldsymbol{x}\|_2^2$, stable LAD-ridge (SLAD-ridge) $\sum_{i\in\mathcal{V}} \|\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{\varepsilon}_i - \boldsymbol{y}_i\|_1 + \delta_{\mathrm{S}_x} \|\boldsymbol{x}\|_2^2 + \delta_{\mathrm{S}_\varepsilon} \|\boldsymbol{\varepsilon}_i\|_2^2$ in the same philosophy of employing $\boldsymbol{\varepsilon}_i$ as D-SORR, Huber's loss $\sum_{i\in\mathcal{V}} {}^{\delta_{\mathrm{H}}}\|\cdot\|_1 (\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i)$, Tukey's biweight loss [2], [54] $\sum_{i\in\mathcal{V}} \sum_{\iota=1}^m \phi_{\delta_{\mathrm{T}}}^{\mathrm{TK}}([\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i]_\iota)$, where

$$\phi_{\delta_{\mathrm{T}}}^{\mathrm{TK}} : \mathbb{R} \ni a \mapsto \begin{cases} \left[1 - \left(1 - (a/\delta_{\mathrm{T}})^2\right)^3\right] \delta_{\mathrm{T}}^2/6, & \text{if } |a| < \delta_{\mathrm{T}}, \\ \delta_{\mathrm{T}}^2/6, & \text{otherwise,} \end{cases}$$

and the fair potential function [6], [55] $\sum_{i\in\mathcal{V}} \sum_{\iota=1}^m \phi_{\delta_{\mathrm{P}}}^{\mathrm{FP}}([\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i]_\iota)$, where $\phi_{\delta_{\mathrm{P}}}^{\mathrm{FP}} : \mathbb{R} \ni a \mapsto \delta_{\mathrm{P}} |a| - \log_{10}(1 + \delta_{\mathrm{P}} |a|)$. Here, $[\cdot]_\iota$ denotes the ιth component of a vector. For reference, the ridge regression $\sum_{i\in\mathcal{V}} \|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i\|_2^2 + \delta_{\mathrm{R}} \|\boldsymbol{x}\|_2^2$, $\delta_{\mathrm{R}} > 0$, is also tested. Comparing SLAD-ridge and D-SORR, both involve the auxiliary vectors $\boldsymbol{\varepsilon}_i$. The difference is, however, that SLAD-ridge utilizes the ℓ_1 norm, whereas D-SORR employs the MC loss function. This implies that SLAD-ridge can be considered as an extreme case of D-SORR as $\gamma \to \infty$. This comparison is explicitly conducted in the experiments to assess the impact of the MC loss and that of the auxiliary vector formulation separately.



(a) SNR $= 0$ dB, outlier density 0.3

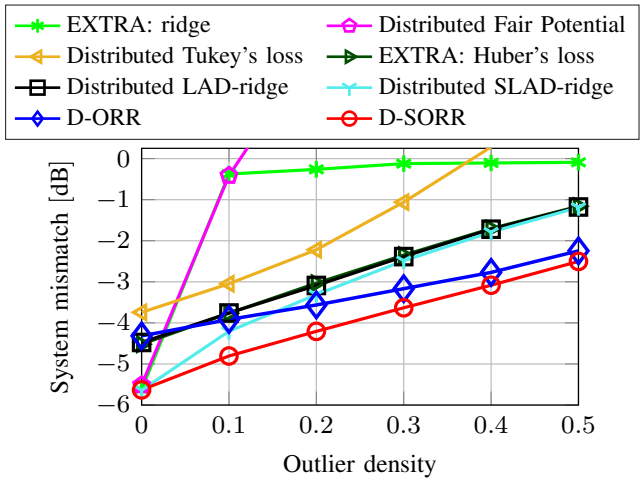
(b) SNR $= 5$ dB, outlier density 0.1

Fig. 5: System mismatch across $\overline{M}_{o_\diamond}$.

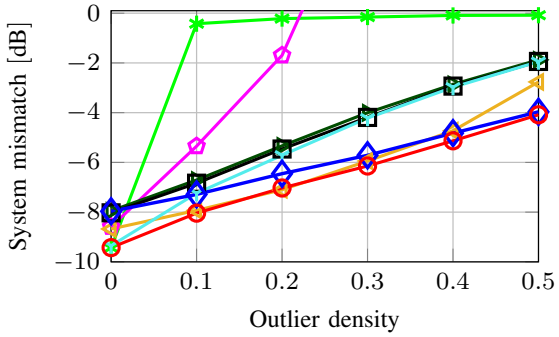
For each method, the delta parameters are tuned by grid search to minimize the system mismatch. Unless stated otherwise, in the following simulations, we consider the "middle" case when the network has $N := 10$ nodes, each of which is given $m := 10$ measurement vectors of dimension $n := 30$.

Fig. 5 shows the performance across $\overline{M}_{o_\diamond}$ for (a) SNR 0 dB with outlier density 0.3 and (b) SNR 5 dB with outlier density 0.1, where larger $\overline{M}_{o_\diamond}$ means larger outlier power. There is remarkably different tendency between the convex and nonconvex approaches. Specifically, in contrast to the monotone behaviors of the convex methods, the nonconvex methods (D-ORR, D-SORR, and distributed Tukey's loss) show "non-monotonic" behaviors — the system mismatch increases up to some point, and it then decreases as the outlier power increases. See Section VI-C for more discussions about this phenomenon.

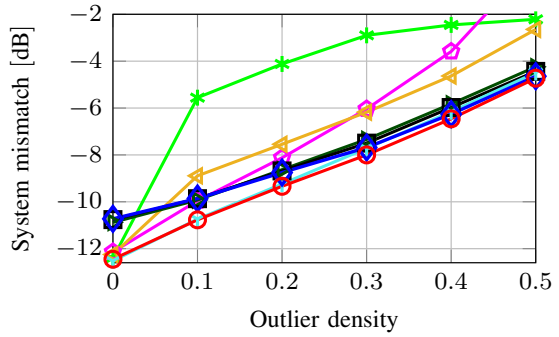
Fig. 6 shows the system mismatch across different outlier densities from 0 to 0.5 under different SNRs, different magnitudes of outliers $\overline{M}_{o_\diamond}$, and different numbers of variables n. It can be seen that D-SORR outperforms the other methods. To distinguish the impacts of the auxiliary vectors (aiming at robustness against Gaussian noise) and the MC loss (aiming at outlier robustness) employed in D-SORR, let us compare LAD-ridge with SLAD-ridge (employing the auxiliary vectors) and with D-ORR (employing the MC function). The SLAD-



(a) SNR $= 0$ dB, $\overline{M}_{o\diamond} = 90$, and $n = 30$



(b) SNR $= 5$ dB, $\overline{M}_{o\diamond} = 90$, and $n = 30$



(c) SNR $= 3$ dB, $\overline{M}_{o\diamond} = 15$, and $n = 10$

Fig. 6: System mismatch across the outlier density.

ridge formulation achieves a lower system mismatch compared to LAD-ridge when there are no or few outliers (outlier density of 0 to 0.1), indicating its effectiveness when Gaussian noise is dominant. On the other hand, D-ORR achieves a lower system mismatch compared to LAD-ridge when there is a large number of outliers (outlier density ≥ 0.2). By combining those advantages of explicit Gaussian noise modeling and robustness to outliers with the MC loss function, D-SORR outperforms the other methods across all outlier densities.

Fig. 7a shows the performance under different levels of Gaussian noise, and Figs. 7b and 7c show the performance with different amounts of measurements available at each node given the fixed network size $N = 10$. (For instance, 20 equations per node means in total 200 equations distributed over ten nodes.) Overall, the proposed method outperforms the other methods significantly. The only exception is the particular case of SNR $= 10$ dB in Fig. 7a and $m = 50$ in Fig. 7b for which "Distributed Turkey's loss" gives slightly better performance than the proposed method. In this case, however, each node has a reasonable amount of information to find a good estimate because the number $m = 50$ of measurements per node is larger than the number $n = 30$ of variables to optimize. For distributed optimization, the case of smaller m (e.g., the case of $m < n$ specifically) is of particular interest, because information exchange among nodes is crucial in such a case to obtain better estimates than using only the local information.

B. Real and synthetic data

We consider the source estimation task in an atmospheric inverse problem [56], which is a real-world example of a regression problem in which those data measured by a network of sensors are contaminated by outliers. For the European Tracer Experiment (ETEX), a tracer gas was released in Monterfil, France, and the gas concentration was measured every three hours for three days by 168 measurement stations across Europe [57]. The task is to predict the release time of the gas by using a linear particle dispersion model, which uses meteorological data of the whole duration of the experiment. The atmospheric inverse problem at each node i can then be written in the form of $\boldsymbol{y}_i = \boldsymbol{A}_i\boldsymbol{x}_i$, where $\boldsymbol{A}_i$ is the linear particle dispersion model, $\boldsymbol{y}_i$ is the gas concentration measurement, and $\boldsymbol{x}_i$ is the predicted time of the gas release. The gas concentration dataset and the linear particle dispersion model used in the present study were made available in the supplemental material of [58]. In the present study, the dataset of the first experiment (ETEX-1) is used, as well as three synthetic datasets made available in [58]. These synthetic datasets are qualitatively similar to the real world dataset, but with a smaller number of parameters and a normalized particle dispersion matrix.

For the simulation, each measurement station from the real world dataset is modeled as one node in the distributed network. All such data points are removed that have all entries of the particle dispersion model and the gas measurement be zero. The total number of data points is 1810 with a varying number of data points per node, and $\boldsymbol{x} \in \mathbb{R}^{112}$. Furthermore, the particle dispersion matrix $\boldsymbol{A}_i$ at each node is normalized by a global scaling factor, which is the largest absolute value entry of all $\boldsymbol{A}_i$'s. When evaluating the system mismatch of the different methods, the result is scaled to the best fit of the true amount of released gas, meaning that only the relative amount of released gas at each time instance is estimated. For the synthetic dataset, no additional normalization and no scaling of the solution is performed, and the shown result is averaged over the three synthetic datasets. For the synthetic dataset, there are five nodes, each of which has four data points, and $\boldsymbol{x} \in \mathbb{R}^{10}$. For both the synthetic datasets and the real-world dataset, the parameters of all methods are optimized by grid search.

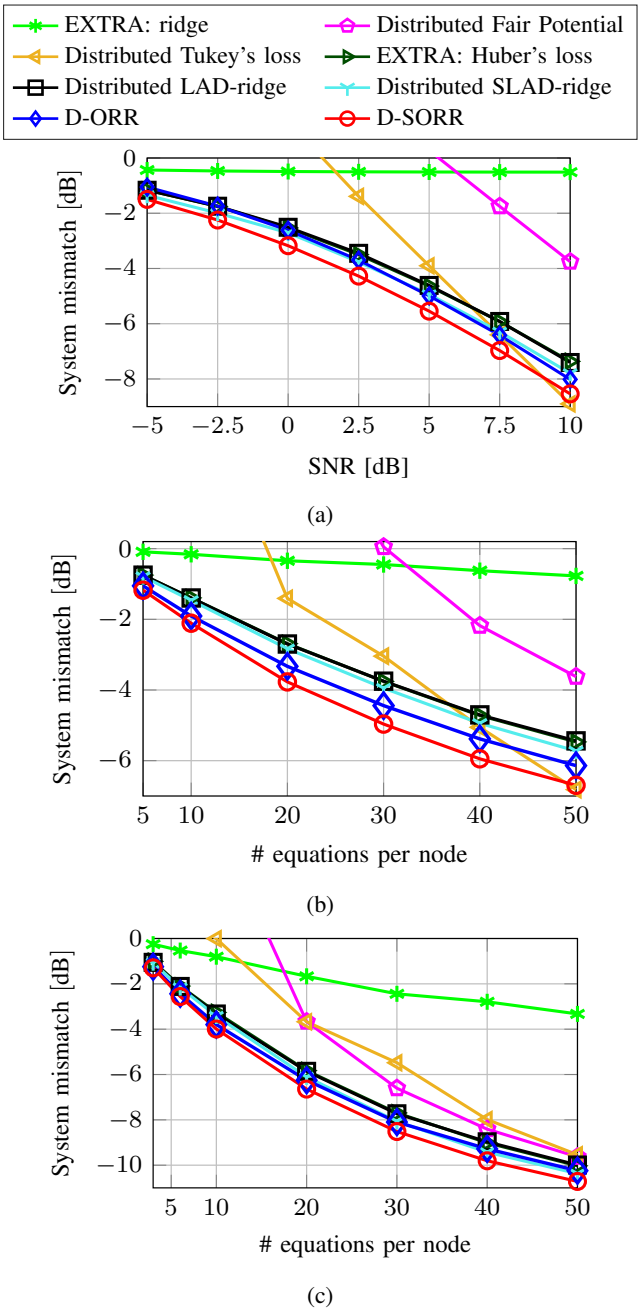


Fig. 7: System mismatch across (a) SNR for outlier density 0.1, $m = 10$, $\overline{M}_{o_\diamond} = 90$, and $n = 50$, (b) the number m of equations per node for outlier density 0.3, SNR $= 0$ dB, $N = 10$, $\overline{M}_{o_\diamond} = 90$, and $n = 50$, and (c) the number m of equations per node for outlier density 0.3, SNR $= 3$ dB, $N = 10$, $\overline{M}_{o_\diamond} = 35$, and $n = 30$.

Fig. 8a shows the system mismatch of the the real-world dataset, and Fig. 8b shows the result of the synthetic datasets. It can be seen that D-SORR outperforms all other methods in both cases.

C. Discussion

The remarkable phenomenon of "non-monotonic" behaviors mentioned in Section VI-A leads us to the hypothesis that the

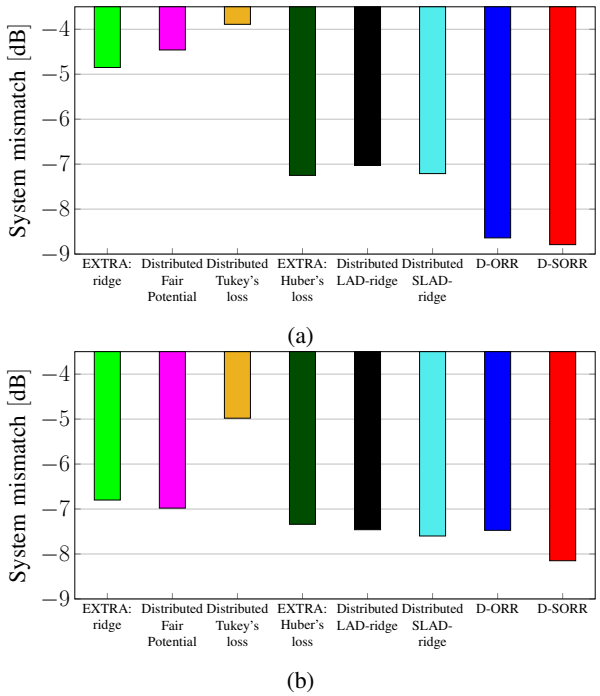


Fig. 8: System mismatch for a simulation (a) with a real world dataset from the European tracer experiment, and (b) with three synthetic datasets.

nonconvex methods implicitly carry out outlier detection. The MC function $\phi_\gamma^{\mathrm{MC}}(x)$ is constant above/below the saturation points which are determined by the parameter γ. (The same applies to Tukey's biweight loss.) Hence, when the residual error is sufficiently large for an outlier measurement, the gradient of the loss function vanishes, and the outlier is rejected accordingly. This would explain why the performance does not degrade for an increase in outlier power, but it does not explain the improvements of the performance.

Intuitively, the detection task is rather difficult when the outlier power is not significantly larger than that of the normal data, while outliers can be identified easily when the outlier power is extremely large. More specifically, larger outliers increase the probability that the magnitude of the error for each outlier measurement lies near or exceeds a fixed threshold γ, meaning that the subgradient of the MC penalty is small for each outlier measurement, and thus has a small impact on the solution. Additionally, in our preliminary experiments it is observed that the optimal threshold γ increases when larger outliers occur, which allows the use of larger regularization parameters μ_i to reduce the bias, and thus decrease the system mismatch further for larger outliers. One may think that γ could be reduced to increase the opportunity of outlier rejection. This, however, makes the regularization parameters μ_i be smaller to satisfy the convexity condition, thereby strengthening the regularization effects undesirably to cause performance degradation. The remarkable robustness of D-SORR discussed above is quite advantageous, because accurate estimates can be obtained in the presence of devastating outliers.

VII. Conclusion

This paper presented the D-ORR and D-SORR formulations for distributed robust signal recovery. Thanks to the weak convexity of the MC loss, the proposed formulations enjoy the two desirable properties simultaneously: (i) significantly high robustness against outliers, and (ii) guarantee of convergence to a solution under convexity of the local objectives. The D-SORR formulation involved an auxiliary vector to model the Gaussianity of noise as well as outliers. We showed the conditions to guarantee convexity of the local and global objectives, respectively, for each formulation. We also showed that the TriPD-Dist algorithm applied to the reformulated versions of the D-ORR and D-SORR problems enjoys linear convergence to a minimizer of each objective under the local convexity condition. The numerical examples showed that our proposed formulations exhibited remarkable robustness under huge outliers as well as outperforming the existing methods. The global convexity condition gave better performance than the local one, but it has the drawback that all data need to be collected at a central node to compute the regularization parameter based on the global convexity condition. It remains an open issue whether the global convexity condition is sufficient to guarantee convergence to a solution.

References

[1] M. H. V. Tillmann and M. Yukawa, "Distributed stable outlier-robust signal recovery using minimax concave loss," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2023, pp. 1–6.

[2] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, May 1974.

[3] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley, 2009.

[4] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: the Approach Based on Influence Functions*. John Wiley & Sons, 2011, vol. 196.

[5] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons, 2019.

[6] J. Li, E. Elhamifar, I.-J. Wang, and R. Vidal, "Consensus with robustness to outliers via distributed optimization," in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 2111–2117.

[7] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations Research*, vol. 43, no. 2, pp. 264–281, Mar.–Apr. 1995.

[8] S. Kar and J. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Selected Topics in Signal Process.*, vol. 5, no. 4, pp. 674–690, 2011.

[9] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge: Cambridge University Press, 2018.

[10] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 2nd ed. London: Academic Press, 2020.

[11] R. Tron and R. Vidal, "Distributed computer vision algorithms," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 32–45, 2011.

[12] H. Wang and C. Li, "Distributed quantile regression over sensor networks," *IEEE Trans. Signal and Inform. Process. over Netw.*, vol. 4, no. 2, pp. 338–348, 2018.

[13] J. Hua and C. Li, "Distributed robust bayesian filtering for state estimation," *IEEE Trans. Signal and Inform. Process. over Netw.*, vol. 5, no. 3, pp. 428–441, 2019.

[14] G. Rajesh and A. Chaturvedi, "Data reconstruction in heterogeneous environmental wireless sensor networks using robust tensor principal component analysis," *IEEE Trans. Signal and Inform. Process. over Netw.*, vol. 7, pp. 539–550, 2021.

[15] S. Modalavalsa, U. K. Sahoo, A. K. Sahoo, and S. Kumar, "Diffusion minimum generalized rank norm over distributed adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal and Inform. Process. over Netw.*, vol. 5, no. 4, pp. 669–683, 2019.

[16] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[17] I. Matei and J. S. Baras, "Performance evaluation of the consensus-based distributed subgradient method under random communication topologies," *IEEE J. Selected Topics in Signal Process.*, vol. 5, no. 4, pp. 754–771, 2011.

[18] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Opt.*, vol. 25, no. 2, pp. 944–966, 2015.

[19] ——, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015.

[20] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. Opt.*, vol. 27, no. 4, pp. 2597–2633, 2017.

[21] P. Latafat, N. M. Freris, and P. Patrinos, "A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization," *IEEE Trans. Automatic Control*, vol. 64, no. 10, pp. 4050–4065, 2019.

[22] H. Li, L. Zheng, Z. Wang, Y. Yan, L. Feng, and J. Guo, "S-DIGing: A stochastic gradient tracking algorithm for distributed optimization," *IEEE Trans. Emerging Topics in Computational Intelligence*, 2020.

[23] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[24] E. J. Candes and P. A. Randall, "Highly robust error correction by convex programming," *IEEE Trans. Inform. Theory*, vol. 54, no. 7, pp. 2829–2840, 2008.

[25] N. H. Nguyen and T. D. Tran, "Robust lasso with missing and grossly corrupted observations," *IEEE Trans. Inform. Theory*, vol. 59, no. 4, pp. 2036–2058, 2013.

[26] K. Yang, J. Huang, Y. Wu, X. Wang, and M. Chiang, "Distributed robust optimization (DRO), part I: framework and example," *Optim. Eng.*, vol. 15, pp. 35–67, 2014.

[27] S. Wang and C. Li, "Distributed robust optimization in networked system," *IEEE Trans. Cybernetics*, vol. 47, no. 8, pp. 2321–2333, 2017.

[28] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[29] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4481–4494, 2017.

[30] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition," *Inverse Problems*, vol. 36, no. 3, p. 035012, 2020.

[31] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *Proc. IEEE Int. Symp. Inf. Theory*, 2010, p. 1518–1522.

[32] M. Yukawa, K. Suzuki, and I. Yamada, "Stable robust regression under sparse outlier and Gaussian noise," in *Proc. EUSIPCO*, 2022, pp. 2236–2240.

[33] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, "Linearly-involved Moreau-enhanced-over-subspace model: Debiased sparse modeling and stable outlier-robust regression," *IEEE Trans. Signal Process.*, vol. 71, pp. 1232–1247, 2023.

[34] K. Komuro, M. Yukawa, and R. L. G. Cavalcante, "Distributed sparse optimization with weakly convex regularizer: Consensus promoting and approximate Moreau enhanced penalties towards global optimality," *IEEE Trans. Signal and Inform. Process. over Netw.*, vol. 8, pp. 514–527, 2022.

[35] H. Bauschke and Y. Lucet, "What is a fenchel conjugate," *Notices of the AMS*, vol. 59, no. 1, pp. 44–46, 2012.

[36] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Advances in Computational Mathematics*, vol. 38, no. 3, pp. 667–681, 2013.

[37] P. Latafat and P. Patrinos, "Primal-dual proximal algorithms for structured convex optimization: A unifying framework," in *Large-Scale and Distributed Optimization*. Springer, 2018, pp. 97–120.

[38] ——, "Asymmetric forward backward adjoint splitting for solving monotone inclusions involving three operators," *Computational Opt. Appl.*, vol. 68, no. 1, pp. 57–93, 2017.

[39] L. Condat, G. Malinovsky, and P. Richtárik, "Distributed proximal splitting algorithms with rates and acceleration," *Frontiers in Signal Process.*, vol. 1, pp. 1–18, Jan. 2022.

[40] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Opt. Theory and Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[41] P. L. Combettes, L. Condat, J.-C. Pesquet, and B. Vũ, "A forward-backward view of some primal-dual optimization methods in image recovery," in *Proc. IEEE ICIP*, 2014, pp. 4141–4145.
[42] T. Chang, M. Hong, H. Wai, X. Zhang, and S. Lu, "Distributed learning in the nonconvex world: From batch data to streaming and beyond," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 26–38, 2020.
[43] H. Yang, P. Antonante, V. Tzoumas, and L. Carlone, "Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1127–1134, 2020.
[44] Y. Tian, Y. Chang, F. Herrera Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2022–2038, 2022.
[45] A. Beznosikov, G. Scutari, A. Rogozin, and A. Gasnikov, "Distributed saddle-point problems under data similarity," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8172–8184.
[46] M. I. Qureshi and U. A. Khan, "Distributed saddle point problems for strongly concave-convex functions," *IEEE Transactions on Signal and Information Processing over Networks*, 2023.
[47] R. Xin, S. Pu, A. Nedić, and U. A. Khan, "A general framework for decentralized optimization with first-order methods," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 1869–1889, 2020.
[48] P. Di Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.
[49] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 3029–3068, 2020.
[50] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York: NY: Springer, 2017.
[51] I. Yamada, M. Yukawa, and M. Yamagishi, *Minimizing Moreau envelope of nonsmooth convex function over the fixed point set of certain quasi-nonexpansive mappings*, ser. Optimization and Its Applications. in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, vol. 49, pp. 345–390.
[52] N. Perraudin, D. Shuman, G. Puy, and P. Vandergheynst, "Unlocbox a matlab convex optimization toolbox using proximal splitting methods," *ArXiv e-prints*, Feb. 2014.
[53] U. Helmke and J. Rosenthal, "Eigenvalue inequalities and Schubert calculus," *Mathematische Nachrichten*, vol. 171, no. 1, pp. 207–225, 1995.
[54] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *Int. J. Computer Vision*, vol. 19, no. 1, pp. 57–91, 1996.
[55] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems," *Inverse Problems*, vol. 23, no. 4, pp. 1495–1518, June 2007.
[56] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust statistics for signal processing*. Cambridge University Press, 2018.
[57] K. Nodop, R. Connolly, and F. Girardi, "The field campaigns of the european tracer experiment (etex): Overview and results," *Atmospheric Environment*, vol. 32, no. 24, pp. 4095–4108, 1998.
[58] O. Tichỳ, V. Šmídl, R. Hofman, and A. Stohl, "Ls-apc v1. 0: a tuning-free method for the linear inverse problem and its application to source-term determination," *Geoscientific Model Development*, vol. 9, no. 11, pp. 4297–4311, 2016.
[59] K. Suzuki and M. Yukawa, "Robust recovery of jointly-sparse signals using minimax concave loss function," *IEEE Trans. Signal Process.*, vol. 69, pp. 669–681, 2020.
[60] A. Lanza, S. Morigi, I. W. Selesnick, and F. Sgallari, "Sparsity-inducing nonconvex nonseparable regularization for convex image processing," *SIAM J. Imaging Sci.*, vol. 12, no. 2, pp. 1099–1134, 2019.
[61] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*. New York: Springer Verlag, 2009.

APPENDIX A
PROOF OF PROPOSITION 1

The proof of Proposition 1 is given based on the following theorem.

Theorem A.1 (The Euclidean case of Corollary 1 in [33])**.** *Let Ψ be a norm defined on $\mathbb{R}^m$. Let $\boldsymbol{L} \in \mathbb{R}^{m\times m}$, $\boldsymbol{D} \in \mathbb{R}^{m\times m}$ be diagonal positive definite matrices, and $\mathscr{A}_1 : \mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{x} \mapsto \boldsymbol{M}_1\boldsymbol{x} + \boldsymbol{c}_1$ and $\mathscr{A}_2 : \mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{x} \mapsto \boldsymbol{M}_2\boldsymbol{x} + \boldsymbol{c}_2$ be affine operators with $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{m\times n}$ and $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathbb{R}^m$. Assume that $\boldsymbol{M}_2$ has full column rank or that $\mathscr{A}_2\boldsymbol{x} = \boldsymbol{0}_m$ for some $\boldsymbol{x} \in \mathbb{R}^n$. Then, for $\mu > 0$, the function*

$$F = \frac{1}{2}\left\|\cdot\right\|_2^2 \circ \mathscr{A}_1 - \frac{\mu}{2}\left\|\cdot\right\|_2^2 \circ \boldsymbol{D}\boldsymbol{L}\mathscr{A}_2 + \mu\, {}^{1}(\Psi^* \circ \boldsymbol{D}) \circ \boldsymbol{D}\boldsymbol{L}\mathscr{A}_2 \tag{A.1}$$

is convex if and only if

$$\boldsymbol{M}_1^\top \boldsymbol{M}_1 - \mu \boldsymbol{M}_2^\top \boldsymbol{L}^\top \boldsymbol{D}^2 \boldsymbol{L} \boldsymbol{M}_2 \succeq \boldsymbol{O}_n. \tag{A.2}$$

Proof of Proposition 1: (a) The classical Moreau decomposition [50] allows to rewrite the MC penalty as [33], [59]

$$\Phi_\gamma^{\text{MC}}(\boldsymbol{x}) = \left\|\boldsymbol{x}\right\|_1 + {}^{\gamma^{-1}}\big(\left\|\cdot\right\|_1^*\big)\left(\gamma^{-1}\boldsymbol{x}\right) - \frac{1}{2\gamma}\left\|\boldsymbol{x}\right\|_2^2. \tag{A.3}$$

Note here that ${}^{\gamma^{-1}}(\left\|\cdot\right\|_1^*)$ is the Moreau envelope of $\left\|\cdot\right\|_1^*$ which is the conjugate function of the ℓ_1 norm $\left\|\cdot\right\|_1$. See Section II-B for the definition of the conjugate function. From (A.3), it follows that

$$F_i^{\text{D-ORR}}(\boldsymbol{x}) = \frac{1}{2\mu_i N}\left\|\boldsymbol{x}\right\|_2^2 - \frac{1}{2\gamma}\left\|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i\right\|_2^2 + {}^{\gamma^{-1}}\big(\left\|\cdot\right\|_1^*\big)\left(\gamma^{-1}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i)\right), \tag{A.4}$$

where the last term is a convex function. Hence, $F_i^{\text{D-ORR}}$ is convex if $\frac{1}{2\mu_i N}\left\|\boldsymbol{x}\right\|_2^2 - \frac{1}{2\gamma}\left\|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i\right\|_2^2$ is so, which gives the condition in (9).

(b) Let $\Psi = \left\|\cdot\right\|_1$, $\boldsymbol{M}_1 = \boldsymbol{I}_n$, $\boldsymbol{c}_1 = \boldsymbol{0}_m$, $\boldsymbol{M}_2 = \boldsymbol{A}_i$, $\boldsymbol{c}_2 = \boldsymbol{y}_i$, $\boldsymbol{D} = \gamma^{-1/2}\boldsymbol{I}_m$, $\boldsymbol{L} = \boldsymbol{I}_m$, and $\mu = \mu_i N$ in (A.1). Then, because it can be verified with [60, Lemma 1] that ${}^{\gamma^{-1}}(\left\|\cdot\right\|_1^*)(\gamma^{-1}\boldsymbol{x}) = {}^{1}(\left\|\cdot\right\|_1^* \circ \gamma^{-\frac{1}{2}}\boldsymbol{I}_n)(\gamma^{-\frac{1}{2}}\boldsymbol{x})$, the right side of (A.1) reduces to $\mu_i N F_i^{\text{D-ORR}}$, which (and thus $F_i^{\text{D-ORR}}$) is thus convex by Theorem A.1 if and only if

$$\frac{1}{\mu_i N}\boldsymbol{I}_n - \frac{1}{\gamma}\boldsymbol{A}_i^\top \boldsymbol{A}_i \succeq \boldsymbol{O}_n, \tag{A.5}$$

which is equivalent to the condition in (9). ■

APPENDIX B
PROOF OF LEMMA 1

We first give some preliminary information to derive the Lipschitz constant $\beta_i^{\text{D-ORR}}$. A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is called *nonexpansive* if

$$\left\|T(\boldsymbol{x}) - T(\boldsymbol{z})\right\|_2^2 \le \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2^2,\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{B.1}$$

In particular, it is called *firmly nonexpansive* if

$$\left\|T(\boldsymbol{x}) - T(\boldsymbol{z})\right\|_2^2 + \left\|(I-T)(\boldsymbol{x}) - (I-T)(\boldsymbol{z})\right\|_2^2 \le \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n,\ \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{B.2}$$

Given a mapping $T : \mathbb{R}^n \to \mathbb{R}^n$, the following statements are equivalent [50], [51]:

1) T is firmly nonexpansive;
2) $I - T$ is firmly nonexpansive;
3) $T = \frac{1}{2}I + \frac{1}{2}\mathcal{N}$ for some nonexpansive mapping $\mathcal{N}$.

The equivalence immediately implies that $I - T = \frac{1}{2}I + \frac{1}{2}\check{\mathcal{N}}$, where $\check{\mathcal{N}} := I - 2T(= -\mathcal{N})$ is a nonexpansive mapping.

Derivation of $\beta_i^{D\text{-}ORR}$: The gradient $\nabla F_i^{\text{D-ORR}}(\boldsymbol{x})$ at a point $\boldsymbol{x} \in \mathbb{R}^n$ is given by

$$\nabla F_i^{\text{D-ORR}}(\boldsymbol{x}) = \frac{1}{\mu_i N}\boldsymbol{x} - \boldsymbol{A}_i^\top \frac{I - \text{prox}_{\gamma\|\cdot\|_1}}{\gamma}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i). \quad \text{(B.3)}$$

Since the proximity operator $\text{prox}_{\gamma\|\cdot\|_1}$ is firmly nonexpansive [50], we have

$$I - \text{prox}_{\gamma\|\cdot\|_1} = \frac{1}{2}I + \frac{1}{2}\mathcal{N}, \quad \text{(B.4)}$$

where the mapping $\mathcal{N} := I - 2\text{prox}_{\gamma\|\cdot\|_1}$ is nonexpansive. Hence, it follows that

$$\begin{aligned}\nabla F_i^{\text{D-ORR}}(\boldsymbol{x}) &= \frac{1}{\mu_i N}\boldsymbol{x} - \frac{1}{2\gamma}\boldsymbol{A}_i^\top(I + \mathcal{N})(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i) \\ &= \frac{1}{\mu_i N}\boldsymbol{x} - \frac{1}{2\gamma}\boldsymbol{A}_i^\top(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i) - \frac{1}{2\gamma}\boldsymbol{A}_i^\top\mathcal{N}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i). \quad \text{(B.5)}\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}&\left\|\nabla F_i^{\text{D-ORR}}(\boldsymbol{x}) - \nabla F_i^{\text{D-ORR}}(\boldsymbol{z})\right\|_2 \\ &\le \left\|\left(\frac{1}{\mu_i N}\boldsymbol{I}_{m_i} - \frac{1}{2\gamma}\boldsymbol{A}_i^\top\boldsymbol{A}_i\right)(\boldsymbol{x} - \boldsymbol{z})\right\|_2 \\ &\quad + \frac{1}{2\gamma}\left\|\boldsymbol{A}_i^\top\left(\mathcal{N}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i) - \mathcal{N}(\boldsymbol{A}_i\boldsymbol{z} - \boldsymbol{y}_i)\right)\right\|_2 \\ &\le \left(\frac{1}{\mu_i N} - \frac{\lambda_{\min}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)}{2\gamma}\right)\|\boldsymbol{x} - \boldsymbol{z}\|_2 + \frac{\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)}{2\gamma}\|\boldsymbol{x} - \boldsymbol{z}\|_2 \\ &\le \beta_i^{\text{D-ORR}}\|\boldsymbol{x} - \boldsymbol{z}\|_2,\end{aligned}$$

where the last inequality is due to the nonexpansivity of $\mathcal{N}$. We remark that $\frac{1}{\mu_i N} - \frac{\lambda_{\min}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)}{2\gamma} \ge \frac{1}{\mu_i N} - \frac{\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)}{2\gamma} \ge \frac{1}{\mu_i N} - \frac{\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)}{\gamma} \ge 0$ from the convexity condition in (9).

APPENDIX C
CONVERGENCE ANALYSIS FOR D-ORR

Theorem C.1. *Assume that (a) the graph is connected, (b) the problem in* (5) *has a solution, and (c) every node satisfies the convexity condition* (9) *and the step size condition*

$$\tau_i < \frac{1}{\beta_i^{\text{D-ORR}}/2 + \varsigma_i\left\|\boldsymbol{A}_i^\top\boldsymbol{A}_i\right\| + \sum_{j\in\mathcal{N}_i}\kappa_{ij}}. \quad \text{(C.1)}$$

Let $(\boldsymbol{x}_i(k))_{k\in\mathbb{N}}$ be the sequence generated by applying the TriPD-Dist algorithm to (10)*. Then, for every node $i \in \mathcal{V}$, $(\boldsymbol{x}_i(k))_{k\in\mathbb{N}}$ converges*[6] *to a common solution $\hat{\boldsymbol{x}}_\star \in \mathbb{R}^n$ of* (5) *R-linearly; i.e., $\|\boldsymbol{x}_i(k) - \hat{\boldsymbol{x}}_\star\| \le v_k$ for some vanishing sequence $(v_k)_{k\in\mathbb{N}} \subset [0, +\infty)$ such that $|v_{k+1}| \le \epsilon\,|v_k|$ for all $k \ge \hat{k}$ for some $\epsilon \in (0, 1)$ and some $\hat{k} \in \mathbb{N}$.*

Proof: We first show that Assumptions 5(i)–(v) and 6(i)–(iii) of [21] are satisfied. Assumptions 5(i)–(ii) and 6(i)–(ii) of [21] are clear from the problem settings of the present study. Assumption 5(iii) is justified by the convexity of $F_i^{\text{D-ORR}}$ ensured by Proposition 1 under Assumption (c) and the Lipschitz continuity of $\nabla F_i^{\text{D-ORR}}$ shown in Lemma 1 of the manuscript. Assumption 5(iv) of [21] corresponds to Assumption (a). Since dom $G_i = \mathbb{R}^n$ and dom $H_i^{\text{D-ORR}} = \mathbb{R}^{m_i}$, it holds that $\boldsymbol{x}_i \in \text{ri dom } G_i(= \mathbb{R}^n)$ and $\boldsymbol{A}_i\boldsymbol{x}_i \in \text{ri dom } H_i^{\text{D-ORR}}(= \mathbb{R}^{m_i})$ for any $\boldsymbol{x}_i$ such that $\boldsymbol{x}_i = \boldsymbol{x}_j$ for $(i, j) \in \mathcal{E}$, where ri$(\cdot)$ stands for *the relative interior* of a set [61]. This together with Assumption (b) justifies Assumption 5(v) of [21]. Assumption 6(iii) can be verified with (C.1) by noting that $\left\|\varsigma_i\boldsymbol{A}_i^\top\boldsymbol{A}_i + \sum_{j\in\mathcal{N}_i}\kappa_{ij}\boldsymbol{I}_n\right\| = \varsigma_i\left\|\boldsymbol{A}_i^\top\boldsymbol{A}_i\right\| + \sum_{j\in\mathcal{N}_i}\kappa_{ij}$.

Now, it suffices to show that the functions $F_i^{\text{D-ORR}}$ and $H_i^{\text{D-ORR}}$ are piecewise linear quadratic (PLQ) functions, where a function $f : \mathbb{R}^n \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ is said to be PLQ if dom f is a union of finitely many polyhedral sets[7], in each of which $f(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{a}^\top\boldsymbol{x} + c$ for some symmetric matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, $\boldsymbol{a} \in \mathbb{R}^n$, and $c \in \mathbb{R}$ [61]. The function $H_i^{\text{D-ORR}} = \|\cdot - \boldsymbol{y}_i\|_1$ is a translation of *a polyhedral norm* $\|\cdot\|_1$, and hence it is clearly PLQ (piecewise linear specifically with $\boldsymbol{Q} := \boldsymbol{O}$) [21]. To verify that $F_i^{\text{D-ORR}}$ is PLQ, we use the following lemmas.

Lemma C.1 (10.22 in [61])**.** *The following calculus rules of PLQ hold.*

1) *Let $f_i : \mathbb{R}^n \to \bar{\mathbb{R}}$ for $i = 1, 2$ be PLQ. Then, $f_1 + f_2$ is also PLQ.*
2) *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be PLQ. Then, $f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is also PLQ for any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.*

Lemma C.2 (11.14 and 12.30 in [61])**.** *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper lower-semicontinuous convex function. Then, the following statements hold.*

1) *f is PLQ if and only if the conjugate f^* is PLQ.*
2) *Let $\gamma > 0$. Then, f is PLQ if and only if the Moreau envelope ${}^\gamma f$ is PLQ.*

Since $\|\cdot\|_1$ is PLQ, it can readily be verified that the last term $\gamma^{-1}\left(\|\cdot\|_1^*\right)\left(\gamma^{-1}(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i)\right)$ of (A.4) is PLQ by combining Lemmas C.1.2 and C.2. In addition, the first two terms of (A.4) are quadratic functions, which are PLQ by definition. Hence, Lemma C.1.1 verifies that $F_i^{\text{D-ORR}}$ is PLQ.

The above arguments justify all assumptions required in [21, Theorem V.1], and thus the assertion is verified. ■

APPENDIX D
PROOF OF LEMMA 2

The derivation is basically the same as in Appendix B. Let $\boldsymbol{\xi}, \boldsymbol{\zeta} \in \mathbb{R}^{n+m_i}$ be arbitrary vectors. Then, it holds that

$$\begin{aligned}&\nabla F_i^{\text{D-SORR}}(\boldsymbol{\xi}) = \\ &\quad \boldsymbol{\Lambda}_i^2\boldsymbol{\xi} - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi} - \boldsymbol{y}_i) - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top\mathcal{N}(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi} - \boldsymbol{y}_i), \quad \text{(D.1)}\end{aligned}$$

[6] In fact, the triplet $(\boldsymbol{x}_i(k), \boldsymbol{y}_i(k), \boldsymbol{w}_i(k))$ of the primal and dual variables converges R-linearly to a primal-dual solution. See [21] for details. Convergence to "a common solution" means that the nodes reach consensus asymptotically.

[7] A set $C \subset \mathbb{R}^n$ is said to be a polyhedral set if it can be expressed as the intersection of finitely many closed halfspaces or hyperplanes. Polyhedral sets are closed convex, and the empty set and the whole space are polyhedral sets [61].

where $\mathcal{N} = I - 2\mathrm{prox}_{\gamma\|\cdot\|_1}$ is nonexpansive. By the triangle inequality, we have

$$\begin{aligned}
&\left\|\nabla F_i^{\text{D-SORR}}(\boldsymbol{\xi}) - \nabla F_i^{\text{D-SORR}}(\boldsymbol{\zeta})\right\|_2 \\
&\leq \left\|\left(\boldsymbol{\Lambda}_i^2 - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i\right)(\boldsymbol{\xi}-\boldsymbol{\zeta})\right\|_2 \\
&\quad + \frac{1}{2\gamma}\left\|\tilde{\boldsymbol{A}}_i^\top\left(\mathcal{N}(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}-\boldsymbol{y}_i) - \mathcal{N}(\tilde{\boldsymbol{A}}_i\boldsymbol{\zeta}-\boldsymbol{y}_i)\right)\right\|_2 \\
&\leq \lambda_{\max}\left(\boldsymbol{\Lambda}_i^2 - \frac{\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i}{2\gamma}\right)\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|_2 + \frac{\lambda_{\max}(\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i)}{2\gamma}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|_2 \\
&\leq \beta_i^{\text{D-SORR}}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|_2,
\end{aligned}$$

where the last inequality is due to the nonexpansivity of $\mathcal{N}$. We remark that $\boldsymbol{\Lambda}_i^2 - \frac{\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i}{2\gamma}$ is positive definite because $\boldsymbol{\Lambda}_i^2 - \frac{\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i}{\gamma}$ is positive semidefinite if and only if the convexity condition in (9) is satisfied (see the proof of Proposition 3 in [33]).

Appendix E
Convergence analysis for D-SORR

Theorem E.1. *Assume that (a) the graph is connected, (b) the problem in* (2) *has a solution, and (c) every node satisfies the convexity condition* (22) *and the step size condition*

$$\tau_i < \frac{1}{\beta_i^{\text{D-SORR}}/2 + \varsigma_i\left\|\boldsymbol{A}_i^\top\boldsymbol{A}_i\right\| + \sum_{j\in\mathcal{N}_i}\kappa_{ij}}. \tag{E.1}$$

Let $(\boldsymbol{\xi}_i(k))_{k\in\mathbb{N}} \subset \mathbb{R}^{n+m_i}$ *be the sequence generated by applying the TriPD-Dist algorithm to* (23)*. Then, for each node* $i \in \mathcal{V}$*,* $(\boldsymbol{\xi}_i(k))_{k\in\mathbb{N}}$ *converges to a solution* $[\hat{\boldsymbol{x}}_\star^\top, \hat{\boldsymbol{\varepsilon}}_{i,\star}^\top]^\top \in \mathbb{R}^{n+m_i}$ *of* (2) *R-linearly, where* $\hat{\boldsymbol{x}}_\star$ *is common to all nodes.*

Proof: The proof is omitted because the assertion can be verified in the same way as the proof of Theorem C.1 in light of Proposition 4. ∎

Appendix F
Proof of Proposition 5

The sum of the terms of the Moreau envelope of the ℓ_1 norm of (27) can be rewritten as

$$\begin{aligned}
&\sum_{i\in\mathcal{V}} {}^{\gamma}\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i) \\
&= \sum_{i\in\mathcal{V}}\min_{\boldsymbol{w}_i\in\mathbb{R}^{m_i}}\left(\|\boldsymbol{w}_i\|_1 + \frac{1}{2\gamma}\|\boldsymbol{w}_i - (\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i)\|_2^2\right) \\
&= \min_{\boldsymbol{w}_i\in\mathbb{R}^{m_i}}\sum_{i\in\mathcal{V}}\left(\|\boldsymbol{w}_i\|_1 + \frac{1}{2\gamma}\|\boldsymbol{w}_i - (\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i)\|_2^2\right).
\end{aligned} \tag{F.1}$$

Let $M := \sum_{i\in\mathcal{V}} m_i$ and

$$\begin{aligned}
\boldsymbol{w} &:= \begin{bmatrix}\boldsymbol{w}_1^\top & \dots & \boldsymbol{w}_N^\top\end{bmatrix}^\top \in \mathbb{R}^M, \\
\boldsymbol{\xi} &:= \begin{bmatrix}\boldsymbol{x}^\top & \boldsymbol{\varepsilon}_1^\top & \dots & \boldsymbol{\varepsilon}_N^\top\end{bmatrix}^\top \in \mathbb{R}^{n+M}, \\
\boldsymbol{c}_2 &:= -\begin{bmatrix}\boldsymbol{y}_1^\top & \dots & \boldsymbol{y}_N^\top\end{bmatrix}^\top \in \mathbb{R}^M, \\
\boldsymbol{M}_2 &:= \begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{I}_{m_1} & \dots & \boldsymbol{O}_{m_1\times m_N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_N & \boldsymbol{O}_{m_N\times m_1} & \dots & \boldsymbol{I}_{m_N}\end{bmatrix} \in \mathbb{R}^{M\times(n+M)}.
\end{aligned}$$

Then, (F.1) reduces to

$$\begin{aligned}
&\min_{\boldsymbol{w}\in\mathbb{R}^M}\left(\|\boldsymbol{w}\|_1 + \frac{1}{2\gamma}\|\boldsymbol{w}-(\boldsymbol{M}_2\boldsymbol{\xi}+\boldsymbol{c}_2)\|_2^2\right) \\
&= {}^{\gamma}\|\cdot\|_1(\boldsymbol{M}_2\boldsymbol{\xi}+\boldsymbol{c}_2) \\
&= -{}^{\gamma^{-1}}\left(\|\cdot\|_1^*\right)\left(\gamma^{-1}(\boldsymbol{M}_2\boldsymbol{\xi}+\boldsymbol{c}_2)\right) + \frac{1}{2\gamma}\|\boldsymbol{M}_2\boldsymbol{\xi}+\boldsymbol{c}_2\|_2^2.
\end{aligned} \tag{F.2}$$

On the other hand, the regularization terms of (27) can be rewritten as

$$\sum_{i\in\mathcal{V}}\left(\frac{\sigma_x^{-2}}{2\mu N}\|\boldsymbol{x}\|_2^2 + \frac{\sigma_\varepsilon^{-2}}{2\mu}\|\boldsymbol{\varepsilon}_i\|_2^2\right) = \frac{1}{2\mu}\|\boldsymbol{M}_1\boldsymbol{\xi}\|_2^2, \tag{F.3}$$

where

$$\boldsymbol{M}_1 := \begin{bmatrix}\sigma_x^{-1}\boldsymbol{I}_n & \boldsymbol{O}_{n\times m_1} & \dots & \boldsymbol{O}_{n\times m_N} \\ \boldsymbol{O}_{m_1\times n} & \sigma_\varepsilon^{-1}\boldsymbol{I}_{m_1} & \dots & \boldsymbol{O}_{m_1\times m_N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O}_{m_N\times n} & \boldsymbol{O}_{m_N\times m_1} & \dots & \sigma_\varepsilon^{-1}\boldsymbol{I}_{m_N}\end{bmatrix}$$

with size $(n+M)\times(n+M)$. Combining (F.1)–(F.3) reduces (27) to the following form:

$$\begin{aligned}
F^{\text{D-SORR}}(\boldsymbol{\xi}) = \frac{1}{2\mu}\|\boldsymbol{M}_1\boldsymbol{\xi}\|_2^2 - \frac{1}{2\gamma}\|\boldsymbol{M}_2\boldsymbol{\xi}+\boldsymbol{c}_2\|_2^2 \\
+ {}^{\gamma^{-1}}\left(\|\cdot\|_1^*\right)\left(\gamma^{-1}(\boldsymbol{M}_2\boldsymbol{\xi}+\boldsymbol{c}_2)\right).
\end{aligned} \tag{F.4}$$

Let $\Psi := \|\cdot\|_1 : \mathbb{R}^{n+M} \to [0,+\infty)$, $\mathscr{A}_1 : \mathbb{R}^{n+M} \to \mathbb{R}^{n+M} : \boldsymbol{x} \mapsto \boldsymbol{M}_1\boldsymbol{x} + \boldsymbol{c}_1$, $\mathscr{A}_2 : \mathbb{R}^{n+M} \to \mathbb{R}^M : \boldsymbol{x} \mapsto \boldsymbol{M}_2\boldsymbol{x} + \boldsymbol{c}_2$, $\boldsymbol{c}_1 := \boldsymbol{0}_{n+M}$, $\boldsymbol{D} := \gamma^{-1/2}\boldsymbol{I}_M$, and $\boldsymbol{L} := \boldsymbol{I}_M$. Then, since ${}^{\gamma^{-1}}(\|\cdot\|_1^*)(\gamma^{-1}\boldsymbol{\xi}) = {}^{1}(\|\cdot\|_1^* \circ \gamma^{-\frac{1}{2}}\boldsymbol{I}_n)(\gamma^{-\frac{1}{2}}\boldsymbol{\xi})$ for every $\boldsymbol{\xi} \in \mathbb{R}^{n+M}$ (see [60, Lemma 1] to verify), we have

$$F^{\text{D-SORR}} = \frac{1}{2\mu}\|\cdot\|_2^2\circ\mathscr{A}_1 - \frac{1}{2}\|\cdot\|_2^2\circ\boldsymbol{D}\boldsymbol{L}\mathscr{A}_2 + {}^{1}(\Psi^*\circ\boldsymbol{D})\circ\boldsymbol{D}\boldsymbol{L}\mathscr{A}_2. \tag{F.5}$$

To apply Theorem A.1 to $\mu F^{\text{D-SORR}}$, we observe that $\boldsymbol{M}_2\boldsymbol{\xi} + \boldsymbol{c}_2 = \boldsymbol{0}_M$ for $\boldsymbol{\xi} = [\boldsymbol{0}_n^\top, -\boldsymbol{c}_2^\top]^\top$, which means that the assumption of the theorem is satisfied in this case. Hence, the theorem verifies that $\mu F^{\text{D-SORR}}$, and thus $F^{\text{D-SORR}}$, is convex if and only if

$$\begin{aligned}
&\boldsymbol{M}_1^\top\boldsymbol{M}_1 - \mu\gamma^{-1}\boldsymbol{M}_2^\top\boldsymbol{M}_2 = \\
&\begin{bmatrix}\sigma_x^{-2}\boldsymbol{I}_n - \mu\gamma^{-1}\boldsymbol{A}^\top\boldsymbol{A} & -\mu\gamma^{-1}\boldsymbol{A}^\top \\ -\mu\gamma^{-1}\boldsymbol{A} & \left(\sigma_\varepsilon^{-2} - \mu\gamma^{-1}\right)\boldsymbol{I}_M\end{bmatrix} \succeq \boldsymbol{O}_{n+M} \\
\Leftrightarrow &\begin{bmatrix}\mu^{-1}\gamma\sigma_x^{-2}\boldsymbol{I}_n - \boldsymbol{A}^\top\boldsymbol{A} & -\boldsymbol{A}^\top \\ -\boldsymbol{A} & \left(\mu^{-1}\gamma\sigma_\varepsilon^{-2} - 1\right)\boldsymbol{I}_M\end{bmatrix} \succeq \boldsymbol{O}_{n+M},
\end{aligned} \tag{F.6}$$

which is identical to [33, Eq. (B.1)]. The positive semidefiniteness condition in (F.6) is equivalent to the following inequality (see [33, Proposition 3]):

$$\mu\gamma^{-1} \leq \frac{1}{\sigma_\varepsilon^2 + \sigma_x^2\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})}, \tag{F.7}$$

which coincides with (28). ∎