2023 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2023, ROME, ITALY

DISTRIBUTED STABLE OUTLIER-ROBUST SIGNAL RECOVERY USING MINIMAX CONCAVE LOSS

Maximilian H. V. Tillmann *Masahiro Yukawa*

Department of Electronics and Electrical Engineering, Keio University, Japan

ABSTRACT

This paper presents a mathematically rigorous framework of remarkably robust signal recovery over networks. The proposed framework is based on the so-called *minimax concave (MC)* loss, which is a "hybrid" between Tukey's biweight loss and Huber's loss in the sense of yielding remarkable outlier-robustness and being able to preserve convexity of the overall cost under an appropriate choice of parameters so that an iterative algorithm could generate a sequence of vectors converging provably to a solution (a global minimizer of the overall cost). We present a formulation which involves an auxiliary vector to accommodate the statistical property of noise explicitly, and we present a condition to guarantee convexity of the local cost. We apply the distributed triangularly preconditioned primal-dual algorithm to our formulation and show by numerical examples that our proposed formulation exhibits remarkable robustness under devastating outliers, and outperforms the existing methods.

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

1. INTRODUCTION

Robust methods in the presence of outliers (or impulsive noise) have widely been studied in signal processing [1] and machine learning as well as many other fields including statistics [2, 3], control [4], and optimization [5]. Outliers frequently occur in wireless communication channels, biomedical sensors, image/video sensors, and other applications. Distributed settings are useful in solving large-scale problems where data volume is too large to store at a single computer. *Decentralized systems (having no central node)* are considered in the present study, which are advantageous in many aspects: no single point of failure, no potential privacy violations by collecting all data at a single node, no need for infrastructures, and suitability for edge computing.

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 [6–8]. The major contributions of this work concern the former aspect basically.

In this paper, we propose the formulation named *distributed stable outlier-robust signal recovery (D-SORR)*, which accommodates statistical properties of Gaussian noise and outliers at the same time. D-SORR uses the nonconvex (weakly convex) minimax concave (MC) loss function at each local node to attain remarkable robustness against outliers. Furthermore, it models Gaussian noise by an auxiliary variable and is thus more "stable" against perturbations caused by Gaussian noise as well as robust large outliers in the sense of [9]. The sensitivity of a loss function to outliers depends on how much the residual error of the loss function grows for larger residual errors. Compared to the ordinary least-square loss function, the least absolute deviation (LAD) loss is relatively insensitive to outliers, as it grows "linearly" (rather than quadratically) for residual errors. However, to achieve robustness to extreme outliers, the loss function needs to stay constant for such residual errors with magnitudes larger than a given threshold. This motivates the use of the MC loss function in the present study.

The two main research questions addressed in this paper are when the proposed formulation is solvable by an iterative algorithm efficiently, and how robust the D-SORR estimator is against outliers compared to existing methods. The first research question is answered by studying the convexity condition for the cost function in (2) using the framework called linearly-involved Moreau-enhanced-over-subspace (LiMES) model developed in [10, 11]. 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 [8] via reformulation using Moreau's decomposition. The second research question stated above is addressed by simulation studies. The numerical examples show that our proposed method leads to remarkable robustness when the data is contaminated by many and/or huge outliers, outperforming the existing methods in a variety of situations.

In the previous work [12], the sparse signal recovery prob-

This work was supported by JSPS KAKENHI Grant Number (22H01492).

Maximilian H. V. Tillmann, with the T.I.M.E. Double Degree, is also with RWTH Aachen University, Germany.

979-8-3503-2411-2/23/$31.00

lem 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 [13] was used. However, as the algorithm cannot be used in the present case because the proximity operator of the ℓ_1 norm involving a linear composition is hardly available, the recently developed solver called distributed triangularly preconditioned primal-dual (TriPD-Dist) algorithm [8] (based on *operator splitting* [14]) is employed in the present study. Other works concerning nonconvex methods for distributed optimization have been studied actively both in signal processing and machine learning communities [15]. Recent developments include most notably the heuristic approach based on the notion of graduated nonconvexity for outlier-robust distributed optimization [16].

2. PRELIMINARIES

We consider *decentralized systems* equipped with a network of N nodes represented by an undirected graph, which is always assumed to be *connected*, with a set of nodes $\mathcal{V}$ and edges $\mathcal{E}$. If node i is connected to j: $(i,j) \in \mathcal{E}$, and the set $\mathcal{N}_j$ contains all neighbors of node j.

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 $\boldsymbol{A}^\top$, the identity matrix is denoted by $\boldsymbol{I}_n$, the zero matrix is denoted by $\boldsymbol{O}_{m\times n}$, and the zero vector is denoted by $\boldsymbol{0}$. The largest eigenvalue of a symmetric matrix $\boldsymbol{A}$ is denoted by $\lambda_{\max}(\boldsymbol{A})$.

In this study, we consider the standard inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \boldsymbol{x}^\top \boldsymbol{y}$ between $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. The induced norm $\|\boldsymbol{x}\|_2 := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}$ is then the ℓ_2 norm. 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 $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper lower-semicontinuous convex function (see [17]). Then, its Fenchel conjugate is defined by $f^*(\boldsymbol{x}) := \sup_{\boldsymbol{y}\in\mathbb{R}^n} (\langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{y}))$ [17], which is again a proper lower-semicontinuous convex function. The proximity operator of f of index $\gamma > 0$ is defined by $\mathrm{prox}_{\gamma f}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathbb{R}^n} \left(f(\boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right)$ [17, 18], and the minimum value ${}^\gamma f(\boldsymbol{x}) := \min_{\boldsymbol{y}\in\mathbb{R}^n} \left(f(\boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right) = f(\mathrm{prox}_{\gamma f}(\boldsymbol{x})) + \frac{1}{2\gamma}\left\|\boldsymbol{x} - \mathrm{prox}_{\gamma f}(\boldsymbol{x})\right\|_2^2$ achieved by the proximity operator is called the Moreau envelope of f [17, 18].

3. DISTRIBUTED STABLE OUTLIER-ROBUST SIGNAL RECOVERY

We present the problem formulation for D-SORR using the MC loss function. We then analyze the convexity of the local objective, discuss the optimization algorithm, and give a remark on parameter design. All results are given without proofs due to the page limitation. An extended version of the present work including all the proofs will be presented elsewhere.

As outliers are typically *sparse*, we consider a linear model where the observation vector at each node i 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. The model in (1) has previously been studied in centralized (non-distributed) settings [19, 20], but it has not been studied well in the distributed settings.

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:

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

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

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

is the MC loss [21, 22] defined with

$$\phi_\gamma^{\mathrm{MC}}(x) := \begin{cases} |x| - x^2/2\gamma, & \text{if } |x| \le \gamma, \\ \gamma/2, & \text{if } |x| > \gamma. \end{cases} \tag{4}$$

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 2 for the definition of the Moreau envelope ${}^\gamma\|\cdot\|_1$ of the ℓ_1 norm.

Each term of the summand in (2) accommodates prior information about the random vectors. Specifically, the first term reflects the sparseness of the outlier $\boldsymbol{o}_{i\diamond} (\approx \boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{\varepsilon}_i - \boldsymbol{y}_i)$, and the second and third terms reflect the Gaussianity of the signal $\boldsymbol{x}_\star$ and the noise $\boldsymbol{\varepsilon}_{i\star}$, respectively. 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.

The derivative of $\phi_\gamma^{\mathrm{MC}}$ at $x \in \mathbb{R}\setminus\{0\}$ is given by $\psi_\gamma^{\mathrm{MC}}(x) := \begin{cases} \mathrm{sign}(x) - x/\gamma, & \text{if } |x| \in (0,\gamma), \\ 0, & \text{if } |x| \ge \gamma. \end{cases}$ Inspecting the behavior of the derivative $\psi_\gamma^{\mathrm{MC}}$, it can be seen that it vanishes for $|x| \ge \gamma$, making the MC loss remarkably robust against huge outliers in analogy with Tukey's biweight loss. 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, meaning that the MC

loss sharply increases by small deviations from zero and thus it would not allow small errors originated by Gaussian noise. For this reason the auxiliary vectors $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{m_i}$ are introduced to model the Gaussian noise explicitly.

3.1. Convexity condition for local objective of D-SORR

The objective function of in (2) can be split into smooth and nonsmooth terms as follows:

$$\min_{\substack{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{\varepsilon}_i\in\mathbb{R}^{m_i}\\ (i\in\mathcal{V})}} \sum_{i\in\mathcal{V}} \Bigg(\underbrace{\|\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i\|_1}_{H_i^{\text{D-SORR}}(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i)} + \underbrace{\frac{\sigma_x^{-2}}{2\mu_i N}\|\boldsymbol{x}\|_2^2+\frac{\sigma_\varepsilon^{-2}}{2\mu_i}\|\boldsymbol{\varepsilon}_i\|_2^2-{}^{\gamma}\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i)}_{F_i^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_i)} \Bigg). \quad (5)$$

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

$$H_i^{\text{D-SORR}}(\boldsymbol{v}) := \|\boldsymbol{v}-\boldsymbol{y}_i\|_1, \quad (6)$$

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 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

$$F_i^{\text{D-SORR}}(\boldsymbol{x},\boldsymbol{\varepsilon}_i) := \frac{\sigma_x^{-2}}{2\mu_i N}\|\boldsymbol{x}\|_2^2+\frac{\sigma_\varepsilon^{-2}}{2\mu_i}\|\boldsymbol{\varepsilon}_i\|_2^2 - {}^{\gamma}\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x}+\boldsymbol{\varepsilon}_i-\boldsymbol{y}_i) \quad (7)$$

is analyzed below.

Proposition 1 (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*

$$\mu_i(\sigma_\varepsilon^2 + N\sigma_x^2\lambda_{\max}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)) \le \gamma. \quad (8)$$

3.2. Distributed convex optimization algorithm for D-SORR: TriPD-Dist

The TriPD-Dist algorithm (from [8]) is a convex analytic solver for distributed optimization problems in the following form [8]:

$$\begin{aligned} \min_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\in\mathbb{R}^n} \quad & \sum_{i\in\mathcal{V}} F_i(\boldsymbol{x}_i)+G_i(\boldsymbol{x}_i)+H_i(\boldsymbol{A}_i\boldsymbol{x}_i) \\ \text{s.t.} \quad & \boldsymbol{B}_{ij}\boldsymbol{x}_i+\boldsymbol{B}_{ji}\boldsymbol{x}_j=\boldsymbol{d}_{ij}, \quad (i,j)\in\mathcal{E}, \end{aligned} \quad (9)$$

where each variable vector $\boldsymbol{x}_i$ is updated at each node. Information exchanges over the given network are allowed only for the variable vector $\boldsymbol{x}_i$ and the edge variable $\boldsymbol{w}_{ij}$ (see Algorithm 1). 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 $(i,j) \in \mathcal{E}$. 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 [23].

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. For convenience, we define $\boldsymbol{\xi}_i := [\boldsymbol{x}_i^\top\ \boldsymbol{\varepsilon}_i^\top]^\top \in \mathbb{R}^{n+m_i}$. 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 nodes. This means that the consensus constraint is required among $\boldsymbol{x}_i$'s, but it is *not* required for $\boldsymbol{\varepsilon}_i$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 $\tilde{\boldsymbol{I}}_i := [\boldsymbol{I}_n\ \boldsymbol{O}_{n\times m_i}] \in \mathbb{R}^{n\times(n+m_i)}$. 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, (5) can be reformulated as follows:

$$\min_{\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_N} \sum_{i\in\mathcal{V}} \left(H_i^{\text{D-SORR}}(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i) + F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i) \right) \quad \text{s.t. } \tilde{\boldsymbol{I}}_i\boldsymbol{\xi}_i = \tilde{\boldsymbol{I}}_j\boldsymbol{\xi}_j,\ \forall i,j = 1,2,\dots,N, \quad (10)$$

where

$$F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i) := \frac{1}{2}\|\boldsymbol{\Lambda}_i\boldsymbol{\xi}_i\|_2^2 - {}^{\gamma}\|\cdot\|_1\left(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i-\boldsymbol{y}_i\right), \quad (11)$$

$$H_i^{\text{D-SORR}}(\boldsymbol{v}) := \|\boldsymbol{v}-\boldsymbol{y}_i\|_1, \ \ \boldsymbol{v}\in\mathbb{R}^{m_i}. \quad (12)$$

Here, $F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i)$ is essentially the same as $F_i^{\text{D-SORR}}(\boldsymbol{x}_i,\boldsymbol{\varepsilon}_i)$. The expression (10) of D-SORR shares the same form as (9) under the following correspondences: $F_i(\boldsymbol{\xi}_i) := F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i)$, $G_i(\boldsymbol{\xi}_i) := 0$, $H_i(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i) := H_i^{\text{D-SORR}}(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i)$, $\boldsymbol{B}_{ij} := \begin{cases}\tilde{\boldsymbol{I}}_i, & \text{if } i<j,\\ -\tilde{\boldsymbol{I}}_i, & \text{otherwise},\end{cases}$ and $\boldsymbol{d}_{ij} := \boldsymbol{0}_n$. The TriPD-Dist algorithm applied to (10) is given in Algorithm 1.

The gradient of $F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i)$ is given as

$$\nabla F_i^{\text{D-SORR}}(\boldsymbol{\xi}_i) = \boldsymbol{\Lambda}_i^2\boldsymbol{\xi}_i - \tilde{\boldsymbol{A}}_i^\top\frac{\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i-\boldsymbol{y}_i-\text{prox}_{\gamma\|\cdot\|_1}(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i-\boldsymbol{y}_i)}{\gamma}. \quad (13)$$

The proximity operator of the ℓ_1 norm in the algorithm can be computed to $\text{prox}_{\alpha\|\cdot\|_1}(\boldsymbol{v}) = \text{soft}_{[-\alpha,\alpha]}(\boldsymbol{v})$ for $\boldsymbol{v}\in\mathbb{R}^m$, with the element wise soft shrinkage operator given for the l-th element as

$$\left[\text{soft}_{[-\alpha,\alpha]}(\boldsymbol{v})\right]_l = \begin{cases} v_l+\alpha, & \text{if } v_l<-\alpha,\\ 0, & \text{if } -\alpha\le v_l\le\alpha,\\ v_l-\alpha, & \text{if } v_l>\alpha.\end{cases} \quad (14)$$

Algorithm 1: Distributed Triangular Preconditioned Primal-Dual (TriPD-Dist) algorithm from [8] for D-SORR

Requirements: step size $\tau_i > 0$, dual step size $\varsigma_i > 0$, link weights $\kappa_{ij} > 0$

Initialisation: $\boldsymbol{\xi}_i(0) = \mathbf{0} \in \mathbb{R}^{n+m_i}$ for $i = 1, 2, \dots, N$,
$\boldsymbol{z}_i(0) = \mathbf{0} \in \mathbb{R}^{m_i}$ for $i = 1, 2, \dots, N$, and
$\boldsymbol{w}_{ij}(0) = \mathbf{0} \in \mathbb{R}^n$ for $(i, j) \in \mathcal{E}$

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

local updates

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

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

end

$\overline{\boldsymbol{z}}_i(k) = \boldsymbol{z}_i(k) + \varsigma_i\left(\tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i(k) - \boldsymbol{y}_i\right) - \varsigma_i \mathrm{prox}_{\varsigma_i^{-1}\|\cdot\|_1}\left(\varsigma_i^{-1}\boldsymbol{z}_i(k) + \tilde{\boldsymbol{A}}_i\boldsymbol{\xi}_i(k) - \boldsymbol{y}_i\right)$

$\boldsymbol{\xi}_i(k+1) = \boldsymbol{\xi}_i(k) - \tau_i\tilde{\boldsymbol{A}}_i^\top\overline{\boldsymbol{z}}_i(k) - \tau_i\sum_{j\in\mathcal{N}_i}\boldsymbol{B}_{ij}\overline{\boldsymbol{w}}_{ij}(k) - \tau_i\nabla F_i(\boldsymbol{\xi}_i(k))$

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

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

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

$$\beta_i^{\text{D-SORR}} := \lambda_{\max}(\boldsymbol{\Lambda}_i^2 - \frac{1}{2\gamma}\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i) + \frac{1}{2\gamma}\lambda_{\max}(\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i). \tag{15}$$

Assume the convexity condition (8) and also assume the primal step-size condition

$$\tau_i < \frac{1}{\beta_i^{\text{D-SORR}}/2 + \varsigma_i\left\|\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i\right\| + \sum_{j\in\mathcal{N}_i}\kappa_{ij}}, \tag{16}$$

where $\left\|\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i\right\|$ is the spectral norm of $\tilde{\boldsymbol{A}}_i^\top\tilde{\boldsymbol{A}}_i$. The problem in (5) may have multiple solutions in general. Under the given assumptions, the sequence $(\boldsymbol{\xi}_i(k))_{k\in\mathbb{N}}$ of $\boldsymbol{\xi}_i(k) := [\boldsymbol{x}_i(k)^\top\ \boldsymbol{\varepsilon}_i(k)^\top]^\top$ generated by the TriPD-Dist algorithm achieves consensus at each node i and converges to a solution[1] of (5).

Remark 1 (Parameter design for D-SORR)**.** *If estimates of $\sigma_{x\star}^2$ and $\sigma_{\varepsilon\star}^2$ are available, tune γ by grid search with μ_i set to its upper bound based on* (8)*. If these estimates are unavailable, one may let $\bar{\mu}_i := \mu_i\sigma_x^2$ and $\varrho := \sigma_x^2/\sigma_\varepsilon^2$ so that the last two terms of* (2) *reduce to $\frac{1}{2\bar{\mu}_i N}\|\boldsymbol{x}\|_2^2 + \frac{\varrho}{2\bar{\mu}_i}\|\boldsymbol{\varepsilon}_i\|_2^2$. In this case, the convexity condition is given by $\bar{\mu}_i(\varrho + N\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})) \le \gamma$. Our recommendation is to tune γ and ϱ by grid search with $\bar{\mu}_i$ set to the upper bound for each given γ and ϱ.*

[1]The existence of a solution is guaranteed by the coercivity of the cost function in (5) in terms of all variable vectors $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}_i$s.

4. SIMULATION RESULTS

We show the efficacy of the proposed method in terms of outlier robustness under various scenarios. 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

$$\text{SNR} := \frac{\sum_{i\in\mathcal{V}}\|\boldsymbol{A}_i\boldsymbol{x}_\star\|_2^2}{\sum_{i\in\mathcal{V}}\|\boldsymbol{\varepsilon}_{i\star}\|_2^2}, \tag{17}$$

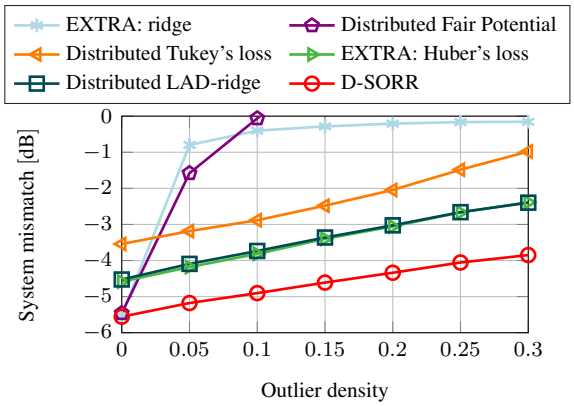
where the temporary vectors are generated from the i.i.d. standard Gaussian distribution. The positions of the ν_i nonzero elements of $\boldsymbol{o}_{i\diamond}$ are chosen randomly, and the nonzero values follow an i.i.d. scaled and shifted uniform distribution. Here, for all simulations, given some prespecified value $\overline{M}_{o_\diamond} > 0$, the interval of uniform distribution is set to $d_{\text{uniform}} := 2/9\overline{M}_{o_\diamond}$ 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 center chosen randomly between 80 and 100 at each independent run. Our primary performance measure is the following:

$$\text{system mismatch} := \frac{1}{N}\sum_{i\in\mathcal{V}}\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_\star\|_2^2}{\|\boldsymbol{x}_\star\|_2^2}. \tag{18}$$

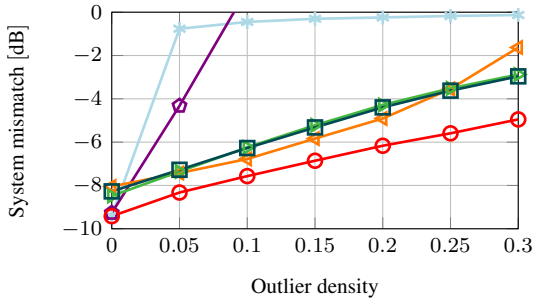
All plots in the figures presented in this section show the averages over 250 independent runs. For D-SORR, we set $\sigma_x^2 := \sigma_{x\star}^2 := 1$ and $\sigma_\varepsilon^2 := \sigma_{\varepsilon\star}^2 := \frac{1}{mN}\sum_{i\in\mathcal{V}}\|\boldsymbol{\varepsilon}_{i\star}\|_2^2$. The algorithm parameters are set to $\varsigma_i := 0.065$, and $\kappa_{ij} := 1$, if $(i, j) \in \mathcal{E}$, and $\kappa_{ij} := 0$, otherwise. For the design of the saturation factor γ and the regularization parameters μ_i for D-SORR, see Remark 1. D-SORR is compared to the following robust loss functions for positive constants $\delta_{\text{L}}, \delta_{\text{H}}, \delta_{\text{T}}, \delta_{\text{P}} > 0$: the LAD-ridge $\sum_{i\in\mathcal{V}}\|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i\|_1 + \delta_{\text{L}}\|\boldsymbol{x}\|_2^2$, Huber's loss $\sum_{i\in\mathcal{V}}{}^{\delta_{\text{H}}}\|\cdot\|_1(\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i)$, Tukey's biweight loss [2] $\sum_{i\in\mathcal{V}}\sum_{\iota=1}^m\phi_{\delta_{\text{T}}}^{\text{TK}}([\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i]_\iota)$, where

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

and the fair potential function [4] $\sum_{i\in\mathcal{V}}\sum_{\iota=1}^m\phi_{\delta_{\text{P}}}^{\text{FP}}([\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{y}_i]_\iota)$, where $\phi_{\delta_{\text{P}}}^{\text{FP}} : \mathbb{R} \ni a \mapsto \delta_{\text{P}}|a| - \log_{10}(1 + \delta_{\text{P}}|a|)$. Here,

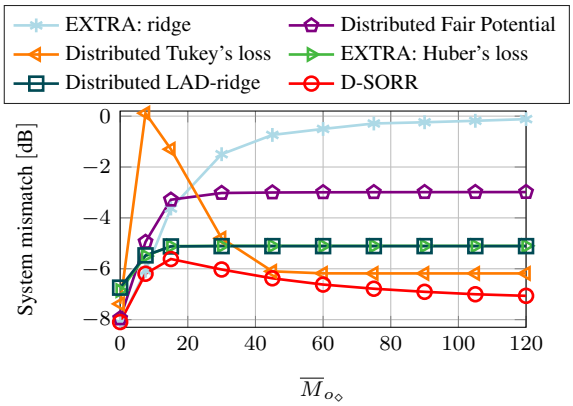


(a) SNR $= 0$ dB, $\overline{M}_{o_\diamond} = 90$, and $n = 30$

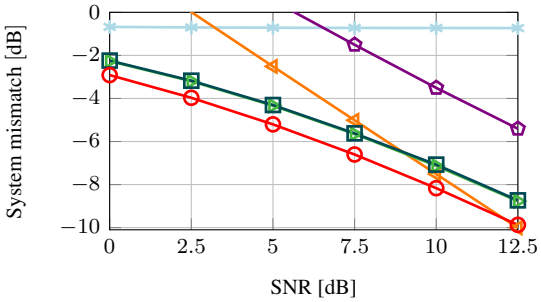


(b) SNR $= 10$ dB, $\overline{M}_{o_\diamond} = 90$, and $n = 60$

Fig. 1: System mismatch across the outlier density.



(a) outlier density 0.2, SNR $= 0$ dB, and $n = 15$



(b) outlier density 0.05, $\overline{M}_{o_\diamond} = 90$, and $n = 60$

Fig. 2: System mismatch across (a) $\overline{M}_{o_\diamond}$, and (b) SNR.

$[\cdot]_\iota$ 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. For each method, the delta parameter is tuned by grid search to minimize the system mismatch. In the following simulations, we consider the case when the network has $N := 5$ nodes, each of which is given $m := 20$ measurement vectors.

Figure 1 shows the system mismatch across different outlier densities from 0 to 0.3 under different SNRs, and different numbers of variables n. Overall, the proposed method outperforms the other methods significantly when outliers are present due to the robustness of the MC loss to outliers. For the case when there are no outliers (outlier density of 0), D-SORR is also able to perform well due to the explicit Gaussian noise modelling with the auxiliary variable. The system mismatch is generally lower in Fig. 1b compared to Fig. 1a due to the higher SNR in the case of Fig. 1b. The larger number of variables ($n = 60$) in case of Fig. 1b compared to the case of Fig. 1a ($n = 30$) increases the overall system mismatch, as more variables must be estimated from the same number of measurements.

Figure 2a shows the performance across $\overline{M}_{o_\diamond}$ for SNR 0 dB, $n = 15$, and an outlier density of 0.2, 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-SORR, and distributed Tukey's loss) show "non-monotonic" behaviors where the system mismatch increases up to some point, and it then decreases as the outlier power increases.

This observation leads us to the hypothesis that the nonconvex methods implicitly carry out outlier detection. Intuitively, it is more likely when the outlier power is larger that the magnitude of the error for each outlier measurement lies near or exceeds a fixed threshold where the gradient of the loss function is zero, i.e. classifying an outlier correctly. Additionally, in our preliminary experiments it is observed that the optimal threshold γ increases when larger outliers occour, which allows to use a larger regularization parameters μ_i to reduce the bias and decrease the system mismatch further for larger outliers for D-SORR.

Figure 2b shows the performance under different levels of Gaussian noise from 0 dB SNR to 12.5 dB SNR with $\overline{M}_{o_\diamond} = 90$, $n = 60$, and an outlier density of 0.05. It can be seen that D-SORR outperforms the other methods. Only for an SNR of 12.5 dB Tukey's biweight loss achieves a similar system mismatch than D-SORR.

5. CONCLUSION

This paper presented the D-SORR formulation for distributed robust signal recovery. Thanks to the weak convexity of the

MC loss, the proposed formulation enjoys 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 condition to guarantee convexity of the local objective of D-SORR and applied the TriPD-Dist algorithm to convergence to a global minimizer of the D-SORR formulation. The numerical examples showed that our proposed formulation exhibited remarkable robustness under huge outliers as well as outperforming the existing methods.

6. REFERENCES

[1] 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.

[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*, Wiley, 2nd edition, 2009.

[4] J. Li, E. Elhamifar, I. 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015.

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

[15] 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.

[16] 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.

[17] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York: NY, 2nd edition, 2017.

[18] I. Yamada, M. Yukawa, and M. Yamagishi, *Minimizing Moreau envelope of nonsmooth convex function over the fixed point set of certain quasi-nonexpansive mappings*, pp. 345–390, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011.

[19] 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.

[20] 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.

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

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

[23] N. Perraudin, D. Shuman, G. Puy, and P. Vandergheynst, "Unlocbox a matlab convex optimization toolbox using proximal splitting methods," *ArXiv e-prints*, Feb. 2014.