Distributed Sparse Optimization Based on Minimax Concave and Consensus Promoting Penalties: Towards Global Optimality

Kei Komuro Masahiro Yukawa
Department of Electronics and Electrical Engineering
Keio University, Japan

Renato L.G. Cavalcante
Fraunhofer Institute for Telecommunications
Heinrich Hertz Institute, Germany

***Abstract*—We propose a distributed optimization framework to generate accurate sparse estimates while allowing an algorithmic solution with guaranteed convergence to a global minimizer. To this end, the proposed problem formulation involves the minimax concave penalty together with an additional penalty called *consensus promoting penalty (CPP)* that induces convexity to the resulting optimization problem. This problem is solved with an exact first-order proximal gradient algorithm, which employs a pair of proximity operators and is referred to as the distributed proximal and debiasing-gradient (DPD) method. Numerical examples show that CPP not only convexifies the whole cost function, but it also accelerates the convergence speed with respect to the system mismatch.**

***Index Terms*—distributed optimization, nonconvex penalty, Moreau envelope, proximity operator, sparseness**

I. Introduction

Distributed optimization [2], [3] is a key component in many applications such as big data analytics [4] and sensor networks [5]. In some scenarios, such as environmental modeling, the estimand can be assumed sparse, so convex penalties have been used in optimization problems to promote sparsity of the solutions. Such convex formulations can be addressed with algorithms such as the proximal distributed gradient descent (Prox-DGD) [6] and the proximal gradient exact first-order algorithm (PG-EXTRA) [7], to name a few. However, despite the tractability of the convergence analysis, convex penalties typically increase the estimation bias to reduce the variance of the estimates.

In the context of nondistributed optimization, nonconvex penalties [8], [9] have widely been studied to reduce estimation biases. *Weakly-convex* penalties [10]–[13], in particular, allow algorithmic solutions with global optimality because the convexity of the overall cost can be preserved if the loss function is strongly convex. In typical distributed settings, however, the amount of data available at each node is limited, so the local loss function is not necessarily strongly convex, and hence each local cost may not be convex.

To address the above challenge, we can use approaches dealing with nonsmooth nonconvex penalties. Most, if not all, of the previous studies on nonconvex distributed approaches [14]–[16] have shown convergence to a stationary point. Examples of approaches of this type include the Prox-DGD [17] and the in-network nonconvex optimization (NEXT) [18]. Although the approaches in [19] and [20] have shown convergence to a second-order optimal point, they require smoothness of the cost functions.

In our recent work [21], the minimax concave (MC) penalty [10], [11] has been used based on the following idea. An application of Moreau's decomposition decouples the MC penalty into a sum of a convex function and a negative squared ℓ_2 norm. The idea therein was to approximate this concave quadratic term by the sum of the squared inner product between the estimate and each input vector, where the approximation relies on the statistical orthogonality assumption of input vectors. The resultant penalty function, which is referred to as the approximate MC (AMC) penalty, convexifies the local cost, thereby ensuring convergence to a global minimizer. Despite global optimality, AMC tends to yield less accurate estimates than the original MC penalty particularly when the underlying assumption is violated. This fact motivates us to devise a better formulation using the MC penalty while guaranteeing global optimality.

In this paper, we propose a different formulation that utilizes the MC penalty without the need for approximations. Specifically, we introduce an additional term which we call *the consensus promoting penalty (CPP)* that makes each local cost function convex *with respect to each local variable*. In addition, we show by simulations that CPP will make the whole "global" cost function convex. The proposed formulation is referred to as the MC-CPP formulation. Numerical examples suggest that the MC-CPP cost function is convex, and the estimates generated by the PG-EXTRA algorithm implementing MC-CPP achieve the same level of errors as the centralized solution at the steady state.

II. Preliminaries

This section introduces notation, convex analytic tools, and the distributed sparse optimization problem.

A. Notation

Let $\mathbb{R}^N$ denote the $N(\in \mathbb{N})$ dimensional Euclidean space, where $\mathbb{N}$ denotes the set of nonnegative integers. The vector in $\mathbb{R}^N$ with all components set to one is denoted by $\mathbf{1}_N$, and the vector with all components set to zero is denoted by $\mathbf{0}_N$. The identity matrix is denoted by $\boldsymbol{I}_N$, and the identity operator is denoted by I. The $N \times N$ matrix with all zero components

This work was supported by JST SICORP Grant Number JPMJSC20C6, Japan. The authors also acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) under grant 01DR21009 and the program "Souverän. Digital. Vernetzt." Joint project 6G-RIC, project identification number: 16KISK020K.

A full version of this work is given in [1].

is denoted by $\boldsymbol{O}_N$. A k sparse vector $\boldsymbol{x} \in \mathbb{R}^N$ is a vector with at most k nonzero components. The ith element of a vector $\boldsymbol{x} \in \mathbb{R}^N$ is denoted by x_i, $i \in \{1, 2, \cdots, N\}$. Similarly, the (i, j) component of a matrix $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ is denoted by x_{ij}, where $i, j \in \mathbb{N}$ and $i \in \{1, 2, \cdots N\}, j \in \{1, 2, \cdots M\}$. Let $\mathrm{tr}(\boldsymbol{A})$ denote the sum of the main diagonal elements of a square matrix. The ℓ_p norm ($p \in \{1, 2, \cdots\}$) of vector $\boldsymbol{x} \in \mathbb{R}^N$ is defined by $\|\boldsymbol{x}\|_p := (\sum_{i=1}^N x_i^p)^{1/p}$. The inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ is defined by $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \sum_{i=1}^N x_i y_i$. The Frobenius norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ is defined by $\|\boldsymbol{A}\|_{\mathrm{F}} := (\sum_{i=1}^N \sum_{j=1}^M a_{ij}^2)^{1/2}$.

B. Convex analytic tools

A function $f : \mathbb{R}^N \to (-\infty, +\infty)$ is convex if $f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \le tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ and every $t \in (0, 1)$. A function is η-weakly convex if $f + \eta\frac{1}{2}\|\cdot\|_2^2$ is convex for some $\eta > 0$. Furthermore, a function is η-strongly convex when $f - \eta\frac{1}{2}\|\cdot\|_2^2$ is convex. We denote by $\Gamma_0(\mathbb{R}^N)$ the class of proper lower semicontinuous convex functions[1] from $\mathbb{R}^N$ to $(-\infty, +\infty]$. The Fenchel conjugate function of $f \in \Gamma_0(\mathbb{R}^N)$ is defined by $f^*(\boldsymbol{x}) := \sup_{\boldsymbol{y}\in\mathbb{R}^N}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{y}))$ [22]. The Moreau envelope ${}^{\gamma}f$ of f of index $\gamma > 0$ is defined by ${}^{\gamma}f(\boldsymbol{x}) := \inf_{\boldsymbol{y}\in\mathbb{R}^N}\left(f(\boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right)$ [22]. The proximity operator prox_f is defined by $\mathrm{prox}_f(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathbb{R}^N}\left(f(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right)$ [22]. The smallest and largest eigenvalues of a symmetric matrix $\boldsymbol{A}$ are denoted by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$.

C. Distributed sparse optimization

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be an undirected graph, where $\mathcal{V} := \{1, 2, \cdots, m\}$ denotes a set of nodes (vertices), and $\mathcal{E}$ denotes a set of edges. Here, $(i, j) \in \mathcal{E}$ means that there is an edge between $i \in \mathcal{V}$ and $j \in \mathcal{V}$. Each node $i \in \mathcal{V}$ is connected to $r_i = |\mathcal{N}_i|$ nodes, where $|\mathcal{N}_i|$ denotes the cardinality of the set $\mathcal{N}_i$ of neighboring nodes of node i. We use the convention that $i \notin \mathcal{N}_i$. The graph is assumed connected; i.e., there exists a path that connects any pair of nodes in a single or multihop way. We consider a situation where each node i has the local sensing matrix $\boldsymbol{U}_i \in \mathbb{R}^{N\times l}$ and the local measurement vector $\boldsymbol{d}_i = \boldsymbol{U}_i^\top \boldsymbol{w}^* + \boldsymbol{n}_i \in \mathbb{R}^l$, where $\boldsymbol{w}^* \in \mathbb{R}^N$ is the sparse unknown vector (which is common to every node), and $\boldsymbol{n}_i \in \mathbb{R}^l$, $i = 1, 2, \cdots, m$, is the additive noise. The objective of distributed sparse optimization is to estimate the sparse vector $\boldsymbol{w}^*$ over the entire network via local updates at each node and information exchanges among nodes.

III. Proposed Approach

We present the proposed problem formulation for distributed sparse optimization based on the consensus promoting penalty, and we then discuss the convexity of the cost function. We finally present a distributed optimization algorithm to solve the proposed formulation.

[1]A convex function $f : \mathbb{R}^N \to (-\infty, +\infty]$ is proper if $f(\boldsymbol{x}) < +\infty$ for some $\boldsymbol{x} \in \mathbb{R}^N$. A convex function $f : \mathbb{R}^N \to (-\infty, +\infty]$ is lower semicontinuous if $\mathrm{lev}_{\le a} f := \{\boldsymbol{x} \in \mathbb{R}^N \mid f(\boldsymbol{x}) \le a\}$ is a closed set for any $a \in \mathbb{R}$.

A. Formulation based on MC penalty

We start with the following penalized formulation:

$$(\mathrm{P}_0)\ \min_{\boldsymbol{w}\in\mathbb{R}^N} \psi(\boldsymbol{w}) := \sum_{i=1}^m \underbrace{\left[\phi_i(\boldsymbol{w}) + \frac{\mu}{m}\pi(\boldsymbol{w})\right]}_{=:\psi_i(\boldsymbol{w})}, \tag{1}$$

where we refer to $\psi(\boldsymbol{w})$ as the *vector global cost* and $\psi_i(\boldsymbol{w})$ as the *vector local cost* of problem (P_0) because its argument $\boldsymbol{w}$ is a vector. Here, $\phi_i(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{U}_i^\top \boldsymbol{w} - \boldsymbol{d}_i\|_2^2$, $\mu > 0$, and

$$\pi(\boldsymbol{w}) := \|\boldsymbol{w}\|_1 - \min_{\boldsymbol{v}\in\mathbb{R}^N}\left(\|\boldsymbol{v}\|_1 + \frac{1}{2\gamma}\|\boldsymbol{w}-\boldsymbol{v}\|_2^2\right) \tag{2}$$

is a weakly convex penalty called the minimax concave (MC) penalty [10], [11], where $\gamma \in (0, +\infty]$. The ℓ_1 norm is separable and proximable (in the sense that the proximity operator can be easily computed). To preserve the overall convexity of the entire cost, we assume that $\eta := \lambda_{\min}(\boldsymbol{U}\boldsymbol{U}^\top) > 0$, so that $\phi(\boldsymbol{w}) := \sum_{i=1}^m \phi_i(\boldsymbol{w})$ is η-strongly convex for $\boldsymbol{U} := [\,\boldsymbol{U}_1\ \boldsymbol{U}_2\ \cdots\ \boldsymbol{U}_m]$. In contrast, each ϕ_i is not strongly convex in typical scenarios.

By virtue of Moreau's decomposition [22] ${}^{\gamma}f + {}^{1/\gamma}(f^*) \circ \gamma^{-1}I = \frac{1}{2\gamma}\|\cdot\|_2^2$ of $f \in \Gamma_0(\mathbb{R}^N)$, the penalty can be rewritten as

$$\pi(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 - \gamma^{-1}\left(\frac{1}{2}\|\boldsymbol{w}\|_2^2 - \gamma\ {}^{\frac{1}{\gamma}}(\|\cdot\|_1^*)(\gamma^{-1}\boldsymbol{w})\right). \tag{3}$$

Substituting (3) into (1) yields

$$\psi(\boldsymbol{w}) = \left(\sum_{i=1}^m \phi_i(\boldsymbol{w})\right) - \frac{\mu}{2\gamma}\|\boldsymbol{w}\|_2^2 + \mu\|\boldsymbol{w}\|_1 + \mu\ {}^{\frac{1}{\gamma}}(\|\cdot\|_1^*)(\gamma^{-1}\boldsymbol{w}). \tag{4}$$

Due to the η-strong convexity of $\phi(\boldsymbol{w})$ and the convexity of the third and fourth terms of (4), the convexity of the entire function ψ is ensured as long as $\eta - \mu/\gamma \ge 0$. However, the local function $\psi_i(\boldsymbol{w})$ cannot be convex unless $\phi_i(\boldsymbol{w})$ is strongly convex. Let $\boldsymbol{W} = [\boldsymbol{w}_1 \boldsymbol{w}_2 \cdots \boldsymbol{w}_m]^\top \in \mathbb{R}^{m\times N}$, where $\boldsymbol{w}_i \in \mathbb{R}^N$ is the estimate of $\boldsymbol{w}_{\mathrm{opt}}^{\psi} := \arg\min_{\boldsymbol{w}\in\mathbb{R}^N} \psi(\boldsymbol{w})$ at node $i \in \mathcal{V}$. Note here that the vector global cost $\psi(\boldsymbol{w})$ possesses a unique minimizer due to the strong convexity of $\psi(\boldsymbol{w})$. Define the consensus subspace $\mathcal{C} := \{\boldsymbol{W} \in \mathbb{R}^{m\times N} \mid \boldsymbol{w}_1 = \boldsymbol{w}_2 = \cdots = \boldsymbol{w}_m\}$. Problem (P_0) can then be formulated equivalently as the following constrained optimization problem:

$$(\mathrm{P}_1)\ \min_{\boldsymbol{W}\in\mathcal{C}} \Psi(\boldsymbol{W}) := \sum_{i=1}^m \underbrace{\left[\phi_i(\boldsymbol{w}_i) + \frac{\mu}{m}\pi(\boldsymbol{w}_i)\right]}_{=\psi_i(\boldsymbol{w}_i)}, \tag{5}$$

where we refer to $\Psi(\boldsymbol{W})$ as *the matrix global cost* and $\psi_i(\boldsymbol{w}_i)$ as the *matrix local cost* of problem (P_1) since the argument $\boldsymbol{W}$ of Ψ is a matrix. Under consideration is the challenging case where $\Psi(\boldsymbol{W})$ is nonconvex. With weak convexity of $\frac{\mu}{m}\Pi(\boldsymbol{W}) := \sum_{i=1}^m \frac{\mu}{m}\pi(\boldsymbol{w}_i)$, the *matrix global cost* $\Psi(\boldsymbol{W})$ can be convex only if $\Phi(\boldsymbol{W}) := \sum_{i=1}^m \phi_i(\boldsymbol{w}_i)$ is strongly convex, which can only be true, due to its separability, if every term $\phi_i(\boldsymbol{w}_i)$ is strongly convex. This implies that the *matrix global cost* $\Psi(\boldsymbol{W})$ is convex only if every node has sufficient information that admits a unique solution. In the following subsection, we present a simple and effective method that

Table I: Types of costs for MC-CPP

	local (each node)	global (entire network)
vector $\boldsymbol{w}$	(i) $\psi_i(\boldsymbol{w})$ vector local cost	(ii) $\psi(\boldsymbol{w})$ vector global cost
matrix $\boldsymbol{W}$	(iii) $\theta_i(\boldsymbol{w}_i)$ matrix local cost	(iv) $\Theta(\boldsymbol{W})$ matrix global cost

Table II: Convexity condition for each cost with MC-CPP

	local	global
vector $\boldsymbol{w}$	(i) $\eta_i > \frac{\mu}{m\gamma}$ (but $\eta_i = 0$ typically)	(ii) $\mu/\gamma < \eta$
matrix $\boldsymbol{W}$	(iii) $\kappa \geq \frac{\mu}{m\gamma}$	(iv) To be investigated (see Figure 2)

allows to deal with the challenging, yet realistic, nonconvex case if $l < N$ (i.e., when the number l of local linear equations is smaller than the number N of local variables). In fact, $\Psi(\boldsymbol{W})$ can be convex only if $l \geq N$ is satisfied.[2]

B. Proposed formulation with consensus promoting penalty: convexity analysis

We introduce a consensus promoting penalty (CPP) term (which is inspired from the incidence matrix [23], [24]) $C(\boldsymbol{W}) := \sum_{i=1}^m \frac{\kappa_i}{2}\|\boldsymbol{w}_i - \overline{\boldsymbol{w}}_i\|_2^2$ to get the following minimax concave CPP (MC-CPP) formulation:

$$(\mathrm{P}_2)\ \min_{\boldsymbol{W}\in\mathcal{C}} \Theta(\boldsymbol{W}) := \Psi(\boldsymbol{W}) + C(\boldsymbol{W}) \tag{6}$$

$$= \sum_{i=1}^m \underbrace{\Big[\underbrace{\phi_i(\boldsymbol{w}_i) + \frac{\mu}{m}\pi(\boldsymbol{w}_i)}_{=\psi_i(\boldsymbol{w}_i)} + \underbrace{\frac{\kappa_i}{2}\|\boldsymbol{w}_i - \overline{\boldsymbol{w}}_i\|_2^2}_{=:C_i(\boldsymbol{W})}\Big]}_{\theta_i(\boldsymbol{w}_i)}, \tag{7}$$

where $\overline{\boldsymbol{w}}_i = \sum_{j\in\mathcal{N}_i}\alpha_{ij}\boldsymbol{w}_j$, $\kappa_i \geq 0, \forall i \in \{1, 2, \cdots, m\}$, with $\alpha_{ij} = \alpha_{ji} \in (0, 1)$ such that $\sum_{j\in\mathcal{N}_i}\alpha_{ij} = 1$. Substituting (2) into (7), $\Theta(\boldsymbol{W})$ can be separated into smooth and nonsmooth parts as $\Theta(\boldsymbol{W}) = S(\boldsymbol{W}) + H(\boldsymbol{W})$, where

$$S(\boldsymbol{W}) = \underbrace{\sum_{i=1}^m \left(\phi_i(\boldsymbol{w}_i) + C_i(\boldsymbol{W})\right) - \frac{\mu}{m\gamma}\sum_{i=1}^m \frac{1}{2}\|\boldsymbol{w}_i\|_2^2}_{=:F(\boldsymbol{W})}$$
$$+ \frac{\mu}{m}\sum_{i=1}^m (^{1/\gamma}\|\cdot\|_1^*)(\gamma^{-1}\boldsymbol{w}_i), \tag{8}$$

and $H(\boldsymbol{W}) := \frac{\mu}{m}\sum_{i=1}^m \|\boldsymbol{w}_i\|_1$. Under certain parameter conditions, convexity of the matrix *local* cost can be ensured. However, this does not directly mean that the matrix *global* cost is convex. This is because CPP is regarded as a function of $\boldsymbol{w}_i$ in the matrix *local* cost, while it is a function of $\boldsymbol{W}$ in the matrix *global* cost. We summarize the four types of cost function in Table I. The vector global cost of the MC-CPP is $\psi(\boldsymbol{w})$ because $C(\boldsymbol{W}) = 0$ for any $\boldsymbol{W} \in \mathcal{C}$. Likewise, the vector local cost is $\psi_i(\boldsymbol{w})$ because $\theta_i(\boldsymbol{w}) = \psi_i(\boldsymbol{w})$ for any $\boldsymbol{W} \in \mathcal{C}$. Since $\boldsymbol{w}_i = \boldsymbol{W}^\top \boldsymbol{e}_i \in \mathbb{R}^N$, we rewrite CPP as $C_i(\boldsymbol{W}) = \frac{\kappa_i}{2}\|\boldsymbol{w}_i - \overline{\boldsymbol{w}}_i\|_2^2 = \frac{\kappa_i}{2}\|\boldsymbol{W}^\top \boldsymbol{v}_i\|_2^2$, where $\boldsymbol{v}_i := \boldsymbol{e}_i - \sum_{j\in\mathcal{N}_i}\alpha_{ij}\boldsymbol{e}_j \in \mathbb{R}^m$. Let us define $\boldsymbol{W} =: [\hat{\boldsymbol{w}}_1\ \hat{\boldsymbol{w}}_2\ \cdots\ \hat{\boldsymbol{w}}_N]$, and $\boldsymbol{V}_\kappa := [\sqrt{\kappa_1}\boldsymbol{v}_1\ \sqrt{\kappa_2}\boldsymbol{v}_2\ \cdots\ \sqrt{\kappa_m}\boldsymbol{v}_m] \in \mathbb{R}^{m\times m}$. The CPP term can then be expressed as

$$C(\boldsymbol{W}) := \sum_{i=1}^m C_i(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^m\sum_{j=1}^N \kappa_i \hat{\boldsymbol{w}}_j^\top \boldsymbol{v}_i \boldsymbol{v}_i^\top \hat{\boldsymbol{w}}_j$$
$$= \frac{1}{2}\sum_{j=1}^N \hat{\boldsymbol{w}}_j^\top \boldsymbol{V}_\kappa \boldsymbol{V}_\kappa^\top \hat{\boldsymbol{w}}_j. \tag{9}$$

Strong convexity of the consensus promoting penalty can thus be analyzed by inspecting strict positivity of the eigenvalues of $\boldsymbol{V}_\kappa \boldsymbol{V}_\kappa^\top$. We now highlight the following equivalence:

$$C(\boldsymbol{W}) = 0 \Leftrightarrow \boldsymbol{W} \in \mathcal{C}. \tag{10}$$

Here, $\Leftarrow$ is clear. To verify the converse $\Rightarrow$, we express $\boldsymbol{V}_\kappa$ as $\boldsymbol{V}_\kappa = \boldsymbol{V}\boldsymbol{\Lambda}_\kappa$ with $\boldsymbol{V} := [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_m] \in \mathbb{R}^{m\times m}$, and $\boldsymbol{\Lambda}_\kappa := \mathrm{diag}(\sqrt{\kappa_1}, \sqrt{\kappa_2}, \cdots, \sqrt{\kappa_m})$, where $\mathrm{diag}(\cdot)$ is the diagonal matrix with its diagonal entries given by the arguments. It is then sufficient, from the nonsingularity of the matrix $\boldsymbol{\Lambda}_\kappa$, to show that $(\mathrm{Ker}\boldsymbol{V}_\kappa^\top =)\mathrm{Ker}\boldsymbol{V}^\top = \mathrm{span}\{\mathbf{1}\}$. Clearly, the matrix $\boldsymbol{Y}^\top := \boldsymbol{I}_m - \boldsymbol{V}^\top$ shares the eigenspace with $\boldsymbol{V}^\top$, and the eigenvalue one of $\boldsymbol{Y}^\top$ and the eigenvalue zero of $\boldsymbol{V}^\top$ both correspond to the same eigenvector $\mathbf{1}_m$. Now, $\boldsymbol{Y}^\top$ is a nonnegative matrix that is irreducible[3] as the graph is assumed connected. The classical Perron Frobenius theory thus tells us that the eigenvalue one corresponding to the positive eigenvector $\mathbf{1}_m$ is simple (has algebraic multiplicity one), which implies the equality of the assertion. By (10), $C(\boldsymbol{W})$ does not change the entire cost on the consensus subspace $\mathcal{C}$, meaning that the global minimizer (the optimal solution) is preserved. More importantly, it also implies that the function $C(\boldsymbol{W})$ is nonzero, and it is actually strongly convex on the orthogonal complement $\mathcal{C}^\perp$ of $\mathcal{C}$. We summarize the important points below. (See [1] for proofs of the lemmas and proposition given in the following.)

Lemma 1. *(a) Each matrix local cost $\theta_i(\boldsymbol{w}_i)$ is convex if $\kappa_i \geq \frac{\mu}{m\gamma}$.*
(b) Each vector local cost $\psi_i(\boldsymbol{w})$ is convex iff $(\lambda_{\min}(\boldsymbol{U}_i\boldsymbol{U}_i^\top) =:)\eta_i \geq \frac{\mu}{m\gamma}$.
(c) The vector global cost $\psi(\boldsymbol{w})$ is convex iff $(\lambda_{\min}(\boldsymbol{U}\boldsymbol{U}^\top) =:)\eta \geq \frac{\mu}{\gamma}$.
(d) $\min_{\boldsymbol{W}\in\mathcal{C}} \Theta(\boldsymbol{W}) = \min_{\boldsymbol{W}\in\mathcal{C}} \Psi(\boldsymbol{W})$.
(e) $\arg\min_{\boldsymbol{W}\in\mathcal{C}} \Theta(\boldsymbol{W}) = \arg\min_{\boldsymbol{W}\in\mathcal{C}} \Psi(\boldsymbol{W})$.

The convexity conditions are summarized in Table II. Note here that $\eta_i = 0$ (i.e., ϕ_i is *not* strongly convex) in typical situations, because each local node is supposed to have no sufficient amount of measurements to identify the solution

[2] In environmental modeling scenarios, for instance, weather sensors may take few measurements and the environment changes slowly compared with the time to apply many iterations of the algorithm.

[3] A square matrix is said to be irreducible if the matrix cannot be reorganized into a block upper-triangular form by simultaneous row/column permutations.

so the nodes need to collaborate with each other (cf. Section III-A).

Lemma 2. *Regarding the costs* (i) – (iv) *presented in Table I, the following implications hold.*

1) *Convexity of* (i) *implies convexity of* (ii), (iii), (iv).
2) *Convexity of* (iv) *implies convexity of* (ii).

Proposition 1. *Let* range $\boldsymbol{U} = \mathbb{R}^N$ *so that the function* $\boldsymbol{w} \mapsto \sum_{i=1}^m \phi_i(\boldsymbol{w})$ *is* η*-strongly convex for* $\eta := \lambda_{\min}(\boldsymbol{U}\boldsymbol{U}^\top)$ *(see Section III-A). Assume that the graph is connected. Then* $\Theta(\boldsymbol{W})$ *is convex only if* $\mu \leq \eta\gamma$.

Most existing algorithms for distributed convex optimization assume convexity of the *vector local cost*, which is not apparent with the use of CPP. In some algorithms such as the PG-EXTRA, however, the convexity of the vector local cost itself is not used explicitly to prove the convergence of the algorithm, and it is only used to ensure convexity of the matrix global cost, which is used directly to prove the convergence. Therefore, convexity of the matrix global cost will be studied by simulations in Section IV-B.

C. Distributed proximal debiasing-gradient (DPD) method

Let $\boldsymbol{W}_k \in \mathbb{R}^{m\times N}$ denote the variable matrix at time $k \in \mathbb{N}$, and $\boldsymbol{M}$, $\tilde{\boldsymbol{M}} \in \mathbb{R}^{m\times m}$ are the mixing matrices.[4] The gradient of $S(\boldsymbol{W})$ is denoted by $\nabla S(\boldsymbol{W}) = [\nabla s_1(\boldsymbol{w}_1)\nabla s_2(\boldsymbol{w}_2)\dots\nabla s_m(\boldsymbol{w}_m)]^\top$. By using the property $\nabla^\gamma f = \gamma^{-1}(I - \mathrm{prox}_{\gamma f})$ for $f \in \Gamma_0(\mathbb{R}^N), \gamma > 0$ [22, Ch. 14] in (8), we have $\nabla s_i(\boldsymbol{w}) = \nabla\phi_i(\boldsymbol{w}) - \frac{\mu}{\gamma m}\big(\boldsymbol{w} - \mathrm{prox}_{\gamma h}(\boldsymbol{w})\big) + \kappa_i(\boldsymbol{w} - \overline{\boldsymbol{w}}_i)$. By the triangle inequality and the (firm) nonexpansivity of $\mathrm{prox}_{\gamma h}$, the following inequality holds:

$$\begin{aligned}&\|\nabla s_i(\boldsymbol{x}) - \nabla s_i(\boldsymbol{y})\|_2\\ &\leq \|\nabla\phi_i(\boldsymbol{x}) - \nabla\phi_i(\boldsymbol{y})\|_2 + \left\|\left(\kappa_i - \frac{\mu}{\gamma m}\right)(\boldsymbol{x}-\boldsymbol{y})\right\|_2\\ &+\frac{\mu}{\gamma m}\left\|\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{x}) - \mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{y})\right\|_2 \leq (L_i + \kappa_i)\|\boldsymbol{x}-\boldsymbol{y}\|_2.\end{aligned}$$

Here, $L_i := \lambda_{\max}^{1/2}(\boldsymbol{U}_i\boldsymbol{U}_i^\top)$ is the Lipschitz constant of $\nabla\phi_i$; i.e., $\|\nabla\phi_i(\boldsymbol{x}) - \nabla\phi_i(\boldsymbol{y})\|_2 \leq L_i\|\boldsymbol{x}-\boldsymbol{y}\|_2$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$.

An application of PG-EXTRA [7] to the proposed formulation (P_2) yields Algorithm 1. Note that other algorithms can be applied to (P_2) that assume convex, proximable penalties. We refer to this particular method utilizing the two proximity operators as the distributed proximal debiasing-gradient method (DPD). Hereafter, we assume that all nodes have the same number of neighbors i.e., $r_1 = r_2 = \cdots = r_m =: r$ for simplicity.

IV. Numerical Examples

We study the performance of the proposed algorithm for $\phi_i(\boldsymbol{w}_i) = \frac{1}{2}\|\boldsymbol{u}_i^\top\boldsymbol{w}_i - d_i\|_2^2$ $(l = 1)$, where $\boldsymbol{u}_i \in \mathbb{R}^N$ is the input vector, distributed according to $\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}_N, \boldsymbol{I}_N)$ $i = 1, 2, \cdots, m$, and $d_i = \boldsymbol{u}_i^\top\boldsymbol{w}^* + n_i$ is the output (see Section 2.2). The normalized error is defined as $\mathrm{Error}(\boldsymbol{x}, \boldsymbol{y}) = 10\log_{10}\left(\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right)$. Furthermore, the components of the mixing matrix $\boldsymbol{M}$ are set to $m_{ij} = \begin{cases}\frac{\zeta}{r+1} & \text{if } j \in \mathcal{N}_i\\ 1 - \frac{\zeta r}{r+1} & \text{if } i = j\\ 0 & \text{otherwise,}\end{cases}$ where $\zeta \in (0, \frac{r+1}{r})$ indicates the strength of the influence of the neighboring estimates, and $\tilde{\boldsymbol{M}} = \frac{\boldsymbol{I}_m + \boldsymbol{M}}{2}$.

[4]The mixing matrices $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are subject to the following conditions [7]: (i) If $i \neq j$ and $(i,j) \notin \mathcal{E}$, then $M_{ij} = \tilde{M}_{ij} = 0$, (ii) $\boldsymbol{M} = \boldsymbol{M}^\top$, $\tilde{\boldsymbol{M}} = \tilde{\boldsymbol{M}}^\top$, (iii) $\mathrm{null}(\boldsymbol{M} - \tilde{\boldsymbol{M}}) = \mathrm{span}\{\boldsymbol{1}_m\}$, $\mathrm{null}(\boldsymbol{I}_m - \tilde{\boldsymbol{M}}) \supseteq \mathrm{span}\{\boldsymbol{1}_m\}$, and (iv) $\tilde{\boldsymbol{M}} \succ 0$ and $\frac{\boldsymbol{I}_m + \boldsymbol{M}}{2} \succeq \tilde{\boldsymbol{M}} \succeq \boldsymbol{M}$.

Algorithm 1 DPD-CPP derived from PG-EXTRA

Require: $\kappa_i > 0$, $i = 1, 2, \cdots, m$, $0 < \beta \leq 2\lambda_{\min}(\tilde{\boldsymbol{M}})/\max\{L_i + \kappa_i\}_{i=1}^m$, $\mu > 0$ and $\gamma > 0$ such that $\mu/\gamma < \min\{\lambda_2(\boldsymbol{V}_\kappa\boldsymbol{V}_\kappa^\top), \eta\}$
$\boldsymbol{W}_0$ is an arbitrary point

$\nabla S(\boldsymbol{W}_0) = \nabla\Phi(\boldsymbol{W}_0) - \frac{\mu}{m\gamma}\boldsymbol{W}_0 + \mathrm{diag}(\kappa_1, \kappa_2, \cdots, \kappa_m)(\boldsymbol{W}_0 - \overline{\boldsymbol{W}}_0) + \frac{\mu}{m\gamma}\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{W}_0)$
$\boldsymbol{W}_{\frac{1}{2}} = \boldsymbol{M}\boldsymbol{W}_0 - \beta\nabla S(\boldsymbol{W}_0)$
$\boldsymbol{W}_1 = \mathrm{prox}_{\frac{\mu\beta}{m}\|\cdot\|_1}(\boldsymbol{W}_0)$
for $k = 0, 1, 2, \cdots$ **do**
  $\nabla S(\boldsymbol{W}_{k+1}) = \nabla\Phi(\boldsymbol{W}_{k+1}) - \frac{\mu}{m\gamma}\boldsymbol{W}_{k+1} + \mathrm{diag}(\kappa_1, \kappa_2, \cdots, \kappa_m)(\boldsymbol{W}_{k+1} - \overline{\boldsymbol{W}}_{k+1}) + \frac{\mu}{m\gamma}\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{W}_{k+1})$
  $\boldsymbol{W}_{k+1+\frac{1}{2}} = \boldsymbol{M}\boldsymbol{W}_{k+1} + \boldsymbol{W}_{k+\frac{1}{2}} - \tilde{\boldsymbol{M}}\boldsymbol{W}_k - \beta[\nabla S(\boldsymbol{W}_{k+1}) - \nabla S(\boldsymbol{W}_k)]$
  $\boldsymbol{W}_{k+2} = \mathrm{prox}_{\frac{\mu\beta}{m}\|\cdot\|_1}(\boldsymbol{W}_{k+1+\frac{1}{2}})$
end for

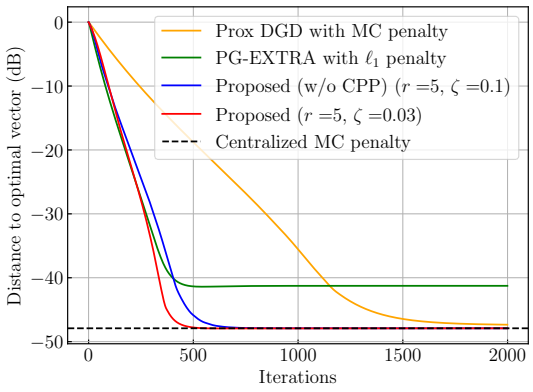

Figure 1: Learning curves in system mismatch of the average estimates.

A. Learning speed of proposed approaches

The dimension of $\boldsymbol{w}^*$ is $N = 10$, the network consists of $m = 100$ nodes with $r = 5$ neighboring nodes, and the signal to noise ratio (SNR) is 30 dB. The estimandum $\boldsymbol{w}^*$ is 30% sparse. For CPP, we let $\boldsymbol{V}^\top = \frac{r+1}{r}(\boldsymbol{I}_m - \boldsymbol{M})$ with $\zeta = 1$. The parameters of the proposed algorithm are set to $\mu = 9.0\times 10^{-3}$, $\gamma = 0.1.5$, $\beta = 0.022$, $(\zeta, \kappa) = (0.15, 0), (0.4, 10)$. The parameters of PG-EXTRA with the ℓ_1 penalty (DPD with $\gamma = +\infty, \kappa = 0$) are set to $\mu = 1.9\times 10^{-3}, \beta = 0.011$, and $\zeta = 1$. Prox-DGD [17] is also tested for reference, where the step size of Prox-DGD is set to 0.014, and the parameters for the MC penalty are set to the same value as DPD.

Figure 1 plots $\mathrm{Error}(\boldsymbol{w}^*, \frac{1}{m}\boldsymbol{W}^\top\boldsymbol{1}_m)$, the system mismatch of an average estimate over nodes, where the dashed line

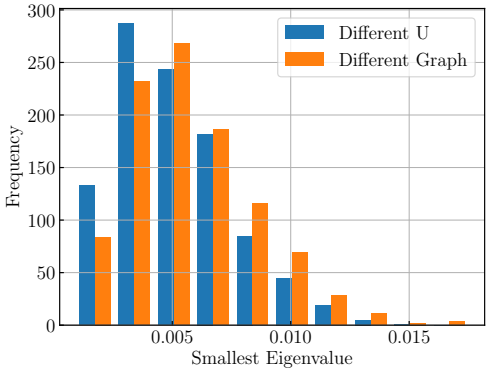

Figure 2: A histogram of the smallest eigenvalue of Hessian of $\Theta(\boldsymbol{W})$ with different $\boldsymbol{U}$ and different graphs.

at the bottom is a benchmark, showing the performance of a centralized approach to problem (P_0). One can see that the proposed MC-based approaches significantly outperform the ℓ_1-based approach, while the Prox-DGD method suffers from slow convergence. One can also see that the proposed method achieves the same error as the centralized solution (the benchmark) even without CPP, which leaves the possibility that PG-EXTRA applied to the minimization problem of $\Psi(\boldsymbol{W})$ would still enjoy provable convergence.

B. Convexity with CPP

We study the convexity of the cost function with the CPP term, which is designed for the cost to become convex when $\boldsymbol{W} \in \mathcal{C}^{\perp}$. However, the assurance of the convexity over the whole space $\mathbb{R}^{N\times m}$ still remains to be proven. Specifically, we inspect strong convexity of the quadratic function $\Phi(\boldsymbol{W}) + C(\boldsymbol{W})$ for $\mu = 0$, or equivalently strict positivity of the smallest eigenvalue of its Hessian matrix. The parameters are set to $N = 10, m = 50$, and $r = 2$.

Figure 2 plots the frequency of the smallest eigenvalue of Θ over 1000 trials. The blue bar in the figure considers the case when $\boldsymbol{U}$ is changed using $\mathrm{numpy.random.randn}$, while the orange bar considers the case when the weights of the graph are fixed to $r = \frac{1}{d}$ with the links of underlying graph selected randomly. In both cases, it is clear that the smallest eigenvalue remains positive. This means that, in this scenario, CPP would serve to convexify the matrix global cost, which may suggest that weakly convex penalties can be used in distributed settings with convergence guarantees to a global minimizer.

V. Concluding Remarks

We introduce a consensus promoting penalty that has the object of convexifying a cost function containing the weakly convex MC penalty. The convexity of the cost is desirable because the formulation can yield accurate sparse estimates with convergence guarantees to a global minimizer. Numerical examples showed that the CPP convexified the entire matrix global cost, and it also accelerated the speed of convergence. The use of the MC penalty in online scenarios has been studied in [25], and an extension of the present work to online settings will be an interesting future work.

References

[1] 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 Transactions on Signal and Information Processing over Networks*, 2022. accepted for publication.

[2] D. Bertsekas and J. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., USA, 1989.

[3] 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 Processing Magazine*, 37(3):26–38, 2020.

[4] I. Notarnicola, Y. Sun, G. Scutari, and G. Notarstefano. Distributed big-data optimization via blockwise gradient tracking. *IEEE Transactions on Automatic Control*, 66(5):2045–2060, 2021.

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

[6] A. I. Chen and A. Ozdaglar. A fast distributed proximal-gradient method. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 601–608, 2012.

[7] W. Shi, Q. Ling, G. Wu, and W. Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, 2015.

[8] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. American Statistical Association*, 96(456):1348–1360, Dec. 2001.

[9] M. Yukawa and S. Amari. ℓ_p-regularized least squares $(0 < p < 1)$ and critical path. *IEEE Trans. Information Theory*, 62(1):488–502, Jan. 2016.

[10] I. Selesnick. Sparse regularization via convex analysis. *IEEE Transactions on Signal Processing*, 65(17):4481–4494, 2017.

[11] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.

[12] J. Abe, M. Yamagishi, and I. Yamada. Linearly involved generalized moreau enhanced models and their proximal splitting algorithm under overall convexity condition. *Inverse Problems*, 36(3):1–36, Nov. 2019.

[13] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada. Linearly-involved Moreau-enhanced-over-subspace model: debiased sparse modeling and stable outlier-robust regression. 2021. [Online]. Available: https://arxiv.org/abs/2201.03235.

[14] P. Lorenzo and G. Scutari. Distributed nonconvex optimization over time-varying networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4124–4128, 2016.

[15] D. Hajinezhad and M. Hong. Nonconvex alternating direction method of multipliers for distributed sparse principal component analysis. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 255–259, 2015.

[16] H. Sun and M. Hong. Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 38–42, 2018.

[17] J. Zeng and W. Yin. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing*, 66(11):2834–2848, 2018.

[18] P. D. Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[19] A. Daneshmand, G. Scutari, and V. Kungurtsev. Second-order guarantees of distributed gradient algorithms. *SIAM Journal on Optimization*, 30(4):3029–3068, 2020.

[20] M. Hong, M. Razaviyayn, and J. Lee. Gradient primal-dual algorithm converges to second-order stationary solution for nonconvex distributed optimization over networks. In *Proc. ICLM*, volume 80, pages 2009–2018, 2018.

[21] K. Komuro, M. Yukawa, and R. L. G. Cavalcante. Distributed sparse optimization with minimax concave regularization. *in Proc. IEEE Statistical Signal Processing Workshop*, 2021.

[22] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Space*. Springer, 2nd edition, 2017.

[23] D. Hajinezhad M. Hong and M. Zhao. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. *in Proc. of the 34th International Conference on Machine Learning*, pages 1529–1538, 2017.

[24] L. Jian, Y. Zhao, J. Hu, and P. Li. Distributed inexact consensus-based admm method for multi-agent unconstrained optimization problem. *IEEE Access*, 7:79311–79319, 2019.

[25] H. Kaneko and M. Yukawa. Normalized least-mean-square algorithms with minimax concave penalty. In *Proc. IEEE ICASSP*, pages 5445–5449, 2020.