Enhancing Convergence of Decentralized Gradient Tracking under the KL Property

Xiaokai Chen, Gesualdo Scutari, *Fellow, IEEE*, and Tianyu Cao

***Abstract*— We study the convergence behavior of the decentralized gradient-tracking algorithm SONATA for composite optimization problems involving a smooth (possibly nonconvex) component and a proper, convex, extended-real-valued nonsmooth term. Assuming that the objective satisfies the Kurdyka–Łojasiewicz (KL) property with exponent $\theta \in [0,1)$, we prove that the sequence generated by SONATA converges to a stationary solution and characterize its convergence rate as a function of θ. In particular, when $\theta \in (0,1/2]$, the convergence is R-linear; when $\theta \in (1/2,1)$, a sublinear rate is established; and when $\theta = 0$, the iterates either converge in finitely many steps or converge R-linearly. The obtained rates match the KL-based convergence behavior of centralized proximal-gradient methods for $\theta \in (0,1)$. Numerical experiments support the theoretical results.**

***Index Terms*— Decentralized optimization, gradient-tracking algorithm, Kurdyka-Łojasiewicz, linear rate.**

I. Introduction

THE paper studies nonconvex, nonsmooth optimization problem over networks of the following form:

$$\min_{x\in\mathbb{R}^d} \quad u(x) := f(x) + r(x), \tag{P}$$

where

$$f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{1}$$

with each $f_i : \mathbb{R}^d \to \mathbb{R}$ being the cost function of agent $i = 1,\dots,m$, assumed to be L_i-smooth (possibly nonconvex) and known only to agent i; $r : \mathbb{R}^d \to \mathbb{R}\cup\{-\infty,+\infty\}$ is a nonsmooth convex (extended-value) function, known to all agents, which can be used to enforce shared constraints or specific structures on the solution (e.g., sparsity). Agents are connected through a communication network, modeled as a fixed, undirected graph.

A vast range of applications belonging to the problem class (P) possess some structural properties, notably in the form of some growth property of u, which can be leveraged to enhance convergence of decentralized solution methods. In this paper we target the so-called *Kurdyka-Łojasiewicz* (KL) property, due to its vast range of applicability. The class of functions satisfying the KL property is ubiquitous, ranging from applications in optimization [1]–[7], machine learning (logistic regression, LASSO, and the Principal Component Analysis (PCA) are notorious examples), deep neural network [8], dynamical systems, and partial differential equations; see [9] and the references therein.

The KL property was originally introduced by Łojasiewicz for real-analytic functions [10]. Kurdyka later extended this framework to functions definable in o-minimal structures, which generalize semialgebraic and subanalytic geometry [11]. Subsequently, Bolte *et al.* extended the KL inequality to nonsmooth subanalytic functions [12].

In its simplest form [3], a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy the KL property with exponent $\theta \in [0,1)$ at a point $\bar{x} \in \mathbb{R}^d$ (possibly critical) if there exist constants $c > 0$ and $\theta \in [0,1)$ such that

$$||\nabla f(x)|| \geq c\,|f(x) - f(\bar{x})|^{\theta}, \tag{2}$$

for all x in a neighborhood of $\bar{x}$. This inequality enforces a lower bound on the gradient magnitude near $\bar{x}$, and therefore rules out arbitrarily flat behavior of f around $\bar{x}$. For the nonsmooth counterpart, we refer the reader to Sec. II-A.

In the recent years, the KL property has been successfully leveraged to enhance convergence of several solution methods for nonconvex, nonsmooth, *centralized* optimization problems; examples include inexact first-order methods [4], alternating (proximal) minimization algorithms [3], and parallel methods [13]. In particular, under the KL, the *entire* sequence is proved to converge to a stationary solution, unlike the weaker subsequence convergence guaranteed otherwise. The convergence rate is governed by the *KL exponent* θ [2]–[7], namely: linear convergence occurs when $\theta \in (0,1/2]$, sublinear rate for $\theta \in (1/2,1)$; and finally, finite iteration convergence for $\theta = 0$.

A natural open question is whether these favorable convergence properties extend to decentralized algorithms. Existing analyses of decentralized algorithms [14]–[18] applicable to problem (P) have not exploited the KL property of u, resulting in overly pessimistic convergence rate predictions compared to centralized methods (see Sec. I-B for a detailed review of the state of the art). For instance, when u in (P) is KL with $\theta \in (0,1/2]$, current decentralized analyses predict sublinear convergence rather than the linear convergence observed in centralized scenarios, contradicting numerical experiments. Two examples are reported in Fig. 1, where the decentralized algorithm SONATA [14] is applied to two nonconvex instances of (P) over a network modeled as an Erdos-Renyi graph–the LASSO problem with SCAD regularizer [7] (cf. Sec. II-A) and the PCA problem. The objective functions of both problems satisfy the KL property with exponent $1/2$. The figures certify sequence *linear* convergence o for both problems. This fact has no theoretical backing. This paper fills this gap.

A. Main contributions

Our technical contributions can be summarized as follow.

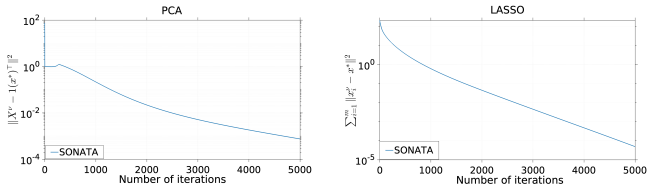

Fig. 1. Distance of the agents' iterates ($\boldsymbol{x}_i^{\boldsymbol{\nu}}, \boldsymbol{i} \in [\boldsymbol{m}]$) from a stationary solution ($\boldsymbol{x}^*$) of the PCA (left panel) and LASSO (right panel) problems, defined as $\sum_{i=1}^{m} \|\boldsymbol{x}_i^{\nu} - \boldsymbol{x}^*\|^2$, versus the iterations $\boldsymbol{\nu}$.

• **Convergence rate under the KL of u:** We establish convergence of the sequence generated by SONATA to stationary solutions of (P), and characterize its convergence rate, for the entire range of the KL exponent $\theta \in [0, 1)$ of u. Specifically,

i) When $\theta \in (0, 1/2]$, the sequence generated by SONATA converges *R-linearly*, requiring communication rounds to reach an ϵ-stationary solution on the order of

$$\tilde{\mathcal{O}}\left(\frac{L}{\kappa^{1/\theta}}\frac{1}{1-\rho}\log(1/\epsilon)\right),$$

where $L := \frac{1}{m}\sum_{i=1}^{m} L_i$, κ is a parameter related with the KL property of u (see Def. 1), and ρ represents the network connectivity (see (6) for the formal definition). The $\tilde{\mathcal{O}}$ notation hides log-dependencies on optimization parameters.

ii) For $\theta \in (1/2, 1)$, convergence of the sequence is sublinear, with rate of the order

$$\mathcal{O}(\epsilon^{-\frac{2\theta-1}{1-\theta}}).$$

iii) When $\theta = 0$, the sequence either converges to a stationary solution in a finite number of iterations or converges at R-linear rate independent of κ.

Notably, the rates in (i) and (ii) match those of the centralized proximal gradient algorithm (up to universal constants).

• **New convergence analysis:** The above rate are certified introducing a new line of analysis of independent interest, which represents the technical contribution of this work. In fact, existing convergence analyses of decentralized algorithms, mostly restricted to unconstrained, smooth instances of (P), either do not exploit the KL property (if present) at all or postulate that the *Lyapunov function* introduced to establish convergence is a KL function. Unfortunately, the KL property of the Lyapunov function or its exponent value *do not transfer to the objective function u of the optimization problem, and vice versa*. This is because the KL property is not closed with respect to the operations used to build the Lyapunov function from the objective function. Consequently, the challenge of establishing convergence guarantees for decentralized algorithms under the KL property of the objective function u–comparable to those certified in centralized settings–remains unresolved.

Our new analysis resolves this issue, leveraging *directly* the the KL property of the objective function u. First, we establish asymptotic convergence of the objective function gap and consensus and tracking errors, along the iterates produced by the SONATA algorithm, through the monotonically decaying trajectory of a newly introduced Lyapunov function. This guarantees that, after a sufficiently large but finite number of iterations–such that the function value gap falls below a critical threshold–the KL property of u can be engaged. Consequently convergence of the entire sequence to a critical point of u is established, at enhanced convergence rate, fully aligned with centralized guarantees.

B. Related works

Centralized setting: The KL property has been extensively used in the convergence analysis of centralized optimization methods. A pioneering work is [19], which appears to be the first to exploit the KL property to prove convergence of iterate sequences to stationary solutions for a broad class of algorithms satisfying suitable descent-type conditions. However, no convergence-rate characterization was provided therein. Subsequent works [2]–[7], [13] established convergence rates under KL assumptions for several important algorithmic frameworks, including the proximal-gradient method [2], alternating proximal minimization [3], the inexact Gauss–Seidel method [4], proximal alternating linearized minimization (PALM) [5], alternating forward–backward splitting [6], and the parallel asynchronous scheme FLEXA [13]. More recent studies [20]–[26] further broaden this line of research by enlarging the class of admissible KL functions and relaxing the algorithmic assumptions, while retaining strong convergence guarantees. In particular, [20] extends the KL framework to symmetric and strong KL variants, whereas [21]–[23] relax standard algorithmic requirements by allowing, e.g., additive errors in the relative-error condition [21] and nonmonotone descent dynamics [22]. Notably, [23] accommodates both features within a unified analysis. In [24]–[26], KL-based analysis is utilized for different accelerated algorithms solving (P).

Decentralized setting: Decentralized algorithms for various instances of Problem (P) have received significant attention in recent years [14], [27]–[34]. Early works [27], [31]–[33] focused on smooth objectives (i.e., $r = 0$), whereas [14], [28]–[30], [34] extended the analysis to constrained and composite formulations, typically under the assumption that the objective (sub)gradients remain bounded along the iterates. This restriction was later removed in [14], [34], and [14] also provided convergence rate guarantees (including over time-varying networks). None of the above works exploits objective-growth properties (when available), such as the KL property, to strengthen convergence guarantees. As a consequence, the resulting guarantees are often *pessimistic*–e.g., asymptotic convergence only, or sublinear rates–in contrast with the sharper results available in the centralized setting and with the numerical behavior observed in Fig. 1 (Sec. I) and Sec.V for decentralized problems.

Some works have explicitly leveraged growth conditions to improve the convergence guarantees of decentralized methods for special instances of (P), including [15], [35]–[39]. In particular, linear convergence has been established under the restricted secant condition [35], [36], the Polyak–Łojasiewicz (PL) condition [37], and the Luo–Tseng error-bound condition [15]. However, in the nonconvex setting, these assumptions are *stronger* than the KL property with exponent $1/2$ [40]. Moreover, the corresponding proof techniques

largely parallel those used under strong convexity, and thus are not directly applicable to the setting considered in this paper.

On the other hand, while studies such as [38], [39] have utilized the KL property in the convergence analysis of some decentralized algorithms, they have not conclusively achieved the desired outcomes. As discussed in Sec I-A , these works postulate the KL property of specifically constructed Lyapunov functions that meet the necessary conditions for convergence, rather than of the original objective functions. The link between the KL property of the objective function and such Lyapunov functions remains unclear, which leaves the question addressed in this paper unresolved.

C. Notation and paper organization

Throughout the paper, we will use the following notation. For any integer m, we write $[m] := \{1, 2, \cdots, m\}$. We use the convention that $0^0 = 0$. We denote by $[c_1 \leq (<)u \leq (<)c_2] := \{x : c_1 \leq (<)u(x) \leq (<)c_2\}$ the level set of the function u at $c \in \mathbb{R}$. We will use capital letters to represent matrices. In particular, 1 denotes the vector of all ones (whose dimensions are clear from the context); $J := 11^\top/m$ is the projection onto the consensus space. We use $\|\cdot\|_p$ to denote the ℓ_p-norm of any input vector ($\|\cdot\|$ will be the Euclidean norm) whereas $\|\cdot\|_p$ represents the operator norm induced by the ℓ_p-norm when the input is a matrix (with $\|\cdot\|$ being the Frobenius norm). Several operators appear in the paper. Given a proper, nonconvex function f, $\partial f(x)$ denotes the (limiting) subdifferential of f at x [41, Def. 8.3(b)]. Given $x \in \mathbb{R}^d$ and $r : \mathbb{R} \to \mathbb{R} \cup \{-\infty, \infty\}$, $\mathtt{prox}_{\alpha r}(x)$ denotes the proximal operator, defined as

$$\mathtt{prox}_{\alpha r}(x) := \mathrm{argmin}_{y \in \mathbb{R}^d} r(y) + \frac{1}{2\alpha}\|x - y\|^2.$$

The rest of the paper is organized as follows. Sec. II introduces the problem formulation along with the underlying assumptions. Sec. III presents the asymptotic convergence analysis of SONATA, which is instrumental to engage the KL property to enhance the convergence rate. Sec. IV contains the main technical result of the paper: the convergence rate analysis of SONATA under the KL property. Finally, some numerical experiments are presented in Sec. V.

II. Problem Setup and Background

We study the Problem (P) over a communication network. Followings are standard assumptions on (P).

Assumption 1 (objective function)**.**

(i) *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and ∇f_i is L_i-Lipschitz, with $L_i < \infty$; define*

$$L_{mx} := \max_{i \in [m]} L_i, \quad \textit{and} \quad L := \frac{1}{m}\sum_{i=1}^m L_i; \tag{3}$$

(ii) *$r : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, \infty\}$ is convex, proper and lower-semicontinuous;*

(iii) *$u : \mathbb{R}^d \to \mathbb{R}$ is lower bounded by some $\underline{u} \in \mathbb{R}$.*

The communication network of the agents is modeled as a time-invariant undirected graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, with the vertex set $\mathcal{V} := \{1, ..., m\}$ and the edge set $\mathcal{E} := \{(i, j)|i, j \in \mathcal{V}\}$ representing the set of agents and the communication links, respectively. Specifically, $(i, j) \in \mathcal{E}$ if and only if there exists a communication link between agent i and j. We make the blanket assumption that $\mathcal{G}$ is connected.

A. The KL property: definitions and illustrative examples

In this section, we formally introduce the KL property and provide examples arising in several applications.

The general definition of the KL can be found in [4]. Here we focus on the class of functions that are sharp up to some reparametrization, characterized through a so-called *desingularizing* function, $\phi : \mathbb{R} \to \mathbb{R}$. Intuitively, ϕ reparametrizes a *singular* region—where the gradients/subgradients may become arbitrarily small–into a *regular* region–where the gradients are bounded away from zero. A widely used choice is $\phi(s) := cs^{1-\theta}$, for some $\theta \in [0, 1)$ and $c > 0$, which goes back to [10]. The KL property associated with this desingularizing function is stated next.

Definition 1 (KL property [7])**.** *A proper closed function u satisfies the KL property at $\bar{x} \in \mathrm{dom}\, u$ with exponent $\theta \in [0, 1)$ if there exists a neighborhood E of $\bar{x}$ and constants $\kappa, \eta \in (0, \infty)$ such that*

$$\|g_x\| \geq \kappa(u(x) - u(\bar{x}))^\theta, \tag{4}$$

for all $x \in E \cap [u(\bar{x}) < u < u(\bar{x}) + \eta]$ and $g_x \in \partial u(x)$. We say that u is a KL function with exponent θ if it satisfies the KL property at every $\bar{x} \in \mathrm{dom}\, \partial u$ with the same exponent θ.

1) Some illustrative examples: We listed some motivating examples of functions u in the form (P) that satisfies the KL property, with specified exponent $\theta \in [0, 1)$.

(i) *ℓ_1-sparse linear regression:* Decentralized sparse linear regression consists of estimating a s-sparse parameter $x^* \in \mathbb{R}^d$ given noisy linear measurements $y = Ax^* + w$. Each agent i owns a subset (y_i, A_i) of the dataset, with $y_i \in \mathbb{R}^n$ and $A_i \in \mathbb{R}^{n \times d}$ such that $y = [y_1^\top, \ldots, y_m^\top]^\top$ and $A = [A_1^\top, \ldots, A_m^\top]^\top$. The decentralized LASSO estimation of $x^\star$ is an instance of (P), with $f_i(x) := (1/2)\|A_i x - b_i\|^2$ and $r(x) := \lambda||x||_1$, $\lambda > 0$. The resulting loss u is KL with exponent $\theta = 1/2$ [7].

(ii) *Sparse linear Regression with SCAD regularization:* To reduce bias in large coefficients typical of LASSO estimators, the smoothly clipped absolute deviation (SCAD) penalty uses a nonconvex regularizer expressible as a difference-of-convex function–see [42] for the specific expression of $r_+(x)$ and $r_-(x)$. The decentralized estimation fits (P), with $f_i(x) := (1/2)\|A_i x - b_i\|^2 - r_-(x)$ and $r(x) := r_+(x)$. The objective function satisfies the KL property with exponent $1/2$ [7].

(iii) *Logistic Regression:* Logistic regression estimates a parameter $x^* \in \mathbb{R}^d$ within a logistic model, where the log-odds of an event $a \in \mathbb{R}$ are modeled as the inner product between the model parameter x^* and the input features b. Specifically, the log-odds of a are given by

$$\log \frac{P_{x^\star}(a)}{1 - P_{x^\star}(a)} = -b^\top x^\star,$$

where $P_{x^\star}(a)$ denotes the predicted probability of $a \in \mathbb{R}$, depending on x^*. Given the data samples $(a_k, b_k)_{k=1}^N$ equally

distributed across the m agents, with each agent i owning the subset $(a_{ij}, b_{ij})_{j=1}^n$, the decentralized estimation of x^* is obtained by maximizing the log-likelihood of P_x (with respect to x). This formulation is an instance of Problem (P), with each $f_i(x) := \frac{1}{n}\sum_{j=1}^n \log\left(1+\exp(a_{ij}b_{ij}^\top x)\right)$ and $r(x) = \lambda\|x\|_1$, $\lambda > 0$. The objective function is KL with exponent $1/2$ [7].

(iv) *Principal Component Analysis (PCA):* The decentralized PCA extracts the leading eigenvector from distributed data sets $\{a_{ij}\}_{j=1}^n$, owned by each agent i, $i = 1, \dots, m$. This task is an instance of (P), with each $f_i(x) = -\frac{1}{n}\sum_{j=1}^n \|a_{ij}^\top x\|^2$, and $r(x) = \iota_{\mathcal{K}}(x)$. Here, $\mathcal{K} := \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$, and $\iota_{\mathcal{K}}(\bullet)$ denotes the indicator function of the convex set $\mathcal{K}$ (hence convex). The objective function is KL with exponent $1/2$ [43].

(v) *Phase Retrieval:* Decentralized phase retrieval estimates a signal $x^* \in \mathbb{R}^d$ from magnitude noisy measurements $y_k = |a_k^\top x^*| + w_k$, $k \in [N]$. Each agent i has local measurements $(y_{ij})_{j=1}^n$ and vectors $(a_{ij})_{j=1}^n$, fitting (P) with $f_i(x) := \frac{1}{2n}\sum_{j=1}^n \left(|a_{ij}^\top x| - y_{ij}\right)^2$ and $r \equiv 0$. The objective function u satisfies the KL property with $\theta = 1/2$ [44].

(vi) *Decentralized training of Neural Network:* Decentralized training of deep neural networks (DNNs) can be cast as a composite optimization problem. Consider a feedforward network with L hidden layers, where layer l has d_l neurons, trained on samples $\{(x_j, y_j)\}_{j=1}^N$, with $x_j \in \mathbb{R}^{d_0}$ and $y_j \in \mathbb{R}^{d_{L+1}}$. Let $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and $b_l \in \mathbb{R}^{d_l}$ denote the weight matrix and bias vector of layer l. Introducing auxiliary variables $\mathbf{a} = \{a_{l,j}\}$ and $\mathbf{u} = \{u_{l,j}\}$, the training of a regularized neural network can be equivalently written as [45]

$$\min_{\mathbf{a},\mathbf{W},\mathbf{b},\mathbf{u}} \quad \widetilde{\Psi}(\mathbf{a},\mathbf{W},\mathbf{b},\mathbf{u}) := \Big\{ \Psi(\mathbf{a},\mathbf{W},\mathbf{b}) + \sum_{j=1}^{N}\sum_{l=1}^{L}\left[\frac{1}{2}\|W_l a_{l-1,j} + b_l - a_{l,j} + u_{l,j}\|^2 + \delta_{\mathcal{C}_l}(a_{l,j})\right]\Big\},$$

where

$$\Psi(\mathbf{a},\mathbf{W},\mathbf{b}) := \sum_{j=1}^{N}\psi(W_{L+1}a_{L,j} + b_{L+1}; x_j, y_j) + \sum_{l=1}^{L+1} r_l(W_l),$$

ψ is a smooth convex loss, $\{r_l\}$ are convex (possibly nonsmooth) regularizers, and $\delta_{\mathcal{C}_l}$ is the indicator of a convex set $\mathcal{C}_l$ induced by the activation ϕ_l (e.g., $\mathcal{C}_l = \mathbb{R}_+^{d_l}$ for ReLU). By [8], $\widetilde{\Psi}$ is a KL function with exponent $\theta \in [0, 1)$.

B. The SONATA algorithm

Our study leverages the SONATA algorithm [14] to solve Problem (P), which we briefly recall next.

Each agent i maintains and updates iteratively a local copy $x_i \in \mathbb{R}^d$ of the global variable x, along with the auxiliary variable $y_i \in \mathbb{R}^d$ that represents the local estimate of ∇f. We stack the iterates and tracking variables into matrices, namely:

$$X := [x_1, x_2, \dots, x_m]^\top, \quad Y := [y_1, y_2, \dots, y_m]^\top.$$

Accordingly, we define the pseudogradient

$$\nabla F(X) := [\nabla f_1(x_1), \nabla f_2(x_2), \cdots, \nabla f_m(x_m)]^\top.$$

At iteration ν, the matrices above will take on the iteration index ν as a superscript, reflecting the corresponding iterates. The SONATA algorithm at iteration ν, reads

$$X^{\nu+1/2} = \mathtt{prox}_{\alpha R}(X^\nu - \alpha Y^\nu), \tag{5a}$$
$$X^{\nu+1} = W X^{\nu+1/2}, \tag{5b}$$
$$Y^{\nu+1} = W\left(Y^\nu + \nabla F(X^{\nu+1}) - \nabla F(X^\nu)\right). \tag{5c}$$

with initialization $x_i^0 \in$ *dom* r and $y_i^0 = \nabla f_i(x_i^0)$, $i \in [m]$. Here, $W = [w_{ij}]_{i,j=1}^m$ is the gossip matrix, satisfying Assumption 2 below; $\alpha > 0$ is an appropriate constant stepsize; and the proximal operator $\mathtt{prox}$ is applied row-wise.

Assumption 2 (Gossip matrix)**.** *Given the graph $\mathcal{G}$ (assumed to be connected), the mixing matrix $W := (w_{ij})_{i,j=1}^m$ satisfies the following:*

(i) $w_{ii} > 0$, *for all* $i \in [m]$;
(ii) $w_{ij} > 0$ *for all* $(i,j) \in \mathcal{E}$, *and* $w_{ij} = 0$ *otherwise;*
(iii) W *is doubly stochastic, i.e.,* $1^\top W = 1^\top$ *and* $W1 = 1$.

Define

$$w_{mx} := \sum_{i=1}^m \max_{j\in[m]} w_{ij} \quad \textit{and} \quad \rho := \|W - 11^\top/m\|_2. \tag{6}$$

Assumption 2 is quite standard in the literature of decentralized algorithms; several rules have been proposed to generate such gossip matrices, see, e.g., [46]–[48].

Notice that, under Assumption 2, it holds $\rho < 1$. Furthermore, since $w_{ij} \neq 0$ only if $(i,j) \in \mathcal{E}$), the SONATA algorithm (5) is fully decentralized, as all the communications are performed only among neighboring agents.

We conclude this section introducing the following notation compliant with the matrix/vector form of SONATA and used in the forthcoming analyses: let

$$D^\nu := X^{\nu+1/2} - X^\nu; \tag{7}$$

and let the consensus and tracking errors in matrix form:

$$X_\perp^\nu := (I-J)X^\nu, \quad X_\perp^{\nu+1/2} := (I-J)X^{\nu+1/2},$$
$$Y_\perp^\nu := (I-J)Y^\nu, \quad \Delta^\nu := [\nabla f(x_1^\nu), \dots, \nabla f(x_m^\nu)]^\top - Y^\nu.$$

Finally, we define the lifted functions

$$U(X) := \sum_{i=1}^m u(x_i) \quad \text{and} \quad R(X) := \sum_{i=1}^m r(x_i).$$

III. Asymptotic Convergence

This section establishes asymptotic convergence of SONATA for Problem (P). Unlike existing analyses (e.g., [14]), whose Lyapunov functions are built on the *average* of the agents' iterates and therefore do not directly enable the use of the KL property of the objective u, we construct a Lyapunov function based on the sum of local objective values, $U(X) = \sum_{i=1}^m u(x_i)$. This choice allows us to exploit the KL property of u directly.

The proof proceeds in two steps: i) establish an inexact descent inequality for $U(X^\nu)$, with perturbation terms due to consensus and tracking errors; and ii) combine such inequality with the consensus/tracking dynamics of SONATA into a single Lyapunov descent relation. Throughout this section, Assumptions 1 and 2 are tacitly assumed.

1) Step 1: inexact descent: The next result (a nonconvex counterpart of Jensen's inequality) is a key tool to control $U(X^\nu)$ along the local iterates rather than their average.

Lemma 1 [49, Thm. 1]**.** *For any L-smooth function $f:\mathbb{R}^d\to\mathbb{R}$, set of weights $\{w_i\}_{i=1}^m$, with $w_i\geq 0$ and $\sum_{i=1}^m w_i=1$, and $x_i\in\mathbb{R}^d$, $i=1,\cdots,m$, the following holds*

$$f\left(\sum_{i=1}^m w_i x_i\right)\leq\sum_{i=1}^m w_i f(x_i)+\frac{L}{2}\sum_{i=1}^m\sum_{j=1}^m w_i\,w_j\|x_j-x_i\|^2.$$

The first result is the one-step descent estimate for $U(X^\nu)$.

Lemma 2 (Inexact descent of U)**.** *For any $\xi>0$, the SONATA iterates satisfy*

$$\begin{aligned}U(X^{\nu+1})\leq U(X^\nu)-\left(\frac{1}{\alpha}-\frac{L}{2}-\frac{\xi}{2}\right)\|D^\nu\|^2+\frac{1}{2\xi}\|\Delta^\nu\|^2\\+4Lw_{mx}\big(\|X_\perp^\nu\|^2+\alpha^2\|Y_\perp^\nu\|^2\big).\end{aligned}\tag{8}$$

Proof. Applying Lemma 1 (row-wise) to the smooth component of $U(X)$ along $X^\nu\to X^{\nu+1}$ yields:

$$\begin{aligned}U(X^{\nu+1})&=\sum_{i=1}^m u(\sum_{j=1}^m w_{ij}x_j^{\nu+1/2})\\&=\sum_{i=1}^m f(\sum_{j=1}^m w_{ij}x_j^{\nu+1/2})+r(x_i^{\nu+1})\\&\leq\underbrace{\sum_{i=1}^m\sum_{j=1}^m w_{ij}u(x_j^{\nu+1/2})}_{\texttt{term I}}\\&\quad+\frac{L}{2}\sum_{i=1}^m\sum_{k=1}^m\sum_{l=1}^m w_{ik}w_{il}\underbrace{\|x_k^{\nu+1/2}-x_l^{\nu+1/2}\|^2}_{\texttt{term II}}\\&\quad+\sum_{i=1}^m\underbrace{\left(r(x_i^{\nu+1})-\sum_{j=1}^m w_{ij}r(x_j^{\nu+1/2})\right)}_{\texttt{term III}}.\end{aligned}\tag{9}$$

The inequality (8) follows from (9) and the following bounds of `term I`-`term III`:

$$\texttt{term I}\leq U(X^\nu)-(\frac{1}{\alpha}-\frac{L}{2}-\frac{\xi}{2})\|D^\nu\|^2+\frac{1}{2\xi}\|\Delta^\nu\|^2,\tag{10}$$

for any given $\xi>0$;

$$\begin{aligned}\texttt{term II}&\overset{(5a)}{=}\|\texttt{prox}(x_k^\nu-\alpha y_k^\nu)-\texttt{prox}(x_l^\nu-\alpha y_l^\nu)\|^2\\&\leq\|x_k^\nu-x_l^\nu-\alpha(y_k^\nu-y_l^\nu)\|^2;\end{aligned}$$

and $\texttt{term III}\leq 0$. □

2) Step 2: Lyapunov function and its descent: We next combine (8) with the consensus/tracking dynamics of SONATA.

First, using [17, Lemma 3.3], the tracking mismatch is bounded as

$$\|\Delta^\nu\|^2\leq 2\|Y_\perp^\nu\|^2+4L_{\rm mx}^2\|X_\perp^\nu\|^2.\tag{11}$$

Substituting (11) into (8) yields

$$\begin{aligned}U(X^{\nu+1})\leq U(X^\nu)-\left(\frac{1}{\alpha}-\frac{L}{2}-\frac{\xi}{2}\right)\|D^\nu\|^2\\+\left(4Lw_{\rm mx}+\frac{2L_{\rm mx}^2}{\xi}\right)\|X_\perp^\nu\|^2+\left(4Lw_{\rm mx}\alpha^2+\frac{1}{\xi}\right)\|Y_\perp^\nu\|^2.\end{aligned}\tag{12}$$

Moreover, adapting [17, Prop. 3.5] to the update (5a), the consensus/tracking errors satisfy

$$\begin{aligned}\|X_\perp^{\nu+1}\|&\leq\rho\|X_\perp^\nu\|+\rho\|D^\nu\|,\\\|Y_\perp^{\nu+1}\|&\leq\rho\|Y_\perp^\nu\|+2\rho L_{\rm mx}\|X_\perp^\nu\|+\rho L_{\rm mx}\|D^\nu\|.\end{aligned}\tag{13}$$

To compactly represent the positive error terms in (12), define

$$\mathcal{E}^\nu:=c_1\|Y_\perp^\nu\|^2+c_2\|X_\perp^\nu\|^2,\qquad c_1,c_2>0.$$

Then (12) can be written as

$$U(X^{\nu+1})\leq U(X^\nu)-\left(\frac{1}{\alpha}-\frac{L}{2}-\frac{\xi}{2}\right)\|D^\nu\|^2+\tilde{c}_1\,\mathcal{E}^\nu,\tag{14}$$

where

$$\tilde{c}_1:=\max\left\{\frac{4Lw_{\rm mx}}{c_2}+\frac{2L_{\rm mx}^2}{c_2\xi},\ \frac{1}{c_1\xi}+\frac{4Lw_{\rm mx}\alpha^2}{c_1}\right\}.$$

Using (13), one obtains the following recursion for $\mathcal{E}^\nu$:

$$\begin{aligned}\mathcal{E}^{\nu+1}\leq&\rho^2\max\left\{3,2+6\frac{c_1}{c_2}L_{\rm mx}^2\right\}\mathcal{E}^\nu\\&+(3\rho^2L_{\rm mx}^2c_1+2\rho^2c_2)\|D^\nu\|^2.\end{aligned}\tag{15}$$

A convenient choice is $c_1=2$ and $c_2=4L_{\rm mx}^2$, which minimizes the contraction factor in (15) and gives

$$\tilde{c}_1=\frac{1}{2\xi}+\max\left\{\frac{L}{L_{\rm mx}^2}w_{\rm mx},\,2L\alpha^2w_{\rm mx}\right\}.\tag{16}$$

We can now combine (14) and (15) into a single Lyapunov descent inequality.

Proposition 1 (Lyapunov descent)**.** *Define*

$$\mathcal{L}^\nu:=U(X^\nu)+\gamma\mathcal{E}^\nu,\tag{17}$$

where $\gamma>0$. Then

$$\mathcal{L}^{\nu+1}\leq\mathcal{L}^\nu-\tilde{c}_2\|D^\nu\|^2-\tilde{c}_3\mathcal{E}^\nu,\tag{18}$$

with

$$\tilde{c}_2:=\frac{1}{\alpha}-\frac{L}{2}-\frac{\xi}{2}-14L_{mx}^2\gamma\,\rho^2,\tag{19}$$

and

$$\tilde{c}_3:=(1-5\rho^2)\gamma-\tilde{c}_1.\tag{20}$$

We are now ready to state the asymptotic convergence result.

Theorem 1. *Consider Problem (P) under Assumption 1. Let $\{X^\nu\}_{\nu\in\mathbb{N}}$ be the sequence generated by the SONATA algorithm (5a)-(5c), under Assumption 2. Suppose*

(i) *$\rho<1/\sqrt{5}$, and*

(ii) *the free parameters $\alpha,\gamma,\xi>0$ are chosen such that*

$$\gamma>\frac{1/(2\xi)+Lw_{mx}/L_{mx}^2}{1-5\rho^2},$$

$$\alpha < \min\left\{\frac{1}{L/2+\xi/2+14L_{mx}^2\gamma\rho^2}, \sqrt{\frac{(1-5\rho^2)\gamma-1/(2\xi)}{2Lw_{mx}}}\right\}.$$

Then, as $\nu\to\infty$,

$$\|D^\nu\|,\ \|X_\perp^\nu\|,\ \|Y_\perp^\nu\|,\ \|\Delta^\nu\|\to 0,$$

and

$$U(X^\nu),\ U(X^{\nu+1/2})\to\mathcal{L}^\infty,$$

for some $\mathcal{L}^\infty\in\mathbb{R}$. *Furthermore, every accumulation point* X^∞ *of* $\{X^\nu\}_{\nu\in\mathbb{N}}$ *is of the form* $X^\infty=1(x^*)^\top$, *with* x^* *being a stationary solution of* (P).

Proof. Under (i)–(ii), one has $\tilde{c}_2>0$ and $\tilde{c}_3>0$ in Proposition 1. Hence, by (18), $\{\mathcal{L}^\nu\}$ is nonincreasing. By Assumption 1, $\mathcal{L}^\nu$ is bounded below; therefore, $\mathcal{L}^\nu\to\mathcal{L}^\infty\in\mathbb{R}$. Summing (18) over ν, yields summability of $\{\|D^\nu\|^2\}$ and $\{\mathcal{E}^\nu\}$, hence $\|D^\nu\|,\ \|X_\perp^\nu\|,\ \|Y_\perp^\nu\|\to 0$. Then $\|\Delta^\nu\|\to 0$ follows from (11). Since $\gamma\mathcal{E}^\nu\to 0$, from (17) and $\mathcal{L}^\nu\to\mathcal{L}^\infty$ we obtain $U(X^\nu)\to\mathcal{L}^\infty$.

Next, the same decomposition used in the proof of Lemma 2 yields a reverse estimate of the form

$$U(X^{\nu+1/2})\geq U(X^{\nu+1})-C\big(\|X_\perp^\nu\|^2+\alpha^2\|Y_\perp^\nu\|^2\big),$$

for some constant $C>0$, which used in conjunction with (8) and $\|X_\perp^\nu\|,\|Y_\perp^\nu\|\to 0$, yield $|U(X^{\nu+1/2})-U(X^{\nu+1})|\to 0$. Because $U(X^{\nu+1})\to\mathcal{L}^\infty$, we conclude $U(X^{\nu+1/2})\to\mathcal{L}^\infty$.

Let X^∞ be an accumulation point of $\{X^\nu\}_{\nu\in\mathbb{N}}$, such that $\lim_{k\to\infty}X^{\nu_k}=X^\infty$, where $\{\nu_k\}_k\subseteq\mathbb{N}$ is a suitable subsequence. It must be

$$X^\infty=1(x^*)^\top,\quad \text{for some } x^*\in\mathbb{R}^d. \tag{21}$$

Further, using again $\{\nu_k\}_k$ without loss of generality (w.l.o.g), we have

$$Y^\infty:=\lim_{k\to\infty}Y^{\nu_k}=\ 1\nabla f(x^*)^\top,$$

$$\lim_{k\to\infty}X^{\nu_k+1/2}\overset{(7)}{=}\lim_{k\to\infty}X^{\nu_k}+D^{\nu_k}=X^\infty.$$

Invoking the continuity of the prox operator, we deduce

$$\begin{aligned}X^\infty&=\lim_{k\to\infty}X^{\nu_k+1/2}\overset{(5a)}{=}\lim_{k\to\infty}\mathtt{prox}_{\alpha R}(X^{\nu_k}-\alpha Y^{\nu_k})\\&=\mathtt{prox}_{\alpha R}(1(x^*)^\top-\alpha\cdot 1\,(\nabla f(x^*))^\top),\end{aligned}$$

which, together with (21), implies $0\in\partial u(x^*)$. Therefore x^* is a stationary point of u in (P). □

The next section establishes sequence convergence along with a convergence rate, under the KL property of u.

IV. Convergence Rate under the KL Property

Through the section, we postulate the setting of Theorem 1 and the KL property of u at its stationary points.

Since $\mathcal{L}^\nu\to\mathcal{L}^\infty$, it is convenient to introduce the Lyapunov gap $\Delta\mathcal{L}^\nu:=\mathcal{L}^\nu-\mathcal{L}^\infty$, and rewrite (18) as

$$(\mathcal{T}^\nu)^2\leq\Delta\mathcal{L}^\nu-\Delta\mathcal{L}^{\nu+1}, \tag{22}$$

where

$$\mathcal{T}^\nu:=\sqrt{\tilde{c}_2\|D^\nu\|^2+\tilde{c}_3\mathcal{E}^\nu}.$$

By Theorem 1, $\mathcal{T}^\nu\to 0$ and $\Delta\mathcal{L}^\nu\to 0$ as $\nu\to\infty$.

A. Sequence convergence

We show that $\{X^\nu\}$ has finite length, which implies that $\{X^\nu\}$ is Cauchy and hence convergent.

By (5) and the non-expansiveness of the proximal operator,

$$\|X^{\nu+1}-X^\nu\|\leq c_3\mathcal{T}^\nu, \tag{23}$$

where

$$c_3:=\sqrt{\max\left\{\frac{\|W-I\|^2}{2\tilde{c}_3L_{\text{mx}}^2},\frac{2\|W\|^2}{\tilde{c}_2}\right\}}.$$

Hence, it suffices to prove the summability of $\{\mathcal{T}^\nu\}$.

Using (22) and the inequality (Lemma 8 in Appendix D)

$$a-b\leq\frac{1}{1-\theta}a^\theta(a^{1-\theta}-b^{1-\theta}),$$

with $a=\Delta\mathcal{L}^\nu$ and $b=\Delta\mathcal{L}^{\nu+1}$, we obtain

$$(\mathcal{T}^\nu)^2\leq\frac{1}{1-\theta}(\Delta\mathcal{L}^\nu)^\theta\underbrace{\left[(\Delta\mathcal{L}^\nu)^{1-\theta}-(\Delta\mathcal{L}^{\nu+1})^{1-\theta}\right]}_{\Delta\mathcal{L}_\theta^\nu}. \tag{24}$$

The key difficulty is that $\Delta\mathcal{L}^\nu$ does not inherit the KL property of u. We therefore exploit the KL property directly on u through the intermediate iterates $X^{\nu+1/2}$, and then transfer the estimate back to $\Delta\mathcal{L}^\nu$. Specifically, from (9) and (15),

$$\Delta\mathcal{L}^{\nu+1}\leq U(X^{\nu+1/2})-\mathcal{L}^\infty+\tilde{c}_4(\mathcal{T}^\nu)^2, \tag{25}$$

where

$$\tilde{c}_4:=\max\left\{\frac{14\gamma\rho^2L_{\text{mx}}^2}{\tilde{c}_2},\frac{5\gamma\rho^2+\tilde{c}_1-1/\xi}{\tilde{c}_3}\right\}. \tag{26}$$

The next proposition gives the required KL control of $U(X^{\nu+1/2})-\mathcal{L}^\infty$.

Proposition 2. *Inherit the setting of Theorem 1. Let* $X^\infty=1(x^\star)^\top$ *be an accumulation point of* $\{X^\nu\}$, *where* $x^\star$ *is some critical point of* u. *Further assume that* u *satisfies the KL property at* $x^\star$, *with exponent* $\theta\in[0,1)$ *and parameters* $\kappa,\eta\in(0,\infty)$. *Then, there exists a neighborhood* $\mathcal{N}_\infty$ *of* X^∞ *and* $\nu_1\geq 0$ *such that*

(i) *the set*

$$V:=\left\{\nu\geq\nu_1: X^{\nu+1/2}\in\mathcal{N}_\infty\right\}\neq\emptyset;$$

(ii) *for every* $\nu\in V$,

$$U(X^{\nu+1/2})-\mathcal{L}^\infty<1,\quad 0\leq\tilde{c}_5\mathcal{T}^\nu<1,$$

and, for $\theta\in(0,1)$,

$$U(X^{\nu+1/2})-\mathcal{L}^\infty<\kappa^{-\frac{1}{\theta}}\left(\tilde{c}_5\mathcal{T}^\nu\right)^{\frac{1}{\theta}},$$

where

$$\tilde{c}_5:=\max\left\{m^{\theta-\frac{1}{2}},1\right\}\cdot\sqrt{\max\left\{\frac{3\left(L^2+\frac{1}{\alpha^2}\right)}{\tilde{c}_2},\frac{3}{\tilde{c}_3}\right\}}. \tag{27}$$

Proof. See Appendix B. □

Combining Proposition 2 with (25) yields the following.

Corollary 1 (Local KL transfer to $\Delta\mathcal{L}^\nu$)**.** *In the setting of Proposition 2, and for* $\nu\in V$, *the following hold:*

(i) *If* $\theta \in (0, \frac{1}{2}]$, *then*

$$\Delta\mathcal{L}^{\nu+1} \le \tilde{c}_6(\mathcal{T}^\nu)^2, \tag{28}$$

with

$$\tilde{c}_6 = \left(\tilde{c}_5^2/\kappa^{\frac{1}{\theta}} + \tilde{c}_4\right), \tag{29}$$

and $\tilde{c}_4$ *and* $\tilde{c}_5$ *given in* (26) *and* (27)*, respectively;*

(ii) *If* $\theta \in (\frac{1}{2}, 1)$, *then*

$$\Delta\mathcal{L}^{\nu+1} \le \tilde{c}_7(\mathcal{T}^\nu)^{1/\theta}, \tag{30}$$

with

$$\tilde{c}_7 = \left((\tilde{c}_5/\kappa)^{\frac{1}{\theta}} + \tilde{c}_4\right); \tag{31}$$

(iii) *If* $\theta = 0$, *then either:*

(a) *there exists* ν_2 *such that* $\Delta\mathcal{L}^{\nu+1} \le \tilde{c}_4(\mathcal{T}^\nu)^2$, *for any* $\nu \in V$, $\nu \ge \nu_2$, *or*

(b) *there exists* ν_2' *such that* $\mathcal{T}^\nu = 0$, *for any* $\nu \in V$, , $\nu \ge \nu_2'$.

Proof. See Appendix C □

Using the relation (22) and Corollary 1, the next result establishes summability of $\{\mathcal{T}^\nu\}$ on consecutive blocks in V.

Proposition 3 (Block summability on V)**.** *Assume the setting of Corollary 1, and w.l.o.g.* $|V| = \infty$. *Then there exists an index* $\nu_0 \ge 0$ *such that: for any integers* $\nu_4 > \nu_3 \ge \nu_0$ *with* $\{\nu_3, \nu_3 + 1, \dots, \nu_4 - 1\} \subset V$, *the following holds for any* $\theta \in [0, 1)$*:*

$$\sum_{\nu=\nu_3}^{\nu_4-1} \mathcal{T}^\nu \le \max\left\{\frac{c_5(\tau)^{\nu_3}}{1-\tau},\ 2\mathcal{T}^{\nu_3} + 2c_6(\Delta\mathcal{L}^{\nu_3})^{1-\theta}\right\}, \tag{32}$$

where $c_5 := \sqrt{c_4}$, $c_6 := (2-2\theta)^{-1}\tilde{c}_7$, *and* $\tau = \sqrt{\frac{\tilde{c}_6}{1+\tilde{c}_6}}$.

Proof. We bound the RHS of (22) for different values of θ.

(i) Let $\theta \in (0, 1/2]$. Chaining (28) and (22), specifically (22) $+ \omega \times$ (28), with a sufficiently small constant $\omega > 0$, we obtain

$$(1+\omega)\Delta\mathcal{L}^{\nu+1} \le \Delta\mathcal{L}^\nu - (1-\omega\tilde{c}_6)\,(\mathcal{T}^\nu)^2, \quad \forall \nu \in V.$$

Choosing $0 < \omega \le 1/\tilde{c}_6$ ensures contraction of $\Delta\mathcal{L}^\nu$:

$$\Delta\mathcal{L}^{\nu+1} \le \frac{1}{1+\omega}\Delta\mathcal{L}^\nu \le c_4\left(\frac{1}{1+\omega}\right)^{\nu+1}, \quad \forall \nu \in V, \tag{33}$$

and some $c_4 > 0$. To optimize the contraction factor, we set $\omega = 1/\tilde{c}_6$. Using (33) in (22) yields

$$\mathcal{T}^\nu \le c_5(\tau)^\nu, \quad \forall \nu \in V, \quad \text{with} \quad \tau = \sqrt{\frac{1}{1+\omega}},\ c_5 = \sqrt{c_4} > 0. \tag{34}$$

(ii) Let $\theta \in (1/2, 1)$. Using (30) in (24) yields

$$\mathcal{T}^{\nu+1} \le \frac{1}{2}\mathcal{T}^\nu + c_6\Delta\mathcal{L}_\theta^{\nu+1}, \quad \forall \nu \in V, \tag{35}$$

where $c_6 = (2-2\theta)^{-1}\tilde{c}_7 > 0$.

Choose ν_0 large enough so that the corresponding case-dependent recursion is valid for all $\nu \in V$ with $\nu \ge \nu_0$ (i.e., (34) if $\theta \in (0, 1/2]$; (35) if $\theta \in (1/2, 1)$; and for $\theta = 0$, cases (a) and (b) in Corollary 1(iii) hold). If $\theta \in (0, 1/2]$, summing (34) over $\nu = \nu_3, \dots, \nu_4 - 1$ yields

$$\sum_{\nu=\nu_3}^{\nu_4-1} \mathcal{T}^\nu \le \sum_{\nu=\nu_3}^{\infty} c_5(\tau)^\nu = \frac{c_5(\tau)^{\nu_3}}{1-\tau}.$$

If $\theta \in (1/2, 1)$, summing (35) from $\nu = \nu_3$ to $\nu_4 - 1$ gives

$$\sum_{\nu=\nu_3}^{\nu_4-1} \mathcal{T}^\nu \le 2\mathcal{T}^{\nu_3} + 2c_6(\Delta\mathcal{L}^{\nu_3})^{1-\theta}.$$

If $\theta = 0$, either $\mathcal{T}^\nu$ vanishes after finitely many indices or (28) holds Combining the cases yields (32). □

To complete the proof–the finite length property of $\{X^\nu\}$–it is sufficient to show that the iterates eventually remain in $\mathcal{N}_\infty$ (equivalently, $\nu \in V$ for all sufficiently large ν). This allows us to extend the summability of $\{\mathcal{T}^\nu\}$ on V (Proposition 3) to *all* indices ν sufficiently large, and therefore concluding by (23) the convergence of $\{X^\nu\}$. Finally, since $D^\nu \to 0$, also $X^{\nu+1/2} = X^\nu + D^\nu \to X^\infty$.

Pick $r > 0$ such that

$$\mathcal{B}_{2r}(X^\infty) := \{X \in \mathbb{R}^{m\times n} : \|X - X^\infty\| < 2r\} \subseteq \mathcal{N}_\infty.$$

Using $D^\nu \to 0$ (Theorem 1) and the fact that X^∞ is an accumulation point of $\{X^\nu\}$ (and V is infinite, without loss of generality), we can pick such $\nu_3 \ge \max\{\nu_2, \nu_2'\}$, $\nu_3 \in V$ so that

$$\|D^\nu\| < \frac{r}{4}, \quad \forall \nu \ge \nu_3, \quad \|X^{\nu_3+1/2} - X^\infty\| < \frac{r}{4}, \quad \text{and}$$
$$c_3 \max\left\{\frac{c_5(\tau)^{\nu_3}}{1-\tau}, 2\mathcal{T}^{\nu_3} + 2c_6(\Delta\mathcal{L}^{\nu_3})^{1-\theta}\right\} < \frac{r}{2}, \tag{36}$$

where c_3 is the constant in (23).

Hence,

$$\|X^{\nu_3} - X^\infty\| < \frac{r}{2}, \quad \text{and thus} \quad X^{\nu_3} \in \mathcal{B}_r(X^\infty).$$

We prove next by contradiction that, for any $\nu \ge \nu_3$, $X^\nu \in \mathcal{B}_r(X^\infty) \subseteq \mathcal{N}_\infty$. Assume the contrary: there exists $\nu_4' > \nu_3$ (the smallest such index) with $\|X^{\nu_4'} - X^\infty\| \ge r$. Then $\|X^\nu - X^\infty\| < r$ for all $\nu_3 \le \nu < \nu_4'$. By (36), for all such ν,

$$\|X^{\nu+1/2} - X^\infty\| \le \|X^\nu - X^\infty\| + \|D^\nu\| < r + r/4 < 2r,$$

hence $X^{\nu+1/2} \in \mathcal{B}_{2r}(X^\infty) \subseteq \mathcal{N}_\infty$, i.e., $\nu \in V$. Therefore, $\{\nu_3, \nu_3 + 1, \dots, \nu_4' - 1\} \subset V$, and Proposition 3 applies replacing $\nu_4 = \nu_4'$. Using (23) and (32), we obtain

$$\|X^{\nu_4'} - X^\infty\| \le \|X^{\nu_3} - X^\infty\| + \sum_{\nu=\nu_3}^{\nu_4'-1} \|X^{\nu+1} - X^\nu\|$$
$$\le \frac{r}{2} + c_3 \sum_{\nu=\nu_3}^{\nu_4'-1} \mathcal{T}^\nu$$
$$\le \frac{r}{2} + c_3 \max\left\{\frac{c_5\tau^{\nu_3}}{1-\tau},\ 2\mathcal{T}^{\nu_3} + 2c_6(\Delta\mathcal{L}^{\nu_3})^{1-\theta}\right\} < r,$$

where the last inequality follows from (36). This contradicts $\|X^{\nu_4'} - X^\infty\| \ge r$.

Since $r > 0$ can be arbitrarily small, we conclude that $\{X^\nu\}$, as well as $\{X^{\nu+1/2}\}$, converge to X^∞.

B. Convergence rate analysis

Since $X^\nu \to X^\infty = 1(x^*)^\top$, we can bound the distance to the limit by the tail length: for any $\bar\nu$ sufficiently large,

$$\|X^{\bar\nu} - X^\infty\| \le \sum_{\nu=\bar\nu}^{\infty} \|X^{\nu+1} - X^\nu\| \le \mathcal{S}^{\bar\nu} := c_3 \sum_{\nu=\bar\nu}^{\infty} \mathcal{T}^\nu, \quad (37)$$

where $c_3 > 0$ is as in (23). The next theorem summarize the rates for $\mathcal{S}^\nu$ (hence for $\|X^\nu - X^\infty\|$) depending on θ.

Theorem 2. *Consider Problem (P) under Assumption 1. Let $\{X^\nu\}_{\nu\in\mathbb{N}}$ be the sequence generated by the SONATA algorithm under Assumption 2 and the tuning of α and γ as in Theorem 1, with $\xi = L$ therein. Further suppose $\rho < 1/\sqrt{5}$.*

Then, if the sequence $\{X^\nu\}$ has an accumulation point X^∞–it must be $X^\infty = 1(x^*)^\top$, for some critical point x^* of u–and u satisfies the KL property at $x^\star$ with exponent $\theta \in [0,1)$ and coefficient κ, then $X^\infty = 1(x^*)^\top$ is unique. Further, the following convergence rates hold:

(i) *If $\theta \in (1/2, 1)$, for all $\nu \in \mathbb{N}_+$ and some $c > 0$,*

$$\|X^\nu - 1(x^*)^\top\| \le c\nu^{-\frac{1-\theta}{2\theta-1}};$$

(ii) *If $\theta \in (0, 1/2]$, then for all $\nu \in \mathbb{N}_+$ and some $c' > 0$,*

$$\|X^\nu - 1(x^*)^\top\| \le c' \left(\frac{1}{1+\omega}\right)^{\nu/2}, \quad (38)$$

where

$$\omega = \mathcal{O}\left(\frac{L}{\kappa^{1/\theta}} + \frac{\rho^2}{1-5\rho^2}\frac{L_{mx}^2}{L^2}\right);$$

(iii) *If $\theta = 0$, then either $\{X^\nu\}$ converges to $1(x^*)^\top$ in a finite number of steps or there exists $c'' > 0$ such that for all $\nu \in \mathbb{N}_+$,*

$$\|X^\nu - 1(x^*)^\top\| \le c'' \left(\frac{1}{1+\omega'}\right)^{\nu/2}, \quad (39)$$

where

$$\omega' = \mathcal{O}\left(\frac{\rho^2}{1-5\rho^2}\frac{L_{mx}^2}{L^2}\right).$$

Proof. **(i)** $\theta \in (1/2, 1)$ **(sublinear).** Using the definition of $\mathcal{S}^\nu$ in (37) along with (24), yields: for all ν sufficiently large,

$$\begin{aligned}\mathcal{S}^\nu \le& c_3 \sum_{t=\nu}^{\infty} \frac{1}{2}\mathcal{T}^{t-1} + c_3c_6 \sum_{t=\nu}^{\infty} \Delta\mathcal{L}_\theta^t \\ \le& \frac{1}{2}\mathcal{S}^{\nu-1} + c_3c_6(\Delta\mathcal{L}^\nu)^{1-\theta} \\ \overset{(30)}{\le}& \frac{1}{2}\mathcal{S}^{\nu-1} + c_3c_6(\tilde c_7)^{1-\theta}\left(\mathcal{T}^{\nu-1}\right)^{\frac{1-\theta}{\theta}}.\end{aligned} \quad (40)$$

Since $\mathcal{S}^{\nu-1} - \mathcal{S}^\nu = c_3\mathcal{T}^{\nu-1}$, the above inequality implies that, for ν large enough,

$$(\mathcal{S}^\nu)^{\frac{\theta}{1-\theta}} \le C'(\mathcal{S}^{\nu-1} - \mathcal{S}^\nu), \quad (41)$$

for some constant $C' > 0$. By Lemma 10 (see Appendix D) and (41), there exists $c > 0$ such that

$$\|X^\nu - 1(x^*)^\top\| \le \mathcal{S}^\nu \le c\,\nu^{-\frac{1-\theta}{2\theta-1}},$$

for large ν. Therefore (possibly enlarging c) the bound holds for all $\nu \ge 1$.

(ii) $\theta \in (0, 1/2]$ **(R-linear).** For ν large enough, (34) holds with $\tau \in (0,1)$. Then (37) yields

$$\|X^\nu - 1(x^*)^\top\| \le \mathcal{S}^\nu \le c_3 \sum_{t=\nu}^{\infty} c_5(\tau)^t = \frac{c_3c_5}{1-\tau}(\tau)^\nu, \qquad \forall \nu \text{ large.}$$

Adjusting the constant to cover the finite prefix gives the desired expression (38).

(iii) $\theta = 0$ (finite termination or R-linear). If Corollary 1(iii)(b) holds, then X^ν reaches $1(x^*)^\top$ in finitely many steps. Otherwise, Corollary 1(iii)(a) holds and the same argument as in (ii) yields (39). □

Next we provide the iteration complexity of the SONATA algorithm (5) under the KL property with $\theta \in (0, 1/2]$.

Corollary 2. *Instate the setting of Theorem 2, with in particular $\theta \in (0, 1/2]$, and additionally*

$$\rho < \min\left\{\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{42w_{mx}+26}} \cdot \frac{L}{L_{mx}}\right\}. \quad (42)$$

Then, the number of iterations needed for $\{X^\nu\}$ to reach an ϵ-stationary solution of (P) is given by

$$\mathcal{O}\left(\frac{L}{\kappa^{1/\theta}}\log(1/\epsilon)\right),$$

where $\kappa \in (0, \infty)$ is the KL coefficient of u at x^*.

C. Discussion

Theorem 2 provides a complete characterization of the convergence behavior of the sequence $\{X^\nu\}$ generated by SONATA under the KL property of u. The resulting rates match the KL-based convergence behavior of centralized proximal-gradient methods [2], while filling a gap in the decentralized optimization literature, where analogous guarantees for gradient-tracking schemes under KL-type growth were not available. The following remarks are in order.

• *On the condition on ρ*: The requirements $\rho < 1/\sqrt{5}$ (Theorem 1) and (42) (Corollary 2) are connectivity conditions on the communication network, quantified by $\rho = \|W - \mathbf{1}\mathbf{1}^\top/m\|$. Such assumptions are standard in decentralized optimization and reflect the need for consensus/tracking errors to contract sufficiently fast relative to the descent in the objective-driven component (cf. the Lyapunov descent argument); see, e.g., [14], [50]. When the graph is fixed and the condition is not met a priori, it can be enforced by performing multiple communication rounds per gradient/proximal evaluation, effectively replacing W by a more contractive mixing operator. With plain repeated gossip (using the same W at each round), the resulting overhead is of order $\tilde{\mathcal{O}}((1-\rho)^{-1})$. This can be reduced to $\tilde{\mathcal{O}}((1-\rho)^{-1/2})$ using accelerated consensus based on Chebyshev or Jacobi polynomials [47], [48], [51], where $\tilde{\mathcal{O}}$ hides logarithmic factors.

A further conceptual point is worth emphasizing: several primal-dual decentralized frameworks admit network-independent stepsize conditions, e.g., [52]–[55], but the resulting complexity bounds typically depend on *local* conditioning

(e.g., through $L_{\max}$ and $\mu_{\min}$) and/or stronger structural assumptions. In contrast, our analysis targets convergence guarantees with dependence aligned with the *centralized* conditioning of the aggregate objective, which requires quantitatively controlling how fast consensus/tracking errors decay; this is precisely where explicit connectivity requirements such as bounds on ρ naturally enter.

• *On the linear convergence and total communication complexity*: For $\theta \in (0, 1/2]$, Theorem 2 yields an R-linear rate in iterations. If one enforces the connectivity condition via $\tilde{\mathcal{O}}((1-\rho)^{-1})$ communication rounds per gradient/proximal evaluation, then for arbitrary $\rho < 1$ the total number of communications required to reach an ϵ-stationary solution of (P) scales as

$$\tilde{\mathcal{O}}\left(\frac{L}{\kappa^{1/\theta}}\frac{1}{1-\rho}\log\left(\frac{1}{\epsilon}\right)\right).$$

Here, $\tilde{\mathcal{O}}$ hides logarithm dependence on L, L_{mx}, κ and θ.

In the special case where f in (P) is μ-strongly convex, u satisfies the KL property with $\theta = 1/2$ and $\kappa = \sqrt{\mu}$; thus $L/\kappa^{1/\theta} = L/\mu$, recovering the classical condition-number dependence reported for SONATA-type methods in strongly convex settings [14].

• *On the finite termination when $\theta = 0$:* Unlike centralized proximal-gradient methods, finite termination when $\theta = 0$ may fail in decentralized settings due to perturbations induced by consensus/tracking errors. Nevertheless, our analysis identifies verifiable conditions under which finite-time convergence occurs. In particular, finite termination is guaranteed if either:

- **(i)** there exists a subsequence of $\{U(X^{\nu+1/2})\}$ such that $U(X^{\nu+1/2}) - \mathcal{L}^{\infty} \geq 0$ along that subsequence (which, in particular, holds when the limit point x^* is a local minimizer of u); or
- **(ii)** u satisfies a slightly stronger variant of the KL property at x^* (see Definition 2 in Appendix A).

Condition **(i)** is consistent with common nonconvex landscapes in machine learning where first-order methods typically avoid strict saddles and converge to local minima under mild assumptions; see, e.g., [39], [56]–[58] and references therein. Condition **(ii)** is satisfied by a broad class of objectives; in particular, it holds for the examples discussed in Section II-A.1, and the associated exponent coincides with that of the standard KL property (see Appendix A).

• *On the existence of accumulation points*: Theorem 2 assumes the existence of an accumulation point of $\{X^{\nu}\}$. This is standard in KL-based analyses because the KL property is local and can be invoked only in a neighborhood of a limit point. For general nonconvex problems, existence of accumulation points cannot be guaranteed even in centralized optimization without additional assumptions (e.g., boundedness of level sets or boundedness of the iterates). Our assumption is weaker than many commonly used sufficient conditions in the literature [3], [4], [13] and is not restrictive, e.g., in the applications considered in Section II-A.1: for instance, coercive objectives ensure bounded level sets (hence bounded iterates and accumulation points), while in constrained formulations boundedness follows from compact feasible sets.

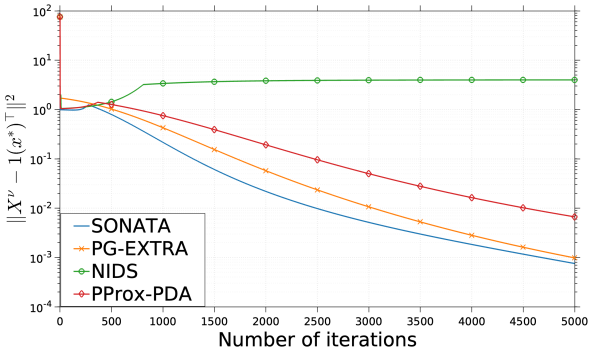

Fig. 2. Decentralized PCA: Distance of the iterates from a stationary solution vs. iterations.

V. Numerical Results

In this section, we present numerical experiments supporting our theoretical findings. All tests are performed over undirected networks with $m = 20$ agents; the communication graph is generated according to the Erdős–Rényi model with edge probability 0.5. The gossip matrix W used by all decentralized methods is constructed via the Metropolis–Hastings rule [47]. We benchmark SONATA against PG-EXTRA [59], NIDS [60], and PProx-PDA [61]. All methods are initialized at the same random point; algorithmic parameters are selected via grid search to obtain the best empirical performance.

A. Decentralized PCA

Consider the decentralized PCA problem, as described in Sec. II-A.1(iv). Each agent i holds a local data matrix $A_i = (a_{ij}^{\top}) \in \mathbb{R}^{20\times 80}$, whose rows $a_{ij}^{\top}$, $j = 1, \ldots, 20$, are i.i.d. samples from $\mathcal{N}(0, \Sigma)$. The covariance matrix Σ is generated as follows: we form an orthogonal matrix U by drawing a square matrix with i.i.d. standard normal entries and extracting its orthonormal factor via QR decomposition; the eigenvalues $\text{diag}(\Lambda)$ are drawn i.i.d. uniformly in $[0, 1]$, and $\Sigma = U\Lambda U^{\top}$.

Figure 2 reports $\log_{10}\|X^{\nu} - \mathbf{1}(x^*)^{\top}\|$ versus ν, where x^* denotes the (common) limit point reached by all methods. The plot highlights the linear convergence of SONATA and its favorable performance relative to the competing schemes. We also remark that, unlike for SONATA, linear convergence guarantees are not available for PG-EXTRA, NIDS, and PProx-PDA in this nonconvex setting.

B. Sparse linear regression with ℓ_1 regularization

We consider the decentralized LASSO problem in Sec. II-A.1(i). The ground-truth vector $x^* \in \mathbb{R}^{350}$ is generated by drawing $x \sim \mathcal{N}(0, I_{350})$ and setting to zero the smallest 70% (in magnitude) of its entries. Each agent observes $y_i = A_i x^* + w_i$, where $A_i \in \mathbb{R}^{20\times 350}$ has i.i.d. $\mathcal{N}(0, 1)$ entries with rows normalized to unit norm, and $w_i \sim \mathcal{N}(0, 0.1\, I)$. The regularization parameter is $\lambda = 0.5$.

Figure 3 shows $\log_{10}\|X^{\nu} - \mathbf{1}(x^*)^{\top}\|$ versus ν. The observed linear rate is consistent with our theory since the LASSO objective is KL with exponent $\theta = 1/2$. Comparable linear-rate guarantees are not available for the other decentralized baselines in this setting.

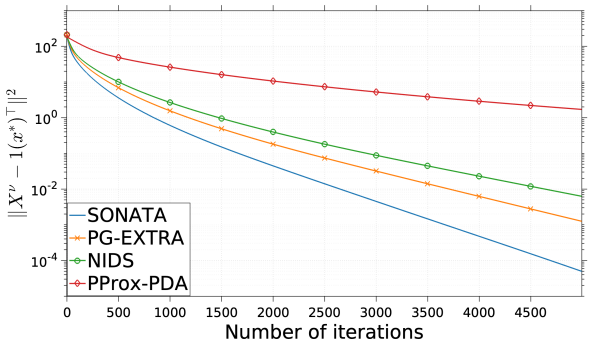

Fig. 3. LASSO with ℓ_1 regularization: distance of the iterates from a solution vs. iterations.

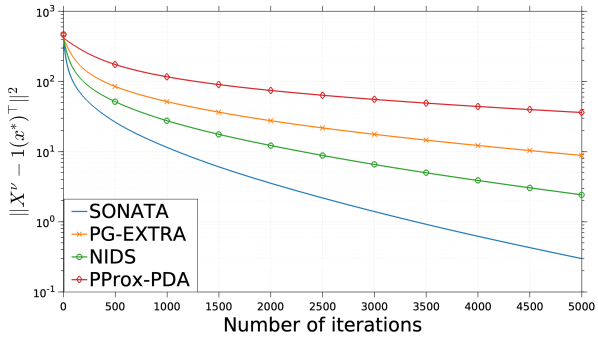

Fig. 4. Sparse linear regression with SCAD regularization: Distance from stationarity vs. iterations.

C. SCAD regularized distributed least squares

We next consider the SCAD-regularized sparse regression problem in Sec. II-A.1(ii). The ground-truth $x^* \in \mathbb{R}^{600}$ is generated as in Sec. V-B (Gaussian draw followed by 70% sparsification). Each agent collects $y_i = A_i x^* + w_i$, where $A_i \in \mathbb{R}^{25\times 600}$ has i.i.d. $\mathcal{N}(0,1)$ entries (with row normalization) and $w_i \sim \mathcal{N}(0, 0.01\, I)$. We set the ℓ_1 parameter $\lambda = 0.2$ and the SCAD parameter $a = 3.7$. Figure 4 reports $\log_{10} \|X^\nu - \mathbf{1}(x^*)^\top\|$ versus ν. All methods converge to the same stationary point, and the empirical behavior is consistent with the KL-based linear-rate regime $\theta = 1/2$.

D. Neural network with ℓ_1 regularization

Finally, we consider the neural-network example in Sec. II-A.1(vi). We train two-layer convolutional model ($L = 1$) on MNIST ($N = 60000$, $d_0 = 784$, $d_{L+1} = 10$). The network consists of one 3×3 convolutional layer with 16 feature maps followed by a fully connected layer with parameters $W_1 \in \mathbb{R}^{10\times 10816}$ and $b_1 \in \mathbb{R}^{10}$, using ReLU activations (function ϕ_l). We adopt the cross-entropy loss and ℓ_1 regularization $r_l(W_l) = \lambda\|W_l\|_1$ with $\lambda = 10^{-4}$. Figure 5 plots the global average loss versus ν. SONATA and PProx-PDA exhibit the expected sublinear behavior, in line with their theoretical guarantees. While PG-EXTRA also decreases the objective in this experiment, theoretical convergence guarantees for NIDS and PG-EXTRA are not available in this nonconvex setting.

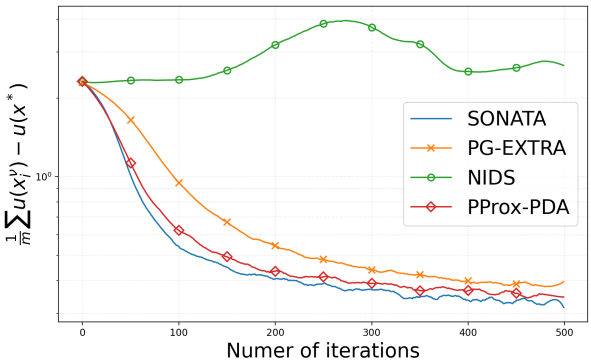

Fig. 5. Neural network with ℓ_1 regularization: Average loss vs. iterations

VI. Concluding Remarks

We provided a new line of analysis for the rate of convergence of the SONATA algorithm, to solve Problem (P), under the KL property of u. The established convergence rates match those obtained in the centralized scenario by the centralized proximal gradient algorithm, for different values of the exponent $\theta \in (0,1)$. When $\theta = 0$, finite time convergence of the sequence (as for the proximal gradient algorithm in the centralized case) cannot always be guarantees, due to the perturbation induced by consensus errors. In such cases we still identified sufficient conditions under which the SONATA algorithm matches the behavior of the centralized proximal gradient. A future direction to address this issue would be devising decentralized designs that escape local maxima.

Appendix

A. A Stronger version of the KL

This appendix introduces a stronger version of KL property than Definition 1, which allows SONATA to match the convergence rate results of the centralized proximal gradient, when the KL exponent of u (according to Definition 2) is zero. Specifically, we have the following.

Definition 2. *(A stronger version of the KL property)[23] A proper closed function u satisfies the KL property at $\bar{x} \in \mathrm{dom} u$ with exponent $\theta \in [0,1)$ if there exists a neighborhood E of $\bar{x}$ and numbers $\kappa, \eta \in (0,\infty)$ such that*

$$\|g_x\| \geq \kappa |u(x) - u(\bar{x})|^\theta, \tag{43}$$

for all $x \in E \cap [u(\bar{x}) - \eta < u < u(\bar{x}) + \eta]$ and $g_x \in \partial u(x)$. We call the function u a KL function with exponent θ if it satisfies the KL property at any point $\bar{x} \in \mathrm{dom}\partial u$ with the same exponent θ.

The relationship with the original KL property in Definition 1 follows readily.

Remark 1. *If a function u is proper and closed. Then u satisfies KL property (Def. 2) at $\bar{x} \in \mathrm{dom}\partial u$ with exponent $\theta \in [0,1)$ if and only if both u and $-u$ satisfy the original KL property (Def. 1) at $\bar{x}$ with some exponents $\theta_1, \theta_2 \in [0,1)$ respectively, where $\theta = \max\{\theta_1, \theta_2\}$.*

Several functions satisfy such a KL property.

Proposition 4. *Let u be a proper and closed sub-analytic function with closed domain. Then, u satisfies the KL property (Def. 2) at every $\bar{x} \in \mathrm{dom}\partial u$ with some exponent $\theta = \theta(\bar{x}) \in [0, 1)$.*

Proof. By in [12, Th. 3.1], both u and $-u$ satisfy the original KL property (Def. 1) at each $x \in \mathrm{dom}\partial u$ with a $\theta \in [0, 1)$. □

Furthermore, such a stronger version of the KL inherits most of the desirable properties of that in Definition 1, including exponent-preserving rules such as composition rule [7, Thm 3.2], minimum rule [7, Thm 3.1] and rule of separable sums[7, Thm 3.3]. To be concrete, with the help of several intermediate properties (Lemma 3, 4, 5) proved below, we provide next two classes of functions satisfying Definition 2 with exponent $\theta = 1/2$ (see Theorem 3 and Theorem 4 below). This also shows that all the illustrative examples discussed in Sec. II-A.1 satisfy Definition 2 with the *same* exponents θ as in Definition 1 (see Remark 2). Unless otherwise specified, in the following when calling the KL property (or a KL function), we refer to the stronger Definition 2.

Lemma 3. *Suppose u is a proper closed function, $\bar{x} \in \mathrm{dom}\,\partial u$ and $\bar{x}$ is not a stationary point ($0 \notin \partial u(\bar{x})$). Then for any $\theta \in [0, 1)$, u satisfies the KL property at $\bar{x}$ with exponent θ.*

Proof. See Lemma 2.1 in [7] with a minor variation. □

Definition 3 (Luo-Tseng Error Bound)**.** *[7, Def 2.1] Consider Problem (P). Let $\mathcal{X}$ be the set of stationary points of u. Suppose $\mathcal{X} \neq \emptyset$. We say that the Luo-Tseng error bound holds if for any $\zeta > \inf u$, there exists $c_1, \epsilon_1 > 0$ such that whenever $\|\mathtt{prox}_r(x - \nabla f(x)) - x\| < \epsilon_1$ and $u(x) < \zeta$,*

$$\mathit{dist}(x, \mathcal{X}) \leq c_1 \|\mathtt{prox}_r(x - \nabla f(x)) - x\|$$

Lemma 4 (Luo-Tseng error bound implies KL)**.** *Consider the problem (P) under Assumption 1. Suppose that the set $\mathcal{X}$ of stationary points of Problem (P) is nonempty and for any $x^* \in \mathcal{X}$, there exists $\delta > 0$ such that $u(x) = u(x^*)$ whenever $x \in \mathcal{X}$ and $\|x - x^*\| \leq \delta$. If the Luo-Tseng error bound holds for u, then u is a KL function with exponent $\theta = 1/2$.*

Proof. It follows similar proof of [7, Th. 4.1]. □

Lemma 5 (Exponent of the minimum of finitely many KL functions)**.** *Let u_i, $1 \leq i \leq r$, be proper closed function with $\mathrm{dom}u_i = \mathrm{dom}\,\partial u_i$ for all i, and $u := \min_{1 \leq i \leq r} u_i$ be continuous on $\mathrm{dom}\partial u$. Suppose further that each u_i is a KL function with exponent $\theta_i \in [0, 1)$ for $1 \leq i \leq r$. Then u is also a KL function with exponent $\theta = \max\{\theta_i : 1 \leq i \leq r\}$.*

Proof. The proof follows similar steps as in [7, Cor. 3.1]. □

We are ready to state the two major results of this section.

Theorem 3 (Convex problems with convex piecewise linear-quadratic regularizers)**.** *Suppose u is a proper closed function taking the form*

$$u(x) = \min_{1 \leq i \leq r} \underbrace{l(Ax) + r_i(x)}_{u_i(x)},$$

where $A \in \mathbb{R}^{m \times n}$, and r_i's are proper closed polyhedral functions, and l satisfies either one of the following conditions:

- *l is a proper closed convex function with an open domain, and is strongly convex on any compact convex subset of $\mathrm{dom}\, l$ and is twice continuously differentiable on $\mathrm{dom}\, l$;*
- *$l(y) = \max_{z \in D} \{\langle y, z\rangle - q(z)\}$ for all $y \in \mathbb{R}^m$, with D being a polyhedron and q being a strongly convex differentiable function with a Lipschitz continuous gradient.*

Suppose in addition that u is continuous on $\mathrm{dom}\ \partial u$, then u is a KL function with exponent $1/2$.

Proof. Following the analysis in [7, Cor. 5.1], one can show that, for each i, u_i satisfies all the assumptions in Proposition 4, and $\mathrm{dom}u_i = \mathrm{dom}\,\partial u_i$. The proof follows then by Proposition 3 (for the case $\mathrm{argmin}\ u_i = \emptyset$), Lemma 4 and Lemma 5. □

Theorem 4 (nonconvex minimum-of-quadratic regularizers)**.** *Consider the following class of functions*

$$u(x) = \min_{1 \leq i \leq r} \underbrace{\left\{\frac{1}{2} x^\top A_i x + b_i^\top x + \beta_i + r_i(x)\right\}}_{u_i(x)},$$

where r_i are proper closed polyhedral functions, $A_i \in \mathbb{R}^{n \times n}$ is symmetric, $b_i \in \mathbb{R}^n$ and $\beta_i \in \mathbb{R}$ for $i = 1, \cdots, r$. Suppose, in addition, that u is continuous on $\mathrm{dom}\,\partial u$. Then u is a KL function with exponent $1/2$.

Proof. Following the analysis in [7, Cor. 5.2], we have that, for each i, $u_i(x)$ satisfies all assumptions in Proposition 4 and $\mathrm{dom}u_i = \mathrm{dom}\partial u_i$. Then, the proof follows combining Proposition 3 (for the case $\mathrm{argmin}\ u_i = \emptyset$), 4 with Lemma 5. □

Remark 2. *If one adopts the symmetric KL property in Definition 2, then the KL exponents reported in Sec. II-A.1 remain valid at the limit points considered in our analysis; cf. Remark 1. In particular, examples (i)–(v) retain exponent $1/2$, while example (vi) admits some exponent $\theta \in [0, 1)$ under Proposition 4.*

B. Proof of Proposition 2

Since $X^\infty = 1(x^*)^\top$ is an accumulation point and $U(X^\nu) \to \mathcal{L}^\infty$ (Theorem 1), by continuity of U we have $\mathcal{L}^\infty = U(X^\infty)$. By the KL assumption on u and Lemma 7, U satisfies the KL property at X^∞ with exponent θ and parameters $(\kappa/\max\{m^{\theta-1/2}, 1\}, 1)$. Hence, there exists a neighborhood $\mathcal{N}_\infty$ of X^∞ such that, for any $X \in \mathcal{N}^\infty$ satisfying $U(X) - \mathcal{L}^\infty < 1$,

$$(U(X) - \mathcal{L}^\infty)^\theta \leq \left(\frac{\max\{m^{\theta-\frac{1}{2}}, 1\}}{\kappa}\right) \|G\|, \quad \forall G \in \partial U(X). \tag{44}$$

By Theorem 1, $\mathcal{T}^\nu \to 0$ and $U(X^{\nu+1/2}) \to \mathcal{L}^\infty$. Therefore, there exists $\nu_1 \in \mathbb{N}_+$ such that

$$\tilde{c}_5 \mathcal{T}^\nu < 1, \quad U(X^{\nu+1/2}) - \mathcal{L}^\infty < 1, \quad \forall \nu \geq \nu_1.$$

Define

$$V := \left\{\nu \geq \nu_1 : X^{\nu+1/2} \in \mathcal{N}_\infty\right\}.$$

We show that $V \neq \emptyset$. Indeed, since X^∞ is an accumulation point of $\{X^\nu\}$, there exists a subsequence $X^{\nu_k} \to X^\infty$. Moreover $D^\nu \to 0$ (Theorem 1), hence $X^{\nu_k+1/2} = X^{\nu_k} + D^{\nu_k} \to X^\infty$, implying $X^{\nu_k+1/2} \in \mathcal{N}_\infty$ for k large enough, and thus $V \neq \emptyset$.

Fix $\nu \in V$. Applying (44) at $X^{\nu+1/2}$ and taking $G_0^\nu \in \partial U(X^{\nu+1/2})$, where G_0^ν is the subgradient provided by Lemma 9 (Appendix D), we have

$$U(X^{\nu+1/2}) - \mathcal{L}^\infty \leq \left(\frac{\max\left\{m^{\theta-\frac{1}{2}}, 1\right\}}{\kappa}\right)^{1/\theta} \|G_0^\nu\|^{1/\theta}, \quad (45)$$

for $\theta \in (0,1)$. By Lemma 9 (Appendix D), for this G_0^ν we have

$$\|G_0^\nu\|^2 \leq 3\left(L^2 + \frac{1}{\alpha^2}\right)\|D^\nu\|^2 + 3\mathcal{E}^\nu. \quad (46)$$

Substitute (46) into (45) and we get our desired result. □

C. Proof of Corollary 1

By (25), for $\theta \in (0,1)$,

$$\Delta\mathcal{L}^{\nu+1} \leq \kappa^{-\frac{1}{\theta}}(\tilde{c}_5\mathcal{T}^\nu)^{\frac{1}{\theta}} + \tilde{c}_4(\mathcal{T}^\nu)^2, \quad \forall \nu \in V; \quad (47)$$

(i) $\theta \in (0, 1/2]$. Using (47), $0 \leq \tilde{c}_5\mathcal{T}^\nu < 1$, and $1/\theta \geq 2$, yields (28).

(ii) $\theta \in (1/2, 1)$. Since $\mathcal{T}^\nu \to 0$ (Theorem 1), by possibly shrinking $\mathcal{N}_\infty$ and increasing ν_1, we may assume (without changing notation) that $0 \leq \mathcal{T}^\nu < 1$, for all $\nu \in V$. Because $1/\theta < 2$, it follows that $(\mathcal{T}^\nu)^2 \leq (\mathcal{T}^\nu)^{1/\theta}$. Using this in (47), we obtain

$$\Delta\mathcal{L}^{\nu+1} \leq \left(\left(\frac{\tilde{c}_5}{\kappa}\right)^{1/\theta} + \tilde{c}_4\right)(\mathcal{T}^\nu)^{1/\theta} = \tilde{c}_7(\mathcal{T}^\nu)^{1/\theta},$$

which proves (30).

(iii) $\theta = 0$. From (25), if there exists ν_2 such that

$$U(X^{\nu+1/2}) - \mathcal{L}^\infty < 0, \qquad \forall \nu \geq \nu_2,\ \nu \in V,$$

then

$$\Delta\mathcal{L}^{\nu+1} \leq \tilde{c}_4(\mathcal{T}^\nu)^2, \qquad \forall \nu \geq \nu_2,\ \nu \in V,$$

which is (a).

Otherwise, there exist infinitely many $\nu' \in V$ such that $U(X^{\nu'+1/2}) - \mathcal{L}^\infty \geq 0$.

Pick such ν' and denote

$$V' := \{\nu' \in V : U(X^{\nu'+1/2}) - \mathcal{L}^\infty \geq 0\}.$$

Then, we have that $0 \leq U(X^{\nu'+1/2}) - \mathcal{L}^\infty < 1$, for all $\nu' \in V'$. By Lemma 9 (AppendixD), we have that for any $\nu' \in V'$ there exists $G^{\nu'} \in \mathbb{R}^{m\times d}$, $G^{\nu'} \in \partial U(X^{\nu'+1/2})$, such that

$$\|G^{\nu'}\|^2 \leq 3\left(L^2 + \frac{1}{\alpha^2}\right)\|D^{\nu'}\|^2 + 3\mathcal{E}^{\nu'}. \quad (48)$$

Notice that each row of $G^{\nu'}$ belongs to the (limiting) subdifferential of $u(x_i^{\nu'+1/2})$.

Further, by Lemma 7 (Appendix D), we have that

$$\kappa|U(X^{\nu'+1/2}) - \mathcal{L}^\infty|^0 < \|G^{\nu'}\|, \quad \forall \nu' \in V'. \quad (49)$$

Combining (48) and (49) yields

$$\mathcal{T}^{\nu'} \geq \frac{\kappa}{\tilde{c}_5} > 0, \quad \forall \nu' \in V', U(X^{\nu'+1/2}) \neq \mathcal{L}^\infty. \quad (50)$$

Using (22), (50), and the fact that $\Delta\mathcal{L}^\nu \to 0$ monotonically, one infers that there exists $\nu_2' \in V'$ such that

$$\mathcal{T}^\nu = 0, \quad \forall \nu \geq \nu_2'. \quad (51)$$

This proves (iii)(b). □

D. Auxiliary Results

Lemma 6. *For $x \in \mathbb{R}^d$ and $\theta \in (0,1)$, the following hold:*

(i) *If $\theta \leq \frac{1}{2}$, then $\|x\|_{1/\theta} \leq \|x\|_2$;*
(ii) *If $\theta > \frac{1}{2}$, then $\|x\|_{1/\theta} \leq d^{\theta-1/2}\|x\|_2$.*

Proof. By the Hölder's inequality, we have for any $0 < r < p$,

$$\|x\|_p \leq \|x\|_r \leq d^{1/r-1/p}\|x\|_p. \quad (52)$$

The inequalities in (i) and (ii) follow from (52), setting therein $p = 1/\theta$ and $r = 2$ for (i), and $p = 2$, $r = 1/\theta$ for (ii). □

Lemma 7 (Exponent for separable sums of KL functions)**.** *Let $F(X) = \sum_{i=1}^m f_i(x_i)$, where f_i is a proper closed function on $\mathbb{R}^d$ and $X = [x_1, x_2, \cdots, x_m]^\top$. Suppose further that for each i, f_i is continuous on $\mathrm{dom}\partial f_i$ and satisfies KL property (Def. 1) at $\bar{x}_i$ with exponent $\theta \in [0,1)$ and parameters (κ_i, η_i). Then F satisfies KL property (Def. 1) at $\bar{X} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m]^\top$ with exponent θ and parameters $(\kappa/\max\{m^{\theta-\frac{1}{2}}), 1\}$, with $\kappa = \min_i \kappa_i$.*

Proof. See the proof of Theorem 3.3 in [7] with minor variation and calculate the explicit expression of KL coefficient of F using Lemma 6. □

Lemma 8. *For $0 \leq b \leq a$, and $\theta < 1$, we have that*

$$a - b \leq \frac{1}{1-\theta}a^\theta(a^{1-\theta} - b^{1-\theta}).$$

Proof. Define $\Phi(x) : x^{1-\theta}$. Then, $\Phi'(x) = (1-\theta)x^{-\theta}$ is decreasing for $x > 0$. By mean value theorem, there exists $c \in [b, a]$ such that

$$\Phi(a) - \Phi(b) = \Phi'(c)(a-b) \geq \Phi'(a)(a-b).$$

$$\Rightarrow a - b \leq \frac{1}{1-\theta}a^\theta(a^{1-\theta} - b^{1-\theta}).$$

□

Lemma 9. *Consider Problem (P) under Assumption 1. Let $\{X^\nu\}_{\nu\in\mathbb{N}}$ be the sequence generated by the SONATA algorithm (5a)-(5c), under Assumption 2. Then there exists $G^\nu \in \partial U(X^{\nu+1/2})$ such that*

$$\|G^\nu\|^2 \leq 3\left(L^2 + \frac{1}{\alpha^2}\right)\|D^\nu\|^2 + 3\mathcal{E}^\nu.$$

Proof. By (5a), there exists $J^\nu \in \partial R(X^{\nu+1/2})$ such that

$$J^\nu + \frac{1}{\alpha}(X^{\nu+1/2} - X^\nu + \alpha Y^\nu) = 0.$$

That is, there exists $G^\nu \in \partial U(X^{\nu+1/2})$ such that

$$\begin{aligned}\|G^\nu\|^2 =& \sum_{i=1}^m \left\|\nabla f(x_i^{\nu+1/2}) - y_i^\nu - \frac{1}{\alpha}(x_i^{\nu+1/2} - x_i^\nu)\right\|^2 \\ \leq& 3\left(L^2 + \frac{1}{\alpha^2}\right)\|D^\nu\|^2 + 3\mathcal{E}^\nu.\end{aligned}$$

□

Lemma 10 (Th. 2 in [2]). *Let $\{\mathcal{S}^\nu\}$ be a nonnegative sequence satisfying*

$$(\mathcal{S}^\nu)^{\frac{\theta}{1-\theta}} \leq c(\mathcal{S}^{\nu-1} - \mathcal{S}^\nu), \quad \forall \nu \geq \nu_0,$$

there exists $c' > 0$ such that

$$\mathcal{S}^\nu \leq c'\nu^{-\frac{1-\theta}{2\theta-1}}, \quad \forall \nu \geq \nu_0.$$

REFERENCES

[1] H. Liu, A. M.-C. So, and W. Wu, "Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods," *Mathematical Programming*, vol. 178, no. 1, pp. 215–262, 2019.

[2] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, pp. 5–16, 2009.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality," *Mathematics of operations research*, vol. 35, no. 2, pp. 438–457, 2010.

[4] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1, pp. 91–129, 2013.

[5] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, 2014.

[6] P. Frankel, G. Garrigos, and J. Peypouquet, "Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates," *Journal of Optimization Theory and Applications*, vol. 165, pp. 874–900, 2015.

[7] G. Li and T. K. Pong, "Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods," *Foundations of computational mathematics*, vol. 18, no. 5, pp. 1199–1232, 2018.

[8] T. T.-K. Lau, J. Zeng, B. Wu, and Y. Yao, "A proximal block coordinate descent algorithm for deep neural network training," *arXiv preprint arXiv:1803.09082*, 2018.

[9] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, "Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity," *Transactions of the American Mathematical Society*, vol. 362, no. 6, pp. 3319–3363, 2010.

[10] S. Lojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," *Les équations aux dérivées partielles*, vol. 117, pp. 87–89, 1963.

[11] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, pp. 769–783, 1998.

[12] J. Bolte, A. Daniilidis, and A. Lewis, "The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.

[13] L. Cannelli, "Asynchronous parallel algorithms for big-data nonconvex optimization,". Ph.D. thesis, Purdue University Graduate School, 2019.

[14] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, pp. 497–544, 2019.

[15] Y. Tian, Y. Sun, and G. Scutari, "Asynchronous decentralized successive convex approximation," *arXiv preprint arXiv:1909.10144*, 2019.

[16] P. Di Lorenzo and G. Scutari, "Distributed nonconvex optimization over networks," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 229–232, IEEE, 2015.

[17] Y. Sun, G. Scutari, and A. Daneshmand, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM Journal on Optimization*, vol. 32, no. 2, pp. 354–385, 2022.

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

[19] P.-A. Absil, R. Mahony, and B. Andrews, "Convergence of the iterates of descent methods for analytic cost functions," *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547, 2005.

[20] G. Bento, B. Mordukhovich, T. Mota, and Y. Nesterov, "Convergence of descent methods under kurdyka-\l ojasiewicz properties," *arXiv preprint arXiv:2407.00812*, 2024.

[21] P. Ochs, Y. Chen, T. Brox, and T. Pock, "ipiano: Inertial proximal algorithm for nonconvex optimization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.

[22] Y. Qian and S. Pan, "Convergence of a class of nonmonotone descent methods for kurdyka–łojasiewicz optimization problems," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 638–651, 2023.

[23] J. Qiu, B. Ma, X. Li, and A. Milzarek, "A kl-based analysis framework with applications to non-descent optimization methods," *arXiv preprint arXiv:2406.02273*, 2024.

[24] A. Barakat and P. Bianchi, "Convergence rates of a momentum algorithm with bounded adaptive step size for nonconvex optimization," in *Asian conference on machine learning*, pp. 225–240, PMLR, 2020.

[25] W. Yang and W. Min, "An accelerated block proximal framework with adaptive momentum for nonconvex and nonsmooth optimization," *arXiv preprint arXiv:2308.12126*, 2023.

[26] X. Yang and L. Xu, "Some accelerated alternating proximal gradient algorithms for a class of nonconvex nonsmooth problems," *Journal of Global Optimization*, vol. 87, no. 2, pp. 939–964, 2023.

[27] M. Hong, S. Zeng, J. Zhang, and H. Sun, "On the divergence of decentralized nonconvex optimization," *SIAM Journal on Optimization*, vol. 32, no. 4, pp. 2879–2908, 2022.

[28] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE transactions on automatic control*, vol. 58, no. 2, pp. 391–405, 2012.

[29] 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.

[30] H.-T. Wai, J. Lafond, A. Scaglione, and E. Moulines, "Decentralized frank–wolfe algorithm for convex and nonconvex problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5522–5537, 2017.

[31] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D2: Decentralized training over decentralized data," in *International Conference on Machine Learning*, pp. 4848–4856, PMLR, 2018.

[32] M. Hong, D. Hajinezhad, and M.-M. Zhao, "Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *International Conference on Machine Learning*, pp. 1529–1538, PMLR, 2017.

[33] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3744–3757, 2017.

[34] M. Zhu and S. Martínez, "An approximate dual subgradient algorithm for multi-agent non-convex optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 6, pp. 1534–1539, 2012.

[35] S. Liang, G. Yin, *et al.*, "Exponential convergence of distributed primal–dual convex optimization algorithm without strong convexity," *Automatica*, vol. 105, pp. 298–306, 2019.

[36] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Exponential convergence for distributed optimization under the restricted secant inequality condition," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 2672–2677, 2020.

[37] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 67, no. 8, pp. 4194–4201, 2021.

[38] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on signal processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[39] 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.

[40] S. Pan and Y. Liu, "Metric subregularity of subdifferential and kl property of exponent 1/2," *arXiv preprint arXiv:1812.00558*, 2019.
[41] R. Rockafellar and R. Wets, "Variational analysis springer-verlag," *New Y ork*, 1998.
[42] M. Ahn, J.-S. Pang, and J. Xin, "Difference-of-convex learning: directional stationarity, optimality, and sparsity," *SIAM Journal on Optimization*, vol. 27, no. 3, pp. 1637–1665, 2017.
[43] Q. Li, Y. Zhou, Y. Liang, and P. K. Varshney, "Convergence analysis of proximal gradient with momentum for nonconvex optimization," in *International Conference on Machine Learning*, pp. 2111–2119, PMLR, 2017.
[44] Y. Zhou and Y. Liang, "Characterization of gradient dominance and regularity conditions for neural networks," *arXiv preprint arXiv:1710.06910*, 2017.
[45] Z. Zhang, Y. Chen, and V. Saligrama, "Efficient training of very deep neural networks for supervised hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1487–1495, 2016.
[46] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, pp. 63–70, IEEE, 2005.
[47] W. Auzinger and J. Melenk, "Iterative solution of large linear systems," *Lecture notes, TU Wien*, 2011.
[48] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *international conference on machine learning*, pp. 3027–3036, PMLR, 2017.
[49] S. Zlobec, "Jensen's inequality for nonconvex functions," *Mathematical Communications*, vol. 9, no. 2, pp. 119–124, 2004.
[50] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, pp. 953–976, September 2018.
[51] R. Berthier, F. Bach, and P. Gaillard, "Accelerated gossip in networks of given dimension using jacobi polynomial iterations," *SIAM J. on Mathematics of Data Science*, vol. 1, pp. 24–47, 2020.
[52] J. Xu, Y. Tian, Y. Sun, and G. Scutari, "Distributed algorithms for composite optimization: Unified framework and convergence analysis," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3555–3570, 2021.
[53] S. A. Alghunaim, E. K. Ryu, K. Yuan, and A. H. Sayed, "Decentralized proximal gradient algorithms with linear convergence rates," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2787–2794, 2020.
[54] L. Guo, X. Shi, J. Cao, and Z. Wang, "Decentralized inexact proximal gradient method with network-independent stepsizes for convex composite optimization," *IEEE Transactions on Signal Processing*, vol. 71, pp. 786–801, 2023.
[55] L. Guo, X. Shi, S. Yang, and J. Cao, "Disa: A dual inexact splitting algorithm for distributed convex composite optimization," *IEEE Transactions on Automatic Control*, vol. 69, no. 5, pp. 2995–3010, 2023.
[56] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on learning theory*, pp. 1246–1257, PMLR, 2016.
[57] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *International conference on machine learning*, pp. 1724–1732, PMLR, 2017.
[58] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Mathematical programming*, vol. 176, pp. 311–337, 2019.
[59] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 22, pp. 6013–6023, 2015.
[60] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4494–4506, 2019.
[61] D. Hajinezhad and M. Hong, "Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization," *Mathematical Programming*, vol. 176, no. 1, pp. 207–245, 2019.

Xiaokai Chen received the Bachelor's degree in applied mathematics in 2018 from the Chinese University of Hong Kong, Shenzhen.

He is currently a Ph.D. candidate in the Edwardson School of Industrial Engineering at Purdue University, West Lafayette, IN, USA. His research interests include continuous (distributed, stochastic) optimization and their applications to machine learning problems.

Gesualdo Scutari (Fellow, IEEE) received the Laurea Degree (cum laude) in electrical engineering in 2002 and the Ph.D. degree in electrical engineering in 2007 both from University of Rome, "La Sapienza", Rome, Italy.

He is the Pedro and Barbara Granadillo Professor in the Edwardson School of Industrial Engineering and the Elmore Family School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. His research interests include continuous (distributed, stochastic) optimization, equilibrium programming, and their applications to signal processing and statistical learning.

He was the recipient of the 2013 NSF CAREER Award, the 2015 IEEE Signal Processing Society Young Author Best Paper Award, and the 2020 IEEE Signal Processing Society Best Paper Award. He serves as an IEEE Signal Processing Distinguished Lecturer (2023-2024). He served on the editorial board of several IEEE journals, and he is currently an Associate Editor of SIAM Journal on Optimization.

Tianyu Cao received the B.Sc. in Mathematics and Applied Mathematics from Chinese University of Hong Kong, Shenzhen in 2020.

He is currently pursuing the Ph.D. degree in the Edwardson School of Industrial Engineering, Purdue University. His research interests lie in theory and application of distributed optimization and statistical machine learning.