For publication in *2018 ACC*

On Consensus-Disagreement Tradeoff in Distributed Optimization

Zhanhong Jiang† Kushal Mukherjee‡ Soumik Sarkar†
zhjiang@iastate.edu kushmukh@in.ibm.com soumiks@iastate.edu
† Department of Mechanical Engineering, Iowa State University, Ames, IA 50011, USA
‡ IBM research, New Delhi, India

***Abstract*— Distributed optimization has been a significantly important topic in recent multi-agent networked systems research for a variety of real-life applications. Most of the previous works are focused on how to find the globally optimal solution under certain assumptions on the objective functions as well as the agent interaction characteristics. However, in many practical problems (specifically, where agents form multiple sub-groups smaller than the overall multi-agent system based on commonalities of objective functions or nature of connectivity), globally optimal solution may not be very useful and quite difficult to achieve. Achieving multiple local optimal solutions for different sub-groups may be more useful in these cases. In this context, this paper presents a new distributed optimization problem formulation by introducing a modified cost function involving a parameter that controls the tradeoff between consensus and disagreement enabling realization of the entire spectrum of globally optimal solution to multiple locally optimal solutions. A distributed generalized consensus-based gradient (DGCG) algorithm is proposed to solve such an optimization problem for strongly convex objective functions. We show the convergence analysis of the proposed algorithm and two illustrative numerical examples for validating the methodology.**

1. Introduction

Multi-agent networked systems have seen considerable attention in recent years as they play a critical role in various application areas such as power network systems, integrated buildings, transportation networks, and mobile robotics [1], [2], [3]. Numerous research works on the (global or local) decision [4], [5], [6], [7], [8], [9] have been performed for distributed optimization associated with multi-agent networked systems with a focus on finding the globally optimal solution. Locally optimal solutions were achieved for the non-convex optimization problems in a distributed manner [10], [11], [12], [13]. However, many practical problems (e.g., distributed resource allocation applications - building-to-grid control and renewable integration), which may or may not be convex, may require multiple local optimal solutions that are useful for local sub-groups of agents. For instance, consider a large multi-source, multi-destination supply-demand optimization problem, where the goal is to find optimal supply rate(s) to satisfy the needs of the demand side agents. However, the demand side agents can form multiple distinct sub-groups based on their similarities/differences in requirement and one globally optimal supply rate may not be useful in order to satisfy drastically different individual agent needs. Therefore, it may be more useful to recognize that partition in the overall system and obtain different optimal supply rates and connect different supply sources to different 'groups' or 'clusters' of demand agents.

In this paper, for addressing such an issue articulated above, we propose a new problem setup by introducing the notion of controlling the tradeoff between the level of consensus and the level of disagreement among agents in deciding the optimal solution(s). The formulation of the proposed cost function has notional similarity with the concept of *augmentability* [14], [15] where a quadratic penalty term (also referred to as the consensus term in literature and used in this paper) is added to the primary cost term (i.e., summation of local objective functions). Such a formulation achieves a weighted consensus but not global variable consensus where a center variable is defined to guide the local workers [16]. A recent publication [17] reported a nonconvex decentralized gradient descent for nonconvex optimization problems in which the authors established a Lyapunov function to obtain a modified cost function that is similar to the cost function proposed in this paper. However, they did not consider the tradeoff between the minimization of the primary cost term and the consensus among agents.

Contributions: This paper studies the tradeoff between consensus and disagreement in distributed optimization by formulating a new problem setup and proposing a distributed generalized consensus-based gradient (DGCG) algorithm to solve it. In this paper, we consider a graph describing multiple sub-groups of agents within the overall networked system and the goal is to obtain different optimal solutions for different sub-groups within the same optimization process. The graph is represented by an agent interaction matrix (or the weight matrix) designed for belief exchange among agents strongly within a sub-group and in a weak manner among the different sub-groups. We show that with constant step size, the algorithm can converge with a linear convergence rate for strongly convex objective functions while the convergence rate becomes sublinear with diminishing step size.

Outline: The organization of the rest of the paper is as follows. In section 2, an illustrative example is provided and

then the problem setup is described. While section 3 states the proposed algorithm, in section 4 we analyze the convergence properties. Section 5 presents two numerical examples for validating the proposed problem setup and algorithm. In section 6, the paper is summarized and concluded with directions for future research.

2. Problem Setup

This section constructs the proposed problem setup in an unconstrained distributed optimization setting. We use an illustrative example to motivate the problem formulation first.

A. Illustrative Example

We consider a typical heating, ventilation, and air-conditioning (HVAC) system (as shown in Figure 1) involving an air handling unit (AHU) - variable air volume (VAV) network. The working mechanism in this networked system is as follows: the AHU provides supply air (conditioned air) for each local zone where based on their zone comfort requirements supply air is reheated in the VAVs before discharging air to the zones. Typically, supply air temperature (SAT) is set low (and constant) to satisfy different zone comfort requirements resulting in large energy wastage. In [18] the authors proposed a distributed optimization algorithm to optimize the energy consumption while maintaining the zone comfort requirements such that the SAT is not constant but time-varying given internal and external loads. However, if the zone requirements are significantly different from each other, a globally optimal SAT set point solution may not be ideal. Instead, different 'groups' or 'clusters' of zones can be considered with different SAT set points. To realize that within a single optimization problem, we consider an underlying graph that strongly connects zones within a 'cluster' (e.g., based on energy requirements of the zones) while connecting different clusters through certain zones (termed here as 'gateway agents') in a much weaker fashion. The weight matrix of the graph represents the strength of the connections. Note, we consider that the graph with intended clusters is static and is provided to the optimization problem. However, in real applications, this graph can be dynamic in nature and can be discovered from data during operation.

B. Problem Setup

Motivated by the illustrative example mentioned above, the proposed problem setup of distributed optimization is presented in this section. Consider a static undirected graph $G = (V, E)$, where V is a vertex set (or node set), E is an edge set. Consequently, if we assume that there exist N nodes in the networked system, $V = \{1, 2, ..., N\}$ while $E \subseteq V \times V$. If $(i, j) \in E$, then agent i can communicate with agent j. A node $i \in V$ has its neighbors $Nb(i) \triangleq \{j \in V : (i, j) \in E\}$. We assume that the network G is connected without loss of generality throughout this paper. We consider a distributed unconstrained optimization problem on the network, modified from [19]:

$$\min J \triangleq \sum_{i=1}^{N} \theta f^i(x^i) + (1-\theta) \sum_{j \in Nb(i)} \pi_{ij} \frac{\|x^i - x^j\|^2}{2} \quad (1)$$

where $x^i \in \mathbb{R}^d$, $f^i : \mathbb{R}^d \to \mathbb{R}, i = 1, 2, \ldots, N$ are local objective functions only known to agents i, $\theta \in (0, 1]$ is the control parameter, π_{ij} is the ith row and jth column entry of the weight matrix Π, where $\Pi \in \mathbb{R}^{N \times N}$ is a row stochastic matrix.

Remark 2.1 *The problem setup in [16] is discussed below to compare the similarity of objective functions with the proposed scheme.*

$$min\ J \triangleq \sum_{i=1}^{N} f^i(x^i) + \frac{\rho \|x^i - \tilde{x}\|^2}{2} \quad (2)$$

where ρ is a penalty term that can prevent local agents from being far away from the center variable $\tilde{x}$. The proposed algorithm in Eq. 1 on the other hand tracks the difference between an agent and other agents in its neighborhood in a weighted manner. Another significant difference is that in the proposed scheme, one can control the level of consensus among agents while global consensus is the target in the problem formulation described by Eq. 2.

To compare the proposed problem setup with traditional constrained optimization problems, we introduce the following:

$$\min J \triangleq \sum_{i=1}^{N} f^i(x^i) \quad (3a)$$

$$\text{s.t. } x^i = x^j, \forall (i, j) \in E \quad (3b)$$

However, the above problem setup still aims to achieve the global optimum while being slightly different from the setup in Eq. 2 in the way that this formulation is leaderless and depends only on the communication among agents. Further, by defining a multiplier μ, Eq. 3 becomes equivalent to the following unconstrained problem:

$$\min J \triangleq \sum_{i=1}^{N} f^i(x^i) + \mu \sum_{j \in Nb(i)} \pi_{ij} \frac{\|x^i - x^j\|^2}{2} \quad (4)$$

Such a problem setup suggests that tuning the parameter μ iteratively can result in the final convergence of solution of Eq. 4 to the solution of Eq. 3. By further replacing μ with α^{-1}, the problem setup becomes same as the objective function in [17], which was proposed for solving the nonconvex optimization problems. Note, this problem still focuses on achieving single global optimization solution.

Remark 2.2 *The proposed problem setup in this paper uses the parameter $\theta \in (0, 1]$ that controls the tradeoff*

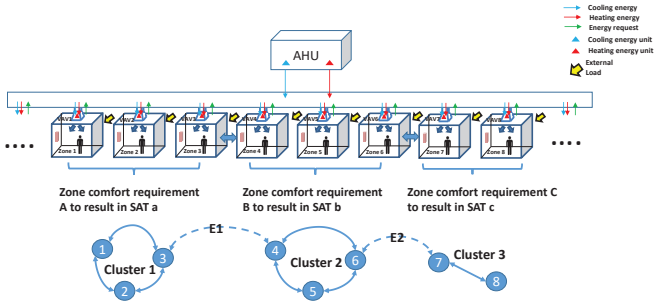


Fig. 1. *A typical AHU-VAV network in a building HVAC system: $E1$ and $E2$ signify the weak connection between different clusters*

between achieving individual objectives (i.e., disagreement) and complete consensus. The objective function is presented in a decentralized form where consensus can be achieved due to the second term and control parameter. Note, the second term in the problem setup is also convex. In the first term, f^is can be convex, possibly Lipschitz continuous or with more stronger conditions. However, in this paper, only the strongly convex case is investigated.

3. Proposed Algorithm

This section mainly presents the proposed algorithm and necessary assumptions for characterizing the main results in the next section.

Let

1) $\mathbf{x} = [x^1 x^2 \dots x^N]^T$ be the states of the agents;
2) $\mathbf{f}(\mathbf{x}) = [f^1(x^1) f^2(x^2) \dots f^N(x^N)]^T$ be the objective functions;
3) $\nabla \mathbf{f}(\mathbf{x}) = [\nabla f^1(x^1) \nabla f^2(x^2) \dots \nabla f^N(x^N)]^T$ be the gradients of the objective functions.

When $d>1$, $\mathbf{x}, \nabla \mathbf{f}(\mathbf{x})$ are all matrices. For simplicity, in this paper, we let $d = 1$. Therefore, Eq. 1 can be equivalent to the following objective function

$$\mathbf{J} = \theta \mathbf{1}^T \mathbf{f}(\mathbf{x}) + (1-\theta)\frac{\|\mathbf{x}\|^2_{I-\Pi}}{2} \tag{5}$$

where $\mathbf{1}$ is row vector with entries being 1, Π is the designed weight matrix, and $\|\cdot\|_{I-\Pi}$ denotes the norm with respect to the PSD matrix $I - \Pi$.

Before calculating the gradient of $\mathbf{J}$ there are several generic definitions and assumptions imposed for this paper.

Definition 3.1 *(γ-Lipschitz differentiable) A function f is γ-Lipschitz differentiable if it satisfies the relation for all $x, y \in \mathbb{R}^d$, $f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{\gamma}{2}\|y-x\|^2$.*

Definition 3.2 *(H-Strongly convex) A function f is H-strongly convex if it satisfies the relation for all $x, y \in \mathbb{R}^d$, $f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{H}{2}\|y-x\|^2$.*

Definition 3.3 *(Coercivity) A function f is coercive if it satisfies the following relation $f(x) \to +\infty$, $\forall \|x\| \to +\infty$.*

Assumption 3.1 *Each local cost function $f^i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}, i = 1, 2, \dots, N$, is Lipschitz differentiable with constant $\gamma_i>0$, proper, and coercive.*

Assumption 3.2 *The weight matrix $\Pi \in \mathbb{R}^{N\times N}$ satisfies the following properties: 1) If $i \neq j$ and $(i,j) \notin E$, then $\pi_{ij} = 0$; 2) the diagonal entries of Π are positive, $\pi_{ii}>0$ for all i; 3) there is a scalar $\epsilon>0$ such that $\pi_{ij} \ge \epsilon$ whenever $\pi_{ij}>0$; 4) $\mathbf{1}^T\Pi = \mathbf{1}^T, \Pi\mathbf{1} = \mathbf{1}$.*

Therefore, based on Assumption 3.2, it can be implied that Π is doubly stochastic.

Assumption 3.3 *(Gradient Upper Bound) $\nabla \mathbf{f}(\mathbf{x})$ is bounded above by some constant $C>0$, i.e., $\|\nabla \mathbf{f}(\mathbf{x})\| \le C$.*

The algorithm distributed generalized consensus-based gradient (**DGCG**) for solving Eq. 1 is proposed as follows:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \underbrace{(\theta \nabla \mathbf{f}(\mathbf{x}_k) + (1-\theta)(I-\Pi)\mathbf{x}_k)}_{\nabla \mathbf{J}(\mathbf{x}_k)} \tag{6}$$

where α_k is the step size.

We also consider the following momentum variant of the proposed algorithm to improve the convergence rate. In that case, we have

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k(\theta \nabla \mathbf{f}(\mathbf{x}_k) + (1-\theta)(I-\Pi)\mathbf{x}_k) + \eta(\mathbf{x}_k - \mathbf{x}_{k-1}) \tag{7}$$

In this study, η is selected as 0.95.

Remark 3.1 *The proposed algorithm shows that in contrast with the regular gradient-based methods [20], $\mathbf{J}(\mathbf{x}_k)$ is a linear combination of $\nabla \mathbf{f}(\mathbf{x}_k)$ and $(I-\Pi)\mathbf{x}_k$ using a parameter θ. Different θ values control the tradeoff between consensus and disagreement among local agents. When the weight matrix is uniform corresponding to a fully connected graph, small θ results in global optimum. On the other hand, a non-uniform weight matrix may lead to a "clustering phenomena", i.e., consensus among agents with high weight connection among themselves, while achieving disagreement among agents loosely or not connected.*

As f^i are assumed to be Lipschitz differentiable, then it is obtained that $\sum_{i=1}^N f^i(x^i)$ is $\gamma_m = \max_{i=1,2,\dots,N}\{\gamma_i\}$-Lipschitz differentiable while similarly being $H_m = \min_{i=1,2,\dots,N}\{H_i\}$-strongly convex when they are strongly convex. Therefore, for the new objective fucntion $\mathbf{J}$, it is $\hat{\gamma}$-Lipschitz differentiable for $\hat{\gamma} = \theta\gamma_m + (1-\theta)(1-\lambda_N)$ and $\hat{H} = \theta H_m + \frac{1}{2}(1-\theta)(1-\lambda_2)$ in strongly convex case, where λ_2 and λ_N are second and Nth largest eigenvalues, respectively. Let $\gamma_m > H_m$ such that $\hat{\gamma} > \hat{H}$.

4. Main Results

This section presents results obtained by using DGCG for the strongly convex case. The initial state is set to 0 throughout the analysis.

A proposition for consensus estimate is first stated to guarantee the convergence for global or clustering optimum scenarios with certain construction of weight matrices.

Proposition 4.1 *(Consensus Estimate) Let Assumptions 3.1, 3.2, and 3.3 hold. For all $k \in \mathbb{N}$, when $\alpha_k \leq \frac{1}{\hat{\gamma}}$ the iterates $\{\mathbf{x}_k\}$ generated by (6) satisfy the following relation*

$$\|\mathbf{x}_k - \mathbf{1}s_k\| \leq \theta C \sum_{l=1}^{k-1} \alpha_l \beta^{k-1-l} \tag{8}$$

where $s_k = \frac{1}{N}\sum_{i=1}^{N} x_k^i$, $\beta<1, B = (1-\theta)(I-\Pi)$.

Remark 4.1 *When the step size is constant, we have that for $k \to \infty, \|\mathbf{x}_k - \mathbf{1}s_k\| \leq \frac{\alpha\theta C}{1-\beta}$. Although this upper bound may not be too tight, this relationship shows that it is proportional to the step size α, the control parameter θ, and the constant C . However, when $\theta = 1$, the agents settle to states that optimize their individual objective functions. On the other hand, when $\theta \to 0$, all the agents (irrespective of the connectivity of Π, where $\lambda_2(\Pi)<1$) achieve consensus over their states. As θ is slowly reduced from 1 to 0, the agents that have connections with higher weights among themselves (i.e., form a 'group' or 'cluster'), achieve consensus within their groups. As θ is further reduced, these groups merge, forming larger groups. The grouping phenomena at given value of θ is also dependent on the similarity of the individual agents' objective functions in addition to the interagent connectivity. When the step size is diminishing, consensus can be eventually achieved with infinite number of iterations. However, θ is an important control parameter to control the consensus levels in both constant and diminishing step sizes. Appropriately chosen θ is able to speed up the convergence rate.*

Based on the definition of strong convexity, it suggests the unique optimizer and a relation between gradient and function value, namely, $2\hat{H}(\mathbf{J}(\mathbf{x}) - \mathbf{J}^*) \leq \|\nabla\mathbf{J}(\mathbf{x})\|^2$ for all $\mathbf{x} \in \mathbb{R}^N$, where $\mathbf{J}^* = \mathbf{J}(\mathbf{x}^*)$, where $\mathbf{x}^*$ is the optimizer. A key lemma is stated as follows to characterize main results.

Lemma 4.1 *Let Assumptions 3.1 and 3.2 hold. For all k, the iterates $\{\mathbf{x}_k\}$ generated by (6) satisfy*

$$\mathbf{J}(\mathbf{x}_{k+1}) - \mathbf{J}(\mathbf{x}_k) \leq -(1 - \frac{\hat{\gamma}\alpha_k}{2})\alpha_k\|\nabla\mathbf{J}(\mathbf{x}_k)\|^2 \tag{9}$$

This lemma states that when the objective function $\mathbf{J}$ is Lipschitz differentiable, under Assumptions 3.1 and 3.2, it can have sufficient descent within the function value when the step size can satisfy the certain condition, which is stated in the following.

With this relation in hand, we are ready to state the following proposition.

Proposition 4.2 *Let Assumptions 3.1 and 3.2 hold. For all k, when the step size $\alpha_k = \alpha \leq \frac{1}{\hat{\gamma}}$, the iterates $\{\mathbf{x}_k\}$ generated by (6) satisfy*

$$\mathbf{J}(\mathbf{x}_k) - \mathbf{J}^* \leq (1 - \alpha\hat{H})^{k-1}(\mathbf{J}(\mathbf{x}_1) - \mathbf{J}^*), \tag{10}$$

which shows that the function value converges to the optimal value with a linear convergence rate.

Remark 4.2 *As $\gamma_m > H_m$, then $\hat{\gamma} > \hat{H}$. Under this condition, it can be obtained that $1 - \alpha\hat{H}<1$. This proposition has a good consequence that when the objective functions are strongly convex and the constant step size satisfies a certain condition, the proposed algorithm has a linear convergence rate. It should be noted that the linear convergence rate is a standard result for the centralized gradient descent method when the objective function is Lipschitz differentiable and strongly convex with a sufficiently small step size.*

Next we discuss the strongly convex case with diminishing step size, which satisfies that $\lim_{k\to\infty} \alpha_k = 0, \sum_{k=0}^{\infty} \alpha_k = \infty$.

Proposition 4.3 *Let Assumptions 3.1 and 3.2 hold. Suppose that $\alpha_k = \frac{1}{k+w}, w>0$ such that $\alpha_1 \leq \frac{1}{\hat{\gamma}}$. Then for all k, the iterates $\{\mathbf{x}_k\}$ generated by (6) satisfy*

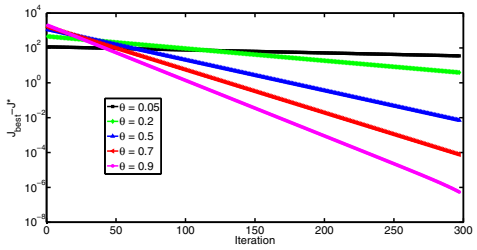
$$\mathbf{J}(\mathbf{x}_k) - \mathbf{J}^* \leq \prod_{l=1}^{k-1}(1 - \alpha_l\hat{H})(\mathbf{J}(\mathbf{x}_1) - \mathbf{J}^*), \tag{11}$$

which shows that the function value converges to the optimal value with a sublinear convergence rate when $k \to \infty$.

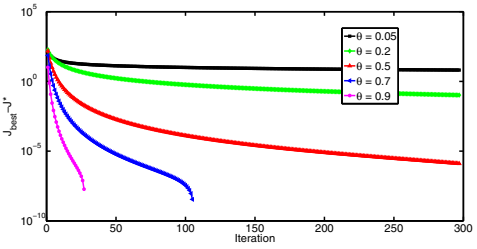
Remark 4.3 *Similarly, based on Remark 4.2, it can be obtained that $1 - \alpha_k\hat{H}<1$. Proposition 4.3 shows that when k approaches infinity, the function value converges to the optimal value and the convergence rate is sublinear.*

5. Numerical Example

For validating the proposed problem setup and algorithm, we use an agent-based building case in the context of optimizing the supply air temperature (SAT) for minimizing energy consumption in a building consisting of 10 zones (agents). In this problem, a general heating, ventilation, and air-conditioning (HVAC) system associated with a building is investigated. Interested readers can see [21] for further details. Each agent consumes cooling energy and reheat energy and for simplicity, each energy consumption function is assumed to be quadratic for strongly convex case. The energy consumption for each zone is $E_i = c(T_{MA} - T_{SA})^2 + h(T_{DA,i} - T_{SA})^2$, where c is the cooling coefficient, T_{MA} is the mixing air temperature, T_{SA} is the supply air



(a)



(b)

Fig. 2. *Fully connected graph, strongly convex case:(a) Convergence rates with different θ when $\alpha = 0.01$ (b) Convergence rates with different θ when step size is diminishing*

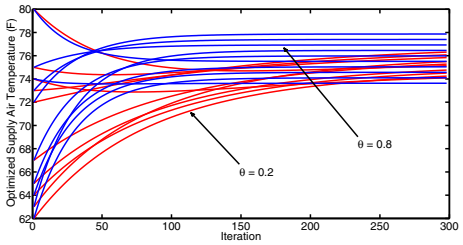


Fig. 3. *Fully connected graph, strongly convex case: optimization solutions with different θ*

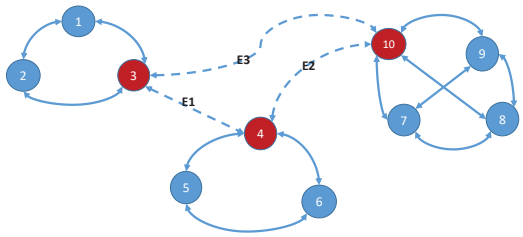
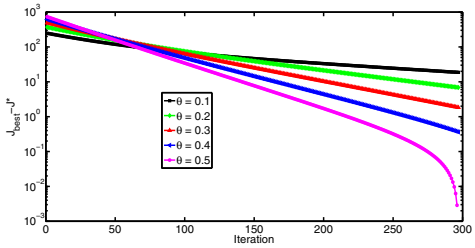
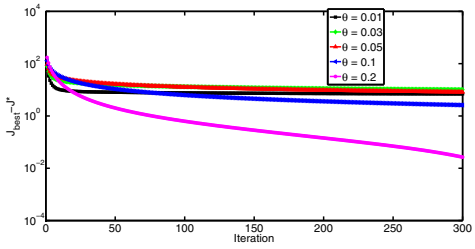


Fig. 4. *Non-fully connected graph with 3 clusters*



(a)



(b)

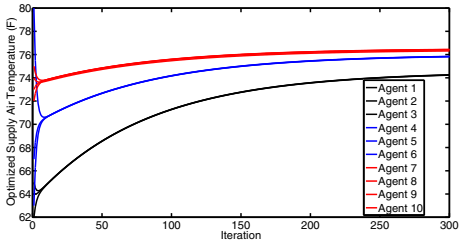
Fig. 5. *Non-fully connected graph, strongly convex case:(a) Convergence rates with different θ when $\alpha = 0.007$ (b) Convergence rates with different θ when step size is diminishing*

temperature, $T_{DA,i}$ is the ith zone discharge air temperature, h is the heating coefficient.

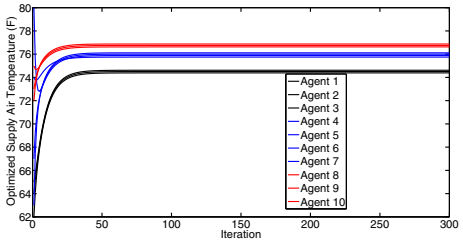
We first consider the fully connected graph case where each agent communicates with every other agent. Figure 2 shows the convergence of function value sequences for the strongly convex case. The results validate the analytical convergence analysis as the convergence rate is observed to be linear when the step size is constant while sublinear for the diminishing step size. Moreover, larger θ value results in better accuracy and faster convergence as the cost function $\mathbf{J}$ has less emphasis on consensus, which is reflected in the optimized variable convergence shown in Fig. 3.

A non-fully connected graph is used to validate the 'clustering' phenomenon of the proposed methodology. Figure 4 shows the non-fully connected graph considered here with 3 clusters and one agent of each cluster (red color) is selected to be the gateway agent to communicate with other clusters. In this scenario, we consider that zones in the same cluster have similar or the same comfort requirements. As shown in Fig. 5, observations still imply the linear convergence rate for constant step size and sublinear convergence rate for diminishing step size. With the designed state transition matrix, Fig. 6 shows the clustering phenomenon. In this case, smaller θ (that generally implies high degree of consensus) shows clear 'clustering' optimal values where individual cluster agents converge. On the other hand, for larger θ, while the convergence is faster, the 'clustering' phenomenon is slightly less pronounced. From the above simulation results, we show the effectiveness of the proposed problem formulation and algorithm.

Figure 7 shows the comparison of algorithm performance between DGCG and momentum variant of DGCG. From the result, it can be observed that DGCG performs the worst with diminishing step size but without any momentum. The best performance comes with constant step size with the momentum term. The approach with diminishing step size and momentum term outperforms the approach with only constant step size at the beginning. However, after around 600 iterations, the opposite phenomenon can be observed as the diminishing step size becomes significantly smaller than the constant step size.



(a)



(b)

Fig. 6. *Non-fully connected graph, strongly convex case:(a) optimization solutions when* $\alpha = 0.3, \theta = 0.01$ *(b) optimization solutions when* $\alpha = 0.3, \theta = 0.1$

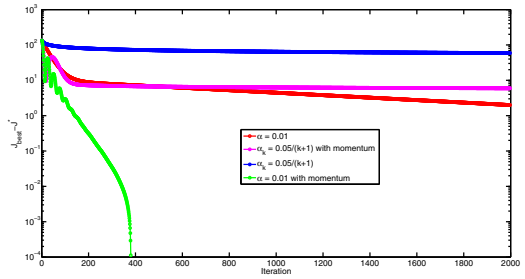


Fig. 7. *Comparison of performance between DGCG and momentum variant of DGCG*

6. Conclusions and Future Work

This paper presents a new distributed optimization problem formulation by introducing the notion of controlling the tradeoff between consensus and disagreement among agents during a distributed optimization process. Convergence analysis for the strongly convex objective function case is presented to show the spectrum from global to locally optimal solutions. The strongly convex case has linear convergence rate with constant step size and sublinear convergence rate with diminishing step size. Two numerical scenarios corresponding to a fully connected graph and a non-fully connected graph based on commercial building HVAC systems are used for validation. Beyond the results reported in this work, several future research directions include: (1) non-strongly convex case study with the current problem setup; (2) analysis of relations among step size, control parameter, and weight matrix; (3) study nonconvex and/or nonsmooth objective functions in the present context.

References

[1] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor," *Automatic Control, IEEE Transactions on*, vol. 48, 2006.

[2] S. Patterson, B. Bamieh, and A. El Abbadi, "Convergence rates of distributed average consensus with stochastic link failures," *Automatic Control, IEEE Transactions on*, vol. 55, pp. 880–892, April 2010.

[3] R. O. Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, pp. 215–233, January, 2007.

[4] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," 2003.

[5] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *Automatic Control, IEEE Transactions on*, vol. 54, pp. 48–61, January, 2009.

[6] A. Nedic and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," *Stochastic Optimization: Algorithms and Applications*, pp. 263–304, 2001.

[7] D. Yuan, S. Xu, and H. Zhao, "Distributed primal-dual subgradient method for multiagent optimization via consensus algorithm," *Systems, Man, and Cybernetics, IEEE Transactions on*, vol. 41, pp. 1715–1724, December, 2011.

[8] Q. Long, C. Wu, and X. Wang, "A system of nonsmooth equations solver based upon subgradient method," *Applied Mathematics and Computation*, pp. 284–299, 2015.

[9] Z. Jiang, S. Sarkar, and K. Mukherjee, "On distributed optimization using generalized gossip," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, pp. 2667–2672, IEEE, 2015.

[10] Y. Sun, G. Scutari, and D. Palomar, "Distributed nonconvex multiagent optimization over time-varying networks," *arXiv preprint arXiv:1607.00249*, 2016.

[11] 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, 2013.

[12] Y. Bian, B. Mirzasoleiman, J. M. Buhmann, and A. Krause, "Guaranteed non-convex optimization: Submodular maximization over continuous domains," *arXiv preprint arXiv:1606.05615*, 2016.

[13] J. Guo, G. Hug, and O. K. Tonguz, "A case for non-convex distributed optimization in large-scale power systems," *IEEE Transactions on Power Systems*, 2016.

[14] M. R. Hestenes, "Optimization theory: the finite dimensional case," *New York*, 1975.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[16] S. Zhang, A. E. Choromanska, and Y. LeCun, "Deep learning with elastic averaging sgd," in *Advances in Neural Information Processing Systems*, pp. 685–693, 2015.

[17] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *arXiv:1608.05766 [math.OC]*, 2017.

[18] Z. Jiang, V. Chinde, A. Kohl, S. Sarkar, and A. Kelkar, "Scalable supervisory control of building energy systems using generalized gossip," in *American Control Conference (ACC), 2016*, pp. 581–586, IEEE, 2016.

[19] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," *in Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2017.

[20] Y.-x. Yuan, "Step-sizes for the gradient method," *AMS IP Studies in Advanced Mathematics*, vol. 42, no. 2, p. 785, 2008.

[21] Y. Ma, G. Anderson, and F. Borrelli, "A distributed predictive control approach to building temperature regulation," *in Proceedings of 2011 American Control Conference, San Francisco, California*, June-July, 2011.