Second-Order Optimality in Non-Convex Decentralized Optimization via Perturbed Gradient Tracking

Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari
Department of Electrical and Computer Engineering
The University of Texas at Austin
{isidoros_13,constantine,mokhtari}@utexas.edu

Abstract

In this paper we study the problem of escaping from saddle points and achieving second-order optimality in a decentralized setting where a group of agents collaborate to minimize their aggregate objective function. We provide a non-asymptotic (finite-time) analysis and show that by following the idea of perturbed gradient descent, it is possible to converge to a second-order stationary point in a number of iterations which depends linearly on dimension and polynomially on the accuracy of second-order stationary point. Doing this in a communication-efficient manner requires overcoming several challenges, from identifying (first order) stationary points in a distributed manner, to adapting the perturbed gradient framework without prohibitive communication complexity. Our proposed Perturbed Decentralized Gradient Tracking (PDGT) method consists of two major stages: (i) a gradient-based step to find a first-order stationary point and (ii) a perturbed gradient descent step to escape from a first-order stationary point, if it is a saddle point with sufficient curvature. As a side benefit of our result, in the case that all saddle points are non-degenerate (strict), the proposed PDGT method finds a local minimum of the considered decentralized optimization problem in a finite number of iterations.

1 Introduction

Recently, we have witnessed an unprecedented increase in the amount of data that is gathered in a distributed fashion and stored over multiple agents (machines). Moreover, the advances in data-driven systems such as Internet of Things, health-care, and multi-agent robotics demand for developing machine learning frameworks that can be implemented in a distributed manner. Simultaneously, convex formulations for training machine learning tasks have been replaced by nonconvex representations such as neural networks. These rapid changes call for the development of a class of communication-efficient algorithms to solve *nonconvex decentralized* learning problems.

In this paper, we focus on a nonconvex decentralized optimization problem where a group of m agents collaborate to minimize their aggregate loss function, while they are allowed to exchange information only with their neighbors. To be more precise, the agents (nodes) aim to solve

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is the objective function of node i which is possibly nonconvex. Finding the global minimizer of this problem, even in the centralized setting where all the functions are available at a single machine, is hard. Given this hardness result, we often settle for finding a stationary point of Problem (1). There have been several lines of work on finding an approximate first-order

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

stationary point of this distributed problem, i.e., finding a set of local solutions $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$ where their average $\tilde{\mathbf{x}}_{avg}$ has a small gradient norm $\|\nabla f(\tilde{\mathbf{x}}_{avg})\|$ and a small consensus error $\sum_{i=1}^m \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{avg}\|$. Achieving first-order optimality, however, in nonconvex settings may not lead to a satisfactory solution as it could be a poor saddle point. Therefore, finding a second-order stationary point could improve the quality of the solution. In fact, when all saddle points are non-degenerate finding a second-order stationary point implies convergence to a local-minimum, and in several problems including matrix completion [1], phase retrieval [2], and dictionary learning [3] local minima are global minima.

While convergence to a second-order stationary point for the centralized setting has been extensively studied in the recent literature, the non-asymptotic complexity analysis of finding such a point for decentralized problems (under standard smoothness assumptions) has thus far evaded solution, in part because of significant additional challenges presented by communication limitations. A major difference between the centralized and the decentralized framework lies in the exchange of information between the nodes. Exchanging Hessian information is, of course, prohibitively expensive. Furthermore, turning to approximating schemes has the potential to create catastrophic problems for the algorithm, as small errors in approximation across the nodes could lead to inconsistent updates that could reverse progress made by prior steps. Moreover, escaping from first-order stationary points requires identifying that the algorithm has reached such a point, and accomplishing even this basic step in a communication-efficient manner presents challenges.

Contributions. In this paper we develop a novel gradient-based method for escaping from saddle points in a decentralized setting and characterize its overall communication cost for achieving a second-order stationary point. The proposed Perturbed Decentralized Gradient Tracking (PDGT) algorithm consists of two major steps: (i) A local decentralized gradient tracking scheme to find a first-order stationary point, while maintaining consensus by averaging over neighboring iterates; (ii) A perturbed gradient tracking scheme to escape from saddle points that are non-degenerate. We show that to achieve an (ϵ, γ, ρ)-second-order stationary point (see Definition 2) the proposed PDGT algorithm requires at most $\tilde{\Theta}\left(\max\left\{\frac{f(\mathbf{x}^0)-f^*}{(1-\sigma)^2\min\{\epsilon^2,\rho^2\}\gamma^3}, \frac{d}{\gamma^6}\right\}\right)$ rounds of communication, where d is dimension, $f(\mathbf{x}^0)$ is the initial objective function value, f^* is the optimal function value, and σ is the second largest eigenvalue of mixing matrix in terms of absolute norm which depends on the connectivity of the underlying graph. To the best of our knowledge, this result provides the first non-asymptotic guarantee for achieving second-order optimality in decentralized optimization under standard smoothness assumptions.

1.1 Related Work

Centralized settings. Convergence to a first-order stationary point for centralized settings has been extensively studied in the nonconvex literature [4–13]. A recent line of work focuses on improving these guarantees and achieving second-order optimality in a finite number of iterations. These schemes can be divided into three categories: (i) fully gradient-based methods which use the perturbation idea for escaping from saddle points once iterates reach a point with small gradient norm [14–16]; (ii) methods which utilize the eigenvector corresponding to the smallest eigenvalue of the Hessian to find an escape direction [5, 6, 17–21]; and (iii) trust-region [22, 23] and cubic regularization algorithms [24–26] which require solving a quadratic or cubic subproblem, respectively, at each iteration. These methods, however, cannot be applied to decentralized settings directly as they require access to the gradient or Hessian of the global objective function.

First-order optimality in decentralized settings. Recently, several iterative methods have been introduced and studied for achieving first-order optimality in decentralized settings. In particular, [27–29] show convergence to a first-order stationary point by leveraging successive convex approximation techniques and using dynamic consensus protocols. Also, a similar guarantee has been established for several well-known decentralized algorithms including distributed gradient descent [30, 31], primal-dual schemes [32–34], gradient tracking methods [35, 36], and decentralized alternating direction method of multipliers (ADMM) [37].

Second-order optimality in decentralized settings. Finding a second-order stationary point in a distributed setting has been studied by several works [38–41], but they all only provide asymptotic guarantees. The most related work to our submission is [42] which studies non-asymptotic convergence of stochastic gradient-based diffusion method for decentralized settings. However, the result of this work is obtained under two relatively less common assumptions. First, it requires a bounded

gradient disagreement condition which ensures that the local gradients ∇f_i are not far from the global gradient ∇f (Assumption 3 in [42]). Second, it assumes that the computed stochastic gradient near a saddle point is such that there is gradient noise present along some descent direction, spanned by the eigenvectors corresponding to the negative eigenvalues of the Hessian, i.e., stochastic gradient leads to an escape direction (Assumption 7 in [42]). Both these assumptions, and, in particular, the second one may not hold in general decentralized settings, and they both significantly simplify the analysis of escaping from saddle points. Unlike [42], the theoretical results presented here do not require assuming these restrictive conditions, and our paper provides the first non-asymptotic guarantee for achieving second-order optimality in decentralized settings, under standard smoothness assumptions. In fact, the conditions that we assume for proving our results are identical to the ones used in [15] for the analysis of perturbed gradient method in the centralized setting.

2 Preliminaries

The problem in (1) is defined over a set of m connected agents (nodes) where each one has access to a component of the objective function. We denote the underlying undirected connectivity graph by $\mathcal{G} = \{V, E\}$, where $V = \{1, \dots, m\}$ is the set of vertices (nodes) and E is the set of edges. As this graph is undirected, if node i can send information to node j, then the reverse communication is also possible. We call two nodes neighbors if there exists an edge between them. We further denote the neighborhood of node i by $\mathcal{N}_i$, which also includes node i itself.

Since the optimization variable $\mathbf{x}$ in (1) appears in each summand of the objective function, this problem is not decomposable into subproblems that can be solved simultaneously over nodes of the network. To make the objective function separable we introduce m local variables $\mathbf{x}_i \in \mathbb{R}^d$, and instead of minimizing $\frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$ in (1), we minimize the objective function $\frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x}_i)$. To ensure that these two problems are equivalent, we enforce the local decision variables to be equal to each other. Since the graph is connected, this condition can be replaced by consensus among neighboring nodes, and therefore the resulting problem can be written as

$$\min_{\underline{\mathbf{x}}=[\mathbf{x}_1;\mathbf{x}_2;\dots;\mathbf{x}_m]\in\mathbb{R}^{md}} F(\underline{\mathbf{x}}) := \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x}_i) \qquad s.t. \quad \mathbf{x}_i = \mathbf{x}_j, \quad \forall (i,j) \in E. \tag{2}$$

Note that in (2) we have introduced the notation $\underline{\mathbf{x}} \in \mathbb{R}^{md}$ to indicate the concatenation of all local variables $\underline{\mathbf{x}} := [\mathbf{x}_1; \mathbf{x}_2; ...; \mathbf{x}_m]$ and defined the function $F : \mathbb{R}^{md} \to \mathbb{R}$ as $F(\underline{\mathbf{x}}) := \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x}_i)$. It can be verified that $\mathbf{x}^*$ is an optimal solution of Problem (1) if and only if $\underline{\mathbf{x}}^* := [\mathbf{x}^*; \dots; \mathbf{x}^*]$ is an optimal solution of Problem (2). In the rest of the paper, therefore, we focus on solving Problem (2) as its objective function is node-separable. We should mention that solving this problem is still challenging as the constraints of this problem are coupled.

In this paper, we only assume standard smoothness conditions for the local objective functions f_i to establish our theoretical guarantees.

Assumption 1. *The local functions f_i have Lipschitz continuous gradient with constant L_1, i.e., for all $i \in \{1, \dots, m\}$ and any $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x}' \in \mathbb{R}^d$ we have $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}')\| \leq L_1 \|\mathbf{x} - \mathbf{x}'\|$.*

Assumption 2. *The local functions f_i have Lipschitz continuous Hessian with constant L_2, i.e., for all $i \in \{1, \dots, m\}$ and any $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x}' \in \mathbb{R}^d$ we have $\left\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{x}')\right\| \leq L_2 \|\mathbf{x} - \mathbf{x}'\|$.*

The gradient Lipschitz continuity condition in Assumption 1 is customary for the analysis of gradient-based methods. The condition in Assumption 2 is also required to ensure that the function is well-behaved near its saddle stationary points.

Finding an optimal solution of (1) or (2) is hard since the local functions f_i are nonconvex. Hence, we settle for finding a stationary point. In the centralized unconstrained case, a first-order stationary point of function f satisfies $\|\nabla f(\hat{\mathbf{x}})\| = 0$, and an approximate ϵ-first-order stationary point is defined as $\|\nabla f(\hat{\mathbf{x}})\| \leq \epsilon$. For the constrained decentralized problem in (2) the notion of first-order stationarity should address both stationarity and feasibility as we state in the following definition.

Definition 1. *A set of vectors $\{\hat{\mathbf{x}}_i\}_{i=1}^m$ is an (ϵ, ρ)-first-order stationary point of Problem* (2) *if*

$$\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\hat{\mathbf{x}}_i)\right\| \leq \epsilon, \qquad \frac{1}{m}\sum_{i=1}^m \left\|\hat{\mathbf{x}}_i - \frac{1}{m}\sum_{j=1}^m \hat{\mathbf{x}}_j\right\| \leq \rho. \tag{3}$$

Algorithm 1: PDGT algorithm

1: **Input:** $\mathbf{x}^0, \nabla f(\mathbf{x}^0), \epsilon, \gamma, \rho, \delta_1, \delta_2$
2: Set $\mathbf{x}_i = \mathbf{x}^0, \quad \mathbf{y}_i = \nabla f(\mathbf{x}^0), \quad T_1 = \tilde{\Theta}\left(\frac{f(\mathbf{x}^0)-f^*}{(1-\sigma)^2 \min\{\epsilon^2,\rho^2\}}\right), \quad T_2 = \tilde{\Theta}\left(\frac{d\log(1/\gamma\delta_2)}{\gamma^3}\right),$
$\eta_1 = \tilde{\Theta}\left((1-\sigma)^2\right), \quad \eta_2 = \tilde{\Theta}\left(\frac{\gamma^2}{d(1-\sigma)}\right), \quad \mathcal{R} = \tilde{\Theta}\left(\gamma^{\frac{3}{2}}\right), \quad B = \tilde{\Theta}\left(\gamma^3\right);$
3: Call ($\tilde{\underline{\mathbf{x}}}$) = PDGT Phase I ($\underline{\mathbf{x}}, \underline{\mathbf{y}}, \eta_1, T_1, \delta_1$);
4: Call ($\hat{\underline{\mathbf{x}}}, \hat{\underline{\mathbf{y}}}, S$) = PDGT Phase II ($\tilde{\underline{\mathbf{x}}}, \eta_2, T_2, \mathcal{R}, B$);
5: **if** $S = 1$ **then**
6: Return $\hat{\underline{\mathbf{x}}}$ as a second-order stationary point and stop;
7: **else**
8: Set $\underline{\mathbf{x}} = \hat{\underline{\mathbf{x}}}$, $\underline{\mathbf{y}} = \hat{\underline{\mathbf{y}}}$ and go to Step 3;
9: **end if**

The first condition in the above definition ensures that the gradient norm is sufficiently small, while the second condition ensures that the iterates are close to their average. It can be shown that if $[\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_m]$ is an (ϵ, ρ)-first-order stationary point of Problem (2), then their average $\hat{\mathbf{x}}_{avg} := \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{x}}_i$ is an $(\epsilon + L_1\rho)$-first-order stationary point of Problem (1), i.e., $\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\hat{\mathbf{x}}_{avg})\| \leq \epsilon + L_1\rho$. The proof of this claim is available in the supplementary material.

The same logic holds for second-order stationary points. In the centralized case, $\mathbf{x}$ is an (ϵ, γ)-second-order stationary point if $\|\nabla f(\hat{\mathbf{x}})\| \leq \epsilon$ and $\nabla^2 f(\hat{\mathbf{x}}) \succeq -\gamma\,\mathbf{I}$. Similarly, we define a second-order stationary point of Problem (2) with an extra condition that enforces consensus approximately.

Definition 2. *A set of vectors $\{\hat{\mathbf{x}}_i\}_{i=1}^m$ is an (ϵ, γ, ρ)-second-order stationary point of Problem* (2) *if*

$$\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\hat{\mathbf{x}}_i)\right\| \leq \epsilon, \qquad \frac{1}{m}\sum_{i=1}^m \nabla^2 f_i(\hat{\mathbf{x}}_i) \succeq -\gamma\,\mathbf{I}, \qquad \frac{1}{m}\sum_{i=1}^m \left\|\hat{\mathbf{x}}_i - \frac{1}{m}\sum_{j=1}^m \hat{\mathbf{x}}_j\right\| \leq \rho. \quad (4)$$

Note that under Assumptions 1 and 2, it can be shown that if the local solutions $[\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_m]$ form an (ϵ, γ, ρ)-second-order stationary point of Problem (2), then their average $\hat{\mathbf{x}}_{avg} := \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{x}}_i$ is an $(\epsilon + L_1\rho, \gamma + L_2\rho)$-second-order stationary point of Problem (1), i.e., $\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\hat{\mathbf{x}}_{avg})\| \leq \epsilon + L_1\rho$ and $\frac{1}{m}\sum_{i=1}^m \nabla^2 f_i(\hat{\mathbf{x}}_{avg}) \succeq -(\gamma + L_2\rho)\,\mathbf{I}$. For proof check the supplementary material.

3 Perturbed Decentralized Gradient Tracking Algorithm

We now present our proposed Perturbed Decentralized Gradient Tracking (PDGT) algorithm. The PDGT method presented in Algorithm 1 can be decomposed into two phases. Phase I of our method uses the gradient tracking ideas proposed in [35,36] to show convergence to some first-order stationary point. Using this scheme for our setup, however, requires overcoming the following hurdle: The nodes do not have access to the global gradient and thus even the task of realizing that they lie close to such a point is not trivial. Moreover, the consensus error is cumulative over the graph and tracking this quantity for each node is an additional challenge. In prior work, it has been shown that there exists an iterate that achieves first-order optimality without explicitly introducing a mechanism for identifying such an iterate. In this paper, we address this issue by utilizing an average consensus protocol as a subroutine of Phase I, which coordinates the nodes and finds with high probability and negligible communication overhead the correct index achieving first-order optimality.

Phase II of PDGT utilizes ideas from centralized perturbed gradient descent developed in [15], in order to escape saddle points. Adapting these ideas to the decentralized setting poses several challenges. A naive use of an approximation scheme could produce further issues as the noise could lead different nodes to take different escaping directions, potentially canceling each other out. Further, in order to control the consensus error and the gradient tracking disagreement we adopt a significantly smaller step size than the one used in the centralized case. Finally, using a common potential function both for Phase I and Phase II derives an interesting tradeoff between the corresponding stepsizes. Taking into account all these challenges we design PDGT to guarantee escaping from strict saddle points. In particular, we show that at the end of the second phase, either a carefully chosen potential function decreases - PDGT escapes from a saddle point - and we go back to Phase I, or an approximate

Algorithm 2: PDGT algorithm: Phase I

Input: $\underline{\mathbf{x}}, \underline{\mathbf{y}}, \eta_1, T_1, \delta_1$
Initialization: $\underline{\mathbf{x}}^0 = \underline{\mathbf{x}}, \quad \underline{\mathbf{y}}^0 = \underline{\mathbf{y}}$;
for $r = 1, \ldots, T_1$ **do**
  Compute $\mathbf{x}_i^r = \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{x}_j^{r-1} - \eta_1 \mathbf{y}_i^{r-1}$; $\forall i = 1, \ldots, m$
  Compute $\mathbf{y}_i^r = \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{y}_j^{r-1} + \nabla f_i(\mathbf{x}_i^r) - \nabla f_i(\mathbf{x}_i^{r-1})$; $\forall i = 1, \ldots, m$
  Exchange $\mathbf{x}_i^r$ and $\mathbf{y}_i^r$ with neighboring nodes; $\forall i = 1, \ldots, m$
end for
for $j = 1 : \log(\frac{1}{\delta_1})$ **do**
  Choose index $\tilde{t}_j \sim [0, T_1]$ uniformly at random and run Consensus Protocol on $\tilde{t}_j$ to find first order stationary point $\underline{\tilde{\mathbf{x}}}$ with small gradient tracking disagreement;
end for
Result: Returns first order stationary point $\underline{\tilde{\mathbf{x}}}$ with probability at least $1 - \delta_1$

second-order stationary point has been reached and the exact iterate is reported. Next, we present the details of both phases of PDGT.

Phase I. Consider $\nabla f_i(\mathbf{x}_i)$, the local gradient of node i, and define $\mathbf{y}_i \in \mathbb{R}^d$ as the variable of node i which is designed to track the global average gradient $\frac{1}{m} \sum_{i=1}^m \nabla f_i(\mathbf{x}_i)$. The algorithm proceeds to update the iterates $\mathbf{x}_i$ based on the directions of $\mathbf{y}_i$. More specifically, at each iteration r, each agent i first updates its local decision variable by averaging its local iterate with the iterates of its neighbors and descending along the negative direction of its gradient estimate $\mathbf{y}_i^{r-1}$, i.e.,

$$\mathbf{x}_i^r = \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{x}_j^{r-1} - \eta_1 \mathbf{y}_i^{r-1}, \tag{5}$$

where η_1 is the stepsize and w_{ij} is the weight that node i assigns to the information that it receives from node j. We assume that $w_{ij} > 0$ only for the nodes j that are in the neighborhood of node i, which also includes node i itself. Further, the sum of these weights is 1, i.e., $\sum_{j \in \mathcal{N}_i} w_{ij} = 1$.

Once the local $\mathbf{x}_i$'s are updated, each agent i computes its local gradient $\nabla f_i(\mathbf{x}_i^r)$ evaluated at its current iterate $\mathbf{x}_i^r$. Then, the nodes use the gradient tracking variable $\mathbf{y}_i^{r-1}$ received from their neighbors in the previous round to update their gradient tracking vector according to the update

$$\mathbf{y}_i^r = \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{y}_j^{r-1} + \nabla f_i(\mathbf{x}_i^r) - \nabla f_i(\mathbf{x}_i^{r-1}), \tag{6}$$

Note that the update in (6) shows that node i computes its new global gradient estimate by combining its previous local estimate with the ones communicated by its neighbors as well as the difference of its two consecutive local gradients. Once the local gradient tracking variables are updated, nodes communicate their local models $\mathbf{x}_i^r$ and local gradient tracking vectors $\mathbf{y}_i^r$ with their neighbors.

After running the updates in (5) and (6) for T_1 rounds, we can ensure that we have visited a set of points $[\mathbf{x}_1, \ldots, \mathbf{x}_m]$ that construct a first-order stationary point of Problem (2) (see Theorem 1); however, nodes are oblivious to the time index of those iterates. To resolve this issue all nodes sample a common time index $r \in \{1, \ldots, T_1\}$ and run an average consensus protocol among themselves to compute the expression $\left\| \frac{1}{m} \sum_{i=1}^m \nabla f_i(\tilde{\mathbf{x}}_i) \right\|^2 + \frac{1}{m} \sum_{i=1}^m \|\tilde{\mathbf{x}}_i - \frac{1}{m} \sum_{j=1}^m \tilde{\mathbf{x}}_j\|^2$ for that time index. By repeating this process at most $\log(\frac{1}{\delta_1})$ times, the output of the process leads to a set of points satisfying first-order optimality with probability at least $1 - \delta_1$. The details of this procedure are provided in the appendix. Note that the consensus procedure is standard and known to be linearly convergent. Hence, the additional cost of running the consensus protocol $\log(\frac{1}{\delta_1})$ times is negligible compared to T_1; see Theorem 1 for more details.

Phase II. In the second phase of PDGT we are given a set of variables denoted by $\underline{\tilde{\mathbf{x}}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m]$ which is a first-order stationary point. The goal is to escape from it, if it is a strict saddle, i.e., the smallest eigenvalue of the Hessian at this point is sufficiently negative. Initialized with a first-order stationary point $\underline{\tilde{\mathbf{x}}}$ the algorithm injects the same noise ξ picked uniformly from a ball of radius $\mathcal{R} = \tilde{O}(\gamma^{\frac{3}{2}})$, to all the local iterates $\tilde{\mathbf{x}}_i$. Thus for all i we have $\mathbf{x}_i^0 = \tilde{\mathbf{x}}_i + \xi$. After initialization

Algorithm 3: PDGT algorithm: Phase II

Input: $\underline{\tilde{\mathbf{x}}}, \eta_2, T_2, \mathcal{R}, B$
All nodes sample a vector $\xi \sim$ uniform ball of radius $\mathcal{R}$ using the same seed;
Set $\mathbf{x}_i^0 = \tilde{\mathbf{x}}_i + \xi$ and run Average Consensus on $\nabla f_i(\mathbf{x}_i^0)$ to set $\mathbf{y}_i^0 = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{x}_i^0)$;
for $r = 1, \ldots, T_2$ **do**
    Compute $\mathbf{x}_i^r = \sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_j^{r-1} - \eta_2 \mathbf{y}_i^{r-1}$; $\quad \forall i = 1, \ldots, m$
    Compute $\mathbf{y}_i^r = \sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{y}_j^{r-1} + \nabla f_i(\mathbf{x}_i^r) - \nabla f_i(\mathbf{x}_i^{r-1})$; $\quad \forall i = 1, \ldots, m$
    Exchange $\mathbf{x}_i^r$ and $\mathbf{y}_i^r$ with neighboring nodes; $\quad \forall i = 1, \ldots, m$
end for
Run Average Consensus Protocol for iterates $\underline{\mathbf{x}}^{T_2}$ and $\underline{\tilde{\mathbf{x}}}$;
if $H(\underline{\mathbf{x}}^{T_2}, \underline{\mathbf{y}}^{T_2}) - H(\underline{\tilde{\mathbf{x}}}, \underline{\tilde{\mathbf{y}}}) > -B$ **then**
    Return approximate second-order stationary point $\underline{\tilde{\mathbf{x}}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m]$ and set $S = 1$;
else
    Return $\underline{\mathbf{x}}^{T_2} = [\mathbf{x}_1^{T_2}, \ldots, \mathbf{x}_m^{T_2}]$, $\quad \underline{\mathbf{y}}^{T_2} = [\mathbf{y}_1^{T_2}, \ldots, \mathbf{y}_m^{T_2}]$ and set $S = 0$;
end if

all nodes follow the updates in (5) and (6) with stepsize η_2, for T_2 rounds. If the initial point was a strict saddle then at the end of this process the iterates escape from it; as a result our properly chosen potential function H (formally defined in (9) in Section 4) decreases substantially and then we revisit Phase I. If the potential function H does not decrease sufficiently, then we conclude that $\underline{\tilde{\mathbf{x}}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m]$ is a second-order stationary point of Problem (2). More precisely, choosing a proper stepsize η_2 and running PDGT for $T_2 = \tilde{O}(d\gamma^{-3})$ iterations decreases the potential function H by at least $B = \tilde{O}(\gamma^3)$, with probability $1 - \delta_2$, where T_2 has only a polylogarithmic dependence on δ_2. If the potential function is not substantially decreased then we confidently report $\underline{\tilde{\mathbf{x}}}$ as an approximate second-order stationary point. Note that S is our indicator, tracking whether we have encountered some approximate second-order stationary point or not. Further, the average consensus protocol is utilized in the second phase both to initialize the gradient tracking variables and to evaluate the potential function H at the iterates $\underline{\mathbf{x}}^{T_2}$ and $\underline{\tilde{\mathbf{x}}}$. Since the communication cost of the average consensus protocol is logarithmic in γ^{-1}, it is negligible compared to T_2. Hence, the number of communication rounds for Phase II is $\tilde{O}(d\gamma^{-3})$. Check Theorem 2 for more details.

4 Theoretical Results

In this section, we study convergence properties of our proposed PDGT method. First, we characterize the number of rounds T_1 required in Phase I of PDGT to find a set of first-order stationary points with high probability. Then, we establish an upper bound for T_2, the number of communication rounds required in the second phase. We further show that each time the algorithm finishes Phase II, a potential function decreases at least by $\tilde{\Theta}(\gamma^3)$. Finally, using these results, we characterize the overall communication rounds between nodes to find a second-order stationary point.

Before stating our result, we first discuss some conditions required for the averaging weights used in (5) and (6). Consider the mixing matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$ where the element of its i-th row and j-th column is w_{ij}. We assume $\mathbf{W}$ satisfies the following conditions.

Assumption 3. *The mixing matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$ satisfies the following:*

$$\mathbf{W} = \mathbf{W}^\top, \qquad \mathbf{W}\mathbf{1} = \mathbf{1}, \qquad \sigma := \max\{|\lambda_2(\mathbf{W})|, |\lambda_m(\mathbf{W})|\} < 1, \tag{7}$$

where $\lambda_i(\mathbf{W})$ denotes the i-th largest eigenvalue of $\mathbf{W}$.

The first condition in Assumption 3 implies that the weight node i assigns to node j equals the weight node j assigns to node i. The second condition means $\mathbf{W}$ is row stochastic, and by symmetry, column stochastic. This condition ensures that the weights that each node i assigns to its neighbors and itself sum up to 1. Further note that the eigenvalues of $\mathbf{W}$ are real and in the interval $[-1, 1]$; in fact they can be sorted in a non-increasing order as $1 = \lambda_1(\mathbf{W}) \geq \lambda_2(\mathbf{W}) \geq \cdots \geq \lambda_m(\mathbf{W}) \geq -1$. The last condition in Assumption 3 ensures that the maximum absolute value of all eigenvalues of $\mathbf{W}$

excluding $\lambda_1(\mathbf{W})$ is strictly smaller than 1. This is required since $\sigma := \max\{|\lambda_2(\mathbf{W})|, |\lambda_m(\mathbf{W})|\}$ indicates the rate of information propagation. For highly connected graphs σ is close to zero, while for less connected graphs it is close to 1. A mixing matrix W satisfying Assumption 3 can be chosen based on local degrees in a variety of ways (e.g., [36]).

Remark 1. *In the appendix we report explicit expressions. To simplify the presentation in the main body, we turn to asymptotic notation and consider sufficiently small η and α, thus hiding constants but preserving the scaling with respect to quantities that capture important elements of our analysis.*

Next, we present our first result, which formally characterizes the choice of parameters for PDGT to find an (ϵ, ρ)-first-order stationary point, as defined in (1), with probability $1 - \delta_1$.

Theorem 1. *Consider Phase I of PDGT presented in Algorithm 2. If Assumptions 1 and 3 hold, and we set $\eta_1 = \Theta((1-\sigma)\sqrt{\alpha})$ where $\alpha = \Theta((1-\sigma)^2)$, and the number of iterations satisfies $T_1 \geq T = \Theta\left(\frac{f(\mathbf{x}^0)-f^*}{\eta_1\epsilon^2}\right) = \Theta\left(\frac{f(\mathbf{x}^0)-f^*}{\sqrt{\alpha}(1-\sigma)\epsilon^2}\right)$, then w.p. at least $1-\delta_1$, the iterates $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$ corresponding to one of the randomly selected time indices $\tilde{t}_1, .., \tilde{t}_{\log(\frac{1}{\delta_1})}$ from $[0 : T_1]$, satisfy*

$$\left\| \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\tilde{\mathbf{x}}_i) \right\|^2 + \frac{1}{m}\sum_{i=1}^{m} \left\| \tilde{\mathbf{x}}_i - \frac{1}{m}\sum_{j=1}^{m} \tilde{\mathbf{x}}_j \right\|^2 \leq \epsilon^2. \tag{8}$$

Theorem 1 shows that after $\Theta\left(\frac{f(\mathbf{x}^0)-f^*}{\sqrt{\alpha}(1-\sigma)\epsilon^2} + \frac{1}{1-\sigma}\log(\frac{1}{\delta_1})\log(\frac{1}{\epsilon})\right)$ rounds of exchanging information with neighboring nodes the goal of Phase I is achieved and we obtain a set of first-order stationary points with small gradient tracking disagreement. Note that the second term $\frac{1}{1-\sigma}\log(\frac{1}{\delta_1})\log(\frac{1}{\epsilon})$ corresponds to the cost of running the average consensus protocol to choose the appropriate iterate among time steps $\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_{\log(\frac{1}{\delta_1})}$. This term is negligible compared to the first term.

Next we present our result for Phase II of PDGT. In particular, we show that if the input of Phase II, which satisfies (8), is a strict saddle meaning it has sufficient negative curvature, then PDGT will escape from it and as a result the following Lyapunov function decreases:

$$H(\underline{\mathbf{x}}, \underline{\mathbf{y}}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}_{avg}) + \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \mathbf{x}_{avg}\|^2 + \frac{\alpha}{m}\sum_{i=1}^{m}\|\mathbf{y}_i - \mathbf{y}_{avg}\|^2, \tag{9}$$

where $\underline{\mathbf{x}} := [\mathbf{x}_1; \ldots; \mathbf{x}_m]$, $\underline{\mathbf{y}} := [\mathbf{y}_1; \ldots; \mathbf{y}_m]$, $\mathbf{x}_{avg} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_j$ and $\mathbf{y}_{avg} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{y}_j$.

Theorem 2. *Consider Phase II of PDGT presented in Algorithm 3, and suppose Assumptions 1-3 hold. Further, suppose we set $\eta_2 = \tilde{\Theta}\left(\frac{\gamma^2}{d(1-\sigma)}\right)$ and $\alpha = \tilde{\Theta}\left((1-\sigma)^2\right)$, and the local perturbed iterates are computed according to $\mathbf{x}_i^0 = \tilde{\mathbf{x}}_i + \xi$, where ξ is drawn from the uniform distribution over the ball of radius $R = \tilde{\Theta}(\gamma^{1.5})$. If the input of the second phase denoted by $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$ satisfies*

$$\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}}_{avg})) \leq -\gamma, \quad \left\| \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\tilde{\mathbf{x}}_i) \right\|^2 \leq \epsilon_1^2, \quad \frac{1}{m}\sum_{i=1}^{m} \left\| \tilde{\mathbf{x}}_i - \frac{1}{m}\sum_{j=1}^{m} \tilde{\mathbf{x}}_j \right\|^2 \leq \epsilon_2^2,$$

where $\epsilon_1^2 = \tilde{\mathcal{O}}(\gamma^3)$ and $\epsilon_2^2 = \tilde{\mathcal{O}}(\frac{\gamma^5}{d})$, then after $T_2 \geq T = \tilde{\Theta}\left(\frac{d\log(1/\gamma\delta_2)}{\gamma^3}\right)$ iterations with probability at least $1-\delta_2$ we have $H(\underline{\mathbf{x}}^{T_2}, \underline{\mathbf{y}}^{T_2}) - H(\underline{\tilde{\mathbf{x}}}, \underline{\tilde{\mathbf{y}}}) = -\tilde{\Omega}(\gamma^3)$.

The result in Theorem 2 shows that if the input of Phase II of PDGT is a first-order stationary point with sufficient negative curvature, then by following the update of PDGT for $\tilde{\Theta}(\frac{d\log(1/\gamma\delta_2)}{\gamma^3})$ iterations with probability at least $1-\delta_2$ the Lyapunov function H decreases by $\tilde{\Omega}(\gamma^3)$. Further in order for the nodes to verify whether enough progress has been made we include two calls on the average consensus protocol on iterates $\underline{\tilde{\mathbf{x}}}$ and $\underline{\mathbf{x}}^{T_2}$ with overall communication complexity $\mathcal{O}(\frac{2}{1-\sigma}\log(\frac{1}{\min\{\epsilon_1,\epsilon_2\}}))$, which is negligible compared to $\tilde{\Theta}(\frac{d\log(1/\gamma\delta_2)}{\gamma^3})$ iterations.

Combining the results of Theorems 1 and 2, and using the fact that the Lyapunov function H is non-increasing in the first phase (proof is available in section 9) we obtain that if the outcome of the first phase has sufficient negative curvature (i.e, is a strict saddle), then the Lyapunov function H after Phase I and Phase II decreases at least by $\tilde{\Theta}(\gamma^3)$. Hence, after at most $\tilde{\Theta}(\gamma^{-3})$ calls to the first and second phase of PDGT, we will find a second-order stationary point of Problem (2).

Theorem 3. *Consider the PDGT method in Algorithm 1, and suppose Assumptions 1-3 hold. If we set the stepsizes as* $\eta_1 = \tilde{\Theta}\left((1-\sigma)^2\right), \eta_2 = \tilde{\Theta}\left(\frac{\gamma^2}{d(1-\sigma)}\right)$ *and the number of iterations as* $T_1 = \tilde{\Theta}\left(\frac{f(\mathbf{x}^0)-f^*}{(1-\sigma)^2\min\{\epsilon^2,\rho^2\}}\right)$ *and* $T_2 = \tilde{\Theta}\left(\frac{d}{\gamma^3}\right)$*, respectively, and we have* $\epsilon^2 = \tilde{\mathcal{O}}\left(\gamma^3\right)$ *and* $\rho^2 = \tilde{\mathcal{O}}\left(\gamma^5/d\right)$*, then after at most* $\tilde{\Theta}\left(\max\left\{\frac{f(\mathbf{x}^0)-f^*}{(1-\sigma)^2\min\{\epsilon^2,\rho^2\}\gamma^3}, \frac{d}{\gamma^6}\right\}\right)$ *communication rounds PDGT finds an* (ϵ,γ,ρ)*-second-order stationary point of Problem* (2)*, with high probability.*

A major difference between the analysis of PDGT and its centralized counterpart in [15] is that as the iterates move away from a first-order stationary point, the consensus error and the gradient tracking disagreement potentially increase exponentially fast blurring the escaping direction. Addressing this issue requires careful selection of the algorithm's parameters and setting appropriate stepsizes finetuning the tradeoff on the number of iterations between the first and the second phase. The aforementioned hurdles and the lack of knowledge regarding when the algorithm iterates lie close to a stationary point lead to an overall slower convergence rate than the one shown in the centralized case.

Recall that if the local solutions $[\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_m]$ form an (ϵ,γ,ρ)-second-order stationary point of Problem (2), then their average $\hat{\mathbf{x}}_{avg} := \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{x}}_i$ is an $(\epsilon + L_1\rho, \gamma + L_2\rho)$-second-order stationary point of Problem (1). Moreover, as discussed earlier, second order stationary points are of paramount importance because when all saddle points are strict, any second-order stationary point is a local minima. We formally state this condition in the following assumption and later show that under this assumption PDGT finds a local minima of Problem (1).

Assumption 4. *Function* $f(\cdot)$ *is* (θ,ζ,ν)*- strict saddle, when for any point* $\mathbf{x}$*, if its gradient norm is smaller than* θ*, then its Hessian satisfies the condition* $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \leq -\zeta$*, unless* $\mathbf{x}$ *is* $\nu-$*close to the set of local minima.*

The strict saddle condition defined in Assumption 4 states that if a function is (θ,ζ,ν)- strict saddle then each point in $\mathbb{R}^d$ belongs to one of these regions: 1) a region where the gradient is large and it is not close to any stationary point; 2) a region where the gradient is small but the Hessian has a significant negative eigenvalue; and 3) the region close to some local minimum. Indeed, under the extra assumption of strict saddle property on function f, PDGT is able to find a local minima in a finite number of iterations as we state in the following corollary.

Corollary 1. *Consider the PDGT method presented in Algorithm 3 and suppose the conditions in Theorem 3 are satisfied. If in addition Assumption 4 holds and the objective function* f *is* (θ,ζ,ν)*-strict saddle point, by setting* $\epsilon + L_1\rho \leq \theta$ *and* $\gamma + L_2\rho \leq \zeta$*, the PDGT will output a point* $\nu-$*close to the set of local minima after* $\tilde{\Theta}\left(\max\left\{\frac{f(\mathbf{x}^0)-f^*}{(1-\sigma)^2\min\{\epsilon^2,\rho^2\}\gamma^3}, \frac{d}{\gamma^6}\right\}\right)$ *communication rounds.*

5 Numerical Experiments

In this section, we compare PDGT with a simple version of D-GET where each node has full knowledge of its local gradient. D-GET is a decentralized gradient tracking method that "does not use the perturbation idea" [36]. Our goal is to show that PDGT escapes quickly from saddle points. We focus on a matrix factorization problem for the MovieLens dataset, where the goal is to find a rank r approximation of a matrix $\mathbf{M} \in \mathcal{M}^{l\times n}$, representing the ratings from 943 users to 1682 movies. Each user has rated at least 20 movies for a total of 9990 known ratings. This problem is given by:

$$(\mathbf{U}^*, \mathbf{V}^*) := \underset{\mathbf{U}\in\mathcal{M}^{l\times r}, \mathbf{V}\in\mathcal{M}^{n\times r}}{\operatorname{argmin}} f(\mathbf{U},\mathbf{V}) = \underset{\mathbf{U}\in\mathcal{M}^{l\times r}, \mathbf{V}\in\mathcal{M}^{n\times r}}{\operatorname{argmin}} \|\mathbf{M} - \mathbf{U}\mathbf{V}^\top\|_F^2. \tag{10}$$

We consider different values of target rank and number of nodes. Both methods are given the same randomly generated connected graph, mixing matrix, and step size. The graph is created using the $\mathrm{G}(n,p)$ model with $p = \frac{\log_2(n)}{n-1}$ enforcing the path $1-2-\ldots-(n-1)-n$ to ensure the connectivity of the graph. Further we utilize the Maximum Degree Weight mixing matrix as is presented in (10) of [36]. The stepsize for D-GET and both phases of PDGT is 3. Finally both methods are initialized at the same point which lies in a carefully chosen neighborhood of a saddle point. Note that in this problem all saddles are escapable and each local min is a global min. Regarding the parameters of PDGT we set the number of rounds during phase I and II to be 1500 and 100, respectively. Further, we set the threshold before we add noise during phase I as presented in (8) to be 10^{-6} and the radius of the noise injected to be 4.

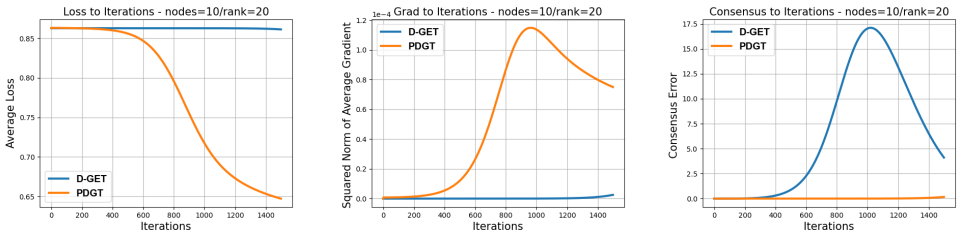

Figure 1: Average loss (left), squared norm of the average gradient (middle), consensus error (right) vs. iteration (10 nodes and target rank 20).

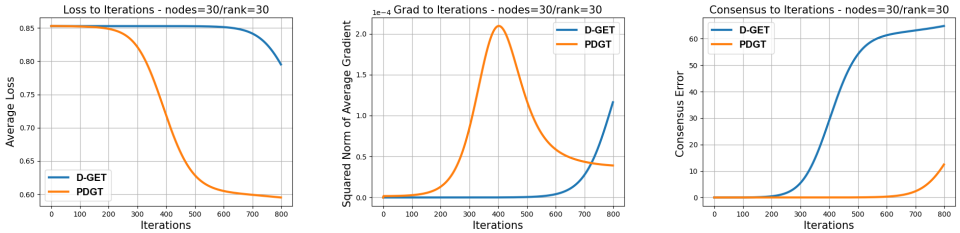

Figure 2: Average loss (left), squared norm of the average gradient (middle), consensus error (right) vs. iteration (30 nodes and target rank 30).

In Fig. 1 the experiment is run for 10 nodes, and the target rank is 20. Initially both algorithms are stuck close to a saddle point and make very little progress. However, since the theoretical criterion for PDGT is satisfied in the very first rounds (small average gradient and consensus error) we have injection of noise. This nudge is sufficient to accelerate substantially the escape of PDGT. As we see in the plot, D-GET remains close to the saddle point at least until iteration 1400 where we can see the gradient increasing somewhat faster. At the same time PDGT escapes the saddle point, decreases the loss and approaches a local minimum. In Fig. 2, the experiment is run for 30 nodes and the target rank is 30. Similarly, PDGT escapes from the saddle point much faster and decreases the loss substantially before it reaches the local minimum. We observe that D-GET also escapes the saddle point eventually following a similar trace to PDGT after spending a lot longer at the saddle. Interestingly, for this experiment, we observed that some parameters such as the stepsize of the first and the second phase, the injected noise and the threshold before we inject noise can afford to be substantially greater than the theoretical propositions casting PDGT useful for a series of practical applications.

6 Conclusion and Future Work

We proposed the Perturbed Decentralized Gradient Tracking (PDGT) algorithm that achieves second-order stationarity in a finite number of iterations, under the assumptions that the objective function gradient and Hessian are Lipschitz. We showed that PDGT finds an (ϵ, γ, ρ)-second-order stationary point, where ϵ and γ indicate the accuracy for first- and second-order optimality, respectively, and ρ shows the consensus error, after $\tilde{\Theta}\left(\max\left\{\frac{f(\mathbf{x}^0)-f^*}{(1-\sigma)^2\min\{\epsilon^2,\rho^2\}\gamma^3}, \frac{d}{\gamma^6}\right\}\right)$ communication rounds, where d is dimension, $f(\mathbf{x}^0) - f^*$ is the initial error, and $1-\sigma$ is related to graph connectivity.

This paper is the first step towards achieving second-order optimality in decentralized settings under standard smoothness assumptions, and several research problems are still unanswered in this area. First, our complexity scales linearly with dimension d, deviating from the poly-logarithmic dependence achieved for centralized perturbed gradient descent [15]. Closing this gap and developing an algorithm that obtains second-order optimality with communication rounds that scale sublinearly or even poly-logarithmically on the dimension is a promising research direction that requires further investigation. Second, in the centralized setting, it has been shown that by using gradient acceleration [16] it is possible to find a second-order stationary point faster than perturbed gradient descent. It would be interesting to see if the same conclusion also holds for decentralized settings. Last, extending the theory developed in this paper to the case that nodes only have access to a noisy estimate of their local gradients is another avenue of research that requires further study.

7 Broader Impact

Over the last couple of years we have witnessed an unprecedented increase in the amount of data collected and processed in order to tackle real life problems. Advances in numerous data-driven system such as the Internet of Things, health-care, multi-agent robotics wherein data are scattered across the agents (e.g., sensors, clouds, robots), and the sheer volume and spatial/temporal disparity of data render centralized processing and storage infeasible or inefficient. Compared to the typical parameter-server type distributed system with a fusion center, decentralized optimization has its unique advantages in preserving data privacy, enhancing network robustness, and improving the computation efficiency. Furthermore, in many emerging applications such as collaborative filtering, federated learning, distributed beamforming and dictionary learning, the data is naturally collected in a decentralized setting, and it is not possible to transfer the distributed data to a central location. Therefore, decentralized computation has sparked considerable interest in both academia and industry. At the same time convex formulations for training machine learning tasks have been replaced by nonconvex representations such as neural networks and a line of significant non convex problems are on the spotlight. Our paper contributes to this line of work and broadens the set of problems that can be successfully solved without the presence of a central coordinating authority in the aforementioned framework. The implications on the privacy of the agents are apparent while rendering the presence of an authority unnecessary has political and economical extensions. Furthermore, numerous applications are going to benefit from our result impacting society in many different ways.

8 Acknowledgments and Disclosure of Funding

The research of I. Tziotis and A. Mokhtari is supported by NSF Award CCF-2007668. C. Caramanis is supported by NSF Awards 1704778, 1646522, and 1609279.

References

[1] R. Ge, J. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *NIPS*, pp. 2973–2981, 2016.

[2] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," in *IEEE International Symposium on Information Theory, ISIT 2016*, pp. 2379–2383, 2016.

[3] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: overview and the geometric picture," *IEEE Trans. Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[4] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2013.

[5] N. Agarwal, Z. Allen Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *STOC*, pp. 1195–1199, 2017.

[6] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for non-convex optimization," *CoRR*, vol. abs/1611.00756, 2016.

[7] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, ""convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions," in *ICML*, pp. 654–663, 2017.

[8] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i," *Mathematical Programming*, pp. 1–50, 2019.

[9] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points ii: First-order methods," *arXiv preprint arXiv:1711.00841*, 2017.

[10] S. J. Reddi, S. Sra, B. Póczos, and A. J. Smola, "Fast incremental method for smooth nonconvex optimization," in *IEEE Conference on Decision and Control, CDC*, pp. 1971–1977, 2016.

[11] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. J. Smola, "Stochastic variance reduction for nonconvex optimization," in *ICML*, pp. 314–323, 2016.

[12] Z. Allen Zhu and E. Hazan, "Variance reduction for faster non-convex optimization," in *ICML*, pp. 699–707, 2016.

[13] L. Lei, C. Ju, J. Chen, and M. I. Jordan, "Non-convex finite-sum optimization via SCSG methods," in *Advances in Neural Information Processing Systems 30*, pp. 2345–2355, 2017.

[14] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points - online stochastic gradient for tensor decomposition," in *COLT*, pp. 797–842, 2015.

[15] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *ICML*, pp. 1724–1732, 2017.

[16] C. Jin, P. Netrapalli, and M. I. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," *CoRR*, vol. abs/1711.10456, 2017.

[17] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than SGD," *CoRR*, vol. abs/1708.08694, 2017.

[18] Y. Xu, R. Jin, and T. Yang, "First-order stochastic algorithms for escaping from saddle points in almost linear time," in *Advances in Neural Information Processing Systems*, pp. 5530–5540, 2018.

[19] C. W. Royer and S. J. Wright, "Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1448–1477, 2018.

[20] S. J. Reddi, M. Zaheer, S. Sra, B. Póczos, F. Bach, R. Salakhutdinov, and A. J. Smola, "A generic approach for escaping saddle points," in *AISTATS*, pp. 1233–1242, 2018.

[21] S. Paternain, A. Mokhtari, and A. Ribeiro, "A newton-based method for nonconvex optimization with fast evasion of saddle points," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 343–368, 2019.

[22] C. Cartis, N. Gould, and P. Toint, "Complexity bounds for second-order optimality in unconstrained optimization," *J. Complexity*, vol. 28, no. 1, pp. 93–108, 2012.

[23] J. M. Martínez and M. Raydan, "Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization," *J. Global Optimization*, vol. 68, no. 2, pp. 367–385, 2017.

[24] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, 2006.

[25] C. Cartis, N. Gould, and P. Toint, "Adaptive cubic regularisation methods for unconstrained optimization. part I: motivation, convergence and numerical results," *Math. Program.*, vol. 127, no. 2, pp. 245–295, 2011.

[26] C. Cartis, N. Gould, and P. Toint, "Adaptive cubic regularisation methods for unconstrained optimization. part II: worst-case function- and derivative-evaluation complexity," *Math. Program.*, vol. 130, no. 2, pp. 295–319, 2011.

[27] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–part i: Theory," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1929–1944, 2016.

[28] G. Scutari and Y. Sun, "Parallel and distributed successive convex approximation methods for big-data optimization," in *Multi-agent Optimization*, pp. 141–308, Springer, 2018.

[29] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, no. 1-2, pp. 497–544, 2019.

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

[31] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," *Advances in Neural Information Processing Systems*, 2019.

[32] D. Hajinezhad, M. Hong, T. Zhao, and Z. Wang, "NESTT: A nonconvex primal-dual splitting method for distributed and stochastic optimization," in *Advances in neural information processing systems*, pp. 3215–3223, 2016.

[33] 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 *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1529–1538, JMLR. org, 2017.

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

[35] 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.

[36] H. Sun, S. Lu, and M. Hong, "Improving the sample and communication complexity for decentralized non-convex optimization: A joint gradient estimation and tracking approach," *arXiv preprint arXiv:1910.05857*, 2019.

[37] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.

[38] M. Hong, J. D. Lee, and M. Razaviyayn, "Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization," *arXiv preprint arXiv:1802.08941*, 2018.

[39] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of gradient algorithms over networks," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 359–365, IEEE, 2018.

[40] B. Swenson, S. Kar, H. V. Poor, and J. M. Moura, "Annealing for distributed global optimization," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3018–3025, IEEE, 2019.

[41] B. Swenson, S. Kar, H. V. Poor, J. M. Moura, and A. Jaech, "Distributed gradient methods for nonconvex optimization: Local and global convergence guarantees," *arXiv preprint arXiv:2003.10309*, 2020.

[42] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments–part ii: Polynomial escape from saddle-points," *arXiv preprint arXiv:1907.01849*, 2019.

Supplementary Material

The first two lemmas introduce a connection between first order and second order stationary points in the centralized and the decentralized regime.

Lemma 1. *Assume $[\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_m]$ is an (ϵ, ρ)-first-order stationary point in the decentralized regime, that is*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_i) \right\| \leq \epsilon, \qquad \frac{1}{m} \sum_{i=1}^{m} \left\| \hat{\mathbf{x}}_i - \frac{1}{m} \sum_{j=1}^{m} \hat{\mathbf{x}}_j \right\| \leq \rho. \tag{11}$$

Then their average $\hat{\mathbf{x}}_{avg} := \frac{1}{m} \sum_{i=1}^{m} \hat{\mathbf{x}}_i$ is an $(\epsilon + L_1\rho)$-first-order stationary point in the centralized regime i.e., $\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_{avg}) \| \leq \epsilon + L_1\rho$.

Proof.

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_{avg}) \right\| \leq \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_{avg}) - \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_i) \right\| + \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_i) \right\| \tag{12}$$

$$\leq \frac{L_1}{m} \sum_{i=1}^{m} \| \hat{\mathbf{x}}_{avg} - \hat{\mathbf{x}}_i \| + \epsilon \tag{13}$$

$$\leq L_1\rho + \epsilon \tag{14}$$

where in the first inequality we add and subtract the same term and in the second one we use smoothness of f. □

Lemma 2. *Assume $[\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_m]$ is an (ϵ, γ, ρ)-second-order stationary point in the decentralized regime, that is*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_i) \right\| \leq \epsilon, \qquad \frac{1}{m} \sum_{i=1}^{m} \nabla^2 f_i(\hat{\mathbf{x}}_i) \succeq -\gamma\, \mathbf{I}, \qquad \frac{1}{m} \sum_{i=1}^{m} \left\| \hat{\mathbf{x}}_i - \frac{1}{m} \sum_{j=1}^{m} \hat{\mathbf{x}}_j \right\| \leq \rho. \tag{15}$$

Then their average $\hat{\mathbf{x}}_{avg} := \frac{1}{m} \sum_{i=1}^{m} \hat{\mathbf{x}}_i$ is an $(\epsilon + L_1\rho, \gamma + L_2\rho)$-second-order stationary point in the centralized regime i.e., i.e., $\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}_{avg}) \| \leq \epsilon + L_1\rho$ and $\frac{1}{m} \sum_{i=1}^{m} \nabla^2 f_i(\hat{\mathbf{x}}_{avg}) \succeq -(\gamma + L_2\rho)\, \mathbf{I}$.

Proof. The first part is identical to Lemma 1. for the second part we work in a similar fashion.

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \nabla^2 f_i(\hat{\mathbf{x}}_{avg}) \right\| \leq \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla^2 f_i(\hat{\mathbf{x}}_{avg}) - \frac{1}{m} \sum_{i=1}^{m} \nabla^2 f_i(\hat{\mathbf{x}}_i) \right\| + \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla^2 f_i(\hat{\mathbf{x}}_i) \right\| \tag{16}$$

$$\leq \frac{L_2}{m} \sum_{i=1}^{m} \| \hat{\mathbf{x}}_{avg} - \hat{\mathbf{x}}_i \| + \gamma \tag{17}$$

$$\leq L_2\rho + \gamma \tag{18}$$

□

where in the first inequality we add and subtract the same term and in the second one we use the Lipschitz continuous Hessian of f. The result follows.

9 Convergence to First Order Stationary Point with Consensus

Initialization of Phase I

$\underline{\mathbf{x}}^0 = \underline{\mathbf{x}}_{input}$

$\underline{\mathbf{y}}^0 = \underline{\mathbf{y}}_{input}$, with $\underline{\mathbf{y}}_{input}$ such that

$\frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i^0 = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}_i^0)$

Recall that the first time we initialize the algorithm the following also hold

$\mathbf{x}_i^0 = \mathbf{x}_j^0, \quad \forall i,j$

$\mathbf{y}_i^0 = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^0)$

Update rule of Gradient Tracking

$$\mathbf{x}_i^r = \sum_{k\in N_i}\mathbf{W}_{ik}\mathbf{x}_k^{r-1} - \eta\mathbf{y}_i^{r-1}$$
$$\mathbf{y}_i^r = \sum_{k\in N_i}\mathbf{W}_{ik}\mathbf{y}_k^{r-1} + \nabla f_i(\mathbf{x}_i^r) - \nabla f_i(\mathbf{x}_i^{r-1})$$

The Update rule of the average iterate

$\hat{\mathbf{x}}^r = \frac{1}{m}\sum_i \mathbf{x}_i^r$

$\hat{\mathbf{y}}^r = \frac{1}{m}\sum_i \mathbf{y}_i^r$

$\hat{\mathbf{x}}^r = \hat{\mathbf{x}}^{r-1} - \eta\hat{\mathbf{y}}^{r-1}$

$\hat{\mathbf{y}}^r = \hat{\mathbf{y}}^{r-1} + \frac{1}{m}\sum_i \nabla f_i(\mathbf{x}_i^r) - \frac{1}{m}\sum_i \nabla f_i(\mathbf{x}_i^{r-1})$

$\hat{\mathbf{y}}^r = \frac{1}{m}\sum_i \nabla f_i(\mathbf{x}_i^r)$

In order to see why the last equality holds notice that $\hat{\mathbf{y}}^0 = \frac{1}{m}\sum_i \nabla f_i(\mathbf{x}_i^0)$ and an induction derives the result.

Also recall that $\sigma := \max\{|\lambda_2(\mathbf{W})|, |\lambda_m(\mathbf{W})|\} < 1$.

First we are going to provide bounds on the iterates proving contraction between consecutive rounds. Consequently, we are going to derive a similar bound on function $P_\alpha(\underline{\mathbf{x}}^r)$ and combining the above we will show that the potential function $H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r)$ is decreasing between consecutive rounds.

Lemma 3 (Bound on consecutive iterates)**.** *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η then we have*

$$\|\underline{\mathbf{x}}^r - \underline{\mathbf{x}}^{r-1}\|^2 \leq 8\|\underline{\mathbf{x}}^{r-1} - \underline{\hat{\mathbf{x}}}^{r-1}\|^2 + 4\eta^2\|\underline{\mathbf{y}}^{r-1} - \underline{\hat{\mathbf{y}}}^{r-1}\|^2 + 4\eta^2\|\underline{\hat{\mathbf{y}}}^{r-1}\|^2 \tag{19}$$

Proof.

$$\begin{aligned}
\|\underline{\mathbf{x}}^r - \underline{\mathbf{x}}^{r-1}\|^2 &= \|\underline{\mathbf{W}}\underline{\mathbf{x}}^{r-1} - \eta\underline{\mathbf{y}}^{r-1} - \underline{\mathbf{x}}^{r-1}\|^2 \\
&\leq 2\|\underline{\mathbf{W}}\underline{\mathbf{x}}^{r-1} - \underline{\mathbf{x}}^{r-1}\|^2 + 2\eta^2\|\underline{\mathbf{y}}^{r-1}\|^2 \\
&\leq 2\|\underline{\mathbf{W}}\underline{\mathbf{x}}^{r-1} - \underline{\mathbf{x}}^{r-1} + \underline{\mathbf{W}}\underline{\hat{\mathbf{x}}}^{r-1} - \underline{\hat{\mathbf{x}}}^{r-1}\|^2 + 2\eta^2\|\underline{\mathbf{y}}^{r-1}\|^2 \\
&\leq 2\|(\underline{\mathbf{W}} - \mathbf{I})(\underline{\mathbf{x}}^{r-1} - \underline{\hat{\mathbf{x}}}^{r-1})\|^2 + 2\eta^2\|\underline{\mathbf{y}}^{r-1} - \underline{\hat{\mathbf{y}}}^{r-1} + \underline{\hat{\mathbf{y}}}^{r-1}\|^2 \\
&\leq 2(\|\underline{\mathbf{W}}\| + \|\mathbf{I}\|)^2\|\underline{\mathbf{x}}^{r-1} - \underline{\hat{\mathbf{x}}}^{r-1}\|^2 + 4\eta^2\|\underline{\mathbf{y}}^{r-1} - \underline{\hat{\mathbf{y}}}^{r-1}\|^2 + 4\eta^2\|\underline{\hat{\mathbf{y}}}^{r-1}\|^2 \\
&\leq 8\|\underline{\mathbf{x}}^{r-1} - \underline{\hat{\mathbf{x}}}^{r-1}\|^2 + 4\eta^2\|\underline{\mathbf{y}}^{r-1} - \underline{\hat{\mathbf{y}}}^{r-1}\|^2 + 4\eta^2\|\underline{\hat{\mathbf{y}}}^{r-1}\|^2
\end{aligned}$$

□

Lemma 4 (Iterate Contraction)**.** *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η and let $\mathbf{v}_i^r = \nabla f_i(\mathbf{x}_i^r)$; then we have*

$$\|\underline{\mathbf{x}}^{r+1} - \hat{\underline{\mathbf{x}}}^{r+1}\|^2 \leq (1+\beta_1)\sigma^2\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + (1+\frac{1}{\beta_1})\eta^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 \tag{20}$$

$$\|\underline{\mathbf{y}}^{r+1} - \hat{\underline{\mathbf{y}}}^{r+1}\|^2 \leq 8L_1^2(1+\frac{1}{\beta_2})\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + ((1+\beta_2)\sigma^2 + \eta^2 4L_1^2(1+\frac{1}{\beta_2}))\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + \eta^2 4L_1^2(1+\frac{1}{\beta_2})\|\hat{\underline{\mathbf{y}}}^r\|^2 \tag{21}$$

Proof.

$$\|\underline{\mathbf{W}}\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\| = \|\underline{\mathbf{W}}(\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r)\| \leq \sigma\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\| \tag{22}$$

To see why the inequality is true notice that $\mathbf{1}^T(\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r) = 0$, i.e. $\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r$ is orthogonal to $\mathbf{1}^T$, which is the eigenvector corresponding to $\lambda_{\max}(\mathbf{W})$. Similarly,

$$\|\underline{\mathbf{W}}\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\| \leq \sigma\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\| \tag{23}$$

$$\begin{aligned}
\|\underline{\mathbf{x}}^{r+1} - \hat{\underline{\mathbf{x}}}^{r+1}\|^2 &= \|\underline{\mathbf{W}}\underline{\mathbf{x}}^r - \eta\underline{\mathbf{y}}^r - (\hat{\underline{\mathbf{x}}}^r - \eta\hat{\underline{\mathbf{y}}}^r)\|^2 \\
&\leq (1+\beta_1)\|\underline{\mathbf{W}}\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + (1+\frac{1}{\beta_1})\eta^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 \\
&\leq (1+\beta_1)\sigma^2\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + (1+\frac{1}{\beta_1})\eta^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2
\end{aligned}$$

the last inequality comes from (22). Also

$$\begin{aligned}
&\|\underline{\mathbf{y}}^{r+1} - \hat{\underline{\mathbf{y}}}^{r+1}\|^2 \\
&= \|\underline{\mathbf{W}}\underline{\mathbf{y}}^r + \underline{\mathbf{v}}^{r+1} - \underline{\mathbf{v}}^r - (\hat{\underline{\mathbf{y}}}^r + \hat{\underline{\mathbf{v}}}^{r+1} - \hat{\underline{\mathbf{v}}}^r)\|^2 \\
&\leq (1+\beta_2)\|\underline{\mathbf{W}}\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + (1+\frac{1}{\beta_2})\|\underline{\mathbf{v}}^{r+1} - \underline{\mathbf{v}}^r - \hat{\underline{\mathbf{v}}}^{r+1} + \hat{\underline{\mathbf{v}}}^r\|^2 \\
&\leq (1+\beta_2)\sigma^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + (1+\frac{1}{\beta_2})\|(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{m})(\underline{\mathbf{v}}^{r+1} - \underline{\mathbf{v}}^r)\|^2 \\
&\leq (1+\beta_2)\sigma^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + (1+\frac{1}{\beta_2})\sum_{i=1}^m\|\mathbf{v}_i^{r+1} - \mathbf{v}_i^r\|^2 \\
&= (1+\beta_2)\sigma^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + (1+\frac{1}{\beta_2})\sum_{i=1}^m\|\nabla f_i(\mathbf{x}_i^{r+1}) - \nabla f_i(\mathbf{x}_i^r)\|^2 \\
&\leq (1+\beta_2)\sigma^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + L_1^2(1+\frac{1}{\beta_2})\sum_{i=1}^m\|\mathbf{x}_i^{r+1} - \mathbf{x}_i^r\|^2 \\
&= (1+\beta_2)\sigma^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + L_1^2(1+\frac{1}{\beta_2})\|\underline{\mathbf{x}}^{r+1} - \underline{\mathbf{x}}^r\|^2 \\
&\leq (1+\beta_2)\sigma^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + L_1^2(1+\frac{1}{\beta_2})(8\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + 4\eta^2\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + 4\eta^2\|\hat{\underline{\mathbf{y}}}^r\|^2) \\
&= 8L_1^2(1+\frac{1}{\beta_2})\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + ((1+\beta_2)\sigma^2 + \eta^2 4L_1^2(1+\frac{1}{\beta_2}))\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 + \eta^2 4L_1^2(1+\frac{1}{\beta_2})\|\hat{\underline{\mathbf{y}}}^r\|^2
\end{aligned}$$

Where the second inequality is from (23) and for the third we use the fact that $\|\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{m}\| \leq 1$. The last inequality is due to Lemma 3. □

In the following lemma an upper bound on function $P_\alpha(\underline{\mathbf{x}}^r)$ is derived which we are going to combine with the iterate contraction lemma to show that a propery constructed potential function decreases between consecutive rounds.

Lemma 5 (Intermediate function)**.** *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η and let $\alpha > 0$. Also let $P_\alpha(\underline{\mathbf{x}}^r) := \frac{1}{m}(\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 + \alpha\|\underline{\mathbf{y}}^r - \underline{\hat{\mathbf{y}}}^r\|^2)$. It follows that*

$$\begin{aligned} P_\alpha(\underline{\mathbf{x}}^{r+1}) - P_\alpha(\underline{\mathbf{x}}^r) \leq & \left((1+\beta_1)\sigma^2 - 1 + 8\alpha L_1^2(1+\frac{1}{\beta_2})\right)\frac{1}{m}\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 \\ & + \left(\alpha((1+\beta_2)\sigma^2 - 1) + \eta^2(1+\frac{1}{\beta_1}) + 4\alpha\eta^2 L_1^2(1+\frac{1}{\beta_2})\right)\frac{1}{m}\|\underline{\mathbf{y}}^r - \underline{\hat{\mathbf{y}}}^r\|^2 \\ & + 4\alpha\eta^2 L_1^2(1+\frac{1}{\beta_2})\|\hat{\mathbf{y}}^r\|^2 \end{aligned} \tag{24}$$

Proof.

$$\begin{aligned} & P_\alpha(\underline{\mathbf{x}}^{r+1}) - P_\alpha(\underline{\mathbf{x}}^r) \\ & \leq \frac{1}{m}\left[\|\underline{\mathbf{x}}^{r+1} - \underline{\hat{\mathbf{x}}}^{r+1}\|^2 + \alpha\|\underline{\mathbf{y}}^{r+1} - \underline{\hat{\mathbf{y}}}^{r+1}\|^2 - \|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 - \alpha\|\underline{\mathbf{y}}^r - \underline{\hat{\mathbf{y}}}^r\|^2\right] \\ & \leq ((1+\beta_1)\sigma^2 - 1 + \alpha 8L_1^2(1+\frac{1}{\beta_2}))\frac{1}{m}\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 \\ & \quad + (\alpha((1+\beta_2)\sigma^2 - 1) + \eta^2(1+\frac{1}{\beta_1}) + \alpha\eta^2 4L_1^2(1+\frac{1}{\beta_2}))\frac{1}{m}\|\underline{\mathbf{y}}^r - \underline{\hat{\mathbf{y}}}^r\|^2 \\ & \quad + \alpha\eta^2 4L_1^2(1+\frac{1}{\beta_2})\frac{1}{m}\|\underline{\hat{\mathbf{y}}}^r\|^2 \\ & = \left((1+\beta_1)\sigma^2 - 1 + 8\alpha L_1^2(1+\frac{1}{\beta_2})\right)\frac{1}{m}\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 \\ & \quad + \left(\alpha\left((1+\beta_2)\sigma^2 - 1\right) + \eta^2(1+\frac{1}{\beta_1}) + 4\alpha\eta^2 L_1^2(1+\frac{1}{\beta_2})\right)\frac{1}{m}\|\underline{\mathbf{y}}^r - \underline{\hat{\mathbf{y}}}^r\|^2 \\ & \quad + 4\alpha\eta^2 L_1^2(1+\frac{1}{\beta_2})\|\hat{\mathbf{y}}^r\|^2 \end{aligned}$$

where the second inequality comes from Lemma 4. □

Below we derive a bound on the function value of consecutive iterates. Notice that it is not strictly decreasing on every round and thus later we are going to focus on a suitable potential function.

Lemma 6 (Function decrease)**.** *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η; we can show the following two bounds hold.*

$$\langle\nabla f(\hat{\mathbf{x}}^r), \hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r\rangle + \frac{L_1}{2}\|\hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r\|^2 \leq \eta\frac{L_1^2}{2m}\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 - (\eta\frac{1}{2} - \eta^2\frac{L_1^2}{2})\|\hat{\mathbf{y}}^r\|^2 \tag{25}$$

$$f(\hat{\mathbf{x}}^{r+1}) - f(\hat{\mathbf{x}}^r) \leq \eta\frac{L_1^2}{2m}\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\|^2 - \eta\left(\frac{1}{2} - \eta\frac{L_1^2}{2}\right)\|\hat{\mathbf{y}}^r\|^2 \tag{26}$$

Proof. For the first one we work as follows

$$\begin{aligned}
\langle \nabla f(\hat{\mathbf{x}}^r), \hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r \rangle + \frac{L_1}{2}\|\hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r\|^2 &\leq -\eta \langle \nabla f(\hat{\mathbf{x}}^r), \hat{\mathbf{y}}^r \rangle + \eta^2 \frac{L_1}{2}\|\hat{\mathbf{y}}^r\|^2 \\
&\leq -\eta \langle \nabla f(\hat{\mathbf{x}}^r) - \hat{\mathbf{y}}^r, \hat{\mathbf{y}}^r \rangle - \eta \|\hat{\mathbf{y}}^r\|^2 + \eta^2 \frac{L_1}{2}\|\hat{\mathbf{y}}^r\|^2 \\
&\leq \frac{\eta}{2}\|\nabla f(\hat{\mathbf{x}}^r) - \hat{\mathbf{y}}^r\|^2 + \frac{\eta}{2}\|\hat{\mathbf{y}}^r\|^2 - \eta\|\hat{\mathbf{y}}^r\|^2 + \eta^2 \frac{L_1}{2}\|\hat{\mathbf{y}}^r\|^2 \\
&= \frac{\eta}{2}\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\hat{\mathbf{x}}^r) - \frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_i^r)\|^2 - (\frac{\eta}{2} - \frac{\eta^2 L_1}{2})\|\hat{\mathbf{y}}^r\|^2 \\
&\leq \frac{\eta}{2}\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\hat{\mathbf{x}}^r) - \nabla f_i(\mathbf{x}_i^r)\|^2 - (\frac{\eta}{2} - \frac{\eta^2 L_1}{2})\|\hat{\mathbf{y}}^r\|^2 \\
&\leq \eta \frac{L_1^2}{2m}\sum_{i=1}^m \|\mathbf{x}_i^r - \hat{\mathbf{x}}^r\|^2 - (\frac{\eta}{2} - \frac{\eta^2 L_1}{2})\|\hat{\mathbf{y}}^r\|^2 \\
&= \eta \frac{L_1^2}{2m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 - (\frac{\eta}{2} - \frac{\eta^2 L_1}{2})\|\hat{\mathbf{y}}^r\|^2
\end{aligned}$$

and thus for the second one we have

$$\begin{aligned}
f(\hat{\mathbf{x}}^{r+1}) &\leq f(\hat{\mathbf{x}}^r) + \langle \nabla f(\hat{\mathbf{x}}^r), \hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r \rangle + \frac{L_1}{2}\|\hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r\|^2 \\
f(\hat{\mathbf{x}}^{r+1}) - f(\hat{\mathbf{x}}^r) &\leq \langle \nabla f(\hat{\mathbf{x}}^r), \hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r \rangle + \frac{L_1}{2}\|\hat{\mathbf{x}}^{r+1} - \hat{\mathbf{x}}^r\|^2 \\
&\leq \eta \frac{L_1^2}{2m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + (\eta\frac{1}{2} - \eta^2\frac{L_1^2}{2})\|\hat{\mathbf{y}}^r\|^2
\end{aligned}$$

□

Lemma 7. *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η. Let us define the potential function $H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) := \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^r) + \frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + \frac{\alpha}{m}\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2$. Then for suitably chosen η, α the potential function is non-increasing over timesteps and specifically there exist positive constants C_1, C_2, C_3 such that*

$$H(\underline{\mathbf{x}}^{r+1}, \underline{\mathbf{y}}^{r+1}) - H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) \leq -C_1\|\hat{\mathbf{y}}^r\|^2 - C_2\frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 - C_3\frac{1}{m}\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 \tag{27}$$

Proof.

$$\begin{aligned}
&H(\underline{\mathbf{x}}^{r+1}, \underline{\mathbf{y}}^{r+1}) - H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) && (28)\\
&= \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^{r+1}) - \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^r) + \frac{1}{m}(\|\underline{\mathbf{x}}^{r+1} - \hat{\underline{\mathbf{x}}}^{r+1}\|^2 - \|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2) && (29)\\
&\quad + \frac{\alpha}{m}(\|\underline{\mathbf{y}}^{r+1} - \hat{\underline{\mathbf{y}}}^{r+1}\|^2 - \|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2) && (30)\\
&= (\frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^{r+1}) - \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^r)) + P_\alpha(\underline{\mathbf{x}}^{r+1}) - P_\alpha(\underline{\mathbf{x}}^r) && (31)\\
&\leq \left((1+\beta_1)\sigma^2 - 1 + \alpha\left(8L_1^2(1+\frac{1}{\beta_2})\right) + \eta\frac{L_1^2}{2}\right)\frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 && (32)\\
&\quad + \left(\alpha((1+\beta_2)\sigma^2 - 1) + \eta^2(1+\frac{1}{\beta_1}) + \alpha\eta^2(4L_1^2(1+\frac{1}{\beta_2}))\right)\frac{1}{m}\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 && (33)\\
&\quad + \left(-\frac{\eta}{2} + \eta^2\frac{L_1^2}{2} + \alpha\eta^2\left(4L_1^2(1+\frac{1}{\beta_2})\right)\right)\|\hat{\mathbf{y}}^r\|^2 && (34)
\end{aligned}$$

The inequality follows from the results of Lemmas 5 and 6. By choosing η and α sufficiently small we can ensure that the Lyapunov function H is decreasing. In particular, we need to ensure that

$$(1+\beta_1)\sigma^2 + \alpha\left(8L_1^2(1+\frac{1}{\beta_2})\right) + \eta\frac{L_1^2}{2} < 1 \tag{35}$$

$$\alpha(1+\beta_2)\sigma^2 + \eta^2\left(1+\frac{1}{\beta_1}\right) + 4\alpha\eta^2 L_1^2\left(1+\frac{1}{\beta_2}\right) < \alpha \tag{36}$$

$$\eta L_1^2 + 8\alpha\eta L_1^2\left(1+\frac{1}{\beta_2}\right) < 1 \tag{37}$$

To simplify these conditions we set $\beta_1 = \beta_2 = \frac{1-\sigma}{\sigma} > 0$ which leads to the following conditions:

$$\sigma + \frac{\alpha 8L_1^2}{1-\sigma} + \eta\frac{L_1^2}{2} < 1 \tag{38}$$

$$\alpha\sigma + \frac{\eta^2}{1-\sigma} + \frac{4\alpha\eta^2 L_1^2}{1-\sigma} < \alpha \tag{39}$$

$$\eta L_1^2 + \frac{8\alpha\eta L_1^2}{1-\sigma} < 1 \tag{40}$$

Indeed, if α and η are sufficiently small these conditions are satisfied. But, to obtain an explicit rate we assume that α satisfies the following inequality

$$\alpha \leq \frac{(1-\sigma)^2}{16L_1^2}, \tag{41}$$

and η as a result satisfies the following conditions

$$\eta \leq \frac{(1-\sigma)}{2L_1^2} \tag{42}$$

$$\eta^2 \leq \frac{\alpha(1-\sigma)^2}{2+8\alpha L_1^2} \tag{43}$$

$$\eta \leq \frac{1-\sigma}{2L_1^2(1-\sigma+8\alpha)} \tag{44}$$

If we assume these four conditions hold and $\beta_1 = \beta_2 = \frac{1-\sigma}{\sigma}$, then we can obtain

$$H(\underline{\mathbf{x}}^{r+1}, \underline{\mathbf{y}}^{r+1}) - H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) \leq -\frac{1-\sigma}{4m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 - \frac{\alpha(1-\sigma)}{2m}\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 - \frac{\eta}{4}\|\hat{\mathbf{y}}^r\|^2 \tag{45}$$

Hence, $C_1 = \frac{\eta}{4}$, $C_2 = \frac{1-\sigma}{4}$, and $C_3 = \alpha\frac{1-\sigma}{2}$ □

Corollary 2. . *Assume the conditions for Lemma 7 hold, then we immediately get*

$$H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^{r+1}, \underline{\mathbf{y}}^{r+1}) \geq C_1\sum_{t=0}^{r}\|\hat{\mathbf{y}}^t\|^2 + C_2\frac{1}{m}\sum_{t=0}^{r}\|\underline{\mathbf{x}}^t - \hat{\underline{\mathbf{x}}}^t\|^2 + C_3\frac{1}{m}\sum_{t=0}^{r}\|\underline{\mathbf{y}}^t - \hat{\underline{\mathbf{y}}}^t\|^2 \tag{46}$$

Theorem 4. *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η_1 such that η_1, α satisfy conditions* (41) - (44)*. Also assume $\tilde{t}$ is sampled from the uniform distribution over $[0, T-1]$. Then we can bound the expectation of the sum of the square of the global gradient estimate and the square of the consensus error as follows:*

$$\mathbb{E}_{\tilde{t}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2 + \frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}} - \hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\right] \leq \frac{4}{\min\{\eta, 1-\sigma\}}\frac{f(\mathbf{x}^0) - f^*}{T} \tag{47}$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{\tilde{t}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}}-\hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\right]\\
&\leq\frac{1}{T}\left(\sum_{t=0}^{T-1}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{t})\right\|^2+\sum_{t=0}^{T-1}\frac{1}{m}\left\|\underline{\mathbf{x}}^{t}-\hat{\underline{\mathbf{x}}}^{t}\right\|^2+\sum_{t=0}^{T-1}\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{t}-\hat{\underline{\mathbf{y}}}^{t}\right\|^2\right)\\
&\leq\frac{1}{TC_0}\left(\frac{\eta_1}{4}\sum_{t=0}^{T-1}\left\|\hat{\underline{\mathbf{y}}}^{t}\right\|^2+\frac{1-\sigma}{4}\sum_{t=0}^{T-1}\frac{1}{m}\left\|\underline{\mathbf{x}}^{t}-\hat{\underline{\mathbf{x}}}^{t}\right\|^2+\frac{1-\sigma}{2}\sum_{t=0}^{T-1}\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{t}-\hat{\underline{\mathbf{y}}}^{t}\right\|^2\right)\\
&\leq\frac{1}{C_0}\frac{H(\underline{\mathbf{x}}^0,\underline{\mathbf{y}}^0)-H(\underline{\mathbf{x}}^T,\underline{\mathbf{y}}^T)}{T}\\
&\leq\frac{1}{C_0}\frac{f(\mathbf{x}^0)-f^*}{T}\\
&=\frac{1}{\min\{\frac{\eta_1}{4},\frac{1-\sigma}{4}\}}\frac{f(\mathbf{x}^0)-f^*}{T}
\end{aligned}
$$

where $C_0:=\min\{\frac{\eta_1}{4},\frac{1-\sigma}{4}\}$. The last inequality holds because

$$H(\underline{\mathbf{x}}^0,\underline{\mathbf{y}}^0):=f(\hat{\mathbf{x}}^0)+\frac{1}{m}\left\|\underline{\mathbf{x}}^0-\hat{\underline{\mathbf{x}}}^0\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^0-\hat{\underline{\mathbf{y}}}^0\right\|^2=f(\hat{\mathbf{x}}^0) \tag{48}$$

$$H(\underline{\mathbf{x}}^r,\underline{\mathbf{y}}^r):=f(\hat{\mathbf{x}}^r)+\frac{1}{m}\left\|\underline{\mathbf{x}}^r-\hat{\underline{\mathbf{x}}}^r\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^r-\hat{\underline{\mathbf{y}}}^r\right\|^2\geq f(\hat{\mathbf{x}}^r)\geq f^* \tag{49}$$

□

Thus we immediately have:

Corollary 3. *To achieve the following ϵ-stationary solution for the separable version of our problem,*

$$\mathbb{E}_{\tilde{t}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}}-\hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\right]\leq\epsilon^2 \tag{50}$$

using the Gradient Tracking Algorithm satisfying the conditions of Theorem 4 we require $T\geq\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\epsilon^2}+1$ communication steps.

Let us call timestep $\tilde{t}$ a **good choice** if $\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}}-\hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\leq\frac{\epsilon^2}{4}$ and a **bad choice** otherwise.

Lemma 8. *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η_1 such that η_1,α satisfy conditions* (41) - (44)*. Let $T\geq 4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\epsilon^2}+1$ Also assume $\tilde{t}$ is sampled from the uniform distribution over $[0,T-1]$. Then*

$$\mathbf{P}\left(\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}}-\hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\right]\geq\frac{\epsilon^2}{4}\right)\leq\frac{1}{e} \tag{51}$$

Proof. From Corollary 3 we have

$$\mathbb{E}_{\tilde{t}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}}-\hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\right]\leq\frac{\epsilon^2}{4e} \tag{52}$$

Then we apply Markov's inequality and derive

$$\mathbf{P}\left(\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{\tilde{t}})\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{\tilde{t}}-\hat{\underline{\mathbf{x}}}^{\tilde{t}}\right\|^2\right]\geq\frac{\epsilon^2}{4}\right)\leq\frac{\frac{\epsilon^2}{4e}}{\frac{\epsilon^2}{4}}=\frac{1}{e}\tag{53}$$

□

Theorem 5. *Assume the iterates $\mathbf{x}_i$ follow the Gradient Tracking Update with stepsize η_1 such that η_1, α satisfy conditions* (41) - (44)*. Let $T \geq 4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\epsilon^2}+1$ and assume we pick i.i.d. random variables $\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_{\log(\frac{1}{\delta_1})}$ sampled from the uniform distribution over $[0, T-1]$. Then the probability that at least one of them is a good choice is at least $1-\delta_1$.*

Proof. From Lemma 8 we know that the probability of picking a bad choice it at most $\frac{1}{e}$. Thus the probability all $\log(\delta_1)$ of them are bad choices is at most $\frac{1}{e^{\log(\frac{1}{\delta_1})}}=\delta_1$. It follows that the probability that we pick at least one good choice is at least $1-\delta_1$. □

Remark 2. *In order to check each of these $\log\left(\frac{1}{\delta_1}\right)$ iterates we invoke the average consensus protocol with communication complexity at most $4\left(\frac{c\log(\frac{1}{\epsilon})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})}+1\right)$ as reported in Corollary 9, with $\mathcal{F}$ defined in* (60)*. Thus the overall number of rounds is $\log\left(\frac{1}{\delta_1}\right)\cdot 4\left(\frac{c\log(\frac{1}{\epsilon})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})}+1\right)$ which is negligible compared to the number of rounds of phase I, which is $4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\epsilon^2}+1$.*

10 Escaping a first order stationary point with negative curvature

Let us denote with $\underline{\mathbf{x}}^{-1}$ the iterate that is returned by the first phase. From here on assume that we have the following bounds:

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{-1})\right\|^2\leq\epsilon_1^2\tag{54}$$

$$\frac{1}{m}\left\|\underline{\mathbf{x}}^{-1}-\hat{\underline{\mathbf{x}}}^{-1}\right\|^2\leq\epsilon_2^2\tag{55}$$

For the first phase to return w.p. $1-\delta_1$ a point that satisfies the condition in (54) we need $4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\epsilon_1^2}+1+\log(\frac{1}{\delta_1})\cdot 4\left(\frac{c\log(\frac{1}{\epsilon})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})}+1\right)$ iterations.

For the first phase to return w.p. $1-\delta_1$ a point that satisfies the condition in (55) we need $4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\epsilon_2^2}+1+\log(\frac{1}{\delta_1})\cdot 4\left(\frac{c\log(\frac{1}{\epsilon})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})}+1\right)$iterations.

Hence

Corollary 4. *Assume the first phase runs Gradient Tracking with η_1, α such that they satisfy conditions* (41) - (44)*. For the first phase to return a point that satisfies conditions* (54) *and* (55) *with probability $1-\delta_1$ we need to run at most*

$$\begin{aligned}T_1=4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1,1-\sigma\}\min\{\epsilon_1^2,\epsilon_2^2\}}+1\\+4\log(\frac{1}{\delta_1})\cdot\left(\frac{c\log(\frac{1}{\epsilon})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})}+1\right)\end{aligned}\tag{56}$$

iterations . Thus

$$T_1=\tilde{\mathcal{O}}\left(\frac{1}{\eta_1\min\{\epsilon_1^2,\epsilon_2^2\}}\right)$$

Notice that if $\underline{\mathbf{x}}^{-1}$ satisfies conditions (54) and (55) then $\hat{\mathbf{x}}$ is either an $(\epsilon_1 + L_1\epsilon_2, \gamma)$-second order stationary point or it has sufficient negative curvature i.e. $\lambda_{\min}\left(\nabla^2 f(\hat{\mathbf{x}}^{-1})\right) \leq -\gamma$. In the former case it suffices to report the iterate; we will now focus on the more involved latter case and specifically we are going to show that after injecting noise to the iterates $\mathbf{x}_i$'s and restarting the gradient tracking algorithm the potential function decreases substantially after a small number of iterations.

We start the second phase of our algorithm by injecting the same noise ξ, uniformly from the ball of radius R, to all local iterates.

$$\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi \tag{57}$$

Then we run the average consensus protocol on $\nabla f_i(\mathbf{x}_i^0)$'s for sufficiently large number of iterations in order to get $\mathbf{y}_i^0$'s such that

$$\hat{\mathbf{y}}^0 = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{x}_i^0) \quad \textit{and} \quad \frac{1}{m}\left\|\underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0\right\|^2 \leq \frac{L_1^2}{2(1-\sigma)}\eta_2\mathcal{F}, \tag{58}$$

η_2 is the stepsize of phase II and $\mathcal{F}$ is defined below. As presented in Lemma 20, the required number of iterations is $\left\lfloor \frac{\log\left(\sqrt{\eta_2\mathcal{F}}\right) - \log\left(\left\|\underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0\right\|\right) + \log\left(\sqrt{m}L_1\right) - \log\left(\sqrt{2(1-\sigma)}\right)}{\log(\sigma)} \right\rfloor$ and is negligible compared to the number of iterations of phase I.

Consequently, the process follows the gradient tracking update for T_{cap} iterations with stepsize η_2 such that η_2, α satisfy conditions (41) - (44). We also have the following useful quantities:

$$H := \nabla^2 f(\hat{\mathbf{x}}^{-1}) \tag{59}$$

$$\mathcal{F} := \frac{|\lambda_{\min}(H)|^3}{\log^3(d\kappa/\delta_2)} \frac{(\sqrt{2}-1)^2}{(24\sqrt{2}L_2\hat{c}^2)^2} \tag{60}$$

$$\mathcal{P} := \frac{|\lambda_{\min}(H)|}{\log(d\kappa/\delta_2)} \frac{\sqrt{2}-1}{(24\sqrt{2}L_2\hat{c}^2)} \tag{61}$$

$$\mathcal{J} := \frac{\log(d\kappa/\delta_2)}{\eta_2|\lambda_{\min}(H)|} \tag{62}$$

$$T_{cap} := \hat{c}\mathcal{J} \tag{63}$$

$$\mathcal{R} := \sqrt{\frac{\mathcal{F}}{L_1}} \tag{64}$$

$$\kappa := \frac{L_1}{\gamma} \tag{65}$$

$$\delta_2 \in \left(0, \frac{d\kappa}{e}\right] \tag{66}$$

Where $\mathcal{F}$ is the target decrease of the potential function, P the bound on the norm of the iterates, T_{cap} the number of iterations in the second phase, R the radius of the ball, κ the condition number, d the dimension and δ_2 the probability of failure; $\hat{c}$ is a positive constant to be defined later.

We proceed in the following lemma to show that if the norm of the global gradient and the consensus errors are small then the norm of the gradient of the average iterate returned by phase 1 is also small.

Lemma 9. *Suppose that conditions* (54)*,* (55) *hold and* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$ *and* $\epsilon_2^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$*. Then we can show that*

$$\left\|\nabla f(\hat{\mathbf{x}}^{-1})\right\| \leq \sqrt{L_1\mathcal{F}}. \tag{67}$$

.

Proof. Adding and subtracting the same term derives:

$$\left\|\nabla f(\hat{\mathbf{x}}^{-1})\right\|^2 \leq 2\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{-1}) - \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\hat{\mathbf{x}}^{-1})\right\|^2 + 2\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{-1})\right\|^2 \tag{68}$$

$$\leq 2\frac{L_1^2}{m}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{-1} - \hat{\mathbf{x}}^{-1}\right\|^2 + 2\epsilon_1^2 \tag{69}$$

$$\leq 2L_1^2\epsilon_2^2 + 2\epsilon_1^2 \tag{70}$$

$$\leq L_1\mathcal{F} \tag{71}$$

Thus

$$\left\|\nabla f(\hat{\mathbf{x}}^{-1})\right\| \leq \sqrt{L_1\mathcal{F}} \tag{72}$$

Where the second inequality comes from (54) and the third from (55). □

Utilizing the previous result we are going to show that by adding perturbation in the worst case we increase the function value at most by $\frac{3}{2}\mathcal{F}$.

Lemma 10. *Suppose that conditions* (54)*,* (55) *hold and let* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$ *and* $\epsilon_2^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$ *and for all* i *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$*. Then*

$$f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) \leq \frac{3}{2}\mathcal{F} \tag{73}$$

Proof. First notice that by Lemma 9 we have $\left\|\nabla f(\hat{\mathbf{x}}^{-1})\right\| \leq \sqrt{L_1\mathcal{F}}$ and thus utilizing smoothness we obtain the bound

$$f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) \leq \langle\nabla f(\hat{\mathbf{x}}^{-1}), \xi\rangle + \frac{L_1}{2}\left\|\xi\right\|^2 \leq \sqrt{L_1\mathcal{F}}\frac{\sqrt{\mathcal{F}}}{\sqrt{L_1}} + \frac{\mathcal{F}}{2} \leq \frac{3}{2}\mathcal{F} \tag{74}$$

□

Below we show that the potential function increases at most by $\frac{7}{4}\mathcal{F}$ after the injection of noise.

Lemma 11. *Suppose that conditions* (54)*,* (55) *and* (58) *hold and* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}, \quad \epsilon_2^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$*. Further let* $\alpha\eta_2 \leq \frac{(1-\sigma)}{2L_1^2}$ *and for all* i *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$*. Then we have the following*

$$H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq \frac{7}{4}\mathcal{F} \tag{75}$$

Proof. Recall that

$$H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) := \frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}^r) + \frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + \frac{\alpha}{m}\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 \tag{76}$$

By the definition of the potential function we get

$$\begin{aligned}
&H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \\
&= f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) + \frac{1}{m}\|\underline{\mathbf{x}}^0 - \hat{\underline{\mathbf{x}}}^0\|^2 - \frac{1}{m}\|\underline{\mathbf{x}}^{-1} - \hat{\underline{\mathbf{x}}}^{-1}\|^2 + \frac{\alpha}{m}\|\underline{\mathbf{y}}^0 - \hat{\underline{\mathbf{y}}}^0\|^2 - \frac{\alpha}{m}\|\underline{\mathbf{y}}^{-1} - \hat{\underline{\mathbf{y}}}^{-1}\|^2 \\
&\leq f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) + \frac{\alpha}{m}\|\underline{\mathbf{y}}^0 - \hat{\underline{\mathbf{y}}}^0\|^2 \\
&\leq \frac{3}{2}\mathcal{F} + \alpha\eta_2\frac{L_1^2}{2(1-\sigma)}\mathcal{F} && (77) \\
&\leq \frac{3}{2}\mathcal{F} + \frac{1}{4}\mathcal{F} && (78) \\
&\leq \frac{7}{4}\mathcal{F}, && (79)
\end{aligned}$$

where the first inequality comes from the fact that the same noise is injected to all local iterates and thus

$$\frac{1}{m}\|\underline{\mathbf{x}}^0 - \hat{\underline{\mathbf{x}}}^0\|^2 - \frac{1}{m}\|\underline{\mathbf{x}}^{-1} - \hat{\underline{\mathbf{x}}}^{-1}\|^2 = 0 \tag{80}$$

for the second inequality we use that $f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) \leq \frac{3}{2}\mathcal{F}$ due to Lemma 10 and the bound from (58). □

Having established the fact that the potential function is not increasing more than $\frac{7}{4}\mathcal{F}$ after the injection of noise, we can proceed to show that after we perturbed the iterates and apply Gradient Tracking Update for T_{cap} iterations, the potential function will decrease substantially. Specifically, we will show that $H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq -\mathcal{F}$. Towards proving this statement we consider two complementary cases.

In the first, more simple case we assume that at least one of the following sums is sufficiently large

$$\sum_{t=0}^{T_{cap}-1} \|\hat{\mathbf{y}}^t\|^2 \geq \frac{12\mathcal{F}}{\eta_2}, \qquad \sum_{t=0}^{T_{cap}-1} \|\underline{\mathbf{x}}^t - \hat{\underline{\mathbf{x}}}^t\|^2 \geq \frac{12m}{(1-\sigma)}\mathcal{F} \tag{81}$$

and proceed to prove the potential function decrease.

Lemma 12. *Suppose that conditions* (54)*,* (55) *and* (58) *hold and* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}, \quad \epsilon_2^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$. *Further let* $\alpha\eta_2 \leq \frac{(1-\sigma)}{4L_1^2}$ *and for all* i *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$. *Assume the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44)*. Finally, let at least one of the following sums be large enough*

$$\sum_{t=0}^{T_{cap}-1} \|\hat{\mathbf{y}}^t\|^2 \geq \frac{12\mathcal{F}}{\eta_2}, \qquad \sum_{t=0}^{T_{cap}-1} \|\underline{\mathbf{x}}^t - \hat{\underline{\mathbf{x}}}^t\|^2 \geq \frac{12m}{1-\sigma}\mathcal{F} \tag{82}$$

Then we can show that

$$H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq -\mathcal{F}$$

Proof. From (46) we get

$$\begin{aligned} H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) &\leq -\frac{\eta_2}{4}\sum_{t=0}^{T_{cap}-1} \|\hat{\mathbf{y}}^t\|^2 - \frac{1-\sigma}{4m}\sum_{t=0}^{T_{cap}-1} \|\underline{\mathbf{x}}^t - \hat{\underline{\mathbf{x}}}^t\|^2 \\ &\quad - \alpha\frac{1-\sigma}{2m}\sum_{t=0}^{T_{cap}-1} \|\underline{\mathbf{y}}^t - \hat{\underline{\mathbf{y}}}\|^2 \qquad (83) \\ &\leq -\frac{\eta_2}{4}\sum_{t=0}^{T_{cap}-1} \|\hat{\mathbf{y}}^t\|^2 - \frac{1-\sigma}{4m}\sum_{t=0}^{T_{cap}-1} \|\underline{\mathbf{x}}^t - \hat{\underline{\mathbf{x}}}^t\|^2 \\ &\leq -3\mathcal{F} \qquad (84) \end{aligned}$$

Thus immediately we get

$$H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) = H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) + H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq -3\mathcal{F} + \frac{7}{4}\mathcal{F} \leq -\mathcal{F} \tag{85}$$

Where in the first inequality we use Lemma 11. □

We are left to deal with the complementary case where both sums are sufficiently small:

$$\sum_{t=0}^{T_{cap}-1} \|\hat{\mathbf{y}}^t\|^2 < \frac{12\mathcal{F}}{\eta_2} \quad \textit{and} \quad \sum_{t=0}^{T_{cap}-1} \|\underline{\mathbf{x}}^t - \hat{\underline{\mathbf{x}}}^t\|^2 < \frac{12m}{1-\sigma}\mathcal{F} \tag{86}$$

The high level idea of the next lemma is the following. For the first T_{cap} iteration either the $\hat{\mathbf{x}}^t$ iterates are going to decrease the function value by at least $3\mathcal{F}$ or the iterates are going to remain in a ball of radius $2\hat{c}\mathcal{P}$ around $\hat{\mathbf{x}}^{-1}$.

Lemma 13. *Assume that* (86) *holds and that we are given* $\underline{\mathbf{x}}^{-1}, \underline{\mathbf{x}}^{-0}$ *such that* $\|\hat{\mathbf{x}}^{-1} - \hat{\mathbf{x}}^0\| \leq 2\mathcal{R}$. *Assume that* $\underline{\mathbf{x}}^t$ *follows the Gradient Tracking Update with stepsize* $\eta_2 \leq \min\{1, \frac{1}{L_1}\}$ *and let* $\hat{c} \geq 36$. *Further, consider the definition of the stopping time* $T_{\hat{\mathbf{x}}}$

$$T_{\hat{\mathbf{x}}} = \min\{\inf_t\{t | f(\hat{\mathbf{x}}^t) - f(\hat{\mathbf{x}}^0) \leq -3\mathcal{F}\}, \hat{c}\mathcal{J}\}$$

Then for all time indices $t < T_{\hat{\mathbf{x}}}$ *we have* $\|\hat{\mathbf{x}}^t - \hat{\mathbf{x}}^0\| \leq \hat{c}\mathcal{P}$ *and as a result* $\|\hat{\mathbf{x}}^t - \hat{\mathbf{x}}^{-1}\| \leq 2\hat{c}\mathcal{P}$.

Proof. For all steps $t \geq 0$ we have

$$f(\hat{\mathbf{x}}^{t+1}) - f(\hat{\mathbf{x}}^t) \tag{87}$$

$$\leq \nabla f(\hat{\mathbf{x}}^t)^T(\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t) + \frac{L_1}{2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{88}$$

$$\leq (\nabla f(\hat{\mathbf{x}}^t) - \hat{\mathbf{y}}^t)^T(\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t) + (\hat{\mathbf{y}}^t)^T(\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t) + \frac{L_1}{2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{89}$$

$$\leq (\frac{1}{m}\sum_{i=1}^m \nabla f_i(\hat{\mathbf{x}}^t) - \frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_i^t))^T(\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t) - \frac{1}{\eta_2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 + \frac{L_1}{2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{90}$$

$$\leq \frac{\eta_2 L_1^2}{m}\|\underline{\mathbf{x}}^t - \underline{\hat{\mathbf{x}}}^t\|^2 + \frac{1}{4\eta_2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 - \frac{1}{2\eta_2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{91}$$

$$\leq \frac{\eta_2 L_1^2}{m}\|\underline{\mathbf{x}}^t - \underline{\hat{\mathbf{x}}}^t\|^2 - \frac{1}{4\eta_2}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{92}$$

In the second inequality we add and subtrack the same term and in the third inequality we utilize the update rule of gradient tracking. In the forth we utilize smoothness and the fact that $\eta_2 \leq \min\{1, \frac{1}{L_1}\}$. Summing up to any $k < T_{\hat{\mathbf{x}}}$ we obtain

$$f(\hat{\mathbf{x}}^k) - f(\hat{\mathbf{x}}^0) \leq \frac{\eta_2 L_1^2}{m}\sum_{t=0}^{k-1}\|\underline{\mathbf{x}}^t - \underline{\hat{\mathbf{x}}}^t\|^2 - \frac{1}{4\eta_2}\sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{93}$$

$$\leq \frac{\eta_2 L_1^2}{m}\frac{12m}{1-\sigma}\mathcal{F} - \frac{1}{4\eta_2}\sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{94}$$

$$\tag{95}$$

the last inequality holds because $T_{\hat{\mathbf{x}}} \leq T_{cap}$ and also due to the second condition in (86).

Utilizing the fact that $f(\hat{\mathbf{x}}^k) - f(\hat{\mathbf{x}}^0) > -3\mathcal{F}$ we derive

$$12\eta_2\mathcal{F} + \frac{48L_1^2}{1-\sigma}\eta_2^2\mathcal{F} \geq \sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{96}$$

$$12\eta_2\mathcal{F}\left(1 + \frac{4L_1^2\eta_2}{1-\sigma}\right) \geq \sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t\|^2 \tag{97}$$

$$\tag{98}$$

By using the Cauchy-Schwartz inequality we can show that $\sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1}-\hat{\mathbf{x}}^t\|$ is bounded above by

$$\sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1}-\hat{\mathbf{x}}^t\| \leq \sqrt{\hat{c}\mathcal{J}\sum_{t=0}^{k-1}\|\hat{\mathbf{x}}^{t+1}-\hat{\mathbf{x}}^t\|^2} \tag{99}$$

$$\leq \sqrt{\hat{c}\frac{\log(d\kappa/\delta_2)}{\eta_2|\lambda_{\min}(H)|}12\eta_2\mathcal{F}\left(1+\frac{4\eta_2 L_1^2}{1-\sigma}\right)} \tag{100}$$

$$= \sqrt{12\hat{c}\left(1+\frac{4\eta_2 L_1^2}{1-\sigma}\right)}\sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \tag{101}$$

$$= \sqrt{12\hat{c}\left(1+\frac{4\eta_2 L_1^2}{1-\sigma}\right)}\mathcal{P} \tag{102}$$

$$\leq \sqrt{36\hat{c}}\mathcal{P} \tag{103}$$

$$\leq \hat{c}\mathcal{P} \tag{104}$$

where the second to last inequality follows from condition (42) and the last from the fact that $\hat{c} \geq 36$.

Next we bound the different between $\hat{\mathbf{x}}^t$ and $\hat{\mathbf{x}}^0$ for all $t \leq k$. In this case we have

$$\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^0\right\| \leq \sum_{i=1}^{t}\left\|\hat{\mathbf{x}}^i-\hat{\mathbf{x}}^{i-1}\right\| \leq \hat{c}\mathcal{P} \tag{105}$$

Hence, so far we have shown that for all $t \leq k$ we have $\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^0\right\| \leq \hat{c}\mathcal{P}$.

Next, we proceed to characterize their distance to $\hat{\mathbf{x}}^{-1}$. To do so, first note that $\forall t \leq k-1$ we can derive the following upper bound

$$\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{-1}\right\| \leq \sum_{i=1}^{t}\left\|\hat{\mathbf{x}}^i-\hat{\mathbf{x}}^{i-1}\right\| + \left\|\hat{\mathbf{x}}^0-\hat{\mathbf{x}}^{-1}\right\| \leq \hat{c}\mathcal{P}+2\mathcal{R} \leq \hat{c}\mathcal{P}+2\mathcal{P} \leq \frac{3}{2}\hat{c}\mathcal{P} \tag{106}$$

where the third inequality holds given that $L_1 \geq |\lambda_{\min}(H)| \geq \frac{|\lambda_{\min}(H)|}{\log(d\kappa/\delta_2)}$. Notice that since $\delta_2 \in \left(0, \frac{d\kappa}{e}\right]$ we have $\log(d\kappa/\delta_2) \geq 1$. The above implies

$$\mathcal{P} := \sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \geq \sqrt{\frac{\mathcal{F}}{L_1}} = \mathcal{R} \tag{107}$$

and the last inequality of (106) holds since $\hat{c} \geq 4$.
Hence, for all $t < T_{\hat{\mathbf{x}}}$ we have $\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{-1}\right\| \leq 2\hat{c}\mathcal{P}$. □

The next lemma is going to be used in Lemma 15. Here we show that since the consensus error was small before the injection of noise it remains relative small with respect to the sequence $\underline{\mathbf{x}}^t$ and $\underline{\mathbf{w}}^t$ for the first T_{cap} iterations.

Lemma 14. *Assume the major condition* (86) *and conditions* (55) *and* (58) *hold. Let* $\epsilon_2^2 \leq \frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}$ *and further* $\forall i$ *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$. *Consider that the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44). *Define the sequence of* $\mathbf{w}_i$*'s similarly to* $\mathbf{x}_i$*'s except* $\forall i \quad \mathbf{w}_i^0 = \mathbf{x}_i^0 + \mu\mathcal{R}\mathbf{e}_1$ *with* $\mathbf{e}_1$ *a unit eigenvector corresponding to the minimum eigenvalue of* $\nabla^2 f(\hat{\mathbf{x}}^{-1})$ *and* $\mu \in [\delta_2/2\sqrt{d}, 1]$. *Finally, consider a sufficiently large positive constant* $c_{new} \geq 14L_1\sqrt{L_1}$.

Then for any $t \leq T_{cap}$ *it holds that*

$$\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\left\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\right\| + \left\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\right\|) \leq c_{new}\sqrt{\frac{\alpha\eta_2}{1-\sigma}}\eta_2\mathcal{R}$$

Proof. We will derive a bound with respect to $\mathbf{x}^t$ since the proof for $\mathbf{w}^t$ is identical.
From the proof of Lemma 5 we know that

$$\frac{1}{m}\left\|\underline{\mathbf{x}}^t-\underline{\hat{\mathbf{x}}}^t\right\|^2-\frac{1}{m}\left\|\underline{\mathbf{x}}^{t-1}-\underline{\hat{\mathbf{x}}}^{t-1}\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^t-\underline{\hat{\mathbf{y}}}^t\right\|^2-\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{t-1}-\underline{\hat{\mathbf{y}}}^{t-1}\right\|^2 \tag{108}$$

$$\leq\left((1+\beta_1)\sigma^2-1+8\alpha L_1^2(1+\frac{1}{\beta_2})\right)\frac{1}{m}\|\underline{\mathbf{x}}^{t-1}-\underline{\hat{\mathbf{x}}}^{t-1}\|^2$$

$$+\left(\alpha\left((1+\beta_2)\sigma^2-1\right)+\eta_2^2(1+\frac{1}{\beta_1})+4\alpha\eta_2^2L_1^2(1+\frac{1}{\beta_2})\right)\frac{1}{m}\|\underline{\mathbf{y}}^{t-1}-\underline{\hat{\mathbf{y}}}^{t-1}\|^2$$

$$+4\alpha\eta_2^2L_1^2(1+\frac{1}{\beta_2})\left\|\hat{\mathbf{y}}^{t-1}\right\|^2 \tag{109}$$

$$\tag{110}$$

For $\beta_2=\frac{1-\sigma}{\sigma}$ and choosing η_2,α to satisfy conditions (41) - (44) we guarantee that the coefficients of $\|\underline{\mathbf{x}}^{t-1}-\underline{\hat{\mathbf{x}}}^{t-1}\|^2$ and $\|\underline{\mathbf{y}}^{t-1}-\underline{\hat{\mathbf{y}}}^{t-1}\|^2$ are non-positive.

Thus, we obtain

$$\frac{1}{m}\left\|\underline{\mathbf{x}}^t-\underline{\hat{\mathbf{x}}}^t\right\|^2-\frac{1}{m}\left\|\underline{\mathbf{x}}^{t-1}-\underline{\hat{\mathbf{x}}}^{t-1}\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^t-\underline{\hat{\mathbf{y}}}^t\right\|^2-\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{t-1}-\underline{\hat{\mathbf{y}}}^{t-1}\right\|^2 \tag{111}$$

$$\leq 4\alpha\eta_2^2L_1^2\frac{1}{1-\sigma}\left\|\hat{\mathbf{y}}^{t-1}\right\|^2 \tag{112}$$

Summing up from 1 to t we get

$$\frac{1}{m}\left\|\underline{\mathbf{x}}^t-\underline{\hat{\mathbf{x}}}^t\right\|^2-\frac{1}{m}\left\|\underline{\mathbf{x}}^0-\underline{\hat{\mathbf{x}}}^0\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^t-\underline{\hat{\mathbf{y}}}^t\right\|^2-\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^0-\underline{\hat{\mathbf{y}}}^0\right\|^2 \tag{113}$$

$$\leq 4\alpha\eta_2^2L_1^2\frac{1}{1-\sigma}\sum_{i=0}^{t-1}\left\|\hat{\mathbf{y}}^i\right\|^2 \tag{114}$$

$$\leq 4\alpha\eta_2^2L_1^2\frac{1}{1-\sigma}\left(12\frac{\mathcal{F}}{\eta_2}\right) \tag{115}$$

$$=48L_1^2\frac{1}{1-\sigma}\alpha\eta_2\mathcal{F} \tag{116}$$

Where the second inequality holds due to (86). The above implies a useful bound for $\frac{1}{m}\left\|\underline{\mathbf{x}}^t-\underline{\hat{\mathbf{x}}}^t\right\|^2$:

$$\frac{1}{m}\left\|\underline{\mathbf{x}}^t-\underline{\hat{\mathbf{x}}}^t\right\|^2\leq\frac{1}{m}\left\|\underline{\mathbf{x}}^0-\underline{\hat{\mathbf{x}}}^0\right\|^2-\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^t-\underline{\hat{\mathbf{y}}}^t\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^0-\underline{\hat{\mathbf{y}}}^0\right\|^2+48L_1^2\frac{1}{1-\sigma}\alpha\eta_2\mathcal{F} \tag{117}$$

$$\leq\frac{1}{m}\left\|\underline{\mathbf{x}}^0-\underline{\hat{\mathbf{x}}}^0\right\|^2+\frac{\alpha}{m}\left\|\underline{\mathbf{y}}^0-\underline{\hat{\mathbf{y}}}^0\right\|^2+48L_1^2\frac{1}{1-\sigma}\alpha\eta_2\mathcal{F} \tag{118}$$

$$\leq\epsilon_2^2+\frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}+48L_1^2\frac{1}{1-\sigma}\alpha\eta_2\mathcal{F} \tag{119}$$

$$\leq 49L_1^2\frac{1}{1-\sigma}\alpha\eta_2\mathcal{F} \tag{120}$$

where the third inequality comes from (55) and (58) and the forth due to the assumption $\epsilon_2^2\leq\frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}$. As immediate corollary we get

$$\frac{1}{\sqrt{m}}\left\|\underline{\mathbf{x}}^t-\underline{\hat{\mathbf{x}}}^t\right\|\leq 7L_1\frac{1}{\sqrt{1-\sigma}}\sqrt{\alpha\eta_2\mathcal{F}}=7\sqrt{L_1}\frac{1}{\sqrt{1-\sigma}}\sqrt{\alpha\eta_2}\mathcal{R} \tag{121}$$

Finally, notice that

$$\eta_2 \frac{L_1}{m} \sum_{i=1}^{m} \left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\| \leq \eta_2 \frac{L_1}{m} \sqrt{m \sum_{i=1}^{m} \left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2} \tag{122}$$

$$\leq \eta_2 \frac{L_1}{\sqrt{m}} \left\|\underline{\mathbf{x}}^t - \underline{\hat{\mathbf{x}}}^t\right\| \tag{123}$$

$$\leq \eta_2 L_1 7\sqrt{L_1} \frac{1}{\sqrt{1-\sigma}} \sqrt{\alpha\eta_2}\mathcal{R} \tag{124}$$

$$= 7L_1\sqrt{L_1} \frac{1}{\sqrt{1-\sigma}} \eta_2\sqrt{\alpha\eta_2}\mathcal{R} \tag{125}$$

$$\leq \frac{c_{new}}{2} \frac{1}{\sqrt{1-\sigma}} \eta_2\sqrt{\alpha\eta_2}\mathcal{R} \tag{126}$$

and the last inequality holds for appropriate constant $c_{new} \geq 14L_1\sqrt{L_1}$. Since $\mathbf{w}$ is a sequence which develops also with the same parameters of Gradient Tracking the same bounds hold for $\eta_2 \frac{L_1}{m} \sum\limits_{i=1}^{m} \|\hat{\mathbf{w}}^t - \mathbf{w}_i^t\|$ as well. The result follows.

□

The next lemma shows that if a sequence $\mathbf{x}^t$ does not a escape from the saddle point after T_{cap} iterations of Gradient Tracking then any other sequence $\mathbf{w}^t$ with the same starting point as $\mathbf{x}^t$ will escape if given a little bit of a nudge towards the direction of negative curvature.

Lemma 15. *Assume the major condition* (86) *and conditions* (55) *and* (58) *hold; Let* $\epsilon_2^2 \leq \frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}$ *and assume* $\mathbf{x}^{-1}$ *such that* $\lambda_{\min}(\nabla^2 f(\hat{\mathbf{x}}^{-1})) \leq -\gamma$ *and further* $\forall i$ *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$. *Consider that the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44) *and further* $\eta_2 \leq \min\{1, \frac{1-\sigma}{L_1}\}$ *and* $\alpha\eta_2 \leq \left(\left(\frac{\delta_2(\sqrt{2}-1)}{8\sqrt{2}\hat{c}\cdot c_{new}}\right)^2 \frac{(1-\sigma)|\lambda_{\min}(H)|^2}{d\log^2\left(\frac{dL_1}{\gamma\delta_2}\right)}\right)$. *Define the sequence of* $\mathbf{w}_i$*'s similarly to* $\mathbf{x}_i$*'s except* $\forall i \quad \mathbf{w}_i^0 = \mathbf{x}_i^0 + \mu\mathcal{R}\mathbf{e}_1$ *with* $\mathbf{e}_1$ *a unit eigenvector corresponding to the minimum eigenvalue of* $\nabla^2 f(\hat{\mathbf{x}}^{-1})$ *and* $\mu \in [\delta_2/2\sqrt{d}, 1]$. *Let* $\mathbf{u}^t = \hat{\mathbf{w}}^t - \hat{\mathbf{x}}^t$ *and consider positive constants* $\hat{c}, c_{new}$ *such that* $\hat{c} \geq 36$ *and* $c_{new} \geq 14L_1\sqrt{L_1}$. *Further, consider the definition of the stopping time* $T_{\hat{\mathbf{w}}}$

$$T_{\hat{\mathbf{w}}} = \min\{\inf_t\{t | f(\hat{\mathbf{w}}^t) - f(\hat{\mathbf{w}}^0) \leq -3\mathcal{F}\}, \hat{c}\mathcal{J}\}$$

If $\left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}^{-1}\right\| \leq 2\hat{c}\mathcal{P}$ *for all* $t < T_{\hat{\mathbf{w}}}$, *then it can be shown that* $T_{\hat{\mathbf{w}}} < \hat{c}\mathcal{J}$.

Proof. From the update rule of the iterates we have

$$\hat{\mathbf{x}}^{t+1} + \mathbf{u}^{t+1} = \hat{\mathbf{w}}^{t+1} \tag{127}$$

$$= \hat{\mathbf{w}}^t - \eta_2 \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\mathbf{w}_i^t) \tag{128}$$

$$= \hat{\mathbf{x}}^t + \mathbf{u}^t - \eta_2 \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}^t + \mathbf{u}^t) + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) \tag{129}$$

$$= \hat{\mathbf{x}}^t + \mathbf{u}^t - \frac{\eta_2}{m} \sum_{i=1}^{m} \nabla f_i(\hat{\mathbf{x}}^t) - \eta_2 \left(\int_0^1 \nabla^2 f(\hat{\mathbf{x}}^t + \theta\mathbf{u}^t) d\theta\right) \mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) \tag{130}$$

where $\Delta^t = \int_0^1 \frac{1}{m} \sum_{i=1}^{m} \nabla f(\hat{\mathbf{x}}^t + \theta \mathbf{u}^t) d\theta - H$ and $H = \nabla^2 f(\hat{\mathbf{x}}^{-1})$. Hence, we obtain that

$$\mathbf{u}^{t+1} = \mathbf{u}^t - \eta_2 (H + \Delta^t)\mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) - \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t)) \tag{131}$$

which can be simplified as

$$\mathbf{u}^{t+1} = (\mathbf{I} - \eta_2 H)\, \mathbf{u}^t - \eta_2 \Delta^t \mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) - \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t)) \tag{132}$$

Consider the decomposition $\mathbf{u}^t = \mathbf{u}_{par}^t + \mathbf{u}_{per}^t$ where $\mathbf{u}_{par}^t$ is parallel with $\mathbf{e}_1$ and $\mathbf{u}_{per}^t$ is perpendicular to $\mathbf{e}_1$. Then, we can show that

$$\begin{aligned} \mathbf{e}_1^T(\mathbf{u}_{par}^{t+1} + \mathbf{u}_{per}^{t+1}) &= \mathbf{e}_1^T\, (\mathbf{I} - \eta_2 H)\, (\mathbf{u}_{par}^t + \mathbf{u}_{per}^t) \\ &\quad + \mathbf{e}_1^T \left(-\eta_2 \Delta^t \mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) - \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t)) \right) \end{aligned} \tag{133}$$

which can be simplified as

$$\begin{aligned} \mathbf{e}_1^T \mathbf{u}_{par}^{t+1} &= \mathbf{e}_1^T (\mathbf{u}_{par}^t + \mathbf{u}_{per}^t) - \eta_2 (\mathbf{e}_1^T H)(\mathbf{u}_{par}^t + \mathbf{u}_{per}^t) \\ &\quad + \mathbf{e}_1^T \left(-\eta_2 \Delta^t \mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) - \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t)) \right) \\ &= \mathbf{e}_1^T \mathbf{u}_{par}^t - \eta_2 (\lambda_{\min}(H) \mathbf{e}_1^T)(\mathbf{u}_{par}^t + \mathbf{u}_{per}^t) \\ &\quad + \mathbf{e}_1^T \left(-\eta_2 \Delta^t \mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) - \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t)) \right) \\ &= (1 - \eta_2 \lambda_{\min}(H)) \mathbf{e}_1^T \mathbf{u}_{par}^t \\ &\quad + \mathbf{e}_1^T \left(-\eta_2 \Delta^t \mathbf{u}^t + \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{w}}^t) - \nabla f_i(\mathbf{w}_i^t)) - \frac{\eta_2}{m} \sum_{i=1}^{m} (\nabla f_i(\hat{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_i^t)) \right) \end{aligned} \tag{134}$$

If we define ψ_t as the norm of the projection of $\mathbf{u}^t$ onto $\mathbf{e}_1$ and define ϕ_t as the norm of the projection on the complementary subspace. Considering the above expression for the sequence $\mathbf{u}^t$, we can show that

$$\psi_{t+1} \geq (1 + \eta_2 |\lambda_{\min}(H)|)\psi_t - \eta_2 \left\|\Delta^t\right\| \left\|\mathbf{u}^t\right\| - \eta_2 \frac{L_1}{m} \sum_{i=1}^{m} (\left\|\hat{\mathbf{w}}^t - \mathbf{w}_i^t\right\| + \left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\|) \tag{135}$$

Next, consider $\mathbf{e}_2, \ldots, \mathbf{e}_d$ as the remaining eigenvectors of H which create the complementary space of $\mathbf{e}_1$. The projection of any vector $\mathbf{v}$ to this subspace is given by $\sum_{j=2}^{d} (\mathbf{v}^T \mathbf{e}_j)\mathbf{e}_j$. Therefore, norm of the projection of vector $\mathbf{u}^{t+1}$ onto this subspace is given by

$$\| \sum_{j=2}^{d} (\mathbf{e}_j^T \mathbf{u}^{t+1}) \mathbf{e}_j \| = \| \sum_{j=2}^{d} (\mathbf{e}_j^T (\mathbf{u}_{par}^{t+1} + \mathbf{u}_{per}^{t+1})) \mathbf{e}_j \| = \| \sum_{j=2}^{d} (\mathbf{e}_j^T \mathbf{u}_{per}^{t+1}) \mathbf{e}_j \| = \| \mathbf{u}_{per}^{t+1} \|$$

as expected by the definition. Using the same argument we can show that $\| \sum_{j=2}^{d} (\mathbf{e}_j^T \mathbf{u}^t) \mathbf{e}_j \| = \| \mathbf{u}_{per}^t \|$. In addition, we can show that

$$\begin{aligned} \| \sum_{j=2}^{d} (\mathbf{e}_j^T (I - \eta_2 H) \mathbf{u}^t) \mathbf{e}_j \| &= \| \sum_{j=2}^{d} ((1 - \eta_2 \lambda_j(H)) \mathbf{e}_j^T \mathbf{u}^t) \mathbf{e}_j \| \\ &\leq (1 - \eta_2 \lambda_{\min}(H)) \| \sum_{j=2}^{d} (\mathbf{e}_j^T \mathbf{u}^t) \mathbf{e}_j \| \\ &= (1 - \eta_2 \lambda_{\min}(H)) \| \mathbf{u}_{per}^t \| \end{aligned} \tag{136}$$

Then, according to (132) we can write

$$\|\mathbf{u}_{per}^{t+1}\| \tag{137}$$

$$= \|\sum_{j=2}^{d}(\mathbf{e}_j^T\mathbf{u}^{t+1})\mathbf{e}_j\| \tag{138}$$

$$= \big\|\sum_{j=2}^{d}(\mathbf{e}_j^T\left(\mathbf{I}-\eta_2 H\right)\mathbf{u}^t)\mathbf{e}_j \tag{139}$$

$$+\sum_{j=2}^{d}(\mathbf{e}_j^T\left(-\eta_2\Delta^t\mathbf{u}^t+\frac{\eta_2}{m}\sum_{i=1}^{m}(\nabla f_i(\hat{\mathbf{w}}^t)-\nabla f_i(\mathbf{w}_i^t))-\frac{\eta_2}{m}\sum_{i=1}^{m}(\nabla f_i(\hat{\mathbf{x}}^t)-\nabla f_i(\mathbf{x}_i^t))\right))\mathbf{e}_j\big\| \tag{140}$$

$$\leq \left\|\sum_{j=2}^{d}(\mathbf{e}_j^T\left(\mathbf{I}-\eta_2 H\right)\mathbf{u}^t)\mathbf{e}_j\right\| \tag{141}$$

$$+\left\|-\eta_2\Delta^t\mathbf{u}^t+\frac{\eta_2}{m}\sum_{i=1}^{m}(\nabla f_i(\hat{\mathbf{w}}^t)-\nabla f_i(\mathbf{w}_i^t))-\frac{\eta_2}{m}\sum_{i=1}^{m}(\nabla f_i(\hat{\mathbf{x}}^t)-\nabla f_i(\mathbf{x}_i^t))\right\| \tag{142}$$

$$\leq (1-\eta_2\lambda_{\min}(H))\|\mathbf{u}_{per}^t\|$$

$$+\left\|-\eta_2\Delta^t\mathbf{u}^t+\frac{\eta_2}{m}\sum_{i=1}^{m}(\nabla f_i(\hat{\mathbf{w}}^t)-\nabla f_i(\mathbf{w}_i^t))-\frac{\eta_2}{m}\sum_{i=1}^{m}(\nabla f_i(\hat{\mathbf{x}}^t)-\nabla f_i(\mathbf{x}_i^t))\right\| \tag{143}$$

Note that ϕ_t is the norm of the projection onto the complementary subspace which is equal to $\|\mathbf{u}_{per}^t\|$. Further, since $\lambda_{\min}(H)\leq-\gamma$ then we have $|\lambda_{\min}(H)|=-\lambda_{\min}(H)$. Considering these points we can write

$$\phi_{t+1}\leq(1+\eta_2|\lambda_{\min}(H)|)\phi_t+\eta_2\left\|\Delta^t\right\|\left\|\mathbf{u}^t\right\|+\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\left\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\right\|+\left\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\right\|) \tag{144}$$

To bound the norm of Δ^t first note that for any $t<T_{\hat{\mathbf{w}}}$

$$\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{-1}\right\|\leq 2\hat{c}\mathcal{P} \tag{145}$$

and further since $\hat{\mathbf{w}}^0$ satisfy the condition of **Lemma 13** we have

$$\left\|\hat{\mathbf{w}}^0-\hat{\mathbf{x}}^{-1}\right\|=\left\|\hat{\mathbf{x}}^0-\hat{\mathbf{x}}^{-1}\right\|+\left\|\mathbf{u}^0\right\|\leq\mathcal{R}+\mu\mathcal{R}\leq 2\mathcal{R}\Rightarrow\left\|\hat{\mathbf{w}}^t-\hat{\mathbf{x}}^{-1}\right\|\leq 2\hat{c}\mathcal{P} \tag{146}$$

It follows that for any $t<T_{\hat{\mathbf{w}}}$

$$\left\|\mathbf{u}^t\right\|=\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{w}}^t\right\|\leq\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{-1}\right\|+\left\|\hat{\mathbf{w}}^t-\hat{\mathbf{x}}^{-1}\right\|\leq 4\hat{c}\mathcal{P}. \tag{147}$$

Now we can show that

$$\left\|\Delta^t\right\|\leq\left\|\nabla^2 f(\hat{\mathbf{x}}^t+\mathbf{u}^t)-\nabla^2 f(\hat{\mathbf{x}}^{-1})\right\|\leq L_2(\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{-1}\right\|+\left\|\mathbf{u}^t\right\|)\leq 6L_2\hat{c}\mathcal{P} \tag{148}$$

Using this upper bound we can show that

$$\psi_{t+1}\geq(1+\eta_2|\lambda_{\min}(H)|)\psi_t-\zeta\sqrt{\psi_t^2+\phi_t^2}-\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\left\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\right\|+\left\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\right\|) \tag{149}$$

and

$$\phi_{t+1}\leq(1+\eta_2|\lambda_{\min}(H)|)\phi_t+\zeta\sqrt{\psi_t^2+\phi_t^2}+\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\left\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\right\|+\left\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\right\|) \tag{150}$$

where

$$\zeta=\eta_2 L_2 6\hat{c}\mathcal{P}=\eta_2 L_2 6\hat{c}\sqrt{\frac{\mathcal{F}\log(d\kappa/\delta_2)}{|\lambda_{\min}(H)|}}. \tag{151}$$

Using induction we are going to prove the following two statements $\forall t < T_{cap}$.

$$\eta_2 \frac{L_1}{m} \sum_{i=1}^{m} (\left\|\hat{\mathbf{w}}^t - \mathbf{w}_i^t\right\| + \left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\|) \leq \zeta \psi_t \tag{152}$$

$$\phi_t \leq 4t\zeta\psi_t \tag{153}$$

Recall the values of the following variables

$$\mathcal{F} = \frac{|\lambda_{\min}(H)|^3}{\log^3(d\kappa/\delta_2)} \frac{(\sqrt{2}-1)^2}{(24\sqrt{2}L_2\hat{c}^2)^2} \tag{154}$$

$$\mathcal{P} = \frac{|\lambda_{\min}(H)|}{\log(d\kappa/\delta_2)} \frac{\sqrt{2}-1}{(24\sqrt{2}L_2\hat{c}^2)} \tag{155}$$

For the base of the induction, note that since $\mathbf{u}^0 = \mu\mathcal{R}\mathbf{e}_1$ then we can conclude that $\psi_0 = \mu\mathcal{R}$ and $\phi_0 = 0$. Then, we have

$$\zeta\psi_0 = \eta_2 L_2 6\hat{c} \sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \psi_0 \tag{156}$$

$$= \eta_2 L_2 6\hat{c} \sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \mathcal{R}\mu \tag{157}$$

$$\geq \eta_2 L_2 6\hat{c} \sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \mathcal{R} \frac{\delta_2}{2\sqrt{d}} \tag{158}$$

$$= 3\hat{c}L_2 \frac{\delta_2}{\sqrt{d}} \eta_2 \sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \mathcal{R} \tag{159}$$

where the inequality follows from the fact that $\mu \geq \frac{\delta_2}{2\sqrt{d}}$. Using this inequality and the result of Lemma 14 we can show that

$$\zeta\psi_0 - \eta_2 \frac{L_1}{m} \sum_{i=1}^{m} (\left\|\hat{\mathbf{w}}^t - \mathbf{w}_i^t\right\| + \left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\|)$$

$$\geq 3\hat{c}L_2 \frac{\delta_2}{\sqrt{d}} \eta_2 \sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}} \mathcal{R} - c_{new} \sqrt{\frac{\alpha\eta_2}{1-\sigma}} \eta_2 \mathcal{R} \tag{160}$$

$$= \eta_2 \mathcal{R} \left(3\hat{c}L_2 \frac{\delta_2}{\sqrt{d}} \sqrt{\frac{\mathcal{F}\log(d\kappa/\delta_2)}{|\lambda_{\min}(H)|}} - c_{new} \sqrt{\frac{\alpha\eta_2}{1-\sigma}} \right) \tag{161}$$

$$= \eta_2 \mathcal{R} \left(3\hat{c}L_2 \frac{\delta_2 \left(\sqrt{2}-1\right)}{24\sqrt{2}L_2\hat{c}^2} \frac{|\lambda_{\min}(H)|}{\sqrt{d}\log(d\kappa/\delta_2)} - c_{new} \sqrt{\frac{\alpha\eta_2}{1-\sigma}} \right) \tag{162}$$

$$= \eta_2 \mathcal{R} \left(\frac{\delta_2 \left(\sqrt{2}-1\right)}{8\sqrt{2}\hat{c}} \frac{|\lambda_{\min}(H)|}{\sqrt{d}\log(d\kappa/\delta_2)} - c_{new} \sqrt{\frac{\alpha\eta_2}{1-\sigma}} \right) \tag{163}$$

$$\geq 0 \tag{164}$$

where the last inequality holds for $\alpha\eta_2 \leq \left(\left(\frac{\delta_2\left(\sqrt{2}-1\right)}{8\sqrt{2}\hat{c}\cdot c_{new}} \right)^2 \frac{(1-\sigma)|\lambda_{\min}(H)|^2}{d\log^2\left(\frac{dL_1}{\gamma\delta_2}\right)} \right)$. Hence, there are α, η_2 properly chosen such that $a\eta_2 = \tilde{\mathcal{O}}\left(\frac{(1-\sigma)\gamma^2}{d}\right)$ that the base of induction for (152) holds. Further, since $\phi_0 = 0$ the second condition (153) is also satisfied for $t = 0$ and the base of the induction is complete.

Now let's assume that the conditions in (152) and (153) hold for time t. Our goal is to show that these conditions also hold for time $t+1$.

From the inductive hypothesis we have $\phi_t \leq 4t\zeta\psi_t$ and also $\zeta\sqrt{\psi_t^2+\phi_t^2} \geq \zeta\psi_t \geq \eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\|+\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\|)$. Thus

$$\begin{aligned}\psi_{t+1} &\geq (1+\eta_2|\lambda_{\min}(H)|)\psi_t-\zeta\sqrt{\psi_t^2+\phi_t^2}-\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\|+\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\|)\\ &\geq (1+\eta_2|\lambda_{\min}(H)|)\psi_t-2\zeta\sqrt{\psi_t^2+\phi_t^2} \end{aligned} \tag{165}$$

And similarly we can show that

$$\phi_{t+1} \leq (1+\eta_2|\lambda_{\min}(H)|)\phi_t+2\zeta\sqrt{\psi_t^2+\phi_t^2} \tag{166}$$

By multiplying both sides of (165) by $4(t+1)\zeta$ we obtain that

$$4(t+1)\zeta\psi_{t+1} \geq 4(t+1)\zeta\left((1+\eta_2|\lambda_{\min}(H)|)\psi_t-2\zeta\sqrt{\psi_t^2+\phi_t^2}\right) \tag{167}$$

And if we replace ϕ_t in the right hand side (166) by its upper bound $4t\zeta\psi_t$ (given by the induction hypothesis), then we obtain

$$\phi_{t+1} \leq (1+\eta_2|\lambda_{\min}(H)|)4t\zeta\psi_t+2\zeta\sqrt{\psi_t^2+\phi_t^2} \tag{168}$$

Considering the inequalities in (167) and (168), to prove that $4(t+1)\zeta\psi_{t+1} \geq \phi_{t+1}$ it suffices to show

$$4(t+1)\zeta\left((1+\eta_2|\lambda_{\min}(H)|)\psi_t-2\zeta\sqrt{\psi_t^2+\phi_t^2}\right) \geq 4t\zeta(1+\eta_2|\lambda_{\min}(H)|)\psi_t+2\zeta\sqrt{\psi_t^2+\phi_t^2} \tag{169}$$

which is equivalent to

$$4(t+1)\left((1+\eta_2|\lambda_{\min}(H)|)\psi_t-2\zeta\sqrt{\psi_t^2+\phi_t^2}\right) \geq 4t(1+\eta_2|\lambda_{\min}(H)|)\psi_t+2\sqrt{\psi_t^2+\phi_t^2} \tag{170}$$

Expanding the left hand side leads to

$$\begin{aligned}&4t(1+\eta_2|\lambda_{\min}(H)|)\psi_t+4(1+\eta_2|\lambda_{\min}(H)|)\psi_t-8t\zeta\sqrt{\psi_t^2+\phi_t^2}-8\zeta\sqrt{\psi_t^2+\phi_t^2}\\ &\geq 4t(1+\eta_2|\lambda_{\min}(H)|)\psi_t+2\sqrt{\psi_t^2+\phi_t^2}\end{aligned} \tag{171}$$

Hence, the conditions in (169), (170), and (171) are equivalent. By regrouping the terms in (171) and dividing both sides by 2 we obtain the following condition

$$2(1+\eta_2|\lambda_{\min}(H)|)\psi_t \geq (1+4(t+1)\zeta)\sqrt{\psi_t^2+\phi_t^2} \tag{172}$$

Indeed, the condition in (172) holds if and only if (171) holds.

Finally to prove the last inequality in (172) notice that

$$\begin{aligned}4(t+1)\zeta &\leq 4\zeta T_{cap}\\ &\leq 4\eta_2 6L_2\hat{c}\mathcal{P}\hat{c}\mathcal{J}\\ &\leq 24\eta_2 L_2\hat{c}^2\mathcal{P}\mathcal{J}\\ &\leq 24\eta_2 L_2\hat{c}^2\frac{|\lambda_{\min}(H)|}{\log(d\kappa/\delta_2)}\frac{\sqrt{2}-1}{(24\sqrt{2}L_2\hat{c}^2)}\frac{\log(d\kappa/\delta_2)}{\eta_2|\lambda_{\min}(H)|}\\ &\leq \sqrt{2}-1\end{aligned} \tag{173}$$

Thus, we can show that

$$(1+4(t+1)\zeta)\sqrt{\psi_t^2+\phi_t^2} \leq \sqrt{2}\sqrt{\psi_t^2+\psi_t^2} \leq \sqrt{2}\sqrt{2\psi_t^2} \leq 2(1+\eta_2|\lambda_{\min}(H)|)\psi_t \tag{174}$$

Hence, the condition in (172) holds, and as a result the condition in (169) holds. As we mentioned, (169) together with (167) and (168) implies that

$$\phi_{t+1} \leq 4t\zeta\psi_{t+1}. \tag{175}$$

Hence, the induction step for (152) is complete.

Next we show that if (153) holds for t it also holds for $t+1$. To do so, note that by considering the fact that $4t\zeta \leq \sqrt{2}-1$, from (173) and the result in (175), we can show that

$$\phi_{t+1} \leq \psi_{t+1} \tag{176}$$

Using the result in (176) as well as the inequality in (165) we can show that

$$\begin{aligned}
\psi_{t+1} &\geq (1+\eta_2|\lambda_{\min}(H)|)\psi_t - 2\sqrt{2}\zeta\psi_t \\
&\geq \psi_t + \eta_2|\lambda_{\min}(H)|\psi_t - 12\sqrt{2}\eta_2 L_2\hat{c}\sqrt{\frac{\mathcal{F}\log(d\kappa/\delta_2)}{|\lambda_{\min}(H)|}}\psi_t \\
&\geq \psi_t + \eta_2|\lambda_{\min}(H)|\psi_t - 12\sqrt{2}\eta_2 L_2\hat{c}\frac{|\lambda_{\min}(H)|}{\log(d\kappa/\delta_2)}\cdot\frac{\sqrt{2}-1}{24\sqrt{2}L_2\hat{c}^2}\psi_t \\
&\geq \psi_t(1+\frac{\eta_2|\lambda_{\min}(H)|}{2})
\end{aligned} \tag{177}$$

Since we showed that $\psi_{t+1} \geq \psi_t$ and thus $\zeta\psi_{t+1} \geq \zeta\psi_t$ it is straight forward to prove the second condition of the inductive step (153) simply by using the same bound for $\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\|+\|\hat{\mathbf{x}}^{t+1}-\mathbf{x}_i^{t+1}\|)$ from Lemma 14. To be more precise, based on the result of Lemma 14 we know that for any $t \leq T_{cap}$

$$\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\|\hat{\mathbf{w}}^t-\mathbf{w}_i^t\|+\|\hat{\mathbf{x}}^t-\mathbf{x}_i^t\|) \leq 3\hat{c}L_2\frac{\delta_2}{\sqrt{d}}\eta_2\sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}}\mathcal{R} \leq \zeta\psi_0 \tag{178}$$

Using this result and the fact that ψ_t is increasing we can show that

$$\eta_2\frac{L_1}{m}\sum_{i=1}^{m}(\|\hat{\mathbf{w}}^{t+1}-\mathbf{w}_i^{t+1}\|+\|\hat{\mathbf{x}}^{t+1}-\mathbf{x}_i^{t+1}\|) \leq 3\hat{c}L_2\frac{\delta_2}{\sqrt{d}}\eta_2\sqrt{\frac{\log(d\kappa/\delta_2)\mathcal{F}}{|\lambda_{\min}(H)|}}\mathcal{R} \leq \zeta\psi_{t+1} \tag{179}$$

and the induction is complete.

Next, using the result of induction in (152) and (153) we show that $T_{\hat{\mathbf{w}}} < \hat{c}\mathcal{J}$. To do so, note that for all $t < T_{\hat{\mathbf{w}}}$ we have

$$\begin{aligned}
4\hat{c}\mathcal{P} &\geq \left\|\mathbf{u}^t\right\| \\
&\geq \psi_t \\
&\geq (1+\frac{\eta_2|\lambda_{\min}(H)|}{2})^t\psi_0 \\
&= (1+\frac{\eta_2|\lambda_{\min}(H)|}{2})^t\mu\mathcal{R} \\
&\geq (1+\frac{\eta_2|\lambda_{\min}(H)|}{2})^t\frac{\delta_2}{2\sqrt{d}}\sqrt{\frac{\mathcal{F}}{L_1}} && (180)\\
&\geq (1+\frac{\eta_2|\lambda_{\min}(H)|}{2})^t\frac{\delta_2}{2\sqrt{d}}\frac{|\lambda_{\min}|}{\log(d\kappa/\delta_2)}\frac{\sqrt{2}-1}{24\sqrt{2}L_2\hat{c}^2}\sqrt{\frac{|\lambda_{\min}|}{L_1\log(d\kappa/\delta_2)}} && (181)\\
&\geq (1+\frac{\eta_2|\lambda_{\min}(H)|}{2})^t\frac{\delta_2\mathcal{P}}{2\sqrt{d}}\sqrt{\frac{|\lambda_{\min}|}{L_1\log(d\kappa/\delta_2)}} && (182)\\
&\geq (1+\frac{\eta_2|\lambda_{\min}(H)|}{2})^t\frac{\delta_2\mathcal{P}}{2\sqrt{d}}\frac{|\lambda_{\min}|}{L_1\log(d\kappa/\delta_2)} && (183)\\
&\geq (1+\eta_2|\lambda_{\min}(H)|/2)^t\frac{\delta_2\mathcal{P}}{2\sqrt{d}\kappa\log(d\kappa/\delta_2)} && (184)
\end{aligned}$$

where the first inequality follows from (147), the second inequality holds since ψ_t is the norm of projection of $\mathbf{u}^t$ onto a subspace, the third inequality holds because of the result in (177), and the second to last inequality holds since $\log(d\kappa/\delta_2) \geq 1$ and $\frac{|\lambda_{\min}|}{L_1} \leq 1$. Hence, we have for $t < T_{\hat{\mathbf{w}}}$

$$\frac{8\hat{c}\sqrt{d}\kappa\log(d\kappa/\delta_2)}{\delta_2} \geq (1+\eta_2|\lambda_{\min}(H)|/2)^t \tag{185}$$

Therefore, this condition should also hold for $t = T_{\hat{\mathbf{w}}} - 1$ and therefore we have

$$T_{\hat{\mathbf{w}}} - 1 \leq \frac{\log(8\frac{\kappa\sqrt{d}}{\delta_2}\hat{c}\log(d\kappa/\delta_2))}{\log(1+\frac{|\lambda_{\min}(H)|\eta_2}{2})} \tag{186}$$

$$\Rightarrow T_{\hat{\mathbf{w}}} < \frac{\log(8\frac{\kappa\sqrt{d}}{\delta_2}\hat{c}\log(d\kappa/\delta_2))}{\log(1+\frac{|\lambda_{\min}(H)|\eta_2}{2})}+1 \leq \frac{5}{2}\frac{\log(8\frac{\kappa\sqrt{d}}{\delta_2}\hat{c}\log(d\kappa/\delta_2))}{1+\frac{|\lambda_{\min}(H)|\eta_2}{2}}+1 \leq \frac{5}{2}(2+\log(8\hat{c}))\mathcal{J}+1 < \hat{c}\mathcal{J} \tag{187}$$

where the last inequality uses the facts that $\delta_2 \in (0, \frac{d\kappa}{e}]$ and $\log(d\kappa/\delta_2) \geq 1$ and $\hat{c}$ such that $\frac{5}{2}(2+\log(8\hat{c})) \leq \hat{c}$. □

Now we are going to use Lemma 15 to show substantial function decrease in a small number of iterations after the noise injection with high probability. Specifically, we are going to show that $f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^{-1}) \leq -\mathcal{F}$ for some $T < \hat{c}\mathcal{J}$ which will be used subsequently consequently to show $H(\mathbf{x}^{\hat{c}\mathcal{J}}, \mathbf{y}^{\hat{c}\mathcal{J}}) - H(\mathbf{x}^{-1}, \mathbf{y}^{-1}) \leq -\mathcal{F}$.

Lemma 16. *Assume the major condition* (86) *and conditions* (54)*,* (55) *and* (58) *hold; let* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$ *and* $\epsilon_2^2 \leq \min\{\frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}, \mathcal{F}\frac{L_1^2}{2+2L_1^2}\}$*. Assume* $\mathbf{x}^{-1}$ *such that* $\lambda_{\min}(\nabla^2 f(\hat{\mathbf{x}}^{-1})) \leq -\gamma$ *and further* $\forall i$ *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$*. Consider that the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44) *and further* $\eta_2 \leq \min\{1, \frac{1-\sigma}{L_1}\}$ *and* $\alpha\eta_2 \leq \left(\left(\frac{\delta_2(\sqrt{2}-1)}{8\sqrt{2}\hat{c}\cdot c_{new}}\right)^2 \frac{(1-\sigma)|\lambda_{\min}(H)|^2}{d\log^2\left(\frac{dL_1}{\gamma\delta_2}\right)}\right)$*. Then with probability at least* $1-\delta_2$ *we have the following for some* $T < \hat{c}\mathcal{J}$

$$f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) \leq -3\mathcal{F} \tag{188}$$

which implies

$$f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^{-1}) \leq -\mathcal{F} \tag{189}$$

Proof. In Lemma 10 by adding perturbation we proved that the function value increases at most by $\frac{3}{2}\mathcal{F}$. Thus we have

$$f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) \leq \frac{3}{2}\mathcal{F} \tag{190}$$

We know that $\hat{\mathbf{x}}^0$ comes from the uniform distribution over $\mathcal{B}_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})$. Let us denote with $\mathcal{X}_{\textit{stuck}} \subset \mathcal{B}_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})$ the set of bad starting points so that if $\hat{\mathbf{x}}^0 \in \mathcal{X}_{\textit{stuck}}$, then the iterates are not going to make substantial progress after at most $\hat{c}\mathcal{J}$ steps i.e. $f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) > -3\mathcal{F}, \ \ \forall T < \hat{c}\mathcal{J}$. On the contrary when $\hat{\mathbf{x}}^0 \in (\mathcal{B}_{\hat{\mathbf{x}}^{-1}}(R) - \mathcal{X}_{\textit{stuck}})$ there exists a T such that $f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) < -3\mathcal{F}$.
By the Lemma 15 we know that when $\hat{\mathbf{x}}^0 \in \mathcal{X}_{\textit{stuck}}$ it is guaranteed that $(\hat{\mathbf{x}}^0 \pm \mu\mathcal{R}\mathbf{e}_1) \notin \mathcal{X}_{\textit{stuck}}$ where $\mu \in \left[\frac{\delta_2}{2\sqrt{d}}, 1\right]$. Denote with $\mathcal{I}_{\mathcal{X}_{\textit{stuck}}}(\cdot)$ the indicator function of being inside set $\mathcal{X}_{\textit{stuck}}$ and vector $\mathbf{v} = \left(v^{(1)}, \mathbf{v}^{(-1)}\right)$, where v^1 is the component along the direction of $\mathbf{e}_1$ and $\mathbf{v}^{(-1)}$ the remaining vector. We are going to derive an upper bound on the volume of $\mathcal{X}_{\textit{stuck}}$.

$$\begin{aligned}
\textit{Vol}(\mathcal{X}_{\textit{stuck}}) &= \int_{\mathcal{B}^d_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})} d\mathbf{x} \cdot \mathcal{I}_{\mathcal{X}_{\textit{stuck}}}(\mathbf{x}) && (191)\\
&\leq \int_{\mathcal{B}^{d-1}_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})} d\mathbf{x}^{(-1)} \cdot 2\frac{\delta_2}{2\sqrt{d}}\mathcal{R} && (192)\\
&= \textit{Vol}(\mathcal{B}^{d-1}_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})) \times \frac{\delta_2\mathcal{R}}{\sqrt{d}} && (193)
\end{aligned}$$

Then we immediately have the ratio:

$$\frac{\textit{Vol}(\mathcal{X}_{\textit{stuck}})}{\textit{Vol}\left(\mathcal{B}^d_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})\right)} \leq \frac{\frac{\delta_2\mathcal{R}}{\sqrt{d}}\textit{Vol}(\mathcal{B}^{d-1}_{\hat{\mathbf{x}}^{-1}}(\mathcal{R}))}{\textit{Vol}\left(\mathcal{B}^d_{\hat{\mathbf{x}}^{-1}}(\mathcal{R})\right)} = \frac{\delta_2}{\sqrt{\pi d}}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{\delta_2}{\sqrt{\pi d}}\sqrt{\frac{d}{2}+\frac{1}{2}} \leq \delta_2 \tag{194}$$

The second to last inequality is by the property of Gamma function that $\frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})} < \sqrt{x+\frac{1}{2}}$ as long as $x \geq 0$. Therefore, with at least probability $1-\delta_2, \hat{\mathbf{x}}^0 \notin \mathcal{X}_{\textit{stuck}}$ i.e.

$$f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) \leq -3\mathcal{F} \tag{195}$$

In this case we have:

$$\begin{aligned}
f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^{-1}) &= f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) + f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) && (196)\\
&\leq -3\mathcal{F} + \frac{3}{2}\mathcal{F} && (197)\\
&\leq -\mathcal{F} && (198)
\end{aligned}$$

□

Lemma 17. *Assume the major condition* (86) *and conditions* (54)*,* (55) *and* (58) *hold; let* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$ *and* $\epsilon_2^2 \leq \min\{\frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}, \mathcal{F}\frac{L_1^2}{2+2L_1^2}\}$*. Assume* $\mathbf{x}^{-1}$ *such that* $\lambda_{\min}(\nabla^2 f(\hat{\mathbf{x}}^{-1})) \leq -\gamma$ *and further* $\forall i$ *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$*. Consider that the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44) *and further* $\eta_2 \leq \min\{1, \frac{1-\sigma}{L_1}\}$ *and* $\alpha\eta_2 \leq \min\left\{\left(\frac{\delta_2(\sqrt{2}-1)}{8\sqrt{2}\hat{c}\cdot c_{new}}\right)^2 \frac{(1-\sigma)|\lambda_{\min}(H)|^2}{d\log^2\left(\frac{dL_1}{\gamma\delta_2}\right)}, \frac{1-\sigma}{192L_1^2}\right\}$*. Then with probability at least* $1-\delta_2$ *we have the following*

$$H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq -\mathcal{F} \tag{199}$$

Proof. Recall that

$$H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) := \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^r) + \frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + \frac{\alpha}{m}\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\|^2 \tag{200}$$

$$\begin{aligned}
&H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \\
&= f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) + \frac{1}{m}\|\underline{\mathbf{x}}^0 - \hat{\underline{\mathbf{x}}}^0\|^2 - \frac{1}{m}\|\underline{\mathbf{x}}^{-1} - \hat{\underline{\mathbf{x}}}^{-1}\|^2 + \frac{\alpha}{m}\|\underline{\mathbf{y}}^0 - \hat{\underline{\mathbf{y}}}^0\|^2 - \frac{\alpha}{m}\|\underline{\mathbf{y}}^{-1} - \hat{\underline{\mathbf{y}}}^{-1}\|^2 \\
&\leq f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) + \frac{\alpha}{m}\|\underline{\mathbf{y}}^0 - \hat{\underline{\mathbf{y}}}^0\|^2 \\
&\leq \frac{3}{2}\mathcal{F} + \alpha\eta_2\frac{L_1^2}{2(1-\sigma)}\mathcal{F} && (201) \\
&\leq \frac{3}{2}\mathcal{F} + \frac{1}{4}\mathcal{F} && (202) \\
&\leq \frac{7}{4}\mathcal{F}, && (203)
\end{aligned}$$

where the first inequality comes from the fact that the same noise is injected to all local iterates and thus

$$\frac{1}{m}\|\underline{\mathbf{x}}^0 - \hat{\underline{\mathbf{x}}}^0\|^2 - \frac{1}{m}\|\underline{\mathbf{x}}^{-1} - \hat{\underline{\mathbf{x}}}^{-1}\|^2 = 0 \tag{204}$$

for the second inequality we use that $f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^{-1}) \leq \frac{3}{2}\mathcal{F}$ due to Lemma 10 and the bound from (58). □

Further, we have for the same $T < \hat{c}\mathcal{J}$ from Lemma 16

$$\begin{aligned}
H(\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T) - H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) &= \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^T) - \frac{1}{m}\sum_{i=1}^m f_i(\hat{\mathbf{x}}^0) + \frac{1}{m}\|\underline{\mathbf{x}}^T - \hat{\underline{\mathbf{x}}}^T\|^2 - \frac{1}{m}\|\underline{\mathbf{x}}^0 - \hat{\underline{\mathbf{x}}}^0\|^2 \\
&\quad + \frac{\alpha}{m}\|\underline{\mathbf{y}}^T - \hat{\underline{\mathbf{y}}}^T\|^2 - \frac{\alpha}{m}\|\underline{\mathbf{y}}^0 - \hat{\underline{\mathbf{y}}}^0\|^2 \\
&= f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) + P(\underline{\mathbf{x}}^T) - P(\underline{\mathbf{x}}^0) \\
&\leq f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) + \alpha\eta_2^2 4L_1^2(1 + \frac{1}{\beta_2})\sum_{t=0}^T \|\hat{\mathbf{y}}^t\|^2 \\
&\leq f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) + \alpha\eta_2^2 4L_1^2\frac{1}{1-\sigma}12\frac{\mathcal{F}}{\eta_2} \\
&= f(\hat{\mathbf{x}}^T) - f(\hat{\mathbf{x}}^0) + \frac{1}{1-\sigma}48L_1^2\alpha\eta_2\mathcal{F} \\
&\leq -3\mathcal{F} + \frac{1}{1-\sigma}48L_1^2\alpha\eta_2\mathcal{F} \\
&\leq -\frac{11}{4}\mathcal{F} && (205)
\end{aligned}$$

where the first inequality comes from Lemma 5, the second by condition (86) and the third by Lemma 16. Finally we get

$$H(\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) = H(\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T) - H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) + H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq \frac{7}{4}\mathcal{F} - \frac{11}{4}\mathcal{F} \leq -\mathcal{F} \tag{206}$$

Since the potential function is non increasing $H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) \leq H(\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T)$ and the result follows.

Combining Lemma 12 and Lemma 17 we derive the following corollary stating that during the second phase, the Gradient Tracking sequence is going to escape with high probability from an initial point of sufficient negative curvature.

Corollary 5. *Assume conditions* (54)*,* (55) *and* (58) *hold; let* $\epsilon_1^2 \leq \mathcal{F}\frac{L_1}{2+2L_1^2}$ *and* $\epsilon_2^2 \leq \min\{\frac{L_1^2}{2(1-\sigma)}\alpha\eta_2\mathcal{F}, \mathcal{F}\frac{L_1}{2+2L_1^2}\}$*. Assume* $\mathbf{x}^{-1}$ *such that* $\lambda_{\min}(\nabla^2 f(\hat{\mathbf{x}}^{-1})) \leq -\gamma$ *and further* $\forall i$ *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$*. Consider that the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44) *and further* $\eta_2 \leq \min\{1, \frac{1-\sigma}{L_1}\}$ *and* $\alpha\eta_2 \leq \min\left\{\left(\frac{\delta_2(\sqrt{2}-1)}{8\sqrt{2}\hat{c}\cdot c_{new}}\right)^2 \frac{(1-\sigma)|\lambda_{\min}(H)|^2}{d\log^2\left(\frac{dL_1}{\gamma\delta_2}\right)}, \frac{1-\sigma}{192L_1^2}\right\}$*. Then with probability at least* $1-\delta_2$ *we have the following*

$$H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) \leq -\mathcal{F} \tag{207}$$

Finally notice that as shown is Theorem 7, we can track whether the second phase succeeded in substantially decreasing the potential function by two runs of the average consensus protocol. One on iterate $\underline{\mathbf{x}}^{-1}$ and one on $\underline{\mathbf{x}}^{T_{cap}}$. If there is substantial decrease then $\underline{\mathbf{x}}^{T_{cap}}$ and $\underline{\mathbf{y}}^{T_{cap}}$ are provided as a starting point for the first phase. If the decrease is not substantial then with probability $1-\delta_2$ the point $\hat{\underline{\mathbf{x}}}^{-1}$ is a $(\epsilon_1 + L_1\epsilon_2, \gamma)-$approximate second order stationary point and we terminate.

10.1 Convergence Rates

Theorem 6. *Let* ϵ, ρ *be the target gradient and consensus error accuracy. Assume condition 58 and let* $\hat{\epsilon} = \min\left\{\epsilon, \rho, \frac{L_1}{\sqrt{2(1-\sigma)}}\sqrt{\alpha\eta_2\mathcal{F}}, \sqrt{\mathcal{F}}\frac{L_1}{\sqrt{2+2L_1^2}}\right\}$ *and assume that in the first phase the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_1 *such that* η_1, α *satisfy conditions* (41) - (44)*. Let* $T_1 = 4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1, 1-\sigma\}\hat{\epsilon}^2} + 1$*. Let* $\underline{\mathbf{x}}^{-1}$ *the point the first phase outputs and assume* $\lambda_{\min}(\nabla^2 f(\hat{\mathbf{x}}^{-1})) \leq -\gamma$*. Further* $\forall i$ *let* $\mathbf{x}_i^0 = \mathbf{x}_i^{-1} + \xi$ *where* ξ *comes from the uniform distribution over the ball of radius* $R = \sqrt{\frac{\mathcal{F}}{L_1}}$*. Consider that the iterates* $\mathbf{x}_i$ *follow the Gradient Tracking Update with stepsize* η_2 *such that* η_2, α *satisfy conditions* (41) - (44) *and further* $\eta_2 \leq \min\{1, \frac{1-\sigma}{L_1}\}$ *and* $\alpha\eta_2 \leq \min\left\{\left(\frac{\delta_2(\sqrt{2}-1)}{8\sqrt{2}\hat{c}\cdot c_{new}}\right)^2 \frac{(1-\sigma)|\lambda_{\min}(H)|^2}{d\log^2\left(\frac{dL_1}{\gamma\delta_2}\right)}, \frac{1-\sigma}{192L_1^2}\right\}$*. Then with probability at least* $(1-\delta_1)(1-\delta_2)$ *we have the*

$$H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{0}, \underline{\mathbf{y}}^{0}) \leq -\mathcal{F} \tag{208}$$

where $\underline{\mathbf{x}}^0$ *is the first iterate of the first phase and* $\underline{\mathbf{x}}^{T_{cap}}$ *is the last iterate of the second phase. Further let us denote the average consensus iterations for phase I and II with* T_{con} *and the total number of communication rounds throughout both phases with* $T_{1,2}$*. Then it holds:*

$$T_{1,2} = 4e\frac{4(f(\mathbf{x}^0)-f^*)}{\min\{\eta_1, 1-\sigma\}\hat{\epsilon}^2} + \hat{c}\frac{\log(d\kappa/\delta)}{\eta_2|\lambda_{\min}(H)|} + T_{con} = \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{\eta_1\hat{\epsilon}^2}, \frac{1}{\eta_2\gamma}\right\}\right) \tag{209}$$

Proof. From Theorem 5 we have $\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_i^{-1})\right\|^2 + \frac{1}{m}\left\|\underline{\mathbf{x}}^{-1} - \hat{\underline{\mathbf{x}}}^{-1}\right\|^2 \leq \hat{\epsilon}^2$ with probability $1-\delta_1$. Since the conditions of Corollary 5 hold we also have that with probability at least $1-\delta_2$

$$H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{0}, \underline{\mathbf{y}}^{0}) \leq H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) + H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) - H(\underline{\mathbf{x}}^{0}, \underline{\mathbf{y}}^{0}) \tag{210}$$

$$\leq -\mathcal{F} + H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) - H(\underline{\mathbf{x}}^{0}, \underline{\mathbf{y}}^{0}) \tag{211}$$

$$\leq -\mathcal{F} \tag{212}$$

where the last inequality comes from the monotonicity of the potential function through the first phase. The total number of communication rounds include the first phase iterations and consensus rounds as well as the second phase iterations and consensus rounds. □

Lemma 18. *Let ϵ, ρ be the target gradient and consensus error accuracy. Further let $\hat{\epsilon} = \min\left\{\epsilon, \rho, \frac{L_1}{\sqrt{2(1-\sigma)}}\sqrt{\alpha\eta_2\mathcal{F}}, \sqrt{\mathcal{F}}\frac{L_1}{\sqrt{2+2L_1^2}}\right\}$. There exist $\alpha = \mathcal{O}\left((1-\sigma)^2\right)$, $\eta_1 = \mathcal{O}\left((1-\sigma)^2\right)$ and $\eta_2 = \tilde{O}\left(\frac{\gamma^2}{d(1-\sigma)}\right)$ such that the conditions of Theorem 6 hold and the communication rounds throughout the first and the second phases is*

$$T_{1,2} = \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{(1-\sigma)^2\hat{\epsilon}^2}, \frac{(1-\sigma)d}{\gamma^3}\right\}\right) \tag{213}$$

Proof. From theorem 6 we have

$$T_{1,2} = \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{\eta_1\hat{\epsilon}^2}, \frac{1}{\eta_2\gamma}\right\}\right) = \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{(1-\sigma)^2\hat{\epsilon}^2}, \frac{d}{\gamma^3}\right\}\right) \tag{214}$$

□

Recall that after each pass of phase II the potential function is decreased at least by $\tilde{\mathcal{O}}(\gamma^3)$ and thus the following corollary captures the overall communication complexity of our algorithm before it reaches some approximate second order stationary point.

Corollary 6. *Assume the conditions of Lemma 18 hold. Then the overall communication rounds performed by our algorithm is at most*

$$T_{total} = \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{(1-\sigma)^2\gamma^3\hat{\epsilon}^2}, \frac{d}{\gamma^6}\right\}\right) \tag{215}$$

If we further assume that the strict saddle property holds then by setting $\hat{\epsilon} \leq \frac{\theta}{1+L_1}$ our algorithm converges to local minima. This claim is an immediate corollary of the following lemma.

Lemma 19. *Assume that conditions* (54)*,* (55) *hold and further $\epsilon_1 + L_1\epsilon_2 < \theta$. Then either $\hat{\mathbf{x}}^{-1}$ is $\nu-$close to some local minimum or $\lambda_{\min}(\nabla^2 f(\hat{\mathbf{x}}^{-1})) \leq -\zeta$.*

Proof. We can bound $\left\|\nabla f(\hat{\mathbf{x}}^{-1})\right\|$ as follows :

$$\begin{aligned}
\left\|\nabla f(\hat{\mathbf{x}}^{-1})\right\| &\leq \left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{-1}) - \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\hat{\mathbf{x}}^{-1})\right\| + \left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^{-1})\right\| && (216)\\
&\leq \frac{L_1}{m}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{-1} - \hat{\mathbf{x}}^{-1}\right\| + \epsilon_1 && (217)\\
&\leq \frac{L_1}{m}\sqrt{m\sum_{i=1}^{m}\left\|\mathbf{x}_i^{-1} - \hat{\mathbf{x}}^{-1}\right\|^2} + \epsilon_1 && (218)\\
&\leq L_1\sqrt{\frac{1}{m}\left\|\underline{\mathbf{x}}^{-1} - \underline{\hat{\mathbf{x}}}^{-1}\right\|^2} + \epsilon_1 && (219)\\
&\leq L_1\epsilon_2 + \epsilon_1 && (220)\\
&< \theta && (221)
\end{aligned}$$

Where the second inequality comes from (54) and the last in equality comes from (55). The result follows from Assumption 4 . □

11 Average Consensus Protocol

We will now present how to utilize the average consensus protocol to achieve the following objectives:

1. Initialize $\underline{\mathbf{y}}^0$ at the beginning of phase II such that
   $\hat{\mathbf{y}}^0 = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_i^0)$ *and* $\frac{1}{m}\left\|\underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0\right\|^2 \leq \frac{L_1^2}{2(1-\sigma)}\eta_2\mathcal{F}$.

2. Coordinate the nodes to pick a phase I iteration r such that $\left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\mathbf{x}_i^r) \right\|^2 + \frac{1}{m} \left\| \underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r \right\|^2 \leq \epsilon^2$.
3. Track the potential function decrease before and at the end of phase II, $H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1})$.

Average Consensus Update Rule for some vector $\underline{\mathbf{x}}^r$

$$\underline{\mathbf{x}}^{r,0} = \underline{\mathbf{x}}^r \tag{222}$$
$$\underline{\mathbf{x}}^{r,t+1} = \underline{\mathbf{W}}\underline{\mathbf{x}}^{r,t} \tag{223}$$

11.1 Initializing the second phase

Towards proving the first of our objectives we present the following lemma where we show that the consensus error diminishes exponentially fast in the number of iterations. Notice that since $\eta_2 \mathcal{F} = \tilde{\mathcal{O}}\left(\frac{\gamma^5}{d}\right)$ the number of iterations of the average consensus protocol have a logarithmic dependence on γ, d and the initial error $\sum_{i=1}^{m} \left\| \nabla f_i(\mathbf{x}_i^0) - \sum_{j=1}^{m} \nabla f_j(\mathbf{x}_j^0) \right\|^2$.

Lemma 20. *Consider the iterates $\mathbf{x}_i^0$'s at the beginning of phase II and let each node, i, set $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0)$. Let $\underline{\mathbf{y}}^{0,0} = \underline{\mathbf{y}}^0$, $\underline{\mathbf{y}}^{0,t+1} = \underline{\mathbf{W}}\underline{\mathbf{y}}^{0,t}$ and $\hat{\mathbf{y}}^0 = \frac{1}{m}\sum_{i=1}^{m} \mathbf{y}_i^0 = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{x}_i^0)$. After $t_{\mathbf{y}} + 1$ rounds of the average consensus protocol on $\mathbf{y}_i^0$'s we have the following guarantee*

$$\frac{1}{m} \left\| \underline{\mathbf{y}}^{0,t+1} - \underline{\hat{\mathbf{y}}}^0 \right\|^2 \leq \frac{L_1^2}{2(1-\sigma)} \eta_2 \mathcal{F} \tag{224}$$

for $t_{\mathbf{y}} = \left\lfloor \frac{\log\left(\sqrt{\eta_2 \mathcal{F}}\right) - \log\left(\left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\|\right) + \log\left(\sqrt{m} L_1\right) - \log\left(\sqrt{2(1-\sigma)}\right)}{\log(\sigma)} \right\rfloor$.

Proof. From the consensus update rule we know

$$\left\| \underline{\mathbf{y}}^{0,t+1} - \underline{\hat{\mathbf{y}}}^0 \right\| \leq \left\| \underline{\mathbf{W}} \left(\underline{\mathbf{y}}^{0,t} - \underline{\hat{\mathbf{y}}}^0 \right) \right\| \leq \sigma \left\| \underline{\mathbf{y}}^{0,t} - \underline{\hat{\mathbf{y}}}^0 \right\| \tag{225}$$

Thus we derive the following

$$\left\| \underline{\mathbf{y}}^{0,t+1} - \underline{\hat{\mathbf{y}}}^0 \right\| \leq \sigma^{t+1} \left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\| \tag{226}$$

Solving for t that guarantees $\sigma^{t+1} \left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\| \leq \frac{L_1\sqrt{m}}{\sqrt{2(1-\sigma)}} \sqrt{\eta_2 \mathcal{F}}$ we get

$$\sigma^{t+1} \leq \frac{L_1\sqrt{m}}{\left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\| \sqrt{2(1-\sigma)}} \sqrt{\eta_2 \mathcal{F}} \tag{227}$$

$$(t+1)\log(\sigma) \leq \log\left(\sqrt{m} L_1\right) + \log\left(\sqrt{\eta_2 \mathcal{F}}\right) - \log\left(\left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\|\right) - \log\left(\sqrt{2(1-\sigma)}\right) \tag{228}$$

$$t \geq \frac{\log\left(\sqrt{m} L_1\right) + \log\left(\sqrt{\eta_2 \mathcal{F}}\right) - \log\left(\left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\|\right) - \log\left(\sqrt{2(1-\sigma)}\right)}{\log(\sigma)} - 1 \tag{229}$$

Thus for $t = \lfloor \frac{\log\left(\sqrt{\eta_2 \mathcal{F}}\right) - \log\left(\left\| \underline{\mathbf{y}}^0 - \underline{\hat{\mathbf{y}}}^0 \right\|\right) + \log\left(\sqrt{m} L_1\right) - \log\left(\sqrt{2(1-\sigma)}\right)}{\log(\sigma)} \rfloor$ we have

$$\left\| \underline{\mathbf{y}}^{0,t+1} - \underline{\hat{\mathbf{y}}}^0 \right\| \leq \frac{L_1\sqrt{m}}{\sqrt{2(1-\sigma)}} \sqrt{\eta_2 \mathcal{F}} \tag{230}$$

which implies

$$\frac{1}{m} \left\| \underline{\mathbf{y}}^{0,t+1} - \underline{\hat{\mathbf{y}}}^0 \right\|^2 \leq \frac{L_1^2}{2(1-\sigma)} \eta_2 \mathcal{F} \tag{231}$$

□

Initializing $\mathbf{y}_i^0 = \mathbf{y}_i^{0,t_{\mathbf{y}+1}}$ we achieve our first objective.

11.2 Choosing a good iterate

In order to achieve our second objective first we provide upper bounds for $\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|$ and $\left\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\right\|$ for any iteration r of our lagorithm.

Lemma 21. *Consider any iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Then*

$$\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\| \leq \sqrt{m(f(\underline{\mathbf{x}}^0) - \underline{f}) + 2m\mathcal{F}} \tag{232}$$

$$\left\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\right\| \leq \sqrt{\frac{m}{\alpha}(f(\underline{\mathbf{x}}^0) - \underline{f}) + \frac{2m}{\alpha}\mathcal{F}} \tag{233}$$

Proof.

$$H(\underline{\mathbf{x}}^0, \underline{\mathbf{y}}^0) - H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) = f(\hat{\mathbf{x}}^0) + 0 + 0 - f(\hat{\mathbf{x}}^r) - \frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 - \frac{\alpha}{m}\left\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\right\|^2 \geq -2\mathcal{F} \tag{234}$$

where the last inequality holds because the potential function is non-increasing throughout a single phase and due to Lemma 11. Thus we get

$$\frac{1}{m}\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 + \frac{\alpha}{m}\left\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\right\|^2 \leq f(\hat{\mathbf{x}}^0) - f(\hat{\mathbf{x}}^r) + 2\mathcal{F} \leq f(\hat{\mathbf{x}}^0) - f^* + 2\mathcal{F} \tag{235}$$

which derives

$$\|\underline{\mathbf{x}}^r - \hat{\underline{\mathbf{x}}}^r\|^2 \leq m(f(\hat{\mathbf{x}}^0) - f^*) + 2m\mathcal{F} \tag{236}$$

$$\left\|\underline{\mathbf{y}}^r - \hat{\underline{\mathbf{y}}}^r\right\|^2 \leq \frac{m}{\alpha}(f(\hat{\mathbf{x}}^0) - f^*) + \frac{2m}{\alpha}\mathcal{F} \tag{237}$$

The result follows after taking the square roots of both bounds. □

In the following lemma we are going to show that the consensus error diminishes exponentially fast in the number of iterations.

Lemma 22. *Consider any iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Let $\underline{\mathbf{y}}^{r,0} = \underline{\mathbf{y}}^r, \quad \underline{\mathbf{y}}^{r,t+1} = \underline{\mathbf{W}}\underline{\mathbf{y}}^{r,t}$ and $\hat{\mathbf{y}}^r = \frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i^r$. After $t_{\mathbf{y}} + 1$ rounds of the average consensus protocol on $\mathbf{y}_i^r$'s we have the following guarantee*

$$\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \hat{\underline{\mathbf{y}}}^r\right\| \leq \epsilon^c \tag{238}$$

for $t_{\mathbf{y}} = \left\lfloor \frac{c\log(\epsilon) + \log(\sqrt{\alpha}) - \log\left(\sqrt{m(f(\mathbf{x}^0) - f^*) + 2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$ and any positive constant c.

Similarly let $\underline{\mathbf{x}}^{r,0} = \underline{\mathbf{x}}^r, \quad \underline{\mathbf{x}}^{r,t+1} = \underline{\mathbf{W}}\underline{\mathbf{x}}^{r,t}$ and $\hat{\mathbf{x}}^r = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^r$. After $t_{\mathbf{x}} + 1$ rounds of the average consensus protocol on on $\mathbf{x}_i^r$'s we have the following guarantee

$$\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \hat{\underline{\mathbf{x}}}^r\right\| \leq \epsilon^c \tag{239}$$

for $t_{\mathbf{x}} = \left\lfloor \frac{c\log(\epsilon) - \log\left(\sqrt{m(f(\mathbf{x}^0) - f^*) + 2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$ and any positive constant c.

Proof. From the consensus update rule we know

$$\left\|\underline{\mathbf{y}}^{r,t+1} - \hat{\underline{\mathbf{y}}}^r\right\| \leq \left\|\underline{\mathbf{W}}\left(\underline{\mathbf{y}}^{r,t} - \hat{\underline{\mathbf{y}}}^r\right)\right\| \leq \sigma\left\|\underline{\mathbf{y}}^{r,t} - \hat{\underline{\mathbf{y}}}^r\right\| \tag{240}$$

Thus we derive the following

$$\left\|\underline{\mathbf{y}}^{r,t+1} - \hat{\underline{\mathbf{y}}}^r\right\| \leq \sigma^{t+1}\left\|\underline{\mathbf{y}}^{r,0} - \hat{\underline{\mathbf{y}}}^r\right\| \leq \sigma^{t+1}\sqrt{\frac{m}{\alpha}(f(\mathbf{x}^0) - f^*) + \frac{2m}{\alpha}\mathcal{F}} \tag{241}$$

where the second inequality comes from Lemma 21. Solving for t that guarantees $\sigma^{t+1}\sqrt{\frac{m}{\alpha}(f(\mathbf{x}^0)-f^*)+\frac{2m}{\alpha}\mathcal{F}} \leq \epsilon^c$ we get

$$\sigma^{t+1} \leq \frac{\sqrt{\alpha}\epsilon^c}{\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}} \tag{242}$$

$$(t+1)\log(\sigma) \leq \log(\sqrt{\alpha}\epsilon^c) - \log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right) \tag{243}$$

$$t \geq \frac{c\log(\epsilon)+\log(\sqrt{\alpha})-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} - 1 \tag{244}$$

Thus for $t = \left\lfloor \frac{c\log(\epsilon)+\log(\sqrt{\alpha})-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$ we have

$$\left\|\underline{\mathbf{y}}^{r,t+1} - \underline{\hat{\mathbf{y}}}^r\right\| \leq \epsilon^c \tag{245}$$

Similarly from the consensus update rule we know

$$\left\|\underline{\mathbf{x}}^{r,t+1} - \underline{\hat{\mathbf{x}}}^r\right\| \leq \left\|\underline{\mathbf{W}}\left(\underline{\mathbf{x}}^{r,t} - \underline{\hat{\mathbf{x}}}^r\right)\right\| \leq \sigma\left\|\underline{\mathbf{x}}^{r,t} - \underline{\hat{\mathbf{x}}}^r\right\| \tag{246}$$

Thus we derive the following

$$\left\|\underline{\mathbf{x}}^{r,t+1} - \underline{\hat{\mathbf{x}}}^r\right\| \leq \sigma^{t+1}\left\|\underline{\mathbf{x}}^{r,0} - \underline{\hat{\mathbf{x}}}^r\right\| \leq \sigma^{t+1}\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}} \tag{247}$$

where the second inequality comes from Lemma 21. Solving for t that guarantees $\sigma^{t+1}\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}} \leq \epsilon^c$ we get

$$\sigma^{t+1} \leq \frac{\epsilon^c}{\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}} \tag{248}$$

$$(t+1)\log(\sigma) \leq c\log(\epsilon) - \log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right) \tag{249}$$

$$t \geq \frac{c\log(\epsilon)-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} - 1 \tag{250}$$

Thus for $t = \left\lfloor \frac{c\log(\epsilon)-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$ we have

$$\left\|\underline{\mathbf{x}}^{r,t+1} - \underline{\hat{\mathbf{x}}}^r\right\| \leq \epsilon^c \tag{251}$$

□

The following corollary suggest that after a logarithmic number of iterations with respect to ϵ, every node is going to have an accurate estimate of the average vector of interest.

Corollary 7. *After* $\left\lfloor \frac{c\log(\epsilon)+\log(\sqrt{\alpha})-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor + 1$ *rounds of the average consensus protocol on* $\mathbf{y}_i^r$*'s we have the following guarantee*

$$\left\|\mathbf{y}_i^{r,t} - \hat{\mathbf{y}}^r\right\| \leq \epsilon^c, \quad \forall i \in [m] \tag{252}$$

After $\left\lfloor \frac{c\log(\epsilon)-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor + 1$ *rounds of the average consensus protocol on* $\mathbf{x}_i^r$*'s we have the following guarantee*

$$\left\|\mathbf{x}_i^{r,t} - \hat{\mathbf{x}}^r\right\| \leq \epsilon^c, \quad \forall i \in [m] \tag{253}$$

The next lemma provides bounds that we will use in order to argue about the number of iterations required when we run the average consensus protocol on $\left\|\mathbf{y}_i^{r,t_\mathbf{y}+1} - \mathbf{y}_i^r\right\|^2$'s and on $\left\|\mathbf{x}_i^{r,t_\mathbf{x}+1} - \mathbf{x}_i^r\right\|^2$'s.

Lemma 23. *Consider any iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Let $\underline{\mathbf{y}}^{r,0} = \underline{\mathbf{y}}^r$ and $\hat{\mathbf{y}}^r = \frac{1}{m}\sum_{i=1}^m \mathbf{y}_i^r$, $\underline{\mathbf{x}}^{r,0} = \underline{\mathbf{x}}^r$ and $\hat{\mathbf{x}}^r = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^r$. Further let $t_{\mathbf{y}} = \left\lfloor \frac{c\log(\epsilon)+\log(\sqrt{\alpha})-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$ and $t_{\mathbf{x}} = \left\lfloor \frac{c\log(\epsilon)-\log\left(\sqrt{m(f(\mathbf{x}^0)-f^*)+2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$. The following bounds hold for $\epsilon \leq \min\{1, \sqrt{\frac{m}{\alpha 16}(f(\mathbf{x}^0)-f^*+2\mathcal{F})}\}$ and $c \geq 1$*

$$\sum_{i=1}^m \left(\left\| \mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r \right\|^2 - \frac{1}{m}\sum_{j=1}^m \left\| \mathbf{y}_j^{r,t_{\mathbf{y}}+1} - \mathbf{y}_j^r \right\|^2 \right)^2 \leq \frac{16m^3}{\alpha^2}(f(\mathbf{x}^0)-f^*+\mathcal{F})^2 \tag{254}$$

$$\sum_{i=1}^m \left(\left\| \mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r \right\|^2 - \frac{1}{m}\sum_{j=1}^m \left\| \mathbf{x}_j^{r,t_{\mathbf{x}}+1} - \mathbf{x}_j^r \right\|^2 \right)^2 \leq 16m^3(f(\mathbf{x}^0)-f^*+\mathcal{F})^2 \tag{255}$$

Proof. First notice that

$$\left\| \mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r \right\|^2 \leq \left\| \mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \hat{\mathbf{y}}^r \right\|^2 + \|\hat{\mathbf{y}}^r - \mathbf{y}_i^r\|^2 + 2\langle \left\| \mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \hat{\mathbf{y}}^r \right\|, \|\hat{\mathbf{y}}^r - \mathbf{y}_i^r\| \rangle \tag{256}$$

$$\leq \epsilon^{2c} + \frac{m}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F}) + 2\epsilon^c\sqrt{\frac{m}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F})} \tag{257}$$

$$\leq \frac{2m}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F}) \tag{258}$$

In the second inequality we use the results from Lemma 21 and Corollary 7. The third inequality holds for sufficiently small $\epsilon \leq \min\{1, \sqrt{\frac{m}{16\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F})}\}$. Thus we have

$$\sum_{i=1}^m \left(\left\| \mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r \right\|^2 - \frac{1}{m}\sum_{j=1}^m \left\| \mathbf{y}_j^{r,t_{\mathbf{y}}+1} - \mathbf{y}_j^r \right\|^2 \right)^2$$

$$\leq \sum_{i=1}^m \left(\left\| \mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r \right\|^2 + \frac{1}{m}\sum_{j=1}^m \left\| \mathbf{y}_j^{r,t_{\mathbf{y}}+1} - \mathbf{y}_j^r \right\|^2 \right)^2 \tag{259}$$

$$\leq \sum_{i=1}^m \left(\frac{2m}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F}) + \frac{1}{m}\sum_{j=1}^m \frac{2m}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F})^2 \right) \tag{260}$$

$$\leq \sum_{i=1}^m \frac{16m^2}{\alpha^2}(f(\mathbf{x}^0)-f^*+2\mathcal{F})^2 \tag{261}$$

$$\leq \frac{16m^3}{\alpha^2}(f(\mathbf{x}^0)-f^*+2\mathcal{F})^2 \tag{262}$$

Notice that

$$\left\| \mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r \right\|^2 \leq \left\| \mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \hat{\mathbf{x}}^r \right\|^2 + \|\hat{\mathbf{x}}^r - \mathbf{x}_i^r\|^2 + 2\langle \left\| \mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \hat{\mathbf{x}}^r \right\|, \|\hat{\mathbf{x}}^r - \mathbf{x}_i^r\| \rangle \tag{263}$$

$$\leq \epsilon^{2c} + m(f(\mathbf{x}^0)-f^*+2\mathcal{F}) + 2\epsilon^c\sqrt{m(f(\mathbf{x}^0)-f^*+2\mathcal{F})} \tag{264}$$

$$\leq 2m(f(\mathbf{x}^0)-f^*+2\mathcal{F}) \tag{265}$$

In the second inequality we use the results from Lemma 21 and Corollary 7. The third inequality holds for sufficiently small $\epsilon \leq \min\{1, \sqrt{\frac{m}{16}(f(\underline{\mathbf{x}}^0) - \underline{f})}\}$. Thus we have

$$\sum_{i=1}^{m}\left(\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2 - \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{x}_j^{r,t_{\mathbf{x}}+1} - \mathbf{x}_j^r\right\|^2\right)^2 \tag{266}$$

$$\leq \sum_{i=1}^{m}\left(\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2 + \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{x}_j^{r,t_{\mathbf{x}}+1} - \mathbf{x}_j^r\right\|^2\right)^2 \tag{267}$$

$$\leq \sum_{i=1}^{m}\left(2m(f(\mathbf{x}^0) - f^* + 2\mathcal{F}) + \frac{1}{m}\sum_{j=1}^{m}2m(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)^2 \tag{268}$$

$$\leq \sum_{i=1}^{m}16m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})^2 \tag{269}$$

$$\leq 16m^3(f(\mathbf{x}^0) - f^* + 2\mathcal{F})^2 \tag{270}$$

□

The next lemma provides an upper bound on the number of iterations required when we run the average consensus protocol on $\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r\right\|^2$'s and on $\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2$'s in order to achieve accuracy ϵ^c.

Lemma 24. *Consider any iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44) *and $\epsilon \leq \min\{1, \sqrt{\frac{m}{16\alpha}(f(\mathbf{x}^0) - f^* + 2\mathcal{F})}\}$, $c \geq 1$. Also assume that the potential function decreases between consecutive first phases. Let $\underline{\mathbf{y}}^{r,0} = \underline{\mathbf{y}}^r$ and $\hat{\mathbf{y}}^r = \frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i^r$, $\underline{\mathbf{x}}^{r,0} = \underline{\mathbf{x}}^r$ and $\hat{\mathbf{x}}^r = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^r$. Further let $t_{\mathbf{y}} = \left\lfloor \frac{c\log(\epsilon) + \log(\sqrt{\alpha}) - \log\left(\sqrt{m(f(\mathbf{x}^0) - f^*) + 2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$ and $t_{\mathbf{x}} = \left\lfloor \frac{c\log(\epsilon) - \log\left(\sqrt{m(f(\mathbf{x}^0) - f^*) + 2m\mathcal{F}}\right)}{\log(\sigma)} \right\rfloor$.*

Define $\mathbf{z}_i^{r,0} := \left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r\right\|^2$, $\hat{\mathbf{z}}^r := \frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r\right\|^2$

and $t_{\mathbf{z}} = \left\lfloor \frac{c\log(\epsilon) + \log\alpha - \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\sigma)} \right\rfloor$.

After $t_{\mathbf{z}} + 1$ rounds of the average consensus protocol on $\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1} - \mathbf{y}_i^r\right\|^2$'s we have the following guarantee

$$\left\|\underline{\mathbf{z}}^{r,t_{\mathbf{z}}+1} - \underline{\hat{\mathbf{z}}}^r\right\| \leq \epsilon^c \tag{271}$$

Define $\mathbf{w}_i^{r,0} := \left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2$, $\hat{\mathbf{w}}^r := \frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2$

and $t_{\mathbf{w}} = \left\lfloor \frac{c\log(\epsilon) - \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\sigma)} \right\rfloor$.

After $t_{\mathbf{w}} + 1$ rounds of the average consensus protocol on $\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2$'s we have the following guarantee

$$\left\|\underline{\mathbf{w}}^{r,t_{\mathbf{w}}+1} - \underline{\hat{\mathbf{w}}}^r\right\| \leq \epsilon^c \tag{272}$$

Proof. From the consensus update rule we know

$$\left\|\underline{\mathbf{z}}^{r,t+1} - \underline{\hat{\mathbf{z}}}^r\right\| \leq \left\|\mathbf{W}\left(\underline{\mathbf{z}}^{r,t} - \underline{\hat{\mathbf{z}}}^r\right)\right\| \leq \sigma\left\|\underline{\mathbf{z}}^{r,t} - \underline{\hat{\mathbf{z}}}^r\right\| \tag{273}$$

Thus we derive the following

$$\left\|\underline{\mathbf{z}}^{r,t+1}-\hat{\underline{\mathbf{z}}}^{r}\right\| \leq \sigma^{t+1}\left\|\underline{\mathbf{z}}^{r,0}-\hat{\underline{\mathbf{z}}}^{r}\right\| \tag{274}$$

$$\leq \sigma^{t+1}\sqrt{\sum_{i=1}^{m}\left(\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}-\mathbf{y}_i^{r}\right\|^2-\frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{y}_j^{r,t_{\mathbf{y}}+1}-\mathbf{y}_j^{r}\right\|^2\right)^2} \tag{275}$$

$$\leq \sigma^{t+1}\sqrt{\frac{16m^3}{\alpha^2}(f(\mathbf{x}^0)-f^*+2\mathcal{F})^2} \tag{276}$$

$$\leq \sigma^{t+1}\frac{4m^2}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F}) \tag{277}$$

where the third inequality comes from Lemma 23. Solving for t that guarantees $\sigma^{t+1}\frac{4m^2}{\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F})\leq\epsilon^c$ we get

$$\sigma^{t+1} \leq \frac{\alpha\epsilon^c}{4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})} \tag{278}$$

$$(t+1)\log(\sigma) \leq c\log(\epsilon)+\log\alpha-\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right) \tag{279}$$

$$t \geq \frac{c\log(\epsilon)+\log\alpha-\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\sigma)}-1 \tag{280}$$

Thus for $t=\lfloor\frac{c\log(\epsilon)+\log\alpha-\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\sigma)}\rfloor$ we have

$$\left\|\underline{\mathbf{z}}^{r,t+1}-\hat{\underline{\mathbf{z}}}^{r}\right\| \leq \epsilon^c \tag{281}$$

Similarly from the consensus update rule we know

$$\left\|\underline{\mathbf{w}}^{r,t+1}-\hat{\underline{\mathbf{w}}}^{r}\right\| \leq \left\|\mathbf{W}\left(\underline{\mathbf{w}}^{r,t}-\hat{\underline{\mathbf{w}}}^{r}\right)\right\| \leq \sigma\left\|\underline{\mathbf{w}}^{r,t}-\hat{\underline{\mathbf{w}}}^{r}\right\| \tag{282}$$

Thus we derive the following

$$\left\|\underline{\mathbf{w}}^{r,t+1}-\hat{\underline{\mathbf{w}}}^{r}\right\| \leq \sigma^{t+1}\left\|\underline{\mathbf{w}}^{r,0}-\hat{\underline{\mathbf{w}}}^{r}\right\| \tag{283}$$

$$\leq \sigma^{t+1}\sqrt{\sum_{i=1}^{m}\left(\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1}-\mathbf{x}_i^{r}\right\|^2-\frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{x}_j^{r,t_{\mathbf{x}}+1}-\mathbf{x}_j^{r}\right\|^2\right)^2} \tag{284}$$

$$\leq \sigma^{t+1}\sqrt{16m^3(f(\mathbf{x}^0)-f^*+2\mathcal{F})^2} \tag{285}$$

$$\leq \sigma^{t+1}4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F}) \tag{286}$$

where the third inequality comes from Lemma 23. Solving for t that guarantees $\sigma^{t+1}4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\leq\epsilon^c$ we get

$$\sigma^{t+1} \leq \frac{\epsilon^c}{4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})} \tag{287}$$

$$(t+1)\log(\sigma) \leq c\log(\epsilon)-\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right) \tag{288}$$

$$t \geq \frac{c\log(\epsilon)-\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\sigma)}-1 \tag{289}$$

Thus for $t=\left\lfloor\frac{c\log(\epsilon)-\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\sigma)}\right\rfloor$ we have

$$\left\|\underline{\mathbf{z}}^{r,t+1}-\hat{\underline{\mathbf{z}}}^{r}\right\| \leq \epsilon^c \tag{290}$$

□

Corollary 8. *Let* $\epsilon \leq \min\{1, \sqrt{\frac{m}{16\alpha}(f(\mathbf{x}^0) - f^* + 2\mathcal{F})}\}$, $c \geq 1$. *The total number of average consensus iterations to achieve sufficient accuracy captured in the following four bounds*

$$\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}} - \underline{\hat{\mathbf{y}}}^{r}\right\| \leq \epsilon^c \tag{291}$$

$$\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}} - \underline{\hat{\mathbf{x}}}^{r}\right\| \leq \epsilon^c \tag{292}$$

$$\left\|\underline{\mathbf{z}}^{r,t_{\mathbf{z}}} - \underline{\hat{\mathbf{z}}}^{r}\right\| \leq \epsilon^c \tag{293}$$

$$\left\|\underline{\mathbf{w}}^{r,t_{\mathbf{w}}} - \underline{\hat{\mathbf{w}}}^{r}\right\| \leq \epsilon^c \tag{294}$$

is at most

$$4\left(\frac{c\log(\frac{1}{\epsilon}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right) \tag{295}$$

Proof. The total number of iterations is at most

$$t_{\mathbf{y}} + t_{\mathbf{x}} + t_{\mathbf{z}} + t_{\mathbf{w}} + 4 \leq 4t_{\mathbf{z}} + 4 \tag{296}$$

$$\leq 4\left(\frac{c\log(\epsilon) + \log\alpha - \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\sigma)} + 1\right) \tag{297}$$

$$\leq 4\left(\frac{c\log(\frac{1}{\epsilon}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right) \tag{298}$$

□

The next lemma shows how far off is the square of the estimated difference $\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^{r}\right\|^2$ (and respectively $\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^{r}\right\|^2$) from the square of the true difference $\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2$(and respectively $\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\|^2$). Similar result for the square of the average gradient estimate is also provided.

Lemma 25. *Consider any iterates* $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ *following Gradient Tracking Update with* η_1, α *that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Let* $\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}$ *and* $\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}$ *as defined in Lemma 22. The following bounds hold:*

$$\frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 + \frac{1}{m}\epsilon^{2c} + \frac{2}{m}\epsilon^c\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\| \geq \frac{1}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^{r}\right\|^2 \geq \frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 - \frac{2}{m}\epsilon^c\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\| \tag{299}$$

$$\frac{1}{m}\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\|^2 + \frac{1}{m}\epsilon^{2c} + \frac{2}{m}\epsilon^c\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\| \geq \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^{r}\right\|^2 \geq \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\|^2 - \frac{2}{m}\epsilon^c\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\| \tag{300}$$

$$\forall i \quad \|\hat{\mathbf{y}}^r\|^2 + \epsilon^{2c} + 2\epsilon^c\|\hat{\mathbf{y}}^r\| \geq \left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 \geq \|\hat{\mathbf{y}}^r\|^2 - 2\epsilon^c\|\hat{\mathbf{y}}^r\| \tag{301}$$

Proof.

$$\frac{1}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^{r}\right\|^2 = \frac{1}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\hat{\mathbf{y}}}^{r} + \underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 \tag{302}$$

$$= \frac{1}{m}\left(\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\hat{\mathbf{y}}}^{r}\right\|^2 + \left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 + 2\langle\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\hat{\mathbf{y}}}^{r}, \underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\rangle\right) \tag{303}$$

and from here we can derive both the upper and the lower bound

$$\frac{1}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^{r}\right\|^2 \geq \frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 - \frac{2}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\hat{\mathbf{y}}}^{r}\right\|\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\| \tag{304}$$

$$\geq \frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 - \frac{2}{m}\epsilon^c\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\| \tag{305}$$

where the last inequality is due to Lemma 22. Also

$$\frac{1}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}-\underline{\mathbf{y}}^{r}\right\|^2 \leq \frac{1}{m}\left(\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}-\underline{\hat{\mathbf{y}}}^{r}\right\|^2+\left\|\underline{\hat{\mathbf{y}}}^{r}-\underline{\mathbf{y}}^{r}\right\|^2+2\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}-\underline{\hat{\mathbf{y}}}^{r}\right\|\left\|\underline{\hat{\mathbf{y}}}^{r}-\underline{\mathbf{y}}^{r}\right\|\right) \tag{306}$$

$$\leq \frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^{r}-\underline{\mathbf{y}}^{r}\right\|^2+\frac{1}{m}\epsilon^{2c}+\frac{2}{m}\epsilon^{c}\left\|\underline{\hat{\mathbf{y}}}^{r}-\underline{\mathbf{y}}^{r}\right\| \tag{307}$$

where again the second inequality comes from Lemma 22.
The proof deriving the bounds for $\frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}-\underline{\hat{\mathbf{x}}}^{r}\right\|^2$ is identical. For the third bound we work as follows

$$\left\|\mathbf{y}_i^{r,t+1}\right\|^2=\left\|\mathbf{y}_i^{r,t+1}-\hat{\mathbf{y}}^{r}+\hat{\mathbf{y}}^{r}\right\|^2=\left\|\mathbf{y}_i^{r,t+1}-\hat{\mathbf{y}}^{r}\right\|^2+\left\|\hat{\mathbf{y}}^{r}\right\|^2+2\langle\mathbf{y}_i^{r,t+1}-\hat{\mathbf{y}}^{r},\hat{\mathbf{y}}^{r}\rangle \tag{308}$$

and thus we derive the bounds

$$\left\|\mathbf{y}_i^{r,t+1}\right\|^2 \geq \left\|\hat{\mathbf{y}}^{r}\right\|^2-2\left\|\mathbf{y}_i^{r,t+1}-\hat{\mathbf{y}}^{r}\right\|\left\|\hat{\mathbf{y}}^{r}\right\| \tag{309}$$

$$\geq \left\|\hat{\mathbf{y}}^{r}\right\|^2-2\epsilon^{c}\left\|\hat{\mathbf{y}}^{r}\right\| \tag{310}$$

$$\tag{311}$$

the last inequality follows from Corollary 7. Also

$$\left\|\mathbf{y}_i^{r,t+1}\right\|^2 \leq \left\|\hat{\mathbf{y}}^{r}\right\|^2+\left\|\mathbf{y}_i^{r,t+1}-\hat{\mathbf{y}}^{r}\right\|^2+2\left\|\mathbf{y}_i^{r,t+1}-\hat{\mathbf{y}}^{r}\right\|\left\|\hat{\mathbf{y}}^{r}\right\| \tag{312}$$

$$\leq \left\|\hat{\mathbf{y}}^{r}\right\|^2+\epsilon^{2c}+2\epsilon^{c}\left\|\hat{\mathbf{y}}^{r}\right\| \tag{313}$$

□

The result derived in the following lemma is utilized as an intermediate step to towards proving second objective.

Lemma 26. *Consider any iterates $\underline{\mathbf{x}}^{r},\underline{\mathbf{y}}^{r}$ following Gradient Tracking Update with η_1,α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Let $\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}$ and $\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}$ as defined in Lemma 22. Also let $\epsilon \leq \min\{\frac{1}{8},\sqrt{\frac{m}{16\alpha}(f(\mathbf{x}^0)-f^*+2\mathcal{F})}\}$ and $c\geq 2$. Assume that*

$$\left\|\hat{\mathbf{y}}^{r}\right\|^2+\frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^{r}-\underline{\mathbf{x}}^{r}\right\|^2 \leq \frac{\epsilon^2}{4} \tag{314}$$

then

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}-\underline{\mathbf{x}}^{r}\right\|^2 \leq \frac{\epsilon^2}{3} \tag{315}$$

Further assume that

$$\left\|\hat{\mathbf{y}}^{r}\right\|^2+\frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^{r}-\underline{\mathbf{x}}^{r}\right\|^2 > \epsilon^2 \tag{316}$$

then

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}-\underline{\mathbf{x}}^{r}\right\|^2 > \frac{7}{8}\epsilon^2 \tag{317}$$

Proof. For the first part of the proof we utilize the upper bounds from Lemma 25:

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2+\frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}-\underline{\mathbf{x}}^{r}\right\|^2 \tag{318}$$

$$\leq \left\|\hat{\mathbf{y}}^{r}\right\|^2+\frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^{r}-\underline{\mathbf{x}}^{r}\right\|^2+\left(1+\frac{1}{m}\right)\epsilon^{2c}+2\epsilon^{c}\left(\left\|\hat{\mathbf{y}}^{r}\right\|^2+\frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^{r}-\underline{\mathbf{x}}^{r}\right\|^2\right) \tag{319}$$

$$\leq \frac{\epsilon^2}{4}+\left(1+\frac{2}{m}\right)\epsilon^{2c}+2\epsilon^{c}\frac{\epsilon^2}{4} \tag{320}$$

$$\leq \epsilon^2\left(\frac{1}{4}+\left(1+\frac{2}{m}\right)\epsilon^{2c-2}+\frac{1}{2}\epsilon^{c}\right) \tag{321}$$

$$\leq \frac{\epsilon^2}{3} \tag{322}$$

where the last inequality holds for $\epsilon \leq \frac{1}{6}$ and $c \geq 2$.
For the second part of the proof we utilize the lower bounds from Lemma 25:

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 \tag{323}$$

$$\geq \|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2 - 2\epsilon^c\left(\|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2\right) \tag{324}$$

$$> \epsilon^2 - 2\epsilon^{c+2} \tag{325}$$

$$\geq \frac{7}{8}\epsilon^2 \tag{326}$$

where the last inequality holds for $\epsilon \leq \frac{1}{6}$ and $c \geq 2$.

□

The following lemma states that if r is a good iteration then for each node the estimation after running the consensus protocol is at most $\frac{\epsilon^2}{2}$. On the other hand if r is an iterate with $\|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2 + \frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^r - \underline{\mathbf{y}}^r\right\|^2 > \epsilon^2$ then each node has an estimation of value at least $\frac{3}{4}\epsilon^2$.

Lemma 27. *Consider any iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Let $\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}$ and $\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}$, $\underline{\mathbf{z}}^{r,t_{\mathbf{z}}+1}$ and $\underline{\mathbf{w}}^{r,t_{\mathbf{w}}+1}$ as defined in Lemma 22 and Lemma 24. Also let $\epsilon \leq \min\{\frac{1}{8}, \sqrt{\frac{m}{16\alpha}(f(\mathbf{x}^0) - f^* + 2\mathcal{F})}\}$ and $c \geq 3$. If it holds that*

$$\|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2 \leq \frac{\epsilon^2}{4} \tag{327}$$

then

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \mathbf{w}_i^{r,t_{\mathbf{w}}+1} \leq \frac{\epsilon^2}{2} \quad \forall i \tag{328}$$

Further if it holds that

$$\|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2 > \epsilon^2 \tag{329}$$

then

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \mathbf{w}_i^{r,t_{\mathbf{w}}+1} > \frac{3}{4}\epsilon^2 \tag{330}$$

Proof. From Lemma 24 we know the following

$$\left\|\underline{\mathbf{w}}^{r,t_{\mathbf{w}}+1} - \underline{\hat{\mathbf{w}}}^r\right\| \leq \epsilon^c \tag{331}$$

$$|\mathbf{w}_i^{r,t_{\mathbf{w}}+1} - \hat{\mathbf{w}}^r| \leq \epsilon^c \tag{332}$$

$$|\mathbf{w}_i^{r,t_{\mathbf{w}}+1} - \frac{1}{m}\sum_{i=1}^m\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2| \leq \epsilon^c \tag{333}$$

$$|\mathbf{w}_i^{r,t_{\mathbf{w}}+1} - \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2| \leq \epsilon^c \tag{334}$$

$$\tag{335}$$

And thus we have

$$\frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 - \epsilon^c \leq \mathbf{w}_i^{r,t_{\mathbf{w}}+1} \leq \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 + \epsilon^c \tag{336}$$

Utilizing the above bounds and Lemma 26 we can show the first claim

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \mathbf{w}_i^{r,t_{\mathbf{w}}+1} \leq \left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 + \epsilon^c \tag{337}$$

$$\leq \frac{\epsilon^2}{3} + \epsilon^c \tag{338}$$

$$\leq \frac{\epsilon^2}{2} \tag{339}$$

where the last inequality holds for $\epsilon \leq \frac{1}{6}$ and $c \geq 3$. The second claim is derived along the same lines:

$$\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \mathbf{w}_i^{r,t_{\mathbf{w}}+1} \geq \left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 - \epsilon^c \tag{340}$$

$$> \frac{7}{8}\epsilon^2 - \epsilon^c \tag{341}$$

$$\geq \frac{3}{4}\epsilon^2 \tag{342}$$

where the last inequality holds for $\epsilon \leq \frac{1}{8}$ and $c \geq 3$. □

The next lemma shows that after a small number of iterations all the nodes can coordinate to either approve or disapprove iteration r.

Lemma 28. *Consider any first phase iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Let $\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1}$, $\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1}$ and $\underline{\mathbf{w}}^{r,t_{\mathbf{w}}+1}$ as defined in Lemma 22 and Lemma 24. Let $\epsilon \leq \min\{\frac{1}{8}, \sqrt{\frac{m}{16\alpha}(f(\mathbf{x}^0) - f^* + 2\mathcal{F})}\}$ and $c \geq 3$.*

Define $ind_i^{r,0} := \mathbb{1}_{\left\{\left\|\mathbf{y}_i^{r,t_{\mathbf{y}}+1}\right\|^2 + \mathbf{w}_i^{r,t_{\mathbf{w}}+1} \leq \frac{\epsilon^2}{2}\right\}}$ and also $\hat{ind}^r := \frac{1}{m}\sum_{i=1}^m ind_i^{r,0}$

Also define $t_{\mathbf{ind}} := \left\lfloor \frac{\log(2m^{\frac{3}{2}})}{\log(\frac{1}{\sigma})} \right\rfloor$

If we run the average consensus protocol on $ind_i^{r,0}$'s for $t_{\mathbf{ind}}$ iterations we are going to achieve the following bound

$$\left\|\underline{\mathbf{ind}}^{r,t_{\mathbf{ind}}+1} - \underline{\hat{\mathbf{ind}}}^r\right\| \leq \frac{1}{2m} \tag{343}$$

Proof. From the update of the average consensus protocol we have

$$\left\|\underline{\mathbf{ind}}^{r,t+1} - \underline{\hat{\mathbf{ind}}}^r\right\| \leq \left\|\mathbf{W}\left(\underline{\mathbf{ind}}^{r,t} - \underline{\hat{\mathbf{ind}}}^r\right)\right\| \leq \sigma\left\|\underline{\mathbf{ind}}^{r,t} - \underline{\hat{\mathbf{ind}}}^r\right\| \tag{344}$$

$$\tag{345}$$

Thus we have

$$\left\|\underline{\mathbf{ind}}^{r,t+1} - \underline{\hat{\mathbf{ind}}}^r\right\| \leq \sigma^{t+1}\left\|\underline{\mathbf{ind}}^{r,0} - \underline{\hat{\mathbf{ind}}}^r\right\| \tag{346}$$

Choosing t such that $\sigma^{t+1}\left\|\underline{\mathbf{ind}}^{r,0} - \underline{\hat{\mathbf{ind}}}^r\right\| \leq \frac{1}{2m}$ and since $\left\|\underline{\mathbf{ind}}^{r,0} - \underline{\hat{\mathbf{ind}}}^r\right\| \leq \sqrt{m}$ we get

$$\sigma^{t+1} \leq \frac{1}{2m^{\frac{3}{2}}} \tag{347}$$

$$(t+1)\log(\sigma) \leq -\log(2m^{\frac{3}{2}}) \tag{348}$$

$$t \leq \frac{\log(2m^{\frac{3}{2}})}{\log(\frac{1}{\sigma})} - 1 \tag{349}$$

Thus choosing $t_{\mathbf{ind}} = \left\lfloor \frac{\log(2m^{\frac{3}{2}})}{\log(\frac{1}{\sigma})} \right\rfloor$ we derive the result. □

Corollary 9. *After running the average consensus protocol on $\mathbf{y}_i^r$'s, $\mathbf{x}_i^r$'s, $\left\|\mathbf{x}_i^{r,t_{\mathbf{x}}+1} - \mathbf{x}_i^r\right\|^2$'s and ind_i^r's each node approves iteration r if $1 - \frac{1}{2m} \leq ind_i^{r,t_{\mathbf{ind}}+1}$. The total number of iterations are at most $t_{tot} = 4\left(\frac{c\log(\frac{1}{\epsilon}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right)$. Further if*

$$\|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2 \leq \frac{\epsilon^2}{4} \tag{350}$$

then all nodes will approve whereas if

$$\|\hat{\mathbf{y}}^r\|^2 + \frac{1}{m}\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\|^2 \geq \epsilon^2 \tag{351}$$

then all nodes will disapprove.

Proof. The second part of the corollary is immediate from Lemma 28. For the number of iterations we have the following:

$$t_{tot} \leq t_{\mathbf{y}} + t_{\mathbf{x}} + t_{\mathbf{w}} + 3 + t_{\mathbf{ind}} + 1 \tag{352}$$

$$\leq 3t_{\mathbf{z}} + 3 + t_{\mathbf{ind}} + 1 \tag{353}$$

$$\leq 3\left(\frac{c\log(\frac{1}{\epsilon}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right) + \frac{\log(2m^{\frac{3}{2}})}{\log(\frac{1}{\sigma})} + 1 \tag{354}$$

$$\leq 4\left(\frac{c\log(\frac{1}{\epsilon}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right) \tag{355}$$

□

11.3 Tracking the Potential Function

Finally, working towards our third objective recall that $H(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = f(\hat{\mathbf{x}}) + \frac{1}{m}\left\|\underline{\mathbf{x}}^r - \underline{\hat{\mathbf{x}}}^r\right\|^2 + \frac{\alpha}{m}\left\|\underline{\mathbf{y}}^r - \underline{\hat{\mathbf{y}}}^r\right\|^2$. We start by utilizing Corollary 8 and thus for sufficiently small $\tilde{\epsilon}$ after $4\left(\frac{c\log(\frac{1}{\tilde{\epsilon}}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right)$ communication rounds we achieve the following accuracy bounds for some iteration r:

$$\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}} - \underline{\hat{\mathbf{y}}}^r\right\| \leq \tilde{\epsilon}^c \tag{356}$$

$$\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}} - \underline{\hat{\mathbf{x}}}^r\right\| \leq \tilde{\epsilon}^c \tag{357}$$

$$\left\|\underline{\mathbf{z}}^{r,t_{\mathbf{z}}} - \underline{\hat{\mathbf{z}}}^r\right\| \leq \tilde{\epsilon}^c \tag{358}$$

$$\left\|\underline{\mathbf{w}}^{r,t_{\mathbf{w}}} - \underline{\hat{\mathbf{w}}}^r\right\| \leq \tilde{\epsilon}^c \tag{359}$$

Further choosing a sufficiently large $\tilde{c}$ and running the consensus protocol for $\mathbf{x}_i^r$'s for $\left(\frac{\tilde{c}\log(\frac{1}{\tilde{\epsilon}}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right)$ rounds guarantees sufficient accuracy on the function value of the average iterate $\frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}})$.

Lemma 29. *Consider any iterates $\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r$ following Gradient Tracking Update with η_1, α that satisfy conditions* (41) - (44)*. Also assume that the potential function decreases between consecutive first phases. Consider a sufficiently large $\tilde{c}$ that guarantees $\max_i \left\|f_i(\mathbf{x}_i^{r,t_{\mathbf{x}}+1}) - f_i(\hat{\mathbf{x}}^r)\right\| \leq \frac{\tilde{\epsilon}^c}{2m}$ after running the consensus protocol on $\mathbf{x}_i^r$'s for $t_{\mathbf{x}} = \left(\frac{\tilde{c}\log(\frac{1}{\tilde{\epsilon}}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 1\right)$ rounds. Define $g_i^{r,0} := f_i(\mathbf{x}_i^{r,t_{\mathbf{x}}+1})$ and $\hat{g}^r := \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}_i^{r,t_{\mathbf{x}}+1})$, let $t_{\mathbf{g}} = \left\lfloor \frac{c\log(\tilde{\epsilon}) - \log\left(2\left\|\underline{\mathbf{g}}^{r,0} - \underline{\hat{\mathbf{g}}}^r\right\|\right)}{\log\sigma} \right\rfloor$ and denote the true target function value by $\hat{g}_{tr}^r := \frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}^r)$.*

Then the following bound holds $\left\|\underline{\mathbf{g}}^{r,t_{\mathbf{g}}+1} - \underline{\hat{\mathbf{g}}}_{tr}^r\right\| \leq \tilde{\epsilon}^c$

Proof. From the update of the consensus protocol we have $\left\|\underline{\mathbf{g}}^{r,t+1} - \underline{\hat{\mathbf{g}}}^r\right\| \leq \sigma^{t+1}\left\|\underline{\mathbf{g}}^{r,0} - \underline{\hat{\mathbf{g}}}^r\right\|$. Thus we solve for t such that

$$\sigma^{t+1}\left\|\underline{\mathbf{g}}^{r,0} - \underline{\hat{\mathbf{g}}}^r\right\| \leq \frac{\tilde{\epsilon}^c}{2} \tag{360}$$

$$t \geq \frac{c\log(\tilde{\epsilon}) - \log\left(2\left\|\underline{\mathbf{g}}^{r,0} - \underline{\hat{\mathbf{g}}}^r\right\|\right)}{\log\sigma} - 1 \tag{361}$$

Thus for $t_{\mathbf{g}} = \left\lfloor \frac{c\log(\tilde{\epsilon}) - \log\left(2\left\|\mathbf{g}^{r,0} - \hat{\mathbf{g}}^r\right\|\right)}{\log\sigma} \right\rfloor$ we have $\left\|\underline{\mathbf{g}}^{r,t_{\mathbf{g}}+1} - \underline{\hat{\mathbf{g}}}^r\right\| \leq \frac{\tilde{\epsilon}^c}{2}$.
Using the assumptions of the lemma we can also show that the estimation of the nodes is not far from the true function value.

$$\max_i \left\| f_i(\mathbf{x}_i^{r,t_{\mathbf{x}}+1}) - f_i(\hat{\mathbf{x}}^r) \right\| \leq \frac{\tilde{\epsilon}^c}{2m} \tag{362}$$

$$\frac{1}{m}\sum_{i=1}^{m} \left\| f_i(\mathbf{x}_i^{r,t_{\mathbf{x}}+1}) - f_i(\hat{\mathbf{x}}^r) \right\| \leq \frac{\tilde{\epsilon}^c}{2m} \tag{363}$$

$$\left\| \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}_i^{r,t_{\mathbf{x}}+1}) - \frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}^r) \right\| \leq \frac{\tilde{\epsilon}^c}{2m} \tag{364}$$

$$\left\| \underline{\hat{\mathbf{g}}}^r - \underline{\hat{\mathbf{g}}}_{tr}^r \right\| \leq \frac{\tilde{\epsilon}^c}{2} \tag{365}$$

which implies that $\left\| \underline{\mathbf{g}}^{r,t_{\mathbf{g}}+1} - \underline{\hat{\mathbf{g}}}_{tr}^r \right\| \leq \tilde{\epsilon}^c$ □

Corollary 10. *Consider the assumption of Lemma 29 hold. Then after* $4\frac{\tilde{c}\log(\frac{1}{\tilde{\epsilon}}) + \log(\frac{1}{\alpha}) + \log\left(4m^2(f(\mathbf{x}^0) - f^* + 2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 4 + \frac{c\log(\tilde{\epsilon}) - \log\left(2\left\|\mathbf{g}^{r,0} - \hat{\mathbf{g}}^r\right\|\right)}{\log\sigma}$ *rounds of consensus protocol we achieve the following accuracy:*

$$\left\| \underline{\mathbf{y}}^{r,t_{\mathbf{y}}} - \underline{\hat{\mathbf{y}}}^r \right\| \leq \tilde{\epsilon}^c \tag{366}$$

$$\left\| \underline{\mathbf{x}}^{r,t_{\mathbf{x}}} - \underline{\hat{\mathbf{x}}}^r \right\| \leq \tilde{\epsilon}^c \tag{367}$$

$$\left\| \underline{\mathbf{z}}^{r,t_{\mathbf{z}}} - \underline{\hat{\mathbf{z}}}^r \right\| \leq \tilde{\epsilon}^c \tag{368}$$

$$\left\| \underline{\mathbf{w}}^{r,t_{\mathbf{w}}} - \underline{\hat{\mathbf{w}}}^r \right\| \leq \tilde{\epsilon}^c \tag{369}$$

$$\left\| g_i^{r,t_{\mathbf{g}}} - \frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}) \right\| \leq \tilde{\epsilon}^c, \forall i \tag{370}$$

Similarly to section 11.2 utilizing Lemma 25 and equation (370) we get the following bounds :

Corollary 11. *Consider the assumptions of Lemma 29 hold. Then the following bounds also hold*

$$\frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^r - \underline{\mathbf{y}}^r\right\|^2 + \frac{1}{m}\tilde{\epsilon}^{2c} + \frac{2}{m}\tilde{\epsilon}^c\left\|\underline{\hat{\mathbf{y}}}^r - \underline{\mathbf{y}}^r\right\| \geq \frac{1}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^r\right\|^2 \geq \frac{1}{m}\left\|\underline{\hat{\mathbf{y}}}^r - \underline{\mathbf{y}}^r\right\|^2 - \frac{2}{m}\tilde{\epsilon}^c\left\|\underline{\hat{\mathbf{y}}}^r - \underline{\mathbf{y}}^r\right\| \tag{371}$$

$$\frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\right\|^2 + \frac{1}{m}\tilde{\epsilon}^{2c} + \frac{2}{m}\tilde{\epsilon}^c\left\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\right\| \geq \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 \geq \frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\right\|^2 - \frac{2}{m}\tilde{\epsilon}^c\left\|\underline{\hat{\mathbf{x}}}^r - \underline{\mathbf{x}}^r\right\| \tag{372}$$

$$\frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}) + \tilde{\epsilon}^c \geq g_i^{r,t_{\mathbf{g}}} \geq \frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}) - \tilde{\epsilon}^c \tag{373}$$

The next lemma is used as an intermediate step in order to derive our final result.

Lemma 30. *Consider the assumptions of Lemma 29 hold. Further let* $\tilde{\epsilon} \leq \min\left\{\frac{1}{8}, \left(4\sqrt{f(\hat{\mathbf{x}}^0) - f^* + 2\mathcal{F}}\right)^{\frac{c}{4}}\right\}$ *and* $c \geq 4$*. Then we can prove the following bounds:*

$$H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) + \tilde{\epsilon}^{\frac{c}{2}} \geq g_i^{r,t_{\mathbf{g}}} + \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^r\right\|^2 + \frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^r\right\|^2 \geq H(\underline{\mathbf{x}}^r, \underline{\mathbf{y}}^r) - \tilde{\epsilon}^{\frac{c}{2}} \tag{374}$$

Proof. Towards proving the upper bound we utilize the results presented in Corollary 10.

$$
\begin{aligned}
& g_i^{r,t_{\mathbf{g}}} + \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^{r}\right\|^2 + \frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^{r}\right\|^2 \\
& \leq \frac{1}{m}\sum_{i=1}^{m} f_i(\hat{\mathbf{x}}) + \tilde{\epsilon}^c + \frac{1}{m}\left\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\right\|^2 + \frac{1}{m}\tilde{\epsilon}^{2c} + \frac{2}{m}\tilde{\epsilon}^c\left\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\right\| \\
& \quad + \frac{\alpha}{m}\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|^2 + \frac{\alpha}{m}\tilde{\epsilon}^{2c} + \frac{2\alpha}{m}\tilde{\epsilon}^c\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\| && (375) \\
& \leq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) + \tilde{\epsilon}^c + \left(\frac{1+\alpha}{m}\right)\tilde{\epsilon}^{2c} + \frac{4}{m}\tilde{\epsilon}^{2c}\left(\left\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\right\| + \alpha\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|\right) && (376) \\
& \leq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) + \tilde{\epsilon}^c + \left(\frac{2}{m}\right)\tilde{\epsilon}^{2c} + 4\tilde{\epsilon}^c\sqrt{f(\hat{\mathbf{x}}^0) - f^* + 2\mathcal{F}} && (377) \\
& \leq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) + \tilde{\epsilon}^{\frac{c}{2}} && (378)
\end{aligned}
$$

where third inequality comes from Lemma 21 aND the last inequality is due to $\tilde{\epsilon} \leq \min\left\{\frac{1}{8}, \left(4\sqrt{f(\hat{\mathbf{x}}^0) - f^* + 2\mathcal{F}}\right)^{\frac{c}{4}}\right\}$ and $c \geq 4$. Similarly for the lower bound we have

$$
\begin{aligned}
& g_i^{r,t_{\mathbf{g}}} + \frac{1}{m}\left\|\underline{\mathbf{x}}^{r,t_{\mathbf{x}}+1} - \underline{\mathbf{x}}^{r}\right\|^2 + \frac{\alpha}{m}\left\|\underline{\mathbf{y}}^{r,t_{\mathbf{y}}+1} - \underline{\mathbf{y}}^{r}\right\|^2 \\
& \geq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) - \tilde{\epsilon}^c - \frac{4}{m}\tilde{\epsilon}^{2c}\left(\left\|\underline{\hat{\mathbf{x}}}^{r} - \underline{\mathbf{x}}^{r}\right\| + \alpha\left\|\underline{\hat{\mathbf{y}}}^{r} - \underline{\mathbf{y}}^{r}\right\|\right) && (379) \\
& \geq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) - \tilde{\epsilon}^{\frac{c}{2}} && (380)
\end{aligned}
$$

□

From the above lemma and inequalities (369) and (368) we can bound the error of the estimation of the potential function by each node i.

Corollary 12. *Assume the conditions of Lemma 30 and inequalities* (369) *and* (368) *hold. Then the following bounds characterize the error on the estimation of the potential function after utilizing the average consensus protocol for* $4\frac{\tilde{c}\log(\frac{1}{\tilde{\epsilon}})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 4 + \frac{c\log(\tilde{\epsilon})-\log\left(2\left\|\underline{\mathbf{g}}^{r,0}-\underline{\hat{\mathbf{g}}}^{r}\right\|\right)}{\log\sigma}$ *rounds.*

$$
H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) + \tilde{\epsilon}^{\frac{c}{2}} + \tilde{\epsilon}^c + \alpha\tilde{\epsilon}^c \geq g_i^{r,t_{\mathbf{g}}} + \mathbf{w}_i^{r,t_{\mathbf{w}}} + \alpha\mathbf{z}_i^{r,t_{\mathbf{z}}} \geq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) - \tilde{\epsilon}^{\frac{c}{2}} - \tilde{\epsilon}^c - \alpha\tilde{\epsilon}^c \tag{381}
$$

And for $\tilde{\epsilon} \leq \frac{1}{8}$ *and* $c \geq 4$ *we also have*

$$
H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) + 2\tilde{\epsilon}^{\frac{c}{2}} \geq g_i^{r,t_{\mathbf{g}}} + \mathbf{w}_i^{r,t_{\mathbf{w}}} + \alpha\mathbf{z}_i^{r,t_{\mathbf{z}}} \geq H(\underline{\mathbf{x}}^{r}, \underline{\mathbf{y}}^{r}) - 2\tilde{\epsilon}^{\frac{c}{2}} \tag{382}
$$

Further after $8\frac{\tilde{c}\log(\frac{1}{\tilde{\epsilon}})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 8 + 2\frac{c\log(\tilde{\epsilon})-\log\left(2\left\|\underline{\mathbf{g}}^{r,0}-\underline{\hat{\mathbf{g}}}^{r}\right\|\right)}{\log\sigma}$ *of the consensus protocol on the iteration before the injection of noise and on the iteration at the end of phase two we have*

$$
H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) - H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) + 4\tilde{\epsilon}^{\frac{c}{2}} \geq g_i^{-1,t_{\mathbf{g}}} + \mathbf{w}_i^{-1,t_{\mathbf{w}}} + \alpha\mathbf{z}_i^{-1,t_{\mathbf{z}}} - \left(g_i^{T_{cap},t_{\mathbf{g}}} + \mathbf{w}_i^{T_{cap},t_{\mathbf{w}}} + \alpha\mathbf{z}_i^{T_{cap},t_{\mathbf{z}}}\right) \tag{383}
$$

$$
H(\underline{\mathbf{x}}^{-1}, \underline{\mathbf{y}}^{-1}) - H(\underline{\mathbf{x}}^{T_{cap}}, \underline{\mathbf{y}}^{T_{cap}}) - 4\tilde{\epsilon}^{\frac{c}{2}} \leq g_i^{-1,t_{\mathbf{g}}} + \mathbf{w}_i^{-1,t_{\mathbf{w}}} + \alpha\mathbf{z}_i^{-1,t_{\mathbf{z}}} - \left(g_i^{T_{cap},t_{\mathbf{g}}} + \mathbf{w}_i^{T_{cap},t_{\mathbf{w}}} + \alpha\mathbf{z}_i^{T_{cap},t_{\mathbf{z}}}\right) \tag{384}
$$

Combining all the above we can achieve our third objective

Theorem 7. *Assume the conditions of Corollary 12 hold and set* $\tilde{\epsilon}^{\frac{c}{2}} = \frac{\mathcal{F}}{40}$. *After* $8\frac{\tilde{c}\log(\frac{1}{\tilde{\epsilon}})+\log(\frac{1}{\alpha})+\log\left(4m^2(f(\mathbf{x}^0)-f^*+2\mathcal{F})\right)}{\log(\frac{1}{\sigma})} + 9 + 2\frac{c\log(\tilde{\epsilon})-\log\left(2\left\|\underline{\mathbf{g}}^{r,0}-\underline{\hat{\mathbf{g}}}^{r}\right\|\right)}{\log\sigma} + \frac{\log(2m^{\frac{3}{2}})}{\log(\frac{1}{\sigma})}$ *iterations of the consensus protocol the nodes decide whether enough progress has been made in phase II.*

Proof. First notice that by setting $\tilde{\epsilon}^{\frac{\epsilon}{2}} = \frac{\mathcal{F}}{40}$ the estimation of each node i regarding the potential function decrease is off at most by $\frac{\mathcal{F}}{10}$. Thus the nodes can distinguish between second phases that achieve decrease at least $\mathcal{F}$ and second phases that achieve decrease less than $\frac{\mathcal{F}}{2}$. To do so we utilize Lemma 28 with $ind_i^0 = \mathbb{1}_{\left\{g_i^{-1,t_{\mathbf{g}}} + \mathbf{w}_i^{-1,t_{\mathbf{w}}} + \alpha \mathbf{z}_i^{-1,t_{\mathbf{z}}} - \left(g_i^{Tcap,t_{\mathbf{g}}} + \mathbf{w}_i^{Tcap,t_{\mathbf{w}}} + \alpha \mathbf{z}_i^{Tcap,t_{\mathbf{z}}}\right) \leq \frac{\mathcal{F}}{2}\right\}}$. Notice that if the potential function decrease in the current phase II is at least $\mathcal{F}$ then all nodes are going to approve and in the case the the current phase II achieves less than $\frac{\mathcal{F}}{2}$ decrease all nodes are going to disapprove. Finally, notice that if the decrease is between $\mathcal{F}$ and $\frac{\mathcal{F}}{2}$ both outcomes are possible; this is acceptable since enough progress have been made in this case as well. □