DECENTRALIZED PROXIMAL GRADIENT METHOD FOR NON-CONVEX COMPOSITE PROBLEMS WITH INEXACT GRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimization problems with composite structure appears in different areas: machine learning, control, signal processing and so on. Gradient-type methods are common approach for such problems. Nevertheless, the exact gradient is not available in many practical applications. Especially, it holds for decentralized case. Therefore, we consider decentralized proximal gradient method with inexact gradient for time-varying graphs. This work contains analysis for problems with functions that satisfy proximal Polyak-Łojasiewicz condition. Thus, there is complexity estimations in terms both of oracle calls, and of communications number. Additionally, we consider stochastic case too. Besides, we provide numerical experiments to demonstrate performance of considered approach.

INTRODUCTION

Let us consider a sum-optimization problem with composite structure:

$$\min F(x) := \sum_{i=1}^{n}(f_i(x) + r_i(x)),$$

where r_i is some smooth function that can be easily optimized. Such problems are common enough in different areas: statistics Parikh & Boyd (2014), machine learning Goodfellow et al. (2016); Drori et al. (2015), signal processing Boyd et al. (2011) etc. Usually, f_i is main part of objective function and r_i are some regularizators. The typical example of such problems is training of models in machine learning with ℓ_1 and ℓ_2 regularization.

Further, note, the common approach for some problems is gradient-type methods with proximal operators. Some of these methods were proposed in Parikh & Boyd (2014). Further, they have been improved in Liang & Schönlieb (2018); Kim & Fessler (2018). Such methods use smoothness of the first part and ”simplicity” of the second. They have different advantages: simple implementation, high performance in practice, well-known theoretical guarantees for wide class of problems. Especially, it is known for convex and strong-convex case in non-distributed case. Nevertheless, non-convex case is not well-researched yet. One of such papers Zhou et al. (2020); Li et al. (2017) is devoted to Kurdyka-Łojasiewicz condition but the obtained rate is sublinear in the worst case and they do not research dependence on the inexactness.

In the case of large amount of data, such problems can be solved in distributed way Li et al. (2020); Lian et al. (2017); Nedic & Ozdaglar (2009). In other words, we have a network where each node knows only its part of sum and can communicate with only neighbors. Such way allow decreasing requirements for devices to solve this problem. On the other hand, its bottleneck is communications that can be carefully designed. There are common approaches based on consensus procedure on each iteration. Because of big data in many practical important applications, researchers are interested in development of distributed methods too. Here, we can emphasize recent works devoted to distributed strong-convex case Rogozin et al. (2023). We will continue research from this paper and generalize its analysis for non-convex case.

Besides distributed methods based on consensus procedure. there is a wide class of methods with several communication steps per iteration. Here, we can find different primal dual methods with proximal operators: ABC Xu et al. (2021), NEXT Di Lorenzo & Scutari (2016), Exact Diffusion Yuan et al. (2018); Xu et al. (2021), NIDS Li et al. (2019), EXTRA Shi et al. (2015), AugDGM Xu et al. (2015). Another approaches are based on dual problems. The first such approach was proposed in Scaman et al. (2017). Nevertheless, the most of such methods have proved convergence only for convex case. Another disadvantage of such approaches is that they usually work only not changing networks. But it is a common situation when some links are lost and some are added Bonawitz et al. (2019); Li et al. (2020); Nedic (2020). Because of that disadvantages, we focus on methods using consensus procedure.

Despite the high complexity of general non-convex problem, there are different generalization of convexity: star-convexity Lee & Valiant (2016), weakly-quasi-convexity Hardt et al. (2016); Guminov & Gasnikov (2017), quadratic growth condition Karimi et al. (2016). One of such generalizations is Polyak-Łojasiewicz (proposed in Polyak (1963)) condition that can be found in many practical problems Belkin (2021); Karimi et al. (2016). Moreover, there is its generalization for convex case - Proximal PL-condition Karimi et al. (2016). It allows proving convergence of different proximal gradient methods for non-convex problems.

Note, that consensus procedure leads to some errors in method. Nevertheless, the gradient of function already contains its own inexactness: problems in Hilbert spaces, distributed systems with compression of gradient etc. Examples of such problems can be found in the following works Devolder et al. (2013); Vasin et al. (2021); Polyak (2020).

Despite the numerous papers, non-convex distributed case is not researched yet. This paper is devoted to the generalization of distributed proximal gradient method for optimization problems with composite structure under proximal pl-condition. Our main contributions are given by the following list:

1. **Convergence for Inexact Proximal Gradient Method.** We present theoretical results for proximal gradient method for non-distributed case under Proximal PL-condition. We estimate sufficient number of iterations to approach required accuracy. Besides, we consider the case of the use of gradient with additive inexactness and stochastic gradient. These results are presented in Section 2

2. **Generalization of Analysis for Distributed Case** We generalize the results from previous point for distributed case with time-varied graphs. In such case, we have additional natural inexactness in directions because of distributed calculations. We provide estimations of complexity in terms of the communications too. You can find our analysis for this case in Section 3.

Besides, we provide numerical experiments in Section 4.

1 Problem Statement

We consider the several problem statements. But all of them has main restriction: exact gradient is not available. We define it in more formally in Section 1.1.

We start from problem of non-distributed composite minimization (see Section 1.1). This section also contains the main assumptions about properties of our non-convex function and inexact gradient.

The next step is the distributed variant of the problem above (see Section 1.2). It also contains assumption about communication matrix.

Notation. Throughout the paper, $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors or matrices. Correspondingly, by $\| \cdot \|$, we denote the 2-norm for vectors or the Frobenius norm for matrices.

1.1 Non-Distributed Case

We start our analysis from centralized case and consider the following problem:

$$\min_{x \in \mathbb{R}^d} \left[F(x) := f(x) + r(x) \right], \tag{1}$$

where f is some smooth possibly non-convex optimization problem and function r is some convex friendly proximal function. The last statement means that proximal operator with respect to the function r can be calculated with low computational cost. The assumption about smoothness means the following assumption.

Assumption 1 (Smoothness)**.** *Function f_i is L_i-smooth i.e. the following condition holds*

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}\|x - y\|^2 \tag{2}$$

for any $x, y \in \mathbb{R}^d$.

We will work with non-convex problems that sutisfy the following proximal Polyak-Łojasiewicz condition (**Proximal PL**):

Assumption 2 (Proximal PL)**.** *Function $F(x) := f(x) + r(x)$ meets proximal-PL condition if function f is L-smooth and sutisfy the following inequality for all $\mathbf{x}$:*

$$F(x) - F(x^\star) \leq \frac{1}{2\mu}\boldsymbol{D}(x, L) \tag{3}$$

where

$$\boldsymbol{D}(x, L) = -2\alpha \min_y \left[\langle \nabla f, x - y \rangle + \frac{\alpha}{2}\|x - y\|^2 + g(y) - g(x)\right]. \tag{4}$$

There are different examples of problems that satisfy this condition:

1. f is composition of strong-convex function and linear function and r is an indicator function of polyhedral set. In particular, such class includes optimization of mean least squares problem over polyhedra.
2. Function F satisfies quadratic growth condition
3. Least squares problem with ℓ_1 regularization
4. Nuclear norm regularization
5. Support Vector Machines
6. Some problems that satisfy KL-condition (see Karimi et al. (2016), Li et al. (2023))

Note, the convergence of Proximal Gradient method for problems with such assumption was researched in Karimi et al. (2016). In our paper, we generalize this analysis for the case of gradient with additive inexactness and the stochastic case. Therefore, let us introduce the following assumptions for gradient inexactness.

Assumption 3 (Inexact Gradient)**.** *Algorithm has access to inexact gradient, i.e. vector $g_{x,\xi}$ is defined at any point x such that the following inequalities hold:*

$$\|g_x - \nabla f(x)\| \leq \delta.$$

The definition above assume that the inexactness has no structure, and we can guarantee only that it is not too large. There are well-known for different optimization method in convex case Polyak (1963); Vasin et al. (2021); Nesterov & Spokoiny (2015) and in even non-convex case Bottou et al. (2018); Ajalloeian & Stich (2020); Stonyakin et al. (2023); Kuruzov & Stonyakin (2021). In particular, there are works demonstrating robustness of simple gradient method to error accumulation Polyak (1963). Note, that the classic accelerated methods usually have not such property (see Vasin et al. (2021)).

On the other hand, stochastic gradient place important role in modern optimization. Besides, it can contain bias too. Further, we introduce assumptions about biased stochastic oracle that are inspired by Ajalloeian & Stich (2020):

Assumption 4 (Stochastic Inexact Gradient)**.** *Algorithm has access to (δ, σ^2) stochastic inexact gradient, i.e. vector $g_{x,\xi}$ is defined at any point x such that the following inequalities hold:*

$$\mathbb{E}_\xi \|g_{x,\xi} - \mathbb{E}_\xi g_{x,\xi}\|^2 \leq \sigma^2$$

and

$$\|\mathbb{E}_\xi g_{x,\xi} - \nabla f(x)\| \leq \delta$$

This assumption cover different practical applications. In particular, there are different methods of zeroth-order methods Vasin et al. (2021); Nesterov & Spokoiny (2015) and gradient compression Beznosikov et al. (2020); Liu et al. (2018). All these methods have bias in gradient and sometimes include stochastic error. Note, the methods of gradient sparsification and compression have important place in decentralized optimization because they allow significantly decreasing complexity of communications.

1.2 Distributed Minimization Problem

Let us introduce the distributed composite problem. In this problem we have a sequence of networks $\{\mathcal{G}_t\}_t$ with the same number of nodes but the set of edges can change. Each communication round can be given through matrices of communications W_t that should satisfy the following conditions:

Assumption 5. *Mixing matrix sequence $\{W^k\}_{k=0}^{\infty}$ satisfies the following properties:*

- *(Decentralized property) If $(i, j) \notin E^k \cup \{i\}$, then $[W^k]_{ij} = 0$.*
- *(Double stochasticity) $W^k \mathbf{1}_n = \mathbf{1}_n$, and $\mathbf{1}_n^\top W^k = \mathbf{1}_n^\top$.*
- *(Contraction property) There exist $\tau \in \mathbb{Z}_{++}$ and $\lambda \in (0, 1)$ such that for every $k \geq \tau - 1$, it holds the following inequality*

  $$\left\| W_\tau^k \mathbf{X} - \overline{\mathbf{X}} \right\| \leq (1 - \lambda) \left\| \mathbf{X} - \overline{\mathbf{X}} \right\|,$$

  where $W_\tau^k = W^k \ldots W^{k-\tau+1}$.

Note, this class of time-varying networks includes static graphs and sequence of connected graphs. At the same time, this sequence may contain disconnected graphs too. In particular, the contraction property is satisfied if each communication is Metropolis weights choice and the sequence is τ-connected graph sequence Nedic et al. (2017).

Algorithm 1 Consensus.

```
Require: Initial point $\mathbf{x}^0 \in \mathbb{R}^{n \times d}$, number of communication rounds $T$.
Take current time moment $t_0$ from global variable.
for $k = 0, \ldots, T - 1$ do
    $\mathbf{z}^{k+1} := W^{t_0+k}\, \mathbf{x}^k$.
Update global variable with current time moment: $t_0 = t_0 + T$. return $\mathbf{x}^T$.
```

Now we can present protocol of communication - Algorithm 1. On each step of communication, each node takes vectors from its neighbors and averages them. Weights of this averaging are given by communication matrix. Note, each matrix can change according to Assumption 5.

Let us introduce the distributed composite problem:

$$\min_{x \in \mathbb{R}^d} \left[F(x) := \frac{1}{n} \sum_{i=1}^{N} (f_i(x) + r_i(x)) \right]$$

where each function r_i is friendly proximal and each function f_i meets Assumption 1. In this problem statement, each node i has access to only its own function f_i and r_i. Besides, during optimization process, each node will have its own variable state x_i. So, we can rewrite optimization problem above in the following way:

$$\min_{\mathbf{x} \in \mathbb{R}^{n \times d} : \frac{1}{n} \mathbf{1}\mathbf{1}^\top = \mathbf{x}} \left[\hat{F}(\mathbf{x}) := \sum_{i=1}^{N} (f_i(x_i) + r_i(x_i)) \right]$$

2 Proximal Gradient Method

In this section, we consider the composite problem in the following form:

$$\min_{x \in \mathbb{R}^d} [F(x) := f(x) + r(x)] \tag{5}$$

where $f(x)$ is smooth part and function F satisfy Assumption 2.

Let us introduce the well-known Karimi et al. (2016); Parikh & Boyd (2014) Proximal Gradient Method (see Algorithm 2). On each iteration we calculate inexact gradient descent and after that method calculate gradient step and step with proximal operator.

Algorithm 2 Proximal Gradient Method

```
Data: step size $\alpha$, number of consensus steps $T$, inexact gradient oracle $g : \mathbb{R}^n \to \mathbb{R}^n$
1: Inexact Gradient Calculation:
       $\mathbf{g}_{\mathbf{x}^k} := g(\mathbf{x}^k)$
2: (S.1) Gradient Step:
       $y^{k+1} = x^k - \alpha g_{x^k}$
3: (S.2) Proximal Step:
       $x^{k+1} = \text{prox}_{\alpha r_i}(z^{k+1})$
4: (S.3) If a termination criterion is not met, then $k \leftarrow k + 1$ and go to step (S.1).
```

Obviously, our algorithm accumulates errors in Step 1. This errors can be deterministic (see Assumption 3) or stochastic 4. Consequently, the results of this section are divided into two corresponding subsections.

Section 2.1 contains Theorem 6 that generalize well-known proximal PL-condition for bounded deterministic inexactness. It is a key result of this paper that allows estimating convergence rate of Proximal Gradient Method for non-convex case.

Furthermore, we have similar results for stochastic case presented in Section 2.2.

2.1 Deterministic Case

In this section, we estimate influence of deterministic inexactness that satisfy Assumption 3.

Firstly, note, that Proximal PL Condition (see Assumption 2) uses exact gradient when we access to only inexact one. The following theorem modificates this condition for this case.

Theorem 6. *Let g_x be an inexact gradient at point x that meets Assumption 3 and function $F(x) = f(x) + r(x)$ meets Assumptions 1 and 2. Then the following inequality holds:*

$$F(x) - F^\star \leq \frac{1}{\mu}\widetilde{\boldsymbol{D}}(x, L/2) + \frac{\delta^2}{\mu}, \tag{6}$$

where $\tilde{L} \geq \frac{1}{n}\sum\limits_{i=1}^{n} L_i$

Proof. Let us consider the following value:

$$\tilde{\mathcal{D}}_g(x, L) = -2L \min_y \left[\langle g_x, y - x\rangle + \frac{L}{2}\|y - x\| + r(y) - r(x)\right]$$

Because of Assumption 3 we have:

$$\langle g_x, x - y\rangle \geq \langle \nabla f(x), x - y\rangle - \frac{\delta^2}{L} - \frac{L}{4}\|y - x\|.$$

It gives us the following estimation:

$$\begin{aligned}
&-2L \min_y \left[\langle g_x, y - x\rangle + \frac{L}{2}\|y - x\| + r(y) - r(x)\right] \\
\geq &-2L \min_y \left[\langle \nabla f(x), y - x\rangle + \frac{L}{4}\|y - x\| + r(y) - r(x)\right] - \frac{\delta^2}{L} \\
= &\frac{1}{2}\mathcal{D}(x, L/2) - \frac{\delta^2}{L}
\end{aligned}$$

We can obtain estimation for $\mathcal{D}(x, L/2)$ through $\tilde{\mathcal{D}}_g(x, L)$. Further, use statement of Assumption 2 gives us the statement of the theorem. □

This theorem allows us to generalize the result from Karimi et al. (2016) to the following result with inexact gradient.

Theorem 7. *Let us consider the problem equation 5 such that function F meets Assumptions 1 and 2, g_x be δ_g-inexact gradient of function f with respect to $\mathbf{x}$ (see Assumption 3). Besides, desired accuracy ε is such that $\varepsilon \geq \frac{3\delta_g^2}{2\mu}$. Then after $N = \left\lceil \frac{4L}{\mu} \ln \frac{F(x^0)-F^*}{\varepsilon} \right\rceil$ iterations of Algorithm 2 one can guarantee $F(x^N) - F^* \leq \varepsilon$.*

The proof of this theorem is similar to analysis of exact PGM in Karimi et al. (2016) and it is placed in Appendix B.1

We can see that this algorithm does not accumulate error even in non-convex case. Moreover, it can approach accuracy $\varepsilon \sim \delta_g^2$ with linear rate.

2.2 Stochastic Case

In this section, we consider the case when g_x is stochastic biased estimation of true gradient. We have discussed the importance of this class in Section 1. Let us start for the generalization of Proximal PL condition for this stochastic case. In similar way to proof Theorem 6 we can obtain the following theorem

Theorem 8. *Let g_x be an inexact gradient at point x that meets Assumption 3 and function F meets Assumptions 1 and 2. Then the following inequality holds:*

$$F(x) - F^\star \leq \frac{1}{\mu}\mathbb{E}\widetilde{\boldsymbol{D}}(x, \frac{\tilde{L}}{2}) + \frac{\delta^2}{\mu} + \frac{\sigma^2}{\mu}. \tag{7}$$

This theorem allows us to generalize the result from Karimi et al. (2016) to the following result with inexact gradient.

Theorem 9. *Let us consider the problem equation 5 such that function F meets Assumptions 1 and 2, g_x be δ_g-inexact gradient of function f with respect to $\mathbf{x}$ (see Assumption 4). Besides, desired accuracy ε is such that $\varepsilon \geq \frac{3\delta_g^2+3\sigma^2}{2\mu}$. Then after $N = \left\lceil \frac{4L}{\mu} \ln \frac{F(x^0)-F^*}{\varepsilon} \right\rceil$ iterations of Algorithm 2 one can guarantee quality $\mathbb{E}F(x^N) - F^* \leq \varepsilon$.*

Proof for this theorem is presented in Appendix B.2. We obtain a natural result that stochastic does not worsen significantly convergence rate, but the quality can be a little worse.

3 Analysis for Distributed Case

Further, let us consider generalization of Algorithm 2 for decentralized case - Decentralized Proximal Gradient Method (see Algorithm 3). This section is devoted to this algorithm in the case of deterministic gradient inexactness. Note, gradient step S.1 and Proximal Step S.2 are made independently for different nodes. Besides, we have here additional step of Consensus (see step S.3). We provide analysis of accumulation of inexactness from gradient and from this consensus procedure.

Let us start from important lemma that demonstrate accumulation of inexactness from gradient and from consensus procedure.

Lemma 10. *Let the number of consensus steps T in Algorithm 3 be such that $\|\mathbf{x}^k - \overline{\mathbf{x}}^k\| \leq \delta_c$. Besides, $\mathbf{g}_{\mathbf{x}^k}$ is inexact gradient of function F with constant δ_g. Then $\overline{g_{x^k}} = \frac{1}{n}\sum\limits_{i=1}^{N} g_{\mathbf{x}_i^k}$ is inexact gradient of function f with respect to $\overline{x^k}$ with parameter $\delta = \frac{\delta_g + \delta_c \max_i L_i}{\sqrt{n}}$.*

Proof. Firstly, because of enough large number of consensus steps and Lipschitz continuous gradient, we have $\|\nabla F(\mathbf{x}^k) - \nabla F(\overline{\mathbf{x}}^k)\| \leq \delta_c \max_i L_i$. On the other hand, $\mathbf{g}_{\mathbf{x}^k}$ meets $\|\nabla F(\mathbf{x}^k) - \mathbf{g}_{\mathbf{x}^k}\| \leq \delta_g$. It gives us the upper bound on difference between inexact gradient and gradient with respect to corresponding mean point:

$$\|\nabla F(\overline{\mathbf{x}}^k) - \mathbf{g}_{\mathbf{x}^k}\| \leq \delta_g + \delta_c \max_i L_i.$$

Algorithm 3 Decentralized Proximal Gradient Method (DPGM)

```
Data: step size $\alpha$, number of consensus steps $T$, inexact gradient oracle $g : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$
Inexact Gradient Calculation:
    $\mathbf{g}_{\mathbf{x}^k} := g(\mathbf{x}^k)$
(S.1) Gradient Step:
    $\mathbf{y}^{k+1} = \mathbf{x}^k - \alpha \mathbf{g}_{\mathbf{x}^k}$
(S.3) Consensus:
    $\mathbf{y}^{k+1} = \text{Consensus}(\mathbf{y}^{k+1}, T)$
(S.2) Proximal Step:
    $x_i^{k+1} = \text{prox}_{\alpha r_i}(z_i^{k+1})$
    for all $i = \overline{1, n}$.
(S.4) If a termination criterion is not met, then $k \leftarrow k+1$ and go to step (S.1).
```

Further, note, that $\nabla f(\overline{x}^k) = \frac{1}{n}\sum\limits_{i=1}^{n}\left[\nabla F(\overline{\mathbf{x}}^k)\right]_i$. Then we have:

$$\|g_{\overline{x}_k} - \nabla f(\overline{x}^k)\| \leq \frac{1}{n}\sum_{i=1}^{n}\|g_{x_i^k} - \left[\nabla F(\overline{\mathbf{x}}^k)\right]_i\| \leq \frac{1}{\sqrt{n}}\|\nabla F(\overline{\mathbf{x}}^k) - \mathbf{g}_{\mathbf{x}^k}\| \leq \frac{\delta_g + \delta_c \max_i L_i}{\sqrt{n}}$$

□

It means that this procedure is still extra similar to usual Proximal Gradient Method with increased inexactness because of communication error.

Finally, let us provide the result for the deterministic case.

Theorem 11. *Let f meet Assumptions 1 and 2, $\mathbf{g}_\mathbf{x}$ be δ_g-inexact gradient of function F with respect to a point $\mathbf{x}$, communication network meets Assumption 5 Besides, desired accuracy ε is such that $\varepsilon \geq \frac{16\delta_g^2}{n\mu}$. Further let us define value $R = \frac{64L^2}{\mu^2}\max[f(\overline{x}^0) - f(x^*), \varepsilon]$. Then $N = \left\lceil \frac{4L}{\mu}\ln\frac{4(f_0 - f^*)}{\varepsilon}\right\rceil$ iterations of Algorithm 3 with $\alpha = \frac{1}{4L}$ and $T = \left\lceil \frac{2\tau}{\lambda}\ln\frac{R}{\varepsilon}\right\rceil$ communications per iteration are enough to approach the point $\mathbf{x}^{k+1}$ such that $F(\overline{x}_k) - F^* \leq \varepsilon$ and $\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^{k+1}\| \leq \frac{\mu\varepsilon}{4L}$.*

Finally, let us provide result for the general case - Decentralized Proximal Gradient Method with Stochastic Biased Gradient.

Theorem 12. *Let F meet Assumptions 1 and 2, $\mathbf{g}_\mathbf{x}$ be (δ_g, σ)-inexact gradient of function F with respect to a point $\mathbf{x}$, communication network meets Assumption 5. Besides, desired accuracy ε is such that $\varepsilon \geq \frac{4\delta_g^2 + 4\sigma^2}{n\mu}$. Further let us define value $R = \frac{500L^2}{\mu^2}\max[f(\overline{x}^0) - f(x^*), \varepsilon]$. Then $N = \left\lceil \frac{4L}{\mu}\ln\frac{4(F_0 - F^*)}{\varepsilon}\right\rceil$ iterations of Algorithm 3 with $\alpha = \frac{1}{4L}$ and $T = \left\lceil \frac{2\tau}{\lambda}\ln\frac{R}{\varepsilon}\right\rceil$ communications per iteration are enough to approach the point $\mathbf{x}^N$ such that $\mathbb{E}F(\overline{x}^n) - F^* \leq \varepsilon$ and $\mathbb{E}\|\mathbf{x}^N - \overline{\mathbf{x}}^N\| \leq \frac{\mu\varepsilon}{4L}$.*

The proof for this theorem is placed in Appendix C.3.

Further, let us give a couple of remarks about obtained results.

Remark 13. *Note, that variance of the stochastic gradient can be decreased through batching technique (see Allen-Zhu (2017)).*

Remark 14. *Theorems 11 and 12 demonstrates that distribution of data does not significantly increase complexity of problem in comparison with deterministic case (see Section 2). The main problem is adding of additional error because of inexact consensus procedure.*

4 Numerical experiments

To provide efficiency of Proximal Gradient Method we will consider the logistic regression problem with ℓ_1 regularization:

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \log(1+\exp(-y_i a_i^\top, x)) + \lambda\|x\|_1, \tag{8}$$

where $a_i \in \mathbb{R}^d$ is feature vectors, $y_i \in \{1, 1\}$ is class of i-th sample, and λ is coefficient of regularization. We will consider this problem on a3a dataset Chang & Lin (2007).

It is non-smooth convex problem. According to lower bounds, it can not be solved faster than with sublinear rate. It holds for proximal methods. On the other hand, Karimi et al. (2016) demonstrates that such problem satisfy proximal PL condition. Nevertheless, it is well-known that such methods are extra effective for such problem Parikh & Boyd (2014). Here we demonstrate that proximal gradient method significantly outperform subgradient method even if the exact gradient is not available.

Another way to solve problem 8 is use of smoothing Nesterov (2003). In particular, we can use Huber function instead of $|x|$. We add consensus procedure for method from Nesterov (2003) but it has not theoretical guarantees for convergence.

To sum up, we will compare proximal gradient method (PGM), subgradient method (GM), accelerated gradient method Nesterov (2003) (AGM) on smoothed version of equation 8. We will consider this methods in centralized and decentralized case. In the first case, we will add artificial noise. In the second case, we will send between nodes gradient with rounded values. For the decentralized case, we will consider some network where at each time moment not more than 0.1 edges will change.

We take the same value T for all methods. We used batched stochastic gradient in each node that provide unbiased noise. Besides, we round it to obtain bias in inexactness. In the table 1 you can find the best approached qualities by different methods and number of iterations.

δ_1	N	$F(\overline{x}^N) - F^*$	$\|\mathbf{x}^N - \overline{\mathbf{x}}^N\|$
PGM	1347	**0.0007**	3.2×10^{-5}
GM	2495	0.0053	6.2×10^{-5}
AGM	**1233**	0.0128	1.2×10^{-5}

Table 1: Results for logistic regression equation 8 with $\delta + \sigma \leq 10^{-4}$ and $\lambda = 10^{-2}$

We can see that PGM approaches the best quality and at the same time its speed is comparable with AGM. At the same time, this method can not theoretical guarantees and PGM's quality is better.

5 Conclusion

In this paper, we propose analysis for Proximal Gradient method for subclass of non-convex non-smooth functions. We prove that this method is robust to both stochastic noise and finite inexactness. Note

Furthermore, these results were generalized for distributed case. We provide sufficient number of communication rounds to decrease the error of communication for sufficient level. Besides, our estimates depend only on start condition, parameter of algorithm and parameters of function but not the sequence $\{\mathbf{x}^k\}_k$ generated by algorithm.

Finally, we presented results of numerical experiments where we compare proximal gradient method with well-known methods for smoothed problem. Our method outperform almost all methods in the speed and approached quality.

Note, that the presented results can be generalized in different ways. In particular, the next step of such research can be word devoted to non-convex saddle-point problems in distributed and non-distributed cases that takes a lot of applications in modern machine learning.

REFERENCES

Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients. 2020. URL https://api.semanticscholar.org/CorpusID:234358812.

Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. pp. 1200–1205, 06 2017. doi: 10.1145/3055399.3055448.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation, 05 2021.

Aleksandr Beznosikov, Samuel Horvath, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning, 02 2020.

K. A. Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé M Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In *SysML 2019*, 2019. URL https://arxiv.org/abs/1902.01046. To appear.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. ISSN 1935-8237. doi: 10.1561/2200000016. URL http://dx.doi.org/10.1561/2200000016.

C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 07 2007.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146, 08 2013. doi: 10.1007/s10107-013-0677-5.

P. Di Lorenzo and G. Scutari. NEXT: In-network nonconvex optimization. *IEEE Trans. Signal Inf. Process. Netw.*, 2(2):120–136, June 2016.

Yoel Drori, Shoham Sabach, and Marc Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Sergey Guminov and Alexander Gasnikov. Accelerated methods for α-weakly-quasi-convex problems. 10 2017.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19, 09 2016.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. 08 2016.

Donghwan Kim and Jeffrey A. Fessler. Another look at the fast iterative shrinkage/thresholding algorithm (fista). *SIAM Journal on Optimization*, 28(1):223–250, 2018. doi: 10.1137/16M108940X. URL https://doi.org/10.1137/16M108940X.

Ilya Kuruzov and Fedor Stonyakin. *Sequential Subspace Optimization for Quasar-Convex Optimization Problems with Inexact Gradient*, pp. 19–33. 12 2021. ISBN 978-3-030-92710-3. doi: 10.1007/978-3-030-92711-0_2.

Jasper Lee and Paul Valiant. Optimizing star-convex functions. pp. 603–614, 10 2016. doi: 10.1109/FOCS.2016.71.

Minghua Li, K. Meng, and Xiaoqi Yang. Variational analysis of kurdyka-lojasiewicz property by way of outer limiting subgradients, 08 2023.

Qunwei Li, Yi Zhou, Yingbin Liang, and P.K. Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimization. 08 2017.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Z. Li, W. Shi, and M. Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17): 4494–4506, 2019.

X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 5330–5340, 2017.

Jingwei Liang and Carola-Bibiane Schönlieb. Improving fista: Faster, smarter and greedier, 11 2018.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization, 05 2018.

Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37:92–101, 05 2020. doi: 10.1109/MSP.2020.2975210.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *Automatic Control, IEEE Transactions on*, 54:48 – 61, 02 2009. doi: 10.1109/TAC.2008.2009515.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Yu Nesterov. Smooth minimization of non-smooth functions. *University catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers*, 103, 01 2003. doi: 10.1007/s10107-004-0552-5.

Yurii Nesterov and Vladimir G. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527 – 566, 2015. URL `https://api.semanticscholar.org/CorpusID:2147817`.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, jan 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL `https://doi.org/10.1561/2400000003`.

Boris Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878, 12 1963. doi: 10.1016/0041-5553(63)90382-3.

Boris Polyak. *Introduction to Optimization.* 07 2020.

Alexander Rogozin, Anton Novitskii, and Alexander Gasnikov. Decentralized proximal optimization method with consensus procedure, 04 2023.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3027–3036. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/scaman17a.html`.

W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM J. on Optimization*, 25(2):944–966, November 2015.

Fedor Stonyakin, Ilya Kuruzov, and Boris Polyak. Stopping rules for gradient methods for non-convex problems with additive noise in gradient. *Journal of Optimization Theory and Applications*, 198:1–21, 06 2023. doi: 10.1007/s10957-023-02245-w.

Artem Vasin, Alexander Gasnikov, and Vladimir Spokoiny. Stopping rules for accelerated gradient methods with additive noise in gradient, 02 2021.

J. Xu, S. Zhu, Y.-C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *Proceedings of the 54th IEEE Conference on Decision and Control*, pp. 2055–2060, 2015.

J. Xu, Y. Tian, Y. Sun, and G. Scutari. Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69: 3555–3570, 2021. ISSN 1053-587X, 1941-0476. doi: 10.1109/TSP.2021.3086579. URL `https://ieeexplore.ieee.org/document/9447939/`.

K. Yuan, B. Ying, X. Zhao, and A. H. Sayed. Exact diffusion for distributed optimization and learning–part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2018.

Yi Zhou, Zhe Wang, Kaiyi Ji, Yingbin Liang, and Vahid Tarokh. Proximal gradient algorithm with momentum and flexible parameter restart for nonconvex optimization, 02 2020.

A Appendix

B Some Proofs for Stochastic Non-decentralized Optimization

B.1 Proof of Theorem 7

Proof. According to definition of function F and its smoothness, we have the following inequality:

$$\begin{aligned} F(x^{k+1}) &= f(x^{k+1}) + r(x^{k+1}) \\ &\le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k + L\|x^{k+1} - x^k\|^2 + r(x^{k+1}) \\ &= F(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 + r(x^{k+1}) - r(x^k) \\ &\le F(x^k) + \langle g_{x^k}, x^{k+1} - x^k \rangle + L\|x^{k+1} - x^k\|^2 + r(x^{k+1}) - r(x^k) + \frac{1}{2L}\delta^2 \\ &\le F(x^k) - \frac{1}{4L}\tilde{\mathcal{D}}_g(x, L) + \frac{1}{2L}\delta^2, \end{aligned}$$

where the last inequality holds because of the definition $\tilde{\mathcal{D}}_g$. Further, we can use Theorem 6 and obtain the following convergence:

$$F(x^{k+1}) - F(x^*) \le \left(1 - \frac{\mu}{4L}\right)(F(x^k) - F(x^*)) + \frac{3}{4L}\delta^2,$$

Iteratively use of the inequality above gives us the following convergence rate:

$$F(x^{k+1}) - F(x^*) \le \left(1 - \frac{\mu}{4L}\right)^k (F(x^0) - F(x^*)) + \frac{3\delta^2}{\mu}.$$

It gives us final convergence rate and allows obtaining statement of Theorem 7. □

B.2 Proof of Theorem 8

Similarly to the proof of Theorem 6 we can consider the following value:

$$\tilde{\mathcal{D}}_g(x, L) = -2L \min_y \left[\langle g_x, y - x\rangle + \frac{L}{2}\|y - x\| + r(y) - r(x)\right]$$

Note, g_x is random value and consequently, $\tilde{\mathcal{D}}_g(x, L)$ is random too. Because of Assumption 3 we have:

$$\mathbb{E}\langle g_x, x - y\rangle \ge \langle \nabla f(x), x - y\rangle - \frac{\delta^2 + \sigma^2}{L} - \frac{L}{4}\|y - x\|.$$

Using the same estimation as in Theorem 6, we can obtain the statement of the theorem.

B.3 Proof of Theorem 9

Proof. According to definition of function F and its smoothness, we have the following inequality:

$$\begin{aligned} F(x^{k+1}) &= f(x^{k+1}) + r(x^{k+1}) \\ &\le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k + L\|x^{k+1} - x^k\|^2 + r(x^{k+1}) \\ &= F(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 + r(x^{k+1}) - r(x^k) \\ &\le F(x^k) + \langle g_{x^k}, x^{k+1} - x^k \rangle + L\|x^{k+1} - x^k\|^2 + r(x^{k+1}) - r(x^k) + \frac{1}{2L}\|\nabla f(x^k) - g_{x^k}\|^2 \\ &\le F(x^k) - \frac{1}{4L}\tilde{\mathcal{D}}_g(x, L) + \frac{1}{2L}\|\nabla f(x^k) - g_{x^k}\|^2, \end{aligned}$$

Further, using Assumption 4 and Theorem 8 we obtain the following result:

$$\mathbb{E}[F(x^{k+1})F(x^*)|x^k] \le (1 - \frac{\mu}{4L})(F(x^k) - F(x^*)) + 2\frac{\delta^2 + \sigma^2}{L}.$$

Iterative use of the inequality above gives us the required statement of Theorem 9. □

C Distributed Case

C.1 Proof of Theorem 11

Proof. Using Lipschitz condition for function f with constant $L = \frac{1}{n}\sum_i L_i$, we get the following result.

$$\begin{aligned} f(\overline{x}^{k+1}) \leq & f(\overline{x}^k) + \left\langle \nabla f(\overline{x}^k), \overline{x}^{k+1} - \overline{x}^k \right\rangle + \frac{L}{2}\|\overline{x}^{k+1} - \overline{x}^k\|^2 + r(\overline{x}^{k+1}) - r(\overline{x}^k) \\ \leq & f(\overline{x}^k) + \left\langle g_{\overline{x}^k}, \overline{x}^{k+1} - \overline{x}^k \right\rangle + L\|\overline{x}^{k+1} - \overline{x}^k\|^2 + r(\overline{x}^{k+1}) - r(\overline{x}^k) + \frac{1}{2L}\delta^2 \end{aligned} \tag{9}$$

Further, let us state the following Lemma:

Lemma 15. *Let us define* $\tilde{g}_{x_i^k} = \left[W^T \mathbf{g}_{x^k}\right]_i$. *Then if for all previous iterations* $\frac{1}{n}\sum\limits_{i=1}^{n} \left\langle g_{\overline{x}^j} - \tilde{g}_{x_i^j}, x_i^{j+1} - \overline{x}^j \right\rangle \leq \delta_r$ *and consensus error not more than* δ_c, *then after* $T = \left\lceil \frac{\tau}{\lambda} \ln \frac{(4+2L)f(\overline{x}^0) - f(x^*) + (2\delta_r + \delta_c)}{\delta_r \mu} \right\rceil$ *communication steps we obtain the same estimation for this iteration:*

$$\frac{1}{n}\sum_{i=1}^{n} \left\langle g_{\overline{x}^k} - \tilde{g}_{x_i^k}, \overline{x}^{k+1} - \overline{x}^k \right\rangle \leq \delta_r$$

and

$$\|\mathbf{x}^{k+1} - \mathbf{1}_n(\overline{x}^{k+1})^\top\| \leq \delta_r$$

It allows us to obtain $\mathcal{D}$ in our further proof.

The last inequality was obtained from $\langle a, b\rangle \leq \frac{1}{2c}\|a\|^2 + \frac{c}{2}\|b\|^2$ for any vectors a, b and constant c. Further, let us express $\mathcal{D}$ value:

$$\begin{aligned} & \left\langle g_{\overline{x}^k}, \overline{x}^{k+1} - \overline{x}^k \right\rangle + L\|\overline{x}^{k+1} - \overline{x}^k\|^2 + r(\overline{x}^{k+1}) - r(\overline{x}^k) \\ \leq & \frac{1}{n}\sum_{i=1}^{n} \left(\langle g_{\overline{x}^k}, x_i^{k+1} - \overline{x}^k \rangle + L\|x_i^{k+1} - \overline{x}^k\|^2 + r(x_i^{k+1}) - r(\overline{x}^k) \right) \\ \leq & \delta_r + \min_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^{n} \left(\langle \left[W^T \mathbf{g}_{\overline{\mathbf{x}}^k}\right]_i, x_i - \overline{x}^k \rangle + L\|x_i - \overline{x}^k\|^2 + r(x_i) - r(\overline{x}^k) \right) \\ \leq & \delta_r + \min_{\mathbf{x}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \left(\langle \left[W^T \mathbf{g}_{\overline{\mathbf{x}}^k}\right]_i, x_i - \overline{x}^k \rangle + L\|x_i - \overline{x}^k\|^2 + r(x_i) - r(\overline{x}^k) \right) \\ = & \delta_r + \min_{x} \left[\langle g_{\overline{x}^k}, x - \overline{x}^k \rangle + L\|x - \overline{x}^k\|^2 + r(x) - r(\overline{x}^k) \right] \\ = & \delta_r - \frac{1}{2L}\widetilde{\mathcal{D}}(\overline{x}^k, 2L) \end{aligned} \tag{10}$$

where $\delta_r = \frac{1}{n}\sum\limits_{i=1}^{n} \left\langle g_{\overline{x}^k} - g_{x_i^k}, \overline{x}_i^{k+1} - \overline{x}^k \right\rangle$. Uniting results above and equation 6, we have:

$$f(\overline{x}^{k+1}) - f^* \leq \left(1 - \frac{\mu}{4L}\right)(f(\overline{x}^k) - f^*) + \delta_r + \frac{2}{nL}\delta^2$$

Using $\delta_r = \frac{\mu}{4L}\frac{\varepsilon}{4}$ and take corresponding T from Lemma 15 we can obtain the following result:

$$\begin{aligned} f(\overline{x}^{k+1}) - f^* & \leq \left(1 - \frac{\mu}{4L}\right)(f(\overline{x}^k) - f^*) + \frac{\mu}{4L}\frac{\varepsilon}{4} + \frac{2\delta_g^2}{nL} \\ & \leq \left(1 - \frac{\mu}{4L}\right)^k (f(\overline{x}^0) - f^*) + \frac{\varepsilon}{4} + \frac{8\delta_g^2}{n\mu} \\ & \leq \left(1 - \frac{\mu}{4L}\right)^k (f(\overline{x}^0) - f^*) + \frac{3\varepsilon}{4} \end{aligned}$$

This inequality gives us sufficient number of iterations to approach quality ε and it finishes our proof. Let us introduce value:

$$R = \frac{64L^2}{\mu^2}\max[f(\overline{x}^0) - f(x^*), \varepsilon] \tag{11}$$

It gives us sufficient number of iterations: $T = \frac{2\tau}{\lambda}\ln\frac{R}{\varepsilon}$ □

C.2 Proof of Lemma 15

Proof. Using convexity of function $(\cdot)^2$, monotonicty of $\sqrt{\cdot}$ and quaratic growth condition, we obtain the following result:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left\langle g_{\overline{x}^k} - \tilde{g}_{x_i^k}, x_i^{k+1} - \overline{x}^k\right\rangle \leq & \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\left\langle g_{\overline{x}^k} - \tilde{g}_{x_i^k}, \overline{x}_i^{k+1} - \overline{x}^k\right\rangle\right|^2} \\
\leq & \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\|g_{\overline{x}^k} - \tilde{g}_{x_i^k}\|^2\right)\left(\sum_{i=1}^{n}\|\overline{x}_i^{k+1} - \overline{x}^k\|^2\right)} \\
\leq & \frac{1}{\sqrt{n}}\|(I - \prod_{k=(k-1)T}^{kT} W^k)\mathbf{g}_{x^k}\| \cdot \|\overline{x}_i^{k+1} - \overline{x}^k\| \\
\leq & \frac{1}{\sqrt{n}}(1-\lambda)^{\frac{T}{k}}\|\mathbf{x}_i^{k+1} - \nVdash_n(\overline{x}^k)^\top\|\|\mathbf{g}_{x^k}\|.
\end{aligned}$$

Note, that Proximal PL satisfy quadratic growth condition (see Karimi et al. (2016). It means that we can write the following estimations:

$$\begin{aligned}
\|\mathbf{g}_{x^k}\| \leq & \delta + \|\nabla f(x^k)\| \leq \delta + \frac{2L}{\mu}(F(x^k) - F(x^*)), \\
\|\mathbf{x}_i^{k+1} - \nVdash_n(\overline{x}^k)^\top\| \leq & \frac{2}{\mu}((F(x^{k+1}) - F(x^*)) + (F(x^k) - F(x^*))) + \|\mathbf{x}^{k+1} - \mathbf{1}_n(\overline{x}^k)^\top\| \\
\leq & \frac{2+2L}{\mu}((F(x^{k+1}) - F(x^*)) + (F(x^k) - F(x^*))) + \delta_r.
\end{aligned}$$

Note, that for the fixed value of consensus error and gradient inexactness we have the following estimation:

$$T = \left\lceil \frac{2\tau}{\lambda}\ln\frac{(8L)\max\left(f(\overline{x}^0) - f(x^*), (2\delta_r + \delta)\right]}{\delta_r\mu} \right\rceil$$

to approach the required quality. □

C.3 Proof of Theorem 12

Let us introduce the following notation for conditional mathematical expectation $\mathbb{E}[X] = \mathbb{E}[X|X_j]$.

Firstly, let us note that we can generalize Lemma 15 for the stochastic case.

Lemma 16. *Let us define* $\tilde{g}_{x_i^k} = \left[\prod_{(k-1)T}^{kT}\mathbf{g}_{x^k}\right]_i$. *Then if for all previous iterations* $\mathbb{E}_{j-1}\frac{1}{n}\sum\limits_{i=1}^{n}\left\langle g_{\overline{x}^j} - \tilde{g}_{x_i^j}, x_i^{j+1} - \overline{x}^j\right\rangle \leq \delta_r$ *and consensus error* $\mathbb{E}_{j-1}\|\mathbf{x}^j - \overline{x}^{j-1}\|$ *not more than* δ_c, *then after* $T = \left\lceil \frac{\tau}{\lambda}\ln\frac{(4+2L)f(\overline{x}^0) - f(x^*) + (2\delta_r + \frac{\delta^2+\sigma^2}{n\mu})}{\delta_r\mu} \right\rceil$ *communication steps we obtain the same estimation for this iteration:*

$$\mathbb{E}_k\frac{1}{n}\sum_{i=1}^{n}\left\langle g_{\overline{x}^k} - \tilde{g}_{x_i^k}, \overline{x}^{k+1} - \overline{x}^k\right\rangle \leq \delta_r$$

and

$$\mathbb{E}_k\|\mathbf{x}^{k+1} - \mathbf{1}_n(\overline{x}^{k+1})^\top\| \leq \delta_r$$

In similar way we can estimate the following value, using the same decomposition as in proof of Theorem 11:

$$\mathbb{E}_k \left\langle g_{\overline{x}^k}, \overline{x}^{k+1} - \overline{x}^k \right\rangle + L\|\overline{x}^{k+1} - \overline{x}^k\|^2 + r(\overline{x}^{k+1}) - r(\overline{x}^k) \le \delta_r - \mathbb{E}_k \frac{1}{2L}\widetilde{\mathcal{D}}(\overline{x}^k, 2L) \tag{12}$$

Uniting the results above and Theorem 8, we have:

$$\mathbb{E}f(\overline{x}^{k+1}) - f^* \le \left(1 - \frac{\mu}{4L}\right)(\mathbb{E}f(\overline{x}^k) - f^*) + \delta_r + \frac{4}{nL}(\delta^2 + \sigma^2)$$

If we take the same $\delta_r = \frac{\mu}{4L}$ then we have that N iterations and T communications per iteration is enough to approach required quality where N is defined in statement of Theorem 12.

C.4 Proof of Lemma 16

Note, that the key estimation holds the same

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle g_{\overline{x}^k} - \tilde{g}_{x_i^k}, x_i^{k+1} - \overline{x}^k \right\rangle \le \frac{1}{\sqrt{n}}(1-\lambda)^{\frac{T}{k}}\|\mathbf{x}_i^{k+1} - \nVdash_n(\overline{x}^k)^\top\|\|\mathbf{g}_{x^k}\|.$$

Further, note that

$$\mathbb{E}\|g_{x^k} - \nabla f(x^k)\|^2 \le 2\delta^2 + 2\sigma^2$$

. In other words, we have the same estimation for multipliers as in proof of Lemma 15 except of additive value:

$$\begin{aligned} \|\mathbf{g}_{x^k}\| \le &\delta + \|\nabla f(x^k)\| \le 2\sigma + 2\delta + \frac{2L}{\mu}(F(x^k) - F(x^*)), \\ \|\mathbf{x}_i^{k+1} - \nVdash_n(\overline{x}^k)^\top\| \le &\frac{2+2L}{\mu}((F(x^{k+1}) - F(x^*)) + (F(x^k) - F(x^*))) + \delta_r + 2\delta + 2\sigma. \end{aligned}$$

It gives us required number of communications round.