Enhancing the convergence estimation of Local SGD for quadratic-like functions / Local SGD converges faster for quadratic-like functions independent of the Hessian

Andrei Sadchikov[1], **Aleksandr Beznosikov**[1, 2], and **Alexander Gasnikov**[1, 3]

[1]MIPT
[2]University Two
[3]University Three

Abstract

One of the challenges of Federated learning is finding the right balance in communication frequency: too infrequent communications lead to worse convergence, while too frequent ones require significant overhead (time for data transmission and network load). Woodworth et al. (2020) proved that for the case where the objective function is quadratic, the communication frequency does not affect the upper bound on the convergence rate. In this work, we focus on generalizing these results, providing an analysis in the case where the objective function is the sum of a quadratic function and an arbitrary remainder.

1 Introduction

1.1 General Words

In the ever-evolving landscape of machine learning, we've witnessed the emergence of enormous models like Gemini, boasting trillions of parameters that push the boundaries of computational capacity. Given the impracticality of training such models on a single device, modern machine learning has embraced Federated Learning, a concept initially introduced by McMahan et al. (2017). This approach involves distributing data across multiple devices, with each device conducting local computations and subsequently communicating to collectively achieve the final result.

However, the challenge of communication frequency remains unresolved in Federated Learning. Insufficient communication may lead to divergence, while overly frequent communication poses its own set of issues. Imagine tackling an optimization problem in a space with a dimension around 10^9, which occurs often in practice (Shahid et al., 2021). Each time we compute a gradient locally, it ends up being gigabytes in size, which makes sending it over for transmission quite a challenge. Thus, striking the delicate balance in communication frequency becomes the key to success in Federated Learning.

To address this challenge, the scientific community has developed sophisticated federated learning frameworks capable of performing well even with infrequent communication. Among the most notable are SCAFFOLD (Karimireddy et al., 2019) and FedAC (Yuan and Ma, 2020). These algorithms are widely utilized in practice since they save a lot of time on communication. They also show efficiency when data distribution varies among computational devices (Darzidehkalani et al., 2022). Unfortunately, due to their complexity, implementation and convergence analysis of such algorithms can be extremely difficult.

On the other hand, the simplest and most classical algorithm that also utilizes the idea of rare communications, named Local SGD (also known as FedAvg or Parallel SGD), has been proven to be as efficient as these intricate algorithms when data distribution is similar (**?**). The core concept of Local SGD can be described as follows: each participating device conducts a few steps of SGD locally, and then the devices communicate with each other by averaging their models' weights (see 1.2 for more details).

Even though the idea is not new and was introduced more than two decades ago by Mangasarian (1995), Local SGD is now again in the center of attention. Its pluses made the algorithm attractive for both practitioners, due to the communication complexity, and theoreticians, due to the simplicity of its analysis.

In environments where the distribution of data remains similar, such as when computations are distributed within a computational cluster or among processor cores within a single device, Local SGD finds extensive application. This classical algorithm remains actively utilized in modern machine learning applications such as Natural Language Processing (NLP) and computer vision. Recent studies (Liu et al., 2024; Do, 2022) have highlighted its efficiency and effectiveness in scenarios with identical data distributions. With its straightforward approach and strong performance, Local SGD continues to play a pivotal role in training large models. In this paper, we concentrate on analyzing convergence rates for Local SGD.

1.2 Problem Formulation

Now, let us formally introduce the task we are solving. Consider a scenario where M devices $1, 2, \ldots, m$, collectively solving an optimization problem to find:

$$x_* := \arg\min_{x \in \mathbb{R}^d} F(x)$$

Here, $F(x)$ is defined as the expected value over a distribution $\mathcal{D}$:

$$F(x) := \mathbb{E}_{z \sim \mathcal{D}}[F(x, z)]$$

Our focus lies on a first-order stochastic oracle, where we have access to the stochastic estimate of ∇F on each node. We denote this estimate as $\nabla F(x, z)$, with z representing a sample from $\mathcal{D}$.

This formulation captures a wide range of practical problems, such as Empirical Risk Minimization (Vapnik, 1991). In our case, F denotes the empirical risk (i.e., the average of losses across some data samples) and z denotes the indices of data samples.

Here it is important to underline the method by which we measure communication complexity for Federated Learning methods. Unlike classical optimization, where complexity was considered as the number of times the oracle is called, i.e., computing local gradients $\nabla F(x^m, z^m)$, we also care about communication complexity. Generally, we define it as the quantity of information we transmit, or the number of communications in our case. So, in this work, we will emphasize reducing the number of communication rounds R or increasing the number of SGD steps between communications, denoted as H, while keeping the total number of SGD steps T unchanged.

Now we are ready to get familiar with the formal definition of Local SGD.

Algorithm 1 Local SGD

```
Input: Initial vector $x_0 = x_0^m$ for all $m \in \{1, \ldots, M\}$; stepsize $\gamma \geqslant 0$
for $t = 1, \ldots, T = R \cdot H$ do
    for $m = 1, \ldots, M$ in parallel do
        Sample $z_t^m \sim \mathcal{D}$.
        Evaluate stochastic gradient $\boldsymbol{g}_t^m = \nabla F(x_t^m, z_t^m)$.
    end for
    if $t + 1 \mod H = 0$ then
        $x_{t+1}^m = \frac{1}{M} \sum_{j=1}^{M} (x_t^j - \gamma \boldsymbol{g}_t^j)$
    else
        $x_{t+1}^m = x_t^m - \gamma \boldsymbol{g}_t^m$
    end if
end for
```

1.3 Related Work

In both scenarios of identical and heterogeneous settings, extensive research has been conducted across various contexts. Li et al. (2020) and Haddadpour et al. (2020) analyzed the non-convex setting under bounded gradient norms. As expected, the further we deviate from the identical and strongly convex settings, the poorer the results become. However, in this work, we will focus solely on convex and strongly convex cases.

The theoretical aspect of the convex case has been extensively studied. Among the most remarkable works, we can mention Stich (2019), who conducted analysis under the assumption of bounded gradient norms. Subsequently, a considerable amount of research was conducted on integrating Local SGD with various Federated Learning techniques, such as momentum, quantization, and restarting (Reisizadeh et al., 2020; Sharma et al., 2020; Yu et al., 2018).

Notably, Khaled et al. (2022) and Woodworth et al. (2020) provided the best-known estimate without any restrictions on the Hessian. In our study, we extend these findings, particularly in cases where the objective function F exhibits proximity to a quadratic form. Importantly, our approach does not assume Hessian Lipschitzness as other works Yuan and Ma (2020), thus representing a generalization of previous research efforts.

It's worth noting that the majority of existing research, with the exception of Spiridonoff et al. (2021), operates under the assumption that the variance of stochastic gradients is uniformly bounded. This assumption posits the existence of a constant σ such that $\|\nabla F(x,z) - \nabla F(x)\|^2 \leqslant \sigma^2$. However, this assumption is often unrealistic in real-world scenarios. For example, it does not hold for Coordinate SGD (Gorbunov et al., 2019), where only a subset of partial derivatives is computed instead of the full gradient. Therefore, one of our objectives was to analyze Local SGD for quadratic-like functions under assumption 2, which seems to be very general and holds almost always.

Table 1: Summary of similar works

Reference	**Not-Lipschitz Hessian**	**Unbounded Gradient**	**Noise model**	**Convexity**	**Convergence rate,** $\mathbb{E}[F(.)] - F(x_=) \leqslant$
Stich (2019)	×(?)	×	Uniform	$\mu = 0$	-
				$\mu > 0$	$\mathcal{O}\left(\frac{D^2}{R^3} + \frac{\sigma^2}{\mu MT} + \frac{\kappa G^2}{\mu R^2}\right)$ (G-note)
Yuan and Ma (2020)	×	✓	Uniform	$\mu = 0$	$\tilde{\mathcal{O}}\left(\frac{LD^2}{T} + \frac{\sigma D}{\sqrt{MT}} + \frac{Q^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{5}{3}}}{T^{\frac{1}{3}}R^{\frac{1}{3}}}\right)$ (Q-note)
				$\mu > 0$	$\tilde{\mathcal{O}}\left(\text{exp. decay} + \frac{\sigma^2}{\mu MT} + \frac{Q^2\sigma^4}{\mu^5T^2R^2}\right)$
Spiridonoff et al. (2021)	✓	✓	Uniform with strong growth (ρ-note)	$\mu = 0$	-
				$\mu > 0$	$\mathcal{O}\left(\frac{(1+\rho\kappa^2\ln(TR^{-2}))D^2}{\kappa^{-2}T^2} + \frac{\kappa\sigma^2}{\mu MT} + \frac{\kappa^2\sigma^2}{\mu TR}\right)$
Khaled et al. (2022)	✓	✓	Uniform	$\mu = 0$	$\mathcal{O}\left(\frac{D^2}{\sqrt{MT}} + \frac{\sigma^2}{L\sqrt{MT}} + \frac{\sigma^2 M}{LR}\right)$ (improve-lr)
				$\mu > 0$	$\tilde{\mathcal{O}}\left(\frac{LD^2}{T^2} + \frac{L\sigma^2}{\mu^2MT} + \frac{L^2\sigma^2}{\mu^3TR}\right)$
This work	✓	✓	Uniform	$\mu = 0$	$\mathcal{O}\left(\frac{D^2}{\sqrt{MT}} + \frac{\sigma^2}{L\sqrt{MT}} + \frac{\boldsymbol{\varepsilon}\sigma^2 M}{LR}\right)$ (ε-note)
				$\mu > 0$	$\tilde{\mathcal{O}}\left(\frac{LD^2}{T^2} + \frac{L\sigma^2}{\mu^2MT} + \frac{\boldsymbol{\varepsilon}L^2\sigma^2}{\mu^3TR}\right)$
	✓	✓	Uniform with strong growth	$\mu = 0$	-
				$\mu > 0$	-

- General notation: $\mathcal{O}$ omits constant factors; $\tilde{\mathcal{O}}$ omits polylogarithmic and constant factors. $D = \|x_0 - x_*\|$ - initial distance from minimum; σ - variance of stochastic gradient, see ass. 2; μ is a strong convexity constant and L is a Lipshitz gradient constant, see ass. 1; M - number of workers; R - number of communication rounds; T - total number of iterations. For more detailed explanation see table 2
- (ρ-note) ρ came from ass. 2
- (improve-lr) Шаг следует выбрать поинтереснее
- (ε-note) ε shows how far function is from quadratic; for all functions $\varepsilon \leqslant 1$ and for quadratic functions $\varepsilon = 0$. For more details see st. 1 and section 4

As evident from this table, our work achieves a clear improvement in the estimates obtained by Khaled et al. (2022).

1.4 Specifics of the Quadratic Case

Woodworth et al. (2020) demonstrated that in the scenario where the objective function is quadratic, the convergence rate of Local SGD remains independent of communication frequency. This important observation aligns with the lower bounds established for Local SGD, indicating that for purely quadratic functions, further enhancement of the convergence rate is unattainable.

The most promising outcomes in the identical convex case, where the function is not quadratic, were previously achieved under the assumption of Lipschitz Hessian (Yuan and Ma, 2020). This observation resonates with the findings of Yuan and Ma (2020) and Spiridonoff et al. (2021), who concluded that closer proximity to a purely quadratic case correlates with improved convergence estimates.

However, a gap still exists between the general case, as analyzed by Khaled et al. (2022), and the challenging assumption of a Lipschitz Hessian. This assumption appears to be invalid for neural networks with ReLU activation functions. Therefore, our primary objective is to offer an analysis that bridges this gap. To achieve this, we introduce the concept of "**quadraticity**", i.e. ε (1), which provides insight into the proximity of a function to a quadratic form without imposing additional assumptions and conduct our analysis with respect to the "quadraticity" of the objective function.

2 Contributions

Our contributions are following:

a) We demonstrate that improvements in convergence rates for Local SGD on quadratic-like functions are achievable not only in cases with a Lipschitz Hessian

b) We are the first to demonstrate that such acceleration holds under assumption 2

In this work, we mainly focus on the improvement in the last term (which represents dependency on communication frequency) of the convergence rates provided by Khaled et al. (2022) in prospect of statement 1.

Theorem 1. Under assumptions 1 and 2, decomposing F as announced in 1 and considering $\mu > 0$ and $\gamma \leqslant \frac{1}{6L}$ gives:

$$\mathbb{E}\left\|\bar{x}_T - x_*\right\|^2 \leqslant (1-\gamma\mu)^T \left\|x_0 - x_*\right\|^2 + \frac{\gamma\sigma^2}{\mu M} + \frac{2\varepsilon\gamma^2\sigma^2 L(H-1)}{\mu} \tag{1}$$

Corollary 1. As it was previously shown in Woodworth et al. (2020), an important special case is is achieved when epsilon is equal to zero. Than from the Theorem 1 it follows that:

$$\mathbb{E}\left\|\bar{x}_T - x_*\right\|^2 \leqslant (1-\gamma\mu)^T \left\|x_0 - x_*\right\|^2 + \frac{\gamma\sigma^2}{\mu M} \tag{2}$$

Thus, if F is a quadratic function, the upper bound on the rate of convergence is independent of the communication frequency.

Corollary 2. Considering γ as a function of t and choosing γ_t as in Theorem 2 from Woodworth et al. (2020), we obtain:

$$\mathbb{E}[F(\bar{x}_t)] - F(x_*) \leqslant c \cdot \left(\exp\left(-\frac{\mu T}{4L}\right) + \frac{\sigma^2}{\mu M T} + \frac{\varepsilon\sigma^2}{\mu^2 T R}\right) \tag{3}$$

Where $c \in \mathbb{R}$ is some universal constant.

Theorem 2. Under assumptions 1 and 2, decomposing F as announced in 1 and considering $\mu = 0$ and $\gamma \leqslant \frac{1}{6L}$ we have:

$$\mathbb{E}[F(\hat{x}_T)] - F(x_*) \leqslant \frac{2}{\gamma T} + \frac{2\gamma\sigma^2}{M} + 12\varepsilon\gamma^2 L\sigma^2(H-1) \tag{4}$$

Where $\hat{x}_T = \frac{1}{T}\sum_{t=1}^{T} \bar{x}_t$

Corollary 3. Choosing $\gamma = \frac{\sqrt{M}}{6L\sqrt{T}}$ and assuming $M \leqslant T$ as in Khaled et al. (2022) yields:

$$\mathbb{E}[F(\hat{x}_T)] - F(x_*) \leqslant \frac{12\left\|x_0 - x_*\right\|^2}{\sqrt{MT}} + \frac{\sigma^2}{3L\sqrt{MT}} + \frac{\varepsilon\sigma^2 M}{3LR} \tag{5}$$

3 Settings

Assumption 1. Assume that F is μ-convex and L-smooth. That is, $\forall x, y \in \mathbb{R}^d$,

$$\frac{\mu}{2}\left\|x-y\right\|^2 \leqslant F(y) - F(x) - \langle \nabla F(x), y-x\rangle \leqslant \frac{L}{2}\left\|x-y\right\|^2$$

Corollary 4. Under assumption 1

$$\frac{1}{2L}\left\|\nabla F(x) - \nabla F(y)\right\|^2 \leqslant F(y) - F(x) - \langle \nabla F(x), y-x\rangle$$

Proof. This is Theorem 2.1.5 in Nesterov (2014)

Assumption 2. Exist such constants σ and ρ that:

$$\mathbb{E}_{z\sim\mathcal{D}} \left\|\nabla F(x,z) - \nabla F(x)\right\|^2 \leqslant \sigma^2 + \rho \left\|\nabla F(x)\right\|^2$$

Assumption 3. The stochastic gradient is an unbiased estimate of its expectation:

$$\mathbb{E}_{z\sim\mathcal{D}}[\nabla(F(x,z)] = F(x)$$

Statement 1. Objective function can be decomposed as $F = Q + R$, where Q is a quadratic function that is μ_Q-convex and L_Q-smooth, and R is μ_R-convex and L_R-smooth. Than we denote parameter $\varepsilon := \frac{L_R}{L}$ which gives us an idea of how far F is from a quadratic function.

Note that such decomposition always takes place beacuse we can take $Q = 0$ and $R = F$.

Corollary 5. In the prospect of statement 1, following takes place:

a) ∇Q is a linear function

b) $\varepsilon \leqslant 1$

c) $\mu_Q + \mu_R \leqslant \mu$

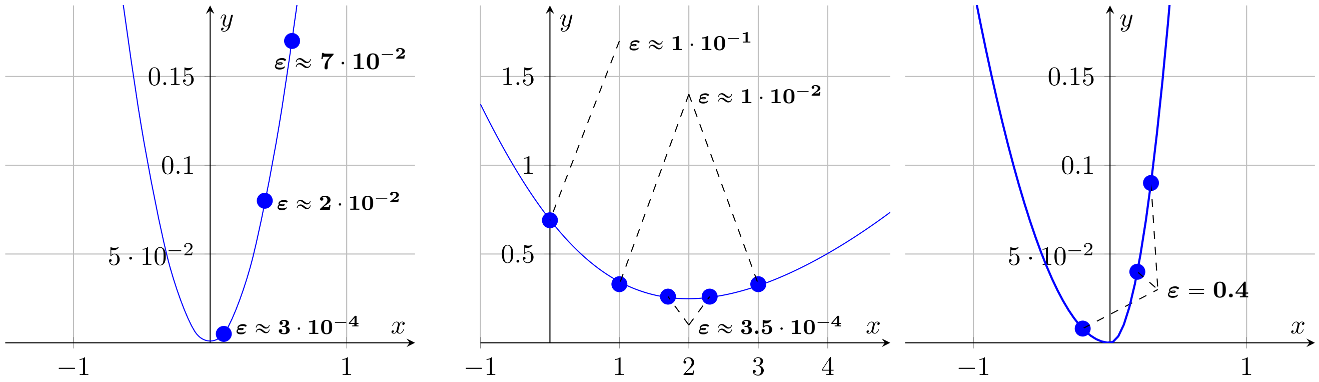
4 Discussion

Things get more interesting in perspective of expressing $\mathbb{E}[\|\bar{x}_{t+1} - x_*\|^2]$ through $\mathbb{E}[\|\bar{x}_t - x_*\|^2]$ instead of solving recurrent relation and analyzing $\mathbb{E}[\|\bar{x}_T - x_*\|^2]$.

As can be observed from 7:

$$\mathbb{E}\left\|\bar{x}_{t+1} - x_*\right\|^2 \leqslant (1-\gamma\mu)\mathbb{E}\left\|\bar{x}_t - x_*\right\|^2 + \frac{\gamma^2\sigma^2}{M} + 10\varepsilon\gamma L(H-1)\gamma^2\sigma^2 \tag{6}$$

Considering ε as a function of x, we can say that for the functions with Lipschitz Hessian we know that the decay rate of ε is fast. Following graphs illustrate decay rate of ε for some functions:



(a) LogCoshLoss: $y = \ln(\cosh(x))$ (b) LogLoss with l_2 reg. (7) (c) Piecewise quadratic func. $\mathcal{F}$ (8)

Figure 1: Decay of ε when getting closer to minima

In point (b) by LogLoss we mean

$$y = -\ln\left(\frac{1}{1+e^{-x}}\right) + 0.03x^2 \tag{7}$$

And in point (c) we analyze well-known function that yields lower bound estimates for Local SGD (Glasgow et al., 2022).

$$\mathcal{F}(x) = \begin{cases} x^2/5 & \text{if } x < 0 \\ x^2 & \text{if } x \geqslant 0 \end{cases} \tag{8}$$

Here we can notice that for functions (a) and (b) ε decreases rapidly (because they satisfy Lipschitz Hessian assumption), and for function (c) it remains constant. This may be considered as some additional insight into why $\mathcal{F}$ yields lower bound for Local SGD.

From this example we can observe that equation 6 provides valuable insights into convergence rate for different types of functions. However, abandoning the assumption of a Lipschitz Hessian means that we cannot estimate the decay rate of epsilon. Therefore conducting a meaningful asymptotic analysis while maintaining generality seems impossible.

Thus our aim is to present our findings within a broader scope, elucidating the physical interpretation of the convergence rate.

References

Erfan Darzidehkalani, Mohammad Ghasemi-rad, and P.M.A. van Ooijen. Federated learning in medical imaging: Part ii: Methods, challenges, and considerations. *Journal of the American College of Radiology*, 19(8):975–982, 2022. ISSN 1546-1440. doi: https://doi.org/10.1016/j.jacr.2022.03.016. URL https://www.sciencedirect.com/science/article/pii/S1546144022002812.

Thanh-Nghi Do. Imagenet challenging classification with the raspberry pi: An incremental local stochastic gradient descent algorithm, 2022.

Margalit Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective, 2022.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent, 2019.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck R. Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for on-device federated learning. *CoRR*, abs/1910.06378, 2019. URL http://arxiv.org/abs/1910.06378.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data, 2022.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020.

Bo Liu, Rachita Chhaparia, Arthur Douillard, Satyen Kale, Andrei A. Rusu, Jiajun Shen, Arthur Szlam, and Marc'Aurelio Ranzato. Asynchronous local-sgd training for language modeling, 2024.

L. O. Mangasarian. Parallel gradient distribution in unconstrained optimization. *SIAM Journal on Control and Optimization*, 33(6):1916–1925, 1995. doi: 10.1137/S0363012993250220. URL https://doi.org/10.1137/S0363012993250220.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization, 2020.

Osama Shahid, Seyedamin Pouriyeh, Reza M. Parizi, Quan Z. Sheng, Gautam Srivastava, and Liang Zhao. Communication efficiency in federated learning: Achievements and challenges, 2021.

Pranay Sharma, Swatantra Kafle, Prashant Khanduri, Saikiran Bulusu, Ketan Rajawat, and Pramod K. Varshney. Parallel restarted spider – communication efficient distributed nonconvex optimization with optimal computation complexity, 2020.

Artin Spiridonoff, Alex Olshevsky, and Ioannis Ch. Paschalidis. Communication-efficient SGD: from local SGD to one-shot averaging. *CoRR*, abs/2106.04759, 2021. URL https://arxiv.org/abs/2106.04759.

Sebastian U. Stich. Local sgd converges fast and communicates little, 2019.

V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd?, 2020.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning, 2018.

Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *CoRR*, abs/2006.08950, 2020. URL https://arxiv.org/abs/2006.08950.

Contents

5 Proofs

5.1 Notation

To begin with, let us introduce useful contractions. By the capital Latin letters $F,\ Q,\ R$ we will denote functions. Corresponding stochastic gradients will be represented by the bold Gothic lowercase Latin letters $\boldsymbol{g},\ \boldsymbol{q},\ \boldsymbol{\mathfrak{r}}$, and the straight lowercase Latin letters $g,\ q,\ r$ will denote the expectations of the gradients.

A bar above the letter (i.e., $\bar{g}$) will indicate that we take the average of this value among all devices.

Symbol	Meaning
M	Number of devices participating in the algorithm
H	Communication frequency - devices communicate and average their weights every H iterations
x_t^m	Local weight on the device m at time t
$\bar{x}_t$	$\frac{1}{M}\sum_{m=1}^{M} x_t^m$ - average of the weights among all devices at time t
$\nabla F(x)$	$\mathbb{E}[\nabla F(x,z)]$ - expectation of a stochastic gradient
$\bar{F}_t,\ \bar{Q}_t,\ \bar{R}_t$	$\frac{1}{M}\sum_{m=1}^{M} F(x_t^m),\ \frac{1}{M}\sum_{m=1}^{M} Q(x_t^m),\ \frac{1}{M}\sum_{m=1}^{M} R(x_t^m)$ respectively
$\boldsymbol{g}_t^m,\ \boldsymbol{q}_t^m,\ \boldsymbol{\mathfrak{r}}_t^m$	$\nabla F(x_t^m, z_t^m),\ \nabla Q(x_t^m, z_t^m),\ \nabla R(x_t^m, z_t^m)$ - corresponding stochastic gradient at time t on device m
$\bar{\boldsymbol{g}}_t,\ \bar{\boldsymbol{q}}_t,\ \bar{\boldsymbol{\mathfrak{r}}}_t$	$\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{g}_t^m,\ \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{q}_t^m,\ \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\mathfrak{r}}_t^m$ - average of stochastic gradients at time t
$g_t^m,\ q_t^m,\ r_t^m$	$\nabla F(x_t^m),\ \nabla Q(x_t^m),\ \nabla R(x_t^m)$ - expected value of stochastic gradients at time t on device m (namely, $\mathbb{E}[\boldsymbol{g}_t^m],\ \mathbb{E}[\boldsymbol{q}_t^m],\ \mathbb{E}[\boldsymbol{\mathfrak{r}}_t^m]$)
$\bar{g}_t,\ \bar{q}_t,\ \bar{r}_t$	$\frac{1}{M}\sum_{m=1}^{M} g_t^m,\ \frac{1}{M}\sum_{m=1}^{M} q_t^m,\ \frac{1}{M}\sum_{m=1}^{M} r_t^m$ - average of expected values of gradients at time t
r_*, q_*, R_*, Q_*	$\nabla R(x_*), \nabla Q(x_*), R(x_*), Q(x_*)$ - values at the optimum
V_t	$\frac{1}{M}\sum_{m=1}^{M} \lVert x_t^m - \bar{x}_t \rVert^2$ - mean deviation of x_t^m from $\bar{x}_t$
$D,\ \lVert r_t \rVert$	$\lVert \bar{x}_t - x_* \rVert$ - distance to the optimum at time t

Table 2: Notation summary

5.2 Technical lemmas

Before demonstrating the claimed facts, let's first establish some technical results. In this section we will follow the path of proving Lemma 3.1 from Stich (2019).

Lemma 1.

$$\lVert \bar{x}_t - x_* - \gamma \bar{g}_t \rVert^2 = \lVert \bar{x}_t - x_* \rVert^2 + \gamma^2 \lVert \bar{q}_t + \bar{r}_t - q_* - r_* \rVert^2 - 2\gamma \langle \bar{x}_t - x_*, \bar{q}_t \rangle - 2\gamma \langle \bar{x}_t - x_*, \bar{r}_t \rangle$$

Proof.

$$\begin{aligned}
\|\bar{x}_t - x_* - \gamma \bar{g}_t\|^2 &= \|\bar{x}_t - x_*\|^2 + \gamma^2 \|\bar{g}_t\|^2 - 2\gamma\langle \bar{x}_t - x_*, \bar{g}_t\rangle \\
&= \|\bar{x}_t - x_*\|^2 + \gamma^2 \|\bar{g}_t\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, \nabla F(x_t^m)\rangle \\
&= \|\bar{x}_t - x_*\|^2 + \gamma^2 \left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F(x_t^m)\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, \nabla F(x_t^m)\rangle \\
&= \|\bar{x}_t - x_*\|^2 + \gamma^2 \left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F(x_t^m) - \nabla F(x_*)\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, \nabla F(x_t^m)\rangle \\
&= \|\bar{x}_t - x_*\|^2 + \gamma^2 \left\|\frac{1}{M}\sum_{m=1}^{M}\nabla Q(x_t^m) + \nabla R(x_t^m) - \nabla Q(x_*) - \nabla R(x_*)\right\|^2 \\
&- \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, \nabla Q(x_t^m)\rangle - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, \nabla R(x_t^m)\rangle \\
&= \|\bar{x}_t - x_*\|^2 + \gamma^2 \|\bar{q}_t + \bar{r}_t - q_* - r_*\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, q_t^m\rangle - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle \bar{x}_t - x_*, r_t^m\rangle \\
&= \|\bar{x}_t - x_*\|^2 + \gamma^2 \|\bar{q}_t + \bar{r}_t - q_* - r_*\|^2 - 2\gamma\langle \bar{x}_t - x_*, \bar{q}_t\rangle - 2\gamma\langle \bar{x}_t - x_*, \bar{r}_t\rangle
\end{aligned}$$

□

Lemma 2. Bounding the gradient norm

$$\|\bar{q}_t + \bar{r}_t - q_* - r_*\|^2 \leqslant 2L_Q(1+\zeta)(Q(\bar{x}_t) - Q_* - \langle q_*, \bar{x}_t - x_*\rangle) + 2L_R(1+\frac{1}{\zeta})(\bar{R}_t - R_* - \langle r_*, \bar{x}_t - x_*\rangle)$$

Proof.

By the generalized Cauchy inequality:

$$\|\bar{q}_t + \bar{r}_t - q_* - r_*\|^2 = \|\bar{q}_t - q_*\|^2 + \|\bar{r}_t - r_*\|^2 + 2\langle \bar{q}_t - q_*, \bar{r}_t - r_*\rangle \tag{9}$$

$$\leqslant \|\bar{q}_t - q_*\|^2 + \|\bar{r}_t - r_*\|^2 + \zeta\|\bar{q}_t - q_*\|^2 + \frac{1}{\zeta}\|\bar{r}_t - r_*\|^2 \tag{10}$$

$$= (1+\zeta)\|\bar{q}_t - q_*\|^2 + (1+\frac{1}{\zeta})\|\bar{r}_t - r_*\|^2 \tag{11}$$

By the L - smoothness (corollary 4):

$$(1+\zeta)\|\bar{q}_t - q_*\|^2 + (1+\frac{1}{\zeta})\|\bar{r}_t - r_*\|^2 \tag{12}$$

$$\leqslant 2L_Q(1+\zeta)(Q(\bar{x}_t) - Q_* - \langle q_*, \bar{x}_t - x_*\rangle) + 2(1+\frac{1}{\zeta})L_R(\bar{R}_t - R_* - \langle r_*, \bar{x}_t - x_*\rangle) \tag{13}$$

Combining together (11) and (13), we obtain:

$$\|\bar{q}_t + \bar{r}_t - q_* - r_*\|^2 \leqslant 2L_Q(1+\zeta)(Q(\bar{x}_t) - Q_* - \langle q_*, \bar{x}_t - x_*\rangle) + 2L_R(1+\frac{1}{\zeta})(\bar{R}_t - R_* - \langle r_*, \bar{x}_t - x_*\rangle) \tag{14}$$

□

Lemma 3.

$$-2\langle \bar{x}_t - x_*, \bar{q}_t\rangle \leqslant 2Q_* - 2Q(\bar{x}_t) - \mu_Q\|\bar{x}_t - x_*\|^2 \tag{15}$$

Proof. By μ_Q - convexity:

$$-2\langle \bar{x}_t - x_*, \bar{q}_t\rangle = 2\langle x_* - \bar{x}_t, \bar{q}_t\rangle \tag{16}$$

$$\leqslant 2Q_* - 2Q(\bar{x}_t) - \mu_Q\|\bar{x}_t - x_*\|^2 \tag{17}$$

□

Lemma 4.

$$-2\langle \bar{x}_t - x_*, \bar{r}_t \rangle \leqslant (\bar{R}_t - R_*)(\frac{1}{p} - 2) + 2pL_R V_t - \mu_R \left\| x_* - \bar{x}_t \right\|^2 - \frac{1}{p}\langle r_*, \bar{x}_t - x_* \rangle \tag{18}$$

Proof.

$$\begin{aligned} -2\langle \bar{x}_t - x_*, \bar{r}_t \rangle &= 2\langle x_* - \bar{x}_t, \bar{r}_t \rangle & (19)\\ &= \frac{2}{M}\sum_{m=1}^{M} \langle x_* - \bar{x}_t, r_t^m \rangle = \frac{2}{M}\sum_{m=1}^{M} \langle x_* - x_t^m + x_t^m - \bar{x}_t, r_t^m \rangle & (20)\\ &= \frac{2}{M}\sum_{m=1}^{M} \langle x_* - x_t^m, r_t^m \rangle + \frac{2}{M}\sum_{m=1}^{M} \langle x_t^m - \bar{x}_t, r_t^m \rangle & (21) \end{aligned}$$

First part, by μ_R - convexity and Jensen's inequality:

$$\frac{2}{M}\sum_{m=1}^{M} \langle x_* - x_t^m, r_t^m \rangle \leqslant \frac{1}{M}\sum_{m=1}^{M} 2R_* - 2R(x_t^m) - \mu_R \left\| x_* - x_t^m \right\|^2 \tag{22}$$

$$\leqslant \frac{1}{M}\sum_{m=1}^{M} 2R_* - 2R(x_t^m) - \mu_R \left\| x_* - \bar{x}_t \right\|^2 = 2R_* - 2\bar{R}_t - \mu_R \left\| x_* - \bar{x}_t \right\|^2 \tag{23}$$

Second part, by the generalized Cauchy inequality:

$$\frac{2}{M}\sum_{m=1}^{M} \langle x_t^m - \bar{x}_t, r_t^m \rangle = \frac{2}{M}\sum_{m=1}^{M} \langle x_t^m - \bar{x}_t, r_t^m - r_* \rangle + \frac{2}{M}\sum_{m=1}^{M} \langle x_t^m - \bar{x}_t, r_* \rangle \tag{24}$$

$$= \frac{2}{M}\sum_{m=1}^{M} \langle x_t^m - \bar{x}_t, r_t^m - r_* \rangle \leqslant \frac{2}{M}\sum_{m=1}^{M} pL_R \left\| x_t^m - \bar{x}_t \right\|^2 + \frac{1}{M}\sum_{m=1}^{M} \frac{1}{2pL_R} \left\| r_t^m - r_* \right\|^2 \tag{25}$$

$$= 2pL_R V_t + \frac{1}{M}\sum_{m=1}^{M} \frac{1}{2pL_R} \left\| r_t^m - r_* \right\|^2 \tag{26}$$

By the corollary 4:

$$\left\| r_t^m - r_* \right\|^2 \leqslant 2L_R(R(x_t^m) - R_* - \langle r_*, x_t^m - x_* \rangle) \tag{27}$$

Substituting (27) into (26), we complete the second part:

$$\frac{2}{M}\sum_{m=1}^{M} \langle x_t^m - \bar{x}_t, r_t^m \rangle \leqslant 2pL_R V_t + \frac{1}{pM}\sum_{m=1}^{M} R(x_t^m) - R_* - \langle r_*, x_t^m - x_* \rangle \tag{28}$$

$$\leqslant 2pL_R V_t + \frac{1}{p}(\bar{R}_t - R_* - \langle r_*, \bar{x}_t - x_* \rangle) \tag{29}$$

Combining together (21), (23), (29), we gain:

$$\begin{aligned} -2\langle \bar{x}_t - x_*, \bar{r}_t \rangle &\leqslant 2R_* - 2\bar{R}_t - \mu_R \left\| x_* - \bar{x}_t \right\|^2 + 2pL_R V_t + \frac{1}{p}(\bar{R}_t - R_* - \langle r_*, \bar{x}_t - x_* \rangle) & (30)\\ &= (\bar{R}_t - R_*)(\frac{1}{p} - 2) + 2pL_R V_t - \mu_R \left\| x_* - \bar{x}_t \right\|^2 - \frac{1}{p}\langle r_*, \bar{x}_t - x_* \rangle & (31) \end{aligned}$$

□

Lemma 5.

$$\begin{aligned} &\left\| \bar{x}_t - x_* - \gamma \bar{g}_t \right\|^2 \leqslant \\ &(1 - \gamma\mu) \left\| \bar{x}_t - x_* \right\|^2 + 2\gamma p L_R V_t \\ &+ 2\gamma(A(\zeta) - 1)(Q(\bar{x}_t) - Q_*) \\ &+ 2\gamma(B(\zeta) - 1)(\bar{R}_t - R_*) \\ &- 2\gamma A(\zeta)\langle q_*, \bar{x}_t - x_* \rangle \\ &- 2\gamma B(\zeta)\langle r_*, \bar{x}_t - x_* \rangle \end{aligned} \tag{32}$$

Where A and B are functions of $\zeta \in \mathbb{R}$ such that:

$$A(\zeta) = \gamma L_Q(1 + \zeta)$$

$$B(\zeta) = \gamma L_R(1 + \frac{1}{\zeta}) + \frac{1}{2p}$$

Proof. Substituting the results of Lemmas 2, 3 and 4 into Lemma 1 and doing some algebraic manipulations:

$$\begin{aligned}
&\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 \leqslant \|\bar{x}_t - x_*\|^2 \\
&+ 2\gamma^2 L_Q(1+\zeta)(Q(\bar{x}_t) - Q_* - \langle q_*, \bar{x}_t - x_*\rangle) \\
&+ 2\gamma^2 L_R(1+\frac{1}{\zeta})(\bar{R}_t - R_* - \langle r_*, \bar{x}_t - x_*\rangle) \\
&+ \gamma(2Q_* - 2Q(\bar{x}_t) - \mu_Q \|\bar{x}_t - x_*\|^2) \\
&+ \gamma(\bar{R}_t - R_*)(\frac{1}{p} - 2) + 2\gamma p L_R V_t - \gamma\mu_R \|x_* - \bar{x}_t\|^2 - \frac{\gamma}{p}\langle r_*, \bar{x}_t - x_*\rangle \\
&= (1 - \gamma\mu_Q - \gamma\mu_R)\|\bar{x}_t - x_*\|^2 + 2\gamma p L_R V_t \\
&+ (Q(\bar{x}_t) - Q_*)\Big[2\gamma^2 L_Q(1+\zeta) - 2\gamma\Big] + (\bar{R}_t - R_*)\Big[2\gamma^2 L_R(1+\frac{1}{\zeta}) - 2\gamma + \frac{\gamma}{p}\Big] \\
&- 2\langle q_*, \bar{x}_t - x_*\rangle\Big[\gamma^2 L_Q(1+\zeta)\Big] - 2\langle r_*, \bar{x}_t - x_*\rangle\Big[\gamma^2 L_R(1+\frac{1}{\zeta}) + \frac{\gamma}{2p}\Big]
\end{aligned} \tag{33}$$

Using the result of 5:

$$(1 - \gamma\mu_Q - \gamma\mu_R)\|\bar{x}_t - x_*\|^2 \leqslant (1 - \gamma\mu)\|\bar{x}_t - x_*\|^2 \tag{34}$$

Combining (34) with (33)

$$\begin{aligned}
&\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 \leqslant (1 - \gamma\mu)\|\bar{x}_t - x_*\|^2 + 2\gamma p L_R V_t \\
&+ (Q(\bar{x}_t) - Q_*)\Big[2\gamma^2 L_Q(1+\zeta) - 2\gamma\Big] + (\bar{R}_t - R_*)\Big[2\gamma^2 L_R(1+\frac{1}{\zeta}) - 2\gamma + \frac{\gamma}{p}\Big] \\
&- 2\langle q_*, \bar{x}_t - x_*\rangle\Big[\gamma^2 L_Q(1+\zeta)\Big] - 2\langle r_*, \bar{x}_t - x_*\rangle\Big[\gamma^2 L_R(1+\frac{1}{\zeta}) + \frac{\gamma}{2p}\Big] \\
&= (1 - \gamma\mu)\|\bar{x}_t - x_*\|^2 + 2\gamma p L_R V_t \\
&+ 2\gamma(Q(\bar{x}_t) - Q_*)\Big[\gamma L_Q(1+\zeta) - 1)\Big] + 2\gamma(\bar{R}_t - R_*)\Big[\gamma L_R(1+\frac{1}{\zeta}) - 1 + \frac{1}{2p}\Big] \\
&- 2\gamma\langle q_*, \bar{x}_t - x_*\rangle\Big[\gamma L_Q(1+\zeta)\Big] - 2\gamma\langle r_*, \bar{x}_t - x_*\rangle\Big[\gamma L_R(1+\frac{1}{\zeta}) + \frac{1}{2p}\Big]
\end{aligned} \tag{35}$$

By substituting the expressions in square brackets with A and B, we obtain the statement of the lemma. □

Lemma 6. Exists ζ_1 such that $A(\zeta_1) = B(\zeta_1)$ and $A(\zeta_1) - 1 \leqslant 0$

Sublemma 6.1. Exists ζ_1 such that $A(\zeta_1) = B(\zeta_1)$

Proof. By equating A and B, we obtain the chain of equalities:

$$\gamma L_Q(1+\zeta) = \gamma L_R(1+\frac{1}{\zeta}) + \frac{1}{2p} \tag{36}$$

$$L_Q(1+\zeta) = L_R(1+\frac{1}{\zeta}) + \frac{1}{2\gamma p} \tag{37}$$

$$L_Q + \zeta L_Q - L_R - \frac{L_R}{\zeta} - \frac{1}{2\gamma p} = 0 \tag{38}$$

$$\zeta L_Q + \zeta^2 L_Q - \zeta L_R - L_R - \frac{2\zeta}{\gamma p} = 0 \tag{39}$$

$$\zeta^2 L_Q + \zeta(L_Q - L_R - \frac{1}{2\gamma p}) - L_R = 0 \tag{40}$$

$$\zeta_1 := \frac{-(L_Q - L_R - \frac{1}{2\gamma p}) + \sqrt{(L_Q - L_R - \frac{1}{2\gamma p})^2 + 4L_Q L_R}}{2L_Q} > 0 \tag{41}$$

ζ_1 is a solution to a quadratic equation. Thus,

$$\gamma L_R(1+\frac{1}{\zeta_1}) + \frac{1}{2p} = \gamma L_Q(1+\zeta_1) \tag{42}$$

□

Sublemma 6.2. For ζ_1 from previous lemma: $A(\zeta_1) - 1 \leqslant -\frac{1}{12} \leqslant 0$

Proof. Let us take $\gamma \leqslant \frac{1}{6L}$, meaning that $L \leqslant \frac{1}{6\gamma}$:

$$\begin{aligned}
A(\zeta_1) - 1 &= \gamma L_Q(1+\zeta_1) - 1 \qquad (43)\\
&= \gamma L_Q\left[1 + \frac{-(L_Q - L_R - \frac{1}{2\gamma p}) + \sqrt{(L_Q - L_R - \frac{1}{2\gamma p})^2 + 4L_QL_R}}{2L_Q}\right] - 1 \qquad (44)\\
&= \frac{\gamma}{2}\left[2L_Q - (L_Q - L_R - \frac{1}{2\gamma p}) + \sqrt{(L_Q - L_R - \frac{1}{2\gamma p})^2 + 4L_QL_R}\right] - 1 \qquad (45)\\
&\leqslant \frac{\gamma}{2}\left[|L_Q + L_R + \frac{1}{2\gamma p}| + |L_Q - L_R - \frac{1}{2\gamma p}| + \sqrt{4L_QL_R}\right] - 1 \qquad (46)\\
&\leqslant \frac{\gamma}{2}\left[L_Q + L_R + \frac{1}{2\gamma p} + L + \frac{1}{2\gamma p} + \sqrt{4L^2}\right] - 1 \qquad (47)\\
&\leqslant \frac{\gamma}{2}\left[5L + \frac{1}{\gamma p}\right] - 1 \leqslant \frac{\gamma}{2}\left[\frac{5}{6\gamma} + \frac{1}{\gamma p}\right] - 1 \qquad (48)\\
&= \frac{5}{12} + \frac{1}{2p} - 1 = \frac{1}{2p} - \frac{7}{12} \qquad (49)
\end{aligned}$$

For $p \geqslant 1$:

$$\gamma L_Q(1+\zeta_1) - 1 \leqslant \frac{1}{2p} - \frac{7}{12} \leqslant \frac{6}{12} - \frac{7}{12} = -\frac{1}{12} < 0 \qquad (50)$$

□

Combining the results of Sublemmas 6.1 and 6.2, we obtain the statement of Lemma 6. Further, let's denote $A(\zeta_1)$ as A_1 and $B(\zeta_1)$ as B_1

Lemma 7. Generalization of Lemma 3.1 from Stich (2019).

$$\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 \leqslant (1-\gamma\mu)\|\bar{x}_t - x_*\|^2 - \frac{\gamma}{6}(F(\bar{x}_t) - F_*) + 2\gamma L_R V_t \qquad (51)$$

Proof. From Lemma 5 we know that:

$$\begin{aligned}
&\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 \leqslant\\
&\qquad (1-\gamma\mu)\|\bar{x}_t - x_*\|^2 + 2\gamma p L_R V_t\\
&\qquad + 2\gamma(A-1)(Q(\bar{x}_t) - Q_*) + 2\gamma(B-1)(\bar{R}_t - R_*)\\
&\qquad - 2\gamma A\langle q_*, \bar{x}_t - x_*\rangle - 2\gamma B\langle r_*, \bar{x}_t - x_*\rangle
\end{aligned} \qquad (52)$$

Using the result of Lemma 6, and substituting ζ_1 into A and B, we obtain that $A(\zeta_1) = B(\zeta_1) \ = A_1 = B_1$:

$$\begin{aligned}
-2\gamma A_1\langle q_*, \bar{x}_t - x_*\rangle - 2\gamma B_1\langle r_*, \bar{x}_t - x_*\rangle &= -2\gamma A_1\langle q_* + r_*, \bar{x}_t - x_*\rangle \qquad (53)\\
&= -2\gamma A_1\langle \nabla F(x_*), \bar{x}_t - x_*\rangle = 0 \qquad (54)
\end{aligned}$$

For $a \geqslant 0$ by Jensen's inequality:

$$-a(\frac{1}{M}\sum_{m=1}^{M} R(x_t^m) - R_*) \leqslant -a(R(\bar{x}_t) - R_*) \qquad (55)$$

Using that $A_1 - 1 \leqslant 0$ allows us to use (55):

$$\begin{aligned}
2\gamma(B_1 - 1)(\bar{R}_t - R_*) &= 2\gamma(A_1 - 1)(\bar{R}_t - R_*) \qquad (56)\\
&= |2\gamma(A_1 - 1)| \cdot (R_* - \bar{R}_t) = |2\gamma(A_1 - 1)| \cdot (R_* - \frac{1}{M}\sum_{m=1}^{M} R(x_t^m)) \qquad (57)\\
&\leqslant |2\gamma(A_1 - 1)| \cdot (R_* - R(\bar{x}_t)) = 2\gamma(A_1 - 1)(R(\bar{x}_t) - R_*) \qquad (58)
\end{aligned}$$

Substituting (54) and (58) into (52):

$$\begin{aligned}
&\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 \qquad (59)\\
&\leqslant (1-\gamma\mu)\|\bar{x}_t - x_*\|^2 + 2\gamma p L_R V_t + 2\gamma(A_1 - 1)(Q(\bar{x}_t - Q_* + R(\bar{x}_t - R_*) \qquad (60)\\
&= (1-\gamma\mu)\|\bar{x}_t - x_*\|^2 + 2\gamma p L_R V_t + 2\gamma(A_1 - 1)(F(\bar{x}_t - F_*) \qquad (61)
\end{aligned}$$

Given that Lemma 6 holds for $p = 1$, we can combine the result of Sublemma 6.2 with the fact that $F(\bar{x}_t) - F_* \geqslant 0$ to further strengthen our argument.

$$\|\bar{x}_t - x_* - \gamma \bar{g}_t\|^2 \leqslant (1 - \gamma\mu)\|\bar{x}_t - x_*\|^2 + 2\gamma L_R V_t + 2\gamma(A_1 - 1)(F(\bar{x}_t - F_*) \tag{62}$$

$$\leqslant (1 - \gamma\mu)\|\bar{x}_t - x_*\|^2 - 2\gamma\frac{1}{12}(F(\bar{x}_t) - F_*) + 2\gamma L_R V_t \tag{63}$$

Thus completing the proof. □

Now we will present the final result of this subsection:

Lemma 8. For $\gamma \leqslant \frac{1}{6L}$:

$$\mathbb{E}\|\bar{x}_{t+1} - x_*\|^2 \leqslant (1 - \gamma\mu)\mathbb{E}\|\bar{x}_t - x_*\|^2 + \gamma^2\mathbb{E}\|\bar{\boldsymbol{g}}_t - \bar{g}_t\|^2 - \frac{\gamma}{6}\mathbb{E}[F(\bar{x}_t) - F_*] + 2\gamma L_R\mathbb{E}[V_t] \tag{64}$$

Proof. Using the update equation (10) we have

$$\|\bar{x}_{t+1} - x_*\|^2 = \|\bar{x}_t - \gamma\bar{\boldsymbol{g}}_t - x_*\|^2 = \|\bar{x}_t - \gamma\bar{\boldsymbol{g}}_t - x_* - \gamma\bar{g}_t + \gamma\bar{g}_t\|^2 \tag{65}$$

$$= \|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 + \gamma^2\|\bar{\boldsymbol{g}}_t - \bar{g}_t\|^2 + 2\gamma\langle\bar{x}_t - x_* - \gamma\bar{g}_t, \bar{\boldsymbol{g}}_t - \bar{g}_t\rangle \tag{66}$$

Taking expectation,

$$\mathbb{E}\|\bar{x}_{t+1} - x_*\|^2 = \mathbb{E}\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 + \gamma^2\mathbb{E}\|\bar{\boldsymbol{g}}_t - \bar{g}_t\|^2 \tag{67}$$

Taking expectiation of result of Lemma 7:

$$\mathbb{E}\|\bar{x}_t - x_* - \gamma\bar{g}_t\|^2 \leqslant (1 - \gamma\mu)\mathbb{E}\|\bar{x}_t - x_*\|^2 - \frac{\gamma}{6}\mathbb{E}[F(\bar{x}_t) - F_*] + 2\gamma L_R\mathbb{E}[V_t] \tag{68}$$

Combination of (67) and (68) yields the claim of the Lemma:

$$\mathbb{E}\|\bar{x}_{t+1} - x_*\|^2 \leqslant (1 - \gamma\mu)\mathbb{E}\|\bar{x}_t - x_*\|^2 + \gamma^2\mathbb{E}\|\bar{\boldsymbol{g}}_t - \bar{g}_t\|^2 - \frac{\gamma}{6}\mathbb{E}[F(\bar{x}_t) - F_*] + 2\gamma L_R\mathbb{E}[V_t] \tag{69}$$

□

5.3 Other Lemmas

Lemma 9.

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\boldsymbol{g}_t^m - g_t^m\|^2 \leqslant \sigma^2 + \rho L^2 V_t + \rho L^2\|r_t\|^2 \tag{70}$$

Proof.

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\boldsymbol{g}_t^m - g_t^m\|^2 \overset{2}{\leqslant} \sigma^2 + \frac{\rho}{M}\sum_{m=1}^{M}\|g_t^m\|^2 \tag{71}$$

$$= \sigma^2 + \frac{\rho}{M}\sum_{m=1}^{M}\|\nabla F(x_t^m) - \nabla F(x_*)\|^2 \tag{72}$$

$$\leqslant \sigma^2 + \frac{\rho L^2}{M}\sum_{m=1}^{M}\|x_t^m - x_*\|^2 \tag{73}$$

$$= \sigma^2 + \frac{\rho L^2}{M}\sum_{m=1}^{M}\|x_t^m - \bar{x}_t + \bar{x}_t - x_*\|^2 \tag{74}$$

$$= \sigma^2 + \frac{\rho L^2}{M}\sum_{m=1}^{M}\left(\|x_t^m - \bar{x}_t\|^2 + \|\bar{x}_t - x_*\|^2 + 2\langle x_t^m - \bar{x}_t, \bar{x}_t - x_*\rangle\right) \tag{75}$$

$$= \sigma^2 + \frac{\rho L^2}{M}\sum_{m=1}^{M}\left(\|x_t^m - \bar{x}_t\|^2 + \|\bar{x}_t - x_*\|^2\right) \tag{76}$$

$$= \sigma^2 + \rho L^2\left(V_t + \|r_t\|^2\right) \tag{77}$$

□

Lemma 10. Variance reduction:

$$\mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 \leqslant \frac{\sigma^2}{M} + \frac{\rho L^2}{M} V_t + \frac{\rho L^2}{M}\left\|r_t\right\|^2 \tag{78}$$

Proof. In the first equality we use that g_t^m on each device are independent, and in the second inequality we use Lemma **??**.

$$\begin{aligned}\mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 &\leqslant \mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{g}_t^m - g_t^m\right\|^2 &(79)\\ &= \frac{1}{M^2}\sum_{m=1}^{M} \mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 &(80)\\ &\leqslant \frac{\sigma^2}{M} + \frac{\rho L^2}{M} V_t + \frac{\rho L^2}{M}\left\|r_t\right\|^2 &(81)\end{aligned}$$

□

Lemma 11.

If $\rho = 0$ and $\gamma \leqslant \frac{1}{6L}$, then:

$$\mathbb{E}[V_t] \leqslant (H-1)\gamma^2\sigma^2 \tag{82}$$

Under following assumptions:

- $\mu > 0$
- $\gamma \leqslant \frac{\mu}{12\rho L^2}$
- For all t it is true that $\left\|r_{t+1}\right\|^2 \geqslant (1 - \frac{\gamma\mu}{12})\left\|r_t\right\|^2$

it is true that

$$\mathbb{E}[V_t] \leqslant (H-1)\gamma^2\sigma^2 + (H-1)\rho\gamma^2 L^2 \mathbb{E}\left\|r_t\right\|^2 \tag{83}$$

Proof. This result is a compilation of results obtained in Sublemmas 11.2 and 11.3 □

Sublemma 11.1.

$$\mathbb{E}[V_{t+1}] \leqslant (1 - \frac{\gamma\mu}{6} + \gamma^2\rho L^2)V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2\left\|r_t\right\|^2 \tag{84}$$

Proof. We follow the of Khaled et al. (2022) in proving their's Lemma 1 but under the boundaries of Assumption 2.

For $t \in \mathbb{N}$ we have $x_{t+1}^m = x_t^m - \gamma\boldsymbol{g}_t^m$ and $\bar{x}_{t+1} = \bar{x}_t - \gamma\bar{\boldsymbol{g}}_t$ if $t+1 \mod H \neq 0$.

Hence for such t and for conditioned expectation it is true that:

$$\begin{aligned}\mathbb{E}\left\|x_{t+1}^m - \bar{x}_{t+1}\right\|^2 &= \left\|x_t^m - \bar{x}_t\right\|^2 + \gamma^2\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 - 2\gamma\mathbb{E}[\langle x_t^m - \bar{x}_t, \boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\rangle] &(85)\\ &= \left\|x_t^m - \bar{x}_t\right\|^2 + \gamma^2\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 - 2\gamma\langle x_t^m - \bar{x}_t, g_t^m\rangle + 2\gamma\langle x_t^m - \bar{x}_t, \bar{\boldsymbol{g}}_t\rangle &(86)\end{aligned}$$

Averaging over m:

$$\begin{aligned}\mathbb{E}[V_{t+1}] &= V_t + \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, g_t^m\rangle + 2\gamma\langle \bar{x}_t - \bar{x}_t, \bar{\boldsymbol{g}}_t\rangle &(87)\\ &= V_t + \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, g_t^m\rangle &(88)\end{aligned}$$

By expanding square:

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 &= \mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{g}_t + \bar{g}_t - \bar{\boldsymbol{g}}_t\right\|^2 &(89)\\ &= \mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{g}_t\right\|^2 + \mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 + 2\mathbb{E}[\langle \boldsymbol{g}_t^m - \bar{g}_t, \bar{g}_t - \bar{\boldsymbol{g}}_t\rangle] &(90)\end{aligned}$$

And again:

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{g}_t\right\|^2 &= \mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m + g_t^m - \bar{g}_t\right\|^2 && (91)\\
&= \mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \left\|g_t^m - \bar{g}_t\right\|^2 + 2\mathbb{E}[\langle \boldsymbol{g}_t^m - g_t^m, g_t^m - \bar{g}_t\rangle] && (92)\\
&= \mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \left\|g_t^m - \bar{g}_t\right\|^2 + 2\langle g_t^m - g_t^m, g_t^m - \bar{g}_t\rangle && (93)\\
&= \mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \left\|g_t^m - \bar{g}_t\right\|^2 && (94)
\end{aligned}$$

Combining (90) and (94) we have:

$$\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 = \mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \left\|g_t^m - \bar{g}_t\right\|^2 + \mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 + 2\mathbb{E}[\langle \boldsymbol{g}_t^m - \bar{g}_t, \bar{g}_t - \bar{\boldsymbol{g}}_t\rangle] \tag{95}$$

By averaging both sides over m:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \bar{g}_t\right\|^2 && (96)\\
&+ \mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 + 2\mathbb{E}[\langle \bar{\boldsymbol{g}}_t - \bar{g}_t, \bar{g}_t - \bar{\boldsymbol{g}}_t\rangle] && (97)\\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \bar{g}_t\right\|^2 && (98)\\
&+ \mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 - 2\mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 && (99)\\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \bar{g}_t\right\|^2 - \mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 && (100)\\
&\leqslant \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \bar{g}_t\right\|^2 && (101)
\end{aligned}$$

We can estimate second term here as follows:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \bar{g}_t\right\|^2 &= \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \nabla F(\bar{x}_t) + \nabla F(\bar{x}_t) - \bar{g}_t\right\|^2 && (102)\\
&= \frac{1}{M}\sum_{m=1}^{M}\Big(\left\|g_t^m - \nabla F(\bar{x}_t)\right\|^2 + \left\|\nabla F(\bar{x}_t) - \bar{g}_t\right\|^2 + 2\langle g_t^m - \nabla F(\bar{x}_t), \nabla F(\bar{x}_t) - \bar{g}_t\rangle\Big) && (103)\\
&= \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \nabla F(\bar{x}_t)\right\|^2 + \left\|\nabla F(\bar{x}_t) - \bar{g}_t\right\|^2 - 2\left\|\nabla F(\bar{x}_t) - \bar{g}_t\right\|^2 && (104)\\
&= \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \nabla F(\bar{x}_t)\right\|^2 - \left\|\nabla F(\bar{x}_t) - \bar{g}_t\right\|^2 && (105)\\
&\leqslant \frac{1}{M}\sum_{m=1}^{M}\left\|g_t^m - \nabla F(\bar{x}_t)\right\|^2 = \frac{1}{M}\sum_{m=1}^{M}\left\|\nabla F(x_t^m) - \nabla F(\bar{x}_t)\right\|^2 && (106)\\
&\overset{4}{\leqslant} \frac{1}{M}\sum_{m=1}^{M}2L(F(\bar{x}_t) - F(x_t^m) - \langle \bar{x}_t - x_t^m, \nabla F(x_t^m)\rangle) && (107)\\
&\overset{Jensen's\ ineq.}{\leqslant} \frac{2L}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle && (108)
\end{aligned}$$

Substituting (108) into (101) and bounding variance:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\| &\leqslant \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - g_t^m\right\|^2 + \frac{2L}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle && (109)\\
&\overset{9}{\leqslant} \sigma^2 + \rho L^2 V_t + \rho L^2\left\|r_t\right\|^2 + \frac{2L}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle && (110)
\end{aligned}$$

Let us substitute this result into (88):

$$\mathbb{E}[V_{t+1}] = V_t + \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\boldsymbol{g}_t^m - \bar{\boldsymbol{g}}_t\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle \tag{111}$$

$$\leqslant V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2 V_t + \gamma^2\rho L^2\left\|r_t\right\|^2 \tag{112}$$

$$+ \frac{2\gamma^2 L}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle - \frac{2\gamma}{M}\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle \tag{113}$$

$$= V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2 V_t + \gamma^2\rho L^2\left\|r_t\right\|^2 - \frac{2\gamma}{M}(1-\gamma L)\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle \tag{114}$$

Now let us analyize last term. We know that $\gamma \leqslant \frac{1}{6L}$, therefore $1-\gamma L \geqslant 0$. Thus, by strong convexity and Jensen's inequality:

$$-\frac{2\gamma}{M}(1-\gamma L)\sum_{m=1}^{M}\langle x_t^m - \bar{x}_t, \nabla F(x_t^m)\rangle = \frac{2\gamma}{M}(1-\gamma L)\sum_{m=1}^{M}\langle \bar{x}_t - x_t^m, \nabla F(x_t^m)\rangle \tag{115}$$

$$\overset{1}{\leqslant} \frac{2\gamma}{M}(1-\gamma L)\sum_{m=1}^{M}\left(F(\bar{x}_t) - F(x_t^m) - \frac{\mu}{2}\left\|x_t^m - \bar{x}_t\right\|^2\right) \tag{116}$$

$$\overset{Jensen's\ ineq.}{\leqslant} -\frac{\gamma}{M}(1-\gamma L)\sum_{m=1}^{M}\mu\left\|x_t^m - \bar{x}_t\right\|^2 = -\gamma(1-\gamma L)\mu V_t \tag{117}$$

Plugging (117) into (114) and again using that $\gamma \leqslant \frac{1}{6L}$:

$$\mathbb{E}[V_{t+1}] \leqslant V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2 V_t + \gamma^2\rho L^2\left\|r_t\right\|^2 - \gamma(1-\gamma L)\mu V_t \tag{118}$$

$$= (1-\gamma(1-\gamma L)\mu)V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2 V_t + \gamma^2\rho L^2\left\|r_t\right\|^2 \tag{119}$$

$$\leqslant (1-\frac{\gamma\mu}{6})V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2 V_t + \gamma^2\rho L^2\left\|r_t\right\|^2 \tag{120}$$

$$= (1-\frac{\gamma\mu}{6} + \gamma^2\rho L^2)V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2\left\|r_t\right\|^2 \tag{121}$$

□

Sublemma 11.2. If $\rho = 0$ and $\gamma \leqslant \frac{1}{6L}$, then

$$\mathbb{E}[V_t] \leqslant (H-1)\gamma^2\sigma^2 \tag{122}$$

Proof. This is Lemma 1 from Khaled et al. (2022). □

Sublemma 11.3. If $\mu > 0$ and $\left\|r_{t+1}\right\|^2 \geqslant (1-\gamma\mu)\left\|r_t\right\|^2$, then

$$\mathbb{E}[V_t] \leqslant (H-1)\gamma^2\sigma^2 + (H-1)\rho\gamma^2 L^2\mathbb{E}\left\|r_t\right\|^2 \tag{123}$$

Proof. Assume that $\gamma \leqslant \frac{\mu}{12\rho L^2}$ which means that $-\frac{\gamma\mu}{6} + \gamma^2\rho L^2 \leqslant -\frac{\gamma\mu}{12}$. By substituting this into the result of Sublemma 11.1 we get:

$$\mathbb{E}[V_{t+1}] \leqslant (1-\frac{\gamma\mu}{12})V_t + \gamma^2\sigma^2 + \gamma^2\rho L^2\left\|r_t\right\|^2 \tag{124}$$

Let us divide t by H, suppose $t = kH + a;\ k, a \in \mathbb{N};\ a < H$. Recalling that $V_{kH} = 0$, recursing (124) and considering $\mathbb{E}$ as a full expectation yields:

$$\mathbb{E}[V_t] \leqslant (1-\frac{\gamma\mu}{12})^a \cdot V_{kH} + \sum_{j=kH}^{kH+a} (1-\frac{\gamma\mu}{12})^{(t-j)} \cdot \left(\rho\gamma^2 L^2 \mathbb{E}\left\|r_j\right\|^2 + \gamma^2\sigma^2\right) \tag{125}$$

$$= \sum_{j=kH}^{kH+a} (1-\frac{\gamma\mu}{12})^{(t-j)} \cdot \left(\rho\gamma^2 L^2 \mathbb{E}\left\|r_j\right\|^2 + \gamma^2\sigma^2\right) \tag{126}$$

$$\leqslant a\gamma^2\sigma^2 + \rho\gamma^2 L^2 \sum_{j=t-a}^{t} (1-\frac{\gamma\mu}{12})^{(t-j)} \cdot \mathbb{E}\left\|r_j\right\|^2 \tag{127}$$

$$\overset{Magic,\ \mu>0}{\leqslant} a\gamma^2\sigma^2 + a\rho\gamma^2 L^2 \mathbb{E}\left\|r_t\right\|^2 \overset{a<H}{\leqslant} (H-1)\gamma^2\sigma^2 + (H-1)\rho\gamma^2 L^2 \mathbb{E}\left\|r_t\right\|^2 \tag{128}$$

□

5.4 Proof of Theorem 1

We will leverage the insights from Lemma 7 to enhance the proofs of Theorem 1 from Khaled et al. (2022), thereby obtaining more precise results.

Proof. Consider $\mu > 0$ and $\gamma \leqslant \min\{\frac{1}{6L},\ \frac{\mu}{12\rho L^2},\ ....\}$ and for all t it is true that $\left\|r_{t+1}\right\|^2 \geqslant (1-\frac{\gamma\mu}{12})\left\|r_t\right\|^2$.

Then let us substitute results of the previous subsection (i.e. Lemmas 10 and 11) into the Lemma 8 and conduct algebraic manipulations.

$$\mathbb{E}\left\|r_{t+1}\right\|^2 \leqslant (1-\gamma\mu)\mathbb{E}\left\|r_t\right\|^2 + \gamma^2\mathbb{E}\left\|\bar{\boldsymbol{g}}_t - \bar{g}_t\right\|^2 - \frac{\gamma}{6}\mathbb{E}[F(\bar{x}_t) - F_*] + 2\gamma L_R \mathbb{E}[V_t] \tag{129}$$

$$\overset{10,11}{\leqslant} (1-\gamma\mu)\mathbb{E}\left\|r_t\right\|^2 + \gamma^2\left(\frac{\sigma^2}{M} + \frac{\rho L^2}{M}V_t + \frac{\rho L^2}{M}\left\|r_t\right\|^2\right) \tag{130}$$

$$+ 2\gamma L_R\left((H-1)\gamma^2\sigma^2 + (H-1)\rho\gamma^2 L^2\mathbb{E}\left\|r_t\right\|^2\right) \tag{131}$$

$$= (1-\gamma\mu)\mathbb{E}\left\|r_t\right\|^2 + \frac{\gamma^2\sigma^2}{M} + \frac{\gamma^2\rho L^2}{M}V_t + \frac{\gamma^2\rho L^2}{M}\left\|r_t\right\|^2 \tag{132}$$

$$+ 2L_R(H-1)\gamma^3\sigma^2 + 2L_R(H-1)\rho\gamma^3 L^2\mathbb{E}\left\|r_t\right\|^2 \tag{133}$$

$$= \left(1-\gamma\mu + \frac{\gamma^2\rho L^2}{M} + 2L_R(H-1)\rho\gamma^3 L^2\right)\mathbb{E}\left\|r_t\right\|^2 \tag{134}$$

$$+ \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 + 2L_R(H-1)\gamma^2\sigma^3 + \frac{\gamma^2\rho L^2}{M}V_t \tag{135}$$

$$\tag{136}$$

By estimating V_t again,

$$\mathbb{E}\left\|r_{t+1}\right\|^2 \overset{11,\ ???\mathbb{E}}{\leqslant} \left(1-\gamma\mu + \frac{\gamma^2\rho L^2}{M} + 2L_R(H-1)\rho\gamma^3 L^2\right)\mathbb{E}\left\|r_t\right\|^2 \tag{137}$$

$$+ \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 + \frac{\gamma^2\rho L^2}{M}\left((H-1)\gamma^2\sigma^2 + (H-1)\rho\gamma^2 L^2\mathbb{E}\left\|r_t\right\|^2\right) \tag{138}$$

$$\leqslant \left(1-\gamma\mu + \frac{\gamma^2\rho L^2}{M} + 2L_R H\rho\gamma^3 L^2\right)\mathbb{E}\left\|r_t\right\|^2 \tag{139}$$

$$+ \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 + \frac{\gamma^2\rho L^2}{M}\left((H-1)\gamma^2\sigma^2 + H\rho\gamma^2 L^2\mathbb{E}\left\|r_t\right\|^2\right) \tag{140}$$

$$\overset{L_R=\varepsilon L}{=} \left(1-\gamma\mu + \frac{\gamma^2\rho L^2}{M} + 2\varepsilon H\rho\gamma^3 L^3 + \frac{\gamma^4 H\rho^2 L^4}{M}\right)\mathbb{E}\left\|r_t\right\|^2 \tag{141}$$

$$+ \frac{\gamma^2\sigma^2}{M} + 2\varepsilon L(H-1)\gamma^3\sigma^2 + \frac{\gamma^4\rho L^2}{M}(H-1)\sigma^2 \tag{142}$$

Due to our restrictions on the stepsize,

$$1-\gamma\mu + \frac{\gamma^2\rho L^2}{M} + 2\varepsilon H\rho\gamma^3 L^3 + \frac{\gamma^4 H\rho^2 L^4}{M} \tag{143}$$

$$\leqslant 1-\gamma\mu + \frac{\gamma\mu}{4} + \frac{\gamma\mu}{4} + \frac{\gamma\mu}{4} \leqslant 1 - \frac{\gamma\mu}{4} \tag{144}$$

Thus,

$$\mathbb{E}\left\|r_{t+1}\right\|^2 \leqslant (1-\frac{\gamma\mu}{4})\mathbb{E}\left\|r_t\right\|^2 + \frac{\gamma^2\sigma^2}{M} + 2\varepsilon L(H-1)\gamma^3\sigma^2 + \frac{\gamma^4\rho L^2}{M}(H-1)\sigma^2 \tag{145}$$

□

5.5 Proof of Theorem 2

Proof. Consider $\gamma \leqslant \frac{1}{6L}$ and $\mu = 0$. In this case, from Lemmas 8 and 11 we obtain:

$$\mathbb{E}\left\|\bar{x}_{t+1} - x_*\right\|^2 \leqslant \mathbb{E}\left\|\bar{x}_t - x_*\right\|^2 + \frac{\gamma^2\sigma^2}{M} - \frac{\gamma}{6}\mathbb{E}[F(\bar{x}_t) - F_*] + 2\gamma L_R(H-1)\gamma^2\sigma^2 \tag{146}$$

Let's denote $r_t = \bar{x}_t - x_*$, then rearranging the above equation we have:

$$\frac{\gamma}{6}\mathbb{E}[F(\bar{x}_t) - F_*] \leqslant \mathbb{E}\left\|r_t\right\|^2 - \mathbb{E}\left\|r_{t+1}\right\|^2 + \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 \tag{147}$$

Averging the above equation over t,

$$\begin{aligned}\frac{\gamma}{6T}\sum_{t=0}^{T-1}\mathbb{E}[F(\bar{x}_t) - F_*] &\leqslant \frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\left\|r_t\right\|^2 - \mathbb{E}\left\|r_{t+1}\right\|^2) + \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 \\ &= \frac{\left\|r_0\right\|^2 - \mathbb{E}\left\|r_T\right\|^2}{T} + \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 \\ &\leqslant \frac{\left\|r_0\right\|^2}{T} + \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2\end{aligned} \tag{148}$$

For $\hat{x}_t = \frac{1}{T}\sum_{t=0}^{T-1}\bar{x}_t$, by Jensen's inequality:

$$\mathbb{E}[F(\hat{x}_t) - F_*] \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[F(\bar{x}_t) - F_*] \tag{149}$$

Plugging (149) into (148),

$$\frac{\gamma}{6}\mathbb{E}[F(\hat{x}_t) - F_*] \leqslant \frac{\left\|r_0\right\|^2}{T} + \frac{\gamma^2\sigma^2}{M} + 2L_R(H-1)\gamma^3\sigma^2 \tag{150}$$

Dividing both sides by $\frac{\gamma}{6}$, we prove the theorem:

$$\mathbb{E}[F(\hat{x}_t) - F_*] \leqslant \frac{6}{\gamma T}\left\|r_t\right\|^2 + \frac{6\gamma\sigma^2}{M} + 12L_R(H-1)\gamma^2\sigma^2 \tag{151}$$

□

Надо сделать упор на 2 вещи:

1. **КРАСИВОЕ улучшение оценок Khaled-а**
2. **Мы работаем в общем случае, а Yuan and Ma в частном**
3. **Упомянуть физ.смысл**