Don't Compress Gradients in Random Reshuffling: Compress Gradient Differences

Abdurakhmon Sadiev[1,2*], **Grigory Malinovsky**[1], **Eduard Gorbunov**[1,2,3,4†‡]
Igor Sokolov[1], **Ahmed Khaled**[5], **Konstantin Burlachenko**[1], **Peter Richtárik**[1]

[1]King Abdullah University of Science and Technology, Saudi Arabia
[2]Moscow Institute of Physics and Technology, Russian Federation
[3]Mohamed bin Zayed University of Artificial Intelligence, UAE
[4]Mila, Université de Montréal, Canada
[5]Princeton University, USA

Abstract

Gradient compression is a popular technique for improving communication complexity of stochastic first-order methods in distributed training of machine learning models. However, the existing works consider only with-replacement sampling of stochastic gradients. In contrast, it is well-known in practice and recently confirmed in theory that stochastic methods based on without-replacement sampling, e.g., Random Reshuffling (RR) method, perform better than ones that sample the gradients with-replacement. In this work, we close this gap in the literature and provide the first analysis of methods with gradient compression and without-replacement sampling. We first develop a distributed variant of random reshuffling with gradient compression (Q-RR), and show how to reduce the variance coming from gradient quantization through the use of control iterates. Next, to have a better fit to Federated Learning applications, we incorporate local computation and propose a variant of Q-RR called Q-NASTYA. Q-NASTYA uses local gradient steps and different local and global stepsizes. Next, we show how to reduce compression variance in this setting as well. Finally, we prove the convergence results for the proposed methods and outline several settings in which they improve upon existing algorithms.

1 Introduction

Distributed learning plays a crucial role in the training of modern Deep Learning (DL) models since distributed approaches are able to significantly reduce training time [Goyal et al., 2017, You et al., 2019]. Moreover, distributed methods are mandatory for such applications as Federated learning (FL) [Konečný et al., 2016, McMahan et al., 2017], where multiple nodes connected over a network collaborate on a learning task. Each node possesses its own dataset and cannot share this data with other nodes or a central server. As a result, algorithms for federated learning often rely on local computation and lack access to the entire dataset of training examples. Federated learning finds applications in diverse fields, including language modeling for mobile keyboards [Liu et al., 2021],

[*]Part of the work was done while A. Sadiev was a research intern at KAUST, working in the Optimization and Machine Learning Lab of P. Richtárik.

[†]Part of the work was done when E. Gorbunov was a researcher at MIPT and Mila & UdeM and also a visiting researcher at KAUST, in the Optimization and Machine Learning Lab of P. Richtárik.

[‡]Corresponding author, eduard.gorbunov@mbzuai.ac.ae

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

healthcare [Antunes et al., 2022], and wireless communications [Yang et al., 2022]. Its applications extend to various other domains [Kairouz et al., 2019].

Distributed learning tasks are often solved through *empirical-risk minimization* (ERM), where the m-th device contributes an empirical loss function $f_m(x)$ representing the average loss of model x on its local dataset, and our goal is to then minimize the average loss over all the nodes:

$$\min_{x\in\mathbb{R}^d}\left[f(x) \stackrel{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} f_m(x)\right], \tag{1}$$

where the function f represents the average loss. Every f_m is an average of sample loss functions f_m^i each representing the loss of model x on the i-th datapoint on the m-th clients' dataset: that is for each $m \in \{1, 2, \ldots, M\}$ we have

$$f_m(x) \stackrel{\text{def}}{=} \frac{1}{n_m}\sum_{i=1}^{n_m} f_m^i(x).$$

For simplicity we shall assume that the datasets on all clients are of equal size: $n_1 = n_2 = \ldots = n_M$, though this assumption is only for convenience and our results easily extend to the case when clients have datasets of unequal sizes. Thus our optimization problem is

$$\min_{x\in\mathbb{R}^d}\left[f(x) = \frac{1}{nM}\sum_{m=1}^{M}\sum_{i=1}^{n} f_m^i(x)\right]. \tag{2}$$

Because d is often very large in practice, the dominant paradigm for solving (2) relies on first-order (gradient) information. Federated learning algorithms have access to two key primitives: (a) local computation, where for a given model $x \in \mathbb{R}^d$ we can compute stochastic gradients $\nabla f_m^i(x)$ locally on client m, and (b) communication, where the different clients can exchange their gradients or models with a central server.

1.1 Communication compression

In practice, communication is more expensive than local computation [Kairouz et al., 2019], and as such one of the chief concerns of algorithms for distributed learning is communication efficiency. Algorithms for distributed/federated learning have thus focused on achieving communication efficiency, with one common ingredient being the use of *gradient compression*, where each client sends a compressed or quantized version of their update instead of the full update vector, potentially saving communication bandwidth by sending fewer bits over the network. There are many operators that can be used for compressing the update vectors: stochastic quantization [Alistarh et al., 2017], random sparsification [Wangni et al., 2018, Stich et al., 2018], and others [Tang et al., 2020]. In this work we consider compression operators satisfying the following assumption:

Assumption 1. *A compression operator is an operator $\mathcal{Q}: \mathbb{R}^d \to \mathbb{R}^d$ such that for some $\omega > 0$, the relations*

$$\mathbb{E}\left[\mathcal{Q}(x)\right] = x \qquad \textit{and} \qquad \mathbb{E}\left[\|\mathcal{Q}(x) - x\|^2\right] \le \omega\|x\|^2 \quad \textit{hold for } x \in \mathbb{R}^d.$$

Unbiased compressors can reduce the number of bits clients communicate per round, but also increases the variance of the stochastic gradients used slowing down overall convergence, see e.g. [Khirirat et al., 2018, Theorem 5.2] and [Stich, 2020, Theorem 1]. By using control iterates, Mishchenko et al. [2019b] developed DIANA—an algorithm that can reduce the variance due to gradient compression with unbiased compression operators, and thus ensure fast convergence. DIANA has been extended and analyzed in many settings [Horváth et al., 2019, Stich, 2020, Safaryan et al., 2021] and forms an important tool in our arsenal for using gradient compression.

1.2 Random Reshuffling

Despite the importance of addressing the communication bottleneck, local computations also significantly affect the training. For simplicity, consider the 1-node scenario. In this case, the update rule of the standard work-horse method in stochastic optimization – stochastic gradient descent (SGD)

Table 1: Summary of known and new complexity results for solving distributed finite-sum optimization problem (2). Column "Complexity" indicates the number of communication rounds to find a solution with accuracy $\varepsilon > 0$. Column "RR?" shows whether an algorithm uses *Random Reshuffling*, "C?" indicates whether a method applies the compression of gradients or difference between the gradients and also whether methods for communication, "H?" means independence from the constant of data heterogeneity in the complexity, "CVX?" indicates whether each loss on the i-th datapoint on the m-th client is convex, but not strongly convex. Notation: $\kappa = L_{\max}/\mu$ and $\widetilde{\kappa} = L_{\max}/\widetilde{\mu}$ are conditional number of problem (2), where $L_{\max}$ = Lipschitz constant, μ and $\widetilde{\mu}$ are the strong convexity constants of f and f_m^i respectively; variances at the solution point $x_\star$: $\sigma_*^2 = \frac{1}{Mn}\sum\limits_{m=1}^{M}\sum\limits_{i=1}^{n}\|\nabla f_m^i(x_\star) - \nabla f_m(x_\star)\|^2$ and $\sigma_{\star,n}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\|\nabla f^i(x_\star)\|^2$; heterogeneity constant $\zeta_\star^2 = \frac{1}{M}\sum\limits_{m=1}^{M}\|\nabla f_m(x_\star)\|^2$. The results of this paper are highlighted in blue.

Method	Complexity	RR?	C?	H?	CVX?
SGD [Gower et al., 2019]	$\kappa + \frac{\sigma_{\star,n}^2}{\mu^2\varepsilon}$	✗	✗	✗[(1)]	✓
RR [Mishchenko et al., 2020]	$\widetilde{\kappa} + \frac{\sigma_{\star,n}}{\widetilde{\mu}}\sqrt{\frac{\widetilde{\kappa}n}{\varepsilon}}$	✓	✗	✗[(1)]	✗
RR [Mishchenko et al., 2020]	$n\kappa + \frac{\sigma_{\star,n}}{\widetilde{\mu}}\sqrt{\frac{\kappa n}{\varepsilon}}$	✓	✗	✗[(1)]	✓
QSGD [Gorbunov et al., 2020]	$\left(1+\frac{\omega}{M}\right)\kappa + \frac{\omega}{M}\frac{\sigma_\star^2+\zeta_\star^2}{\mu^2\varepsilon} + \frac{\sigma_\star^2}{M\mu^2\varepsilon}$[(2)]	✗	✓	✗	✓
Q-RR Corollary 1	$\left(1+\frac{\omega}{M}\right)\widetilde{\kappa} + \frac{\omega}{M}\frac{\sigma_\star^2+\zeta_\star^2}{\widetilde{\mu}^2\varepsilon} + \frac{\sigma_{\star,n}}{\widetilde{\mu}}\sqrt{\frac{\widetilde{\kappa}n}{\varepsilon}}$	✓	✓	✗	✗
Q-RR Corollary 6	$\left(n+\frac{\omega}{M}\right)\kappa + \frac{\omega}{M}\frac{\sigma_\star^2+\zeta_\star^2}{\mu^2\varepsilon} + \frac{\rho_\star}{\mu}\sqrt{\frac{\kappa n}{\varepsilon}}$[(3)]	✓	✓	✗	✓
DIANA [Mishchenko et al., 2019a]	$\left(1+\frac{\omega}{M}\right)\kappa + \frac{\omega}{M}\frac{\sigma_\star^2}{\mu^2\varepsilon} + \frac{\sigma_\star^2}{M\mu^2\varepsilon}$	✗	✓	✓	✓
DIANA-RR Corollary 2	$n(1+\omega) + \left(1+\frac{\omega}{M}\right)\widetilde{\kappa} + \frac{\sigma_{\star,n}}{\widetilde{\mu}}\sqrt{\frac{\widetilde{\kappa}n}{\varepsilon}}$	✓	✓	✓	✗
DIANA-RR Corollary 8	$n(1+\omega) + \left(n+\frac{\omega}{M}\right)\kappa + \frac{\sigma_{\star,n}}{\mu}\sqrt{\frac{\kappa n}{\varepsilon}}$	✓	✓	✓	✓

[(1)] In the case of SGD, RR we use ✗ in "H?" to show that the complexity of these methods is provided in the non-distributed setup.
[(2)] The following inequality is useful for the comparison of complexities: $\sigma_{\star,n}^2 \leq \sigma_\star^2$.
[(3)] We denote $\rho_\star^2 = \frac{\omega}{M}(\sigma_\star^2 + \zeta_\star^2) + \sigma_{\star,n}^2$.

[Robbins and Monro, 1951] – can be written as follows: $x_{t+1} = x_t - \gamma\nabla f^j(x_t)$, where j is sampled from $\{1,\ldots,n\}$ uniformly at random. This procedure thus uses *with-replacement sampling* in order to select the stochastic gradient used at each step from the dataset. However, in the training of DL models, *without-replacement sampling* is used much more often: that is, at the beginning of each *epoch* we choose a permutation $\pi_1, \pi_2, \ldots, \pi_n$ of $\{1, 2, \ldots, n\}$ and do the i-th update using the π_i-th gradient: $x_t^{i+1} = x_t^i - \gamma\nabla f^{\pi_i}(x_t^i)$. Without-replacement sampling SGD, also known as Random Reshuffling (RR) [Bottou, 2009], typically achieves better asymptotic convergence rates compared to with-replacement SGD and can improve upon it in many settings as shown by recent theoretical progress [Mishchenko et al., 2020, Ahn et al., 2020, Rajput et al., 2020, Safran and Shamir, 2021]. While with-replacement SGD achieves an error proportional to $\mathcal{O}\left(\frac{1}{T}\right)$ after T steps [Stich, 2019], Random Reshuffling achieves an error of $\mathcal{O}\left(\frac{n}{T^2}\right)$ after T steps, faster than SGD when the number of steps T is large.

1.3 Can Communication Compression and Random Reshuffling be Friends?

As we described earlier, Random Reshuffling and communication compression are two important tools for training modern DL models, and both techniques are relatively well understood. However, there are no papers that study Random Reshuffling and communication compression in combination. This leads us to the natural question: *how these techniques should be combined to improve the convergence speed of existing distributed methods?*

1.4 Contributions

In this paper, we aim to develop methods for Distributed and Federated Learning that combine gradient compression and random reshuffling. While each of these techniques can aid in reducing the communication complexity of distributed optimization, their combination is under-explored. Thus our goal is to design methods that improve upon existing algorithms in convergence rates and in practice. We summarize our contributions as follows.

- ◇ **The issue: naïve combination has no improvements.** As a natural step towards our goal, we propose and study a new algorithm, Q-RR (Algorithm 1), that combines random reshuffling with gradient compression at every communication round. However, for Q-RR our theoretical results do not show any improvement upon QSGD when the compression level is reasonable (see Table 1). Moreover, we observe similar performance of Q-RR and QSGD in various numerical experiments. Therefore, we conclude that this phenomenon is not an artifact of our analysis but rather an issue of Q-RR: communication compression adds an additional noise that dominates the one coming from the stochastic gradients sampling.
- ◇ **The remedy: reduction of compression variance.** To remove the additional variance added by the compression and unleash the potential of Random Reshuffling in distributed learning with compression, we propose DIANA-RR (Algorithm 2), a combination of Q-RR and the DIANA algorithm. We derive the convergence rates of the new method and show that it improves upon the convergence rates of Q-RR, QSGD, and DIANA (see Table 1). We point out that to achieve such results we use n shift-vectors per worker in DIANA-RR unlike DIANA that uses only 1 shift-vector.
- ◇ **Extensions to the local steps.** Inspired by the NASTYA algorithm of Malinovsky et al. [2022], we propose a variant of NASTYA, Q-NASTYA (Algorithm 3), that naïvely mixes quantization, local steps with random reshuffling, and uses different local and server stepsizes. Although it improves in per-round communication cost over NASTYA but, similar to Q-RR, we show that Q-NASTYA suffers from added variance due to gradient quantization. To overcome this issue, we propose another algorithm, DIANA-NASTYA (Algorithm 4), that adds DIANA-style variance reduction to Q-NASTYA and removes the additional variance.

Finally, to illustrate our theoretical findings we conduct experiments on federated logistic regression tasks and on distributed training of neural networks.

2 Algorithms and convergence theory

We will primarily consider the setting of strongly-convex and smooth optimization. We assume that the average function f is strongly convex:

Assumption 2. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, i.e., for all $x, y \in \mathbb{R}^d$,*

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2}\|x - y\|^2, \tag{3}$$

and functions $f_1^i, f_2^i, \ldots, f_M^i : \mathbb{R}^d \to \mathbb{R}$ are convex for all $i = 1, \ldots, n$.

Examples of objectives satisfying Assumption 2 include ℓ_2-regularized linear and logistic regression. Throughout the paper, we assume that f has the unique minimizer $x_\star \in \mathbb{R}^d$. We also use the assumption that each individual loss f_m^i is smooth, i.e. has Lipschitz-continuous first-order derivatives:

Assumption 3. *Function $f_m^i : \mathbb{R}^d \to \mathbb{R}$ is $L_{i,m}$-smooth for every $i \in [n]$ and $m \in [M]$, i.e., for all $x, y \in \mathbb{R}^d$ and for all $m \in [M]$ and $i \in [n]$,*

$$\|\nabla f_m^i(x) - \nabla f_m^i(y)\| \leq L_{i,m}\|x - y\|. \tag{4}$$

We denote the maximal smoothness constant as $L_{\max} \overset{\text{def}}{=} \max_{i,m} L_{i,m}$.

For some methods, we shall additionally impose the assumption that *each* function is strongly convex:

Assumption 4. *Each function $f_m^i : \mathbb{R}^d \to \mathbb{R}$ is $\widetilde{\mu}$-strongly convex.*

The *Bregman divergence* associated with a convex function h is defined for all $x, y \in \mathbb{R}^d$ as

$$D_h(x, y) \overset{\text{def}}{=} h(x) - h(y) - \langle \nabla h(y), x - y \rangle .$$

Note that the inequality (3) defining strong convexity can be written as $D_f(x, y) \geq \frac{\mu}{2}\|x - y\|^2$.

2.1 Algorithm Q-RR

The first method we introduce is Q-RR (Algorithm 1). Q-RR is a straightforward combination of distributed random reshuffling and gradient quantization. This method can be seen as the stochastic without-replacement analogue of the distributed quantized gradient method of Khirirat et al. [2018].

Algorithm 1 Q-RR: Distributed Random Reshuffling with Quantization

Input: x_0 – starting point, $\gamma > 0$ – stepsize

for $t = 0, 1, \ldots, T-1$ **do**
  Receive x_t from the server and set $x_{t,m}^0 = x_t$
  Sample random permutation of $[n]$: $\pi_m = (\pi_m^0, \ldots, \pi_m^{n-1})$
  for $i = 0, 1, \ldots, n-1$ **do**
    for $m = 1, \ldots, M$ in parallel **do**
      Receive x_t^i from the server, compute and send $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)$ back
    end for
    Compute and send $x_t^{i+1} = x_t^i - \gamma \frac{1}{M} \sum_{m=1}^M \mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)$ to the workers
  end for
  $x_{t+1} = x_t^n$
end for

Output: x_T

We shall use the notion of *shuffling radius* defined by Mishchenko et al. [2021] for the analysis of distributed methods with random reshuffling:

Definition 2.1. *Define the iterate sequence* $x_\star^{i+1} = x_\star^i - \frac{\gamma}{M} \sum_{m=1}^M \nabla f_m^{\pi_m^i}(x_\star)$. *Then the shuffling radius is the quantity*

$$\sigma_{rad}^2 \overset{\text{def}}{=} \max_i \left\{ \frac{1}{\gamma^2 M} \sum_{m=1}^M \mathbb{E} D_{f_m^{\pi^i}}(x_\star^i, x_\star) \right\}.$$

We provide clarifications regarding this term in Appendix C.1. To compare our subsequent results with known ones, we introduce bounds on the shuffling radius. The following lemma demonstrates that these bounds are independent of the stepsize γ, even though γ is used in Definition 2.1.

Lemma 2.1 ([Mishchenko et al., 2020])**.** *Let Assumptions 3, 4 hold. Then the shuffling radius* σ_{rad}^2 *satisfies the following inequlity*

$$\frac{\widetilde{\mu} n}{8} \sigma_{\star,n}^2 \leq \sigma_{rad}^2 \leq \frac{L_{\max} n}{4} \sigma_{\star,n}^2,$$

where $\sigma_{\star,n}^2 \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \|\nabla f^i(x_\star)\|^2$, *and* $f^i = \frac{1}{M} \sum_{m=1}^M f_m^i$.

We now state the main convergence theorem for Algorithm 1:

Theorem 2.1. *Let Assumptions 1, 3, 4 hold and let the stepsize satisfy* $0 < \gamma \leq \frac{1}{\left(1+2\frac{\omega}{M}\right) L_{\max}}$. *Then, for all* $T \geq 0$ *the iterates produced by* Q-RR *(Algorithm 1) satisfy*

$$\mathbb{E}\|x_T - x_\star\|^2 \leq (1 - \gamma\widetilde{\mu})^{nT} \|x_0 - x_\star\|^2 + \frac{2\gamma^2 \sigma_{rad}^2}{\widetilde{\mu}} + \frac{2\gamma\omega}{\widetilde{\mu} M} (\zeta_\star^2 + \sigma_\star^2), \tag{5}$$

where $\zeta_\star^2 \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^M \|\nabla f_m(x_\star)\|^2$, *and* $\sigma_\star^2 \overset{\text{def}}{=} \frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \|\nabla f_m^i(x_\star) - \nabla f_m(x_\star)\|^2$.

All proofs are relegated to the appendix. By choosing the stepsize γ properly, we can obtain the communication complexity (number of communication rounds) needed to find an ε-approximate solution as follows:

Corollary 1. *Under the same conditions as Theorem 2.1 and for Algorithm 1, there exists a stepsize* $\gamma > 0$ *such that the number of communication rounds* nT *to find a solution with accuracy* $\varepsilon > 0$ *(i.e.* $\mathbb{E}\|x_T - x_\star\|^2 \leq \epsilon$*) is equal to* $\widetilde{\mathcal{O}}\left(\left(1 + \frac{\omega}{M}\right) \frac{L_{\max}}{\widetilde{\mu}} + \frac{\omega(\zeta_\star^2 + \sigma_\star^2)}{M\widetilde{\mu}^2\varepsilon} + \frac{\sigma_{rad}}{\sqrt{\widetilde{\mu}^3 \varepsilon}}\right)$, *where* $\widetilde{\mathcal{O}}(\cdot)$ *hides constants and logarithmic factors.*

The complexity of Quantized SGD (QSGD) is [Gorbunov et al., 2020]: $\widetilde{\mathcal{O}}\left(\left(1 + \frac{\omega}{M}\right) \frac{L_{\max}}{\mu} + \frac{\left(\omega\zeta_\star^2 + (1+\omega)\sigma_\star^2\right)}{M\mu^2\varepsilon}\right)$. For simplicity, let us neglect the differences between

Algorithm 2 DIANA-RR

Input: x_0 – starting point, $\{h_{0,m}^i\}_{m,i=1,1}^{M,n}$ – initial shift-vectors, $\gamma > 0$ – stepsize, $\alpha > 0$ – stepsize for learning the shifts

for $t = 0, 1, \ldots, T-1$ **do**
  Receive x_t from the server and set $x_{t,m}^0 = x_t$
  Sample random permutation of $[n]$: $\pi_m = (\pi_m^0, \ldots, \pi_m^{n-1})$
  for $i = 0, 1, \ldots, n-1$ **do**
    for $m = 1, 2, \ldots, M$ in parallel **do**
      Receive x_t^i from the server, compute and send $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right)$ back
      Set $\hat{g}_{t,m}^{\pi_m^i} = h_{t,m}^{\pi_m^i} + \mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_{t,m}^i) - h_{t,m}^{\pi_m^i}\right)$
      Set $h_{t+1,m}^{\pi_m^i} = h_{t,m}^{\pi_m^i} + \alpha\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_{t,m}^i) - h_{t,m}^{\pi_m^i}\right)$
    end for
    Compute $x_t^{i+1} = x_t^i - \gamma\frac{1}{M}\sum_{m=1}^M \hat{g}_{t,m}^{\pi_m^i}$ and send x_t^{i+1} to the workers
  end for
  $x_{t+1} = x_t^n$
end for

Output: x_T

μ and $\widetilde{\mu}$. First, when $\omega = 0$ we recover the complexity of FedRR [Mishchenko et al., 2021] which is known to be better than the one of SGD as long as ε is sufficiently small as we have ${}^{n\mu\sigma_{\star,n}^2}/_8 \leq \sigma_{\text{rad}}^2 \leq {}^{nL\sigma_{\star,n}^2}/_4$ from Lemma 2.1. Next, when $M = 1$ and $\omega = 0$ (single node, no compression) our results recovers the rate of RR [Mishchenko et al., 2020].

However, it is more interesting to compare Q-RR and QSGD when $M > 1$ and $\omega > 1$, which is typically the case. In these settings, Q-RR and QSGD have *the same complexity* since the $\mathcal{O}({}^1/_\varepsilon)$ term dominates the $\mathcal{O}({}^1/_{\sqrt{\varepsilon}})$ one if ε is sufficiently small. That is, the derived result for Q-RR has no advantages over the known one for QSGD unless ω is very small, which means that there is almost no compression at all. We also observe this phenomenon in the experiments.

The main reason for that is the variance appearing due to compression. Indeed, even if the current point is the solution of the problem ($x_t^i = x_*$), the update direction $-\gamma\frac{1}{M}\sum_{m=1}^M \mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)$ has the compression variance

$$\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{\gamma}{M}\sum_{m=1}^M\left(\mathcal{Q}(\nabla f_m^{\pi_m^i}(x_\star)) - \nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\right] \leq \frac{\gamma^2\omega}{M^2}\sum_{m=1}^M\|\nabla f_m^{\pi_m^i}(x_\star)\|^2.$$

This upper bound is tight and non-zero in general. Moreover, it is proportional to γ^2 that creates the term proportional to γ in (5) like in the convergence results for QSGD/SGD, while the RR-variance is proportional to γ^2 in the same bound. Therefore, during the later stages of the convergence Q-RR behaves similarly to QSGD when we decrease the stepsize.

2.2 Algorithm DIANA-RR

To reduce the additional variance caused by compression, we apply DIANA-style shift sequences [Mishchenko et al., 2019b, Horváth et al., 2019]. Thus we obtain DIANA-RR (Algorithm 2), which applies compression to the differences between the gradients and learnable shifts. Since the shifts are updated using the past gradients information, one can see DIANA-RR as a method with compression of gradient differences. We notice that unlike DIANA, DIANA-RR has n shift-vectors on each node.

Theorem 2.2. *Let Assumptions 1, 3, 4 hold and suppose that the stepsizes satisfy* $\gamma \leq \min\left\{\frac{\alpha}{2n\widetilde{\mu}}, \frac{1}{\left(1+\frac{6\omega}{M}\right)L_{\max}}\right\}$, *and* $\alpha \leq \frac{1}{1+\omega}$. *Define the following Lyapunov function for every* $t \geq 0$

$$\Psi_{t+1} \overset{\text{def}}{=} \|x_{t+1} - x_\star\|^2 + \frac{4\omega\gamma^2}{\alpha M^2}\sum_{m=1}^M\sum_{j=0}^{n-1}(1-\gamma\mu)^j\|\Delta_{t+1,m}^j\|^2, \tag{6}$$

Algorithm 3 Q-NASTYA

Input: x_0 – starting point, $\gamma > 0$ – local stepsize, $\eta > 0$ – global stepsize

for $t = 0, 1, \ldots, T-1$ **do**
  for $m \in [M]$ in parallel **do**
    Receive x_t from the server and set $x_{t,m}^0 = x_t$
    Sample random permutation of $[n]$: $\pi_m = (\pi_m^0, \ldots, \pi_m^{n-1})$
    for $i = 0, 1, \ldots, n-1$ **do**
      Set $x_{t,m}^{i+1} = x_{t,m}^i - \gamma \nabla f_m^{\pi_m^i}(x_{t,m}^i)$
    end for
    Compute $g_{t,m} = \frac{1}{\gamma n}\left(x_t - x_{t,m}^n\right)$ and send $\mathcal{Q}_t(g_{t,m})$ to the server
  end for
  Compute $g_t = \frac{1}{M}\sum_{m=1}^M \mathcal{Q}_t(g_{t,m})$
  Compute $x_{t+1} = x_t - \eta g_t$ and send x_{t+1} to the workers
end for

Output: x_T

where $\Delta_{t+1,m}^j = h_{t+1,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_\star)$ *Then, for all* $T \geq 0$ *the iterates produced by* DIANA-RR *(Algorithm 2) satisfy*

$$\mathbb{E}\left[\Psi_T\right] \leq (1 - \gamma\widetilde{\mu})^{nT}\,\Psi_0 + \frac{2\gamma^2\sigma_{rad}^2}{\widetilde{\mu}}$$

Corollary 2. *Under the same conditions as Theorem 2.2 and for Algorithm 2, there exists stepsizes* $\gamma, \alpha > 0$ *such that the number of communication rounds* nT *to find a solution with accuracy* $\varepsilon > 0$ *is* $\widetilde{\mathcal{O}}\left(n(1+\omega) + \left(1 + \frac{\omega}{M}\right)\frac{L_{\max}}{\widetilde{\mu}} + \frac{\sigma_{rad}}{\sqrt{\varepsilon\widetilde{\mu}^3}}\right)$.

Unlike Q-RR/QSGD/DIANA, DIANA-RR does not have a $\widetilde{\mathcal{O}}(1/\varepsilon)$-term, which makes it superior to Q-RR/QSGD/DIANA for small enough ε. However, the complexity of DIANA-RR has an additive $\widetilde{\mathcal{O}}(n(1+\omega))$ term arising due to learning the shifts $\{h_{t,m}^i\}_{m\in[M],i\in[n]}$. Nevertheless, this additional term is not the dominating one when ε is small enough. Next, we elaborate a bit more on the comparison between DIANA and DIANA-RR. That is, DIANA has $\widetilde{\mathcal{O}}\left(\left(1 + \frac{\omega}{M}\right)\frac{L_{\max}}{\mu} + \frac{(1+\omega)\sigma_\star^2}{M\mu^2\varepsilon}\right)$ complexity [Gorbunov et al., 2020]. Neglecting the differences between μ and $\widetilde{\mu}$, we observe a similar relation between DIANA-RR and DIANA as between RR and SGD: instead of the term $\mathcal{O}\left((1+\omega)\sigma_\star^2/(M\mu^2\varepsilon)\right)$ appearing in the complexity of DIANA, DIANA-RR has $\mathcal{O}\left(\sigma_{\text{rad}}/\sqrt{\varepsilon\widetilde{\mu}^3}\right)$ term much better depending on ε. To the best of our knowledge, our result is the only known one establishing the theoretical superiority of RR to regular SGD in the context of distributed learning with gradient compression. Moreover, when $\omega = 0$ (no compression) we recover the rate of FedRR and when additionally $M = 1$ (single worker) we recover the rate of RR.

2.3 Algorithms with Local Steps

In this subsection, we study a new variant of NASTYA, Q-NASTYA (Algorithm 3), that unifies quantization, local steps with random reshuffling, and uses different local and server stepsizes. Although it improves in per-round communication cost over NASTYA but, similar to Q-RR, we show that Q-NASTYA suffers from added variance due to gradient quantization. To overcome this issue, we propose another algorithm, DIANA-NASTYA (Algorithm 4), that adds DIANA-style variance reduction to Q-NASTYA and removes the additional variance.

Theorem 2.3. *Let Assumptions 1, 2, 3 hold. Let the stepsizes* γ, η *satisfy* $0 < \eta \leq \frac{1}{16L_{\max}\left(1+\frac{\omega}{M}\right)}$, $0 < \gamma \leq \frac{1}{5nL_{\max}}$. *Then, for all* $T \geq 0$ *the iterates produced by* Q-NASTYA *(Algorithm 3) satisfy*

$$\mathbb{E}\left[\|x_T - x_\star\|^2\right] \leq \left(1 - \frac{\eta\mu}{2}\right)^T \|x_0 - x_\star\|^2 + 8\frac{\eta\omega}{\mu M}\zeta_\star^2 + \frac{9}{2}\frac{\gamma^2 n L_{\max}}{\mu}\left((n+1)\zeta_\star^2 + \sigma_\star^2\right).$$

Corollary 3. *Under the same conditions as Theorem E.1 and for Algorithm 3, there exist stepsizes* $\gamma = \eta/n$ *and* $\eta > 0$ *such that the number of communication rounds* T *to find a solution with accuracy*

Algorithm 4 DIANA-NASTYA

Input: x_0 – starting point, $\{h_{0,m}\}_{m=1}^M$ – initial shift-vectors, $\gamma > 0$ – local stepsize, $\eta > 0$ – global stepsize, $\alpha > 0$ – stepsize for learning the shifts

for $t = 0, 1, \ldots, T-1$ **do**
  for $m = 1, \ldots, M$ in parallel **do**
    Receive x_t from the server and set $x_{t,m}^0 = x_t$
    Sample random permutation of $[n]$: $\pi_m = (\pi_m^0, \ldots, \pi_m^{n-1})$
    for $i = 0, 1, \ldots, n-1$ **do**
      Set $x_{t,m}^{i+1} = x_{t,m}^i - \gamma \nabla f_m^{\pi_m^i}(x_{t,m}^i)$
    end for
    Compute $g_{t,m} = \frac{1}{\gamma n}\left(x_t - x_{t,m}^n\right)$ and send $\mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$ to the server
    Set $h_{t+1,m} = h_{t,m} + \alpha \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
    Set $\hat{g}_{t,m} = h_{t,m} + \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
  end for
  $h_{t+1} = \frac{1}{M}\sum_{m=1}^M h_{t+1,m} = h_t + \frac{\alpha}{M}\sum_{m=1}^M \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
  $\hat{g}_t = \frac{1}{M}\sum_{m=1}^M \hat{g}_{t,m} = h_t + \frac{1}{M}\sum_{m=1}^M \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
  $x_{t+1} = x_t - \eta \hat{g}_t$
end for

Output: x_T

$\varepsilon > 0$ *is* $\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1 + \frac{\omega}{M}\right) + \frac{\omega}{M}\frac{\zeta_\star^2}{\varepsilon\mu^3} + \sqrt{\frac{L_{\max}}{\varepsilon\mu^3}}\sqrt{\zeta_\star^2 + \frac{\sigma_\star^2}{n}}\right)$. *If* $\gamma \to 0$, *one can choose* $\eta > 0$ *such that the above complexity bound improves to* $\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1 + \frac{\omega}{M}\right) + \frac{\omega}{M}\frac{\zeta_\star^2}{\varepsilon\mu^3}\right)$.

We emphasize several differences with the known theoretical results. First, the FedCOM method of Haddadpour et al. [2021] was analyzed in the homogeneous setting only, i.e., $f_m(x) = f(x)$ for all $m \in [M]$, which is an unrealistic assumption for FL applications. In contrast, our result holds in the fully heterogeneous case. Next, the analysis of FedPAQ of Reisizadeh et al. [2020] uses a bounded variance assumption, which is also known to be restrictive. Nevertheless, let us compare to their result. Reisizadeh et al. [2020] derive the following complexity for their method: $\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1 + \frac{\omega}{M}\right) + \frac{\omega}{M}\frac{\sigma^2}{\mu^2\varepsilon} + \frac{\sigma^2}{M\mu^2\varepsilon}\right)$. This result is inferior to the one we show for Q-NASTYA: when ω is small, the main term in the complexity bound of FedPAQ is $\widetilde{\mathcal{O}}\left(1/\varepsilon\right)$, while for Q-NASTYA the dominating term is of the order $\widetilde{\mathcal{O}}\left(1/\sqrt{\varepsilon}\right)$ (when ω and ε are sufficiently small). We also highlight that FedCRR [Malinovsky and Richtárik, 2022] does not converge if $\omega > M^2\gamma\mu\varepsilon/(2\|x_{*,m}^n\|^2)$, while Q-NASTYA does for any $\omega \geq 0$. Finally, when $\omega = 0$ (no compression) we recover NASTYA as a special case, and using $\gamma = \eta/n$, we recover the rate of FedRR [Mishchenko et al., 2021].

Theorem 2.4. *Let Assumptions 1, 2, 3 hold. Suppose the stepsizes* γ, η, α *satisfy* $0 < \gamma \leq \frac{1}{16L_{\max}n}$, $0 < \eta \leq \min\left\{\frac{\alpha}{2\mu}, \frac{1}{16L_{\max}\left(1+\frac{9\omega}{M}\right)}\right\}$, *and* $\alpha \leq \frac{1}{1+\omega}$. *Define the following Lyapunov function:*

$$\Psi_{t+1} \overset{\text{def}}{=} \|x_{t+1} - x_\star\|^2 + \frac{8\omega\eta^2}{\alpha M^2}\sum_{m=1}^M \|h_{t+1,m} - h_m^\star\|^2. \tag{7}$$

Then, for all $T \geq 0$ *the iterates produced by* DIANA-NASTYA *(Algorithm 4) satisfy*

$$\mathbb{E}\left[\Psi_T\right] \leq \left(1 - \frac{\eta\mu}{2}\right)^T \Psi_0 + \frac{9}{2}\frac{\gamma^2 nL}{\mu}\left((n+1)\zeta_\star^2 + \sigma_\star^2\right). \tag{8}$$

Corollary 4. *Under the same conditions as Theorem E.2 and for Algorithm 4, there exist stepsizes* $\gamma = \eta/n$, $\eta > 0$, $\alpha > 0$ *such that the number of communication rounds* T *to find a solution with accuracy* $\varepsilon > 0$ *is* $\widetilde{\mathcal{O}}\left(\omega + \frac{L_{\max}}{\mu}\left(1 + \frac{\omega}{M}\right) + \sqrt{\frac{L_{\max}}{\varepsilon\mu^3}}\sqrt{\zeta_\star^2 + \frac{\sigma_\star^2}{n}}\right)$. *If* $\gamma \to 0$, *one can choose* $\eta > 0$ *such that the above complexity bound improves to* $\widetilde{\mathcal{O}}\left(\omega + \frac{L_{\max}}{\mu}\left(1 + \frac{\omega}{M}\right)\right)$.

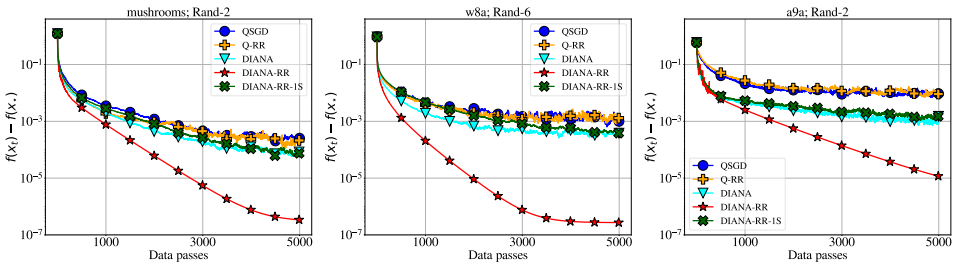

Figure 1: Non-local methods

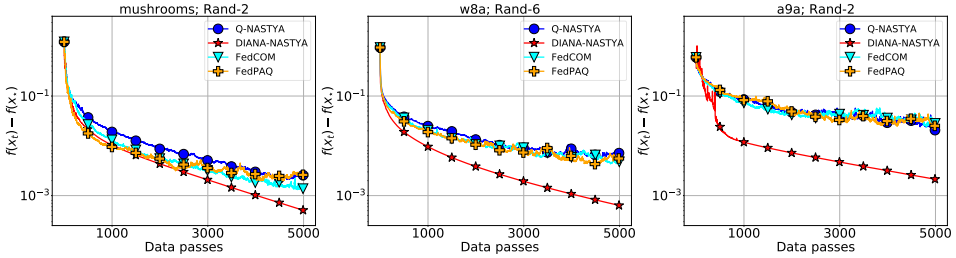

Figure 2: Local methods

Figure 3: The comparison of the proposed methods (Q-NASTYA, DIANA-NASTYA, Q-RR, DIANA-RR), DIANA-RR-1S (a modification of DIANA-RR), and existing baselines (QSGD, DIANA, FedCOM, FedPAQ). All methods use tuned stepsizes and the Rand-k compressor.

In contrast to Q-NASTYA, DIANA-NASTYA does not suffer from the $\widetilde{\mathcal{O}}(1/\varepsilon)$ term in the complexity bound. This shows the superiority of DIANA-NASTYA to Q-NASTYA. Next, FedCRR-VR [Malinovsky and Richtárik, 2022] has the rate $\widetilde{\mathcal{O}}\left(\frac{(\omega+1)\left(1-\frac{1}{\kappa}\right)^n}{\left(1-\left(1-\frac{1}{\kappa}\right)^n\right)^2}+\frac{\sqrt{\kappa}(\zeta_\star+\sigma_\star)}{\mu\sqrt{\varepsilon}}\right)$, which depends on $\widetilde{\mathcal{O}}\left(1/\sqrt{\varepsilon}\right)$. However, the first term is close to $\widetilde{\mathcal{O}}\left((\omega+1)\kappa^2\right)$ for a large condition number. FedCRR-VR-2 utilizes variance reduction technique from Malinovsky et al. [2021] and it allows to get rid of permutation variance. This method has $\widetilde{\mathcal{O}}\left(\frac{(\omega+1)\left(1-\frac{1}{\kappa\sqrt{\kappa n}}\right)^{\frac{n}{2}}}{\left(1-\left(1-\frac{1}{\kappa\sqrt{\kappa n}}\right)^{\frac{n}{2}}\right)^2}+\frac{\sqrt{\kappa}\zeta_\star}{\mu\sqrt{\varepsilon}}\right)$ complexity, but it requires additional assumption on number of functions n and thus not directly comparable with our result. Note that if we have no compression ($\omega=0$), DIANA-NASTYA recovers rate of NASTYA.

In Appendix J, we provide versions of Q-NASTYA and DIANA-NASTYA with partial participation of clients, which is another important aspect of FL, and derive the convergence results for them.

3 Experiments

We evaluated our methods for solving logistic regression problems and training neural networks in three parts: (i) Comparison of the proposed non-local methods with existing baselines; (ii) Comparison of the proposed local methods with existing baselines; (iii) Comparison of the proposed non-local methods in training `ResNet-18` on `CIFAR10`.

Logistic Regression. To confirm our theoretical results we conducted several numerical experiments on binary classification problem with L2 regularized logistic regression of the form

$$\min_{x\in\mathbb{R}^d}\left[f(x)\overset{\text{def}}{=}\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m}f_{m,i}\right], \tag{9}$$

where $f_{m,i}\overset{\text{def}}{=}\log\left(1+\exp(-y_{mi}a_{mi}^\top x)\right)+\lambda\|x\|_2^2$ $(a_{mi},y_{mi})\in\mathbb{R}^d\times\in\{-1,1\}, i=1,\ldots,n_m$ are the training data samples stored on machines $m=1,\ldots,M$, and $\lambda>0$ is a regularization parameter. In all experiments, for each method, we used the largest stepsize allowed by its theory

multiplied by some individually tuned constant multiplier. For better parallelism, each worker m uses mini-batches of size $\approx 0.1 n_m$. In all algorithms, as a compression operator $\mathcal{Q}$, we use Rand-k [Beznosikov et al., 2020] with fixed compression ratio $k/d \approx 0.02$, where d is the number of features in the dataset.

In our first experiment (see Figure 1), we compare Q-RR, DIANA-RR, and DIANA-RR-1S with classical baselines (QSGD [Alistarh et al., 2017], DIANA [Mishchenko et al., 2019b]) that use a with-replacement mini-batch SGD estimator. DIANA-RR-1S is a memory-friendly version of DIANA-RR that stores and uses a single shift $h_{t,m}$ on the worker side rather than n individual shifts $h_{t,m}^{\pi_m^i}$. Figure 1 illustrates that Q-RR exhibits similar behavior to QSGD, with both methods being slower than DIANA methods across all considered datasets. DIANA-RR-1S and DIANA show comparable convergence rates, indicating that random reshuffling alone, without introducing additional shifts, does not make a significant difference. Finally, DIANA-RR achieves the best rate among all considered non-local methods, efficiently reducing the variance and reaching the lowest functional sub-optimality tolerance. These experimental results align perfectly with our theoretical analysis.

The second experiment shows that DIANA-based method can significantly outperform in practice when one applies it to local methods as well. In particular, whereas Q-NASTYA shows comparative behavior as existing methods FedCOM [Haddadpour et al., 2021], FedPAQ [Reisizadeh et al., 2020] in all considered datasets, DIANA-NASTYA noticeably outperforms other methods.

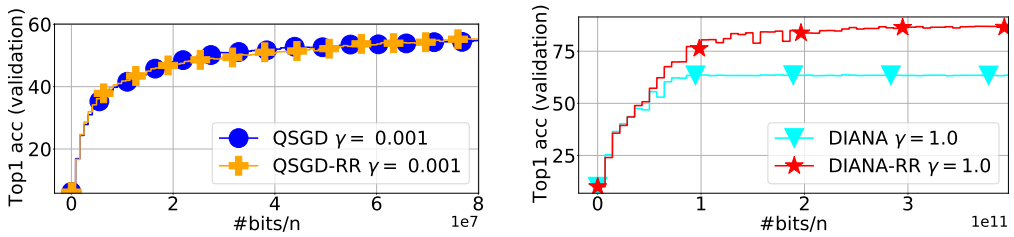

Figure 4: The comparison of Q-RR, QSGD, DIANA, and DIANA-RR on the task of training `ResNet-18` on `CIFAR-10` with $n = 10$ workers. Top-1 accuracy on test set is reported. Stepsizes were tuned and workers used Rand-k compressor with $k/d \approx 0.05$.

Training Deep Neural Network model: ResNet-18 on CIFAR-10. Since random reshuffling is a very popular technique in training neural networks, it is natural to test the proposed methods on such problems. Therefore, in the second set of experiments, we consider training `ResNet-18` [He et al., 2016] model on the `CIFAR10` dataset Krizhevsky and Hinton [2009]. To conduct these experiments we use `FL_PyTorch` simulator [Burlachenko et al., 2021].

The main goal of this experiment is to verify the phenomenon observed in Experiment 1 on the training of a deep neural network. That is, we tested Q-RR, QSGD, DIANA, and DIANA-RR in the distributed training of `ResNet-18` on `CIFAR10`, see Figure 4. As in the logistic regression experiments, we observe that (i) Q-RR and QSGD behave similarly and (ii) DIANA-RR outperforms DIANA. For further experimental results and details, we refer to Appendix B.

4 Conclusion

In this work, we provide the first study of distributed random reshuffling with communication compression. Our theoretical and empirical findings illustrate the inefficiency of naïve combination of random reshuffling and communication compression. We also show how this issue can be resolved via the usage of shifts for communication compression. Finally, we develop and analyze methods with random reshuffling, communication compression, and local steps. It is worth mentioning that although our theoretical results are obtained for strongly convex problems, the considered methods perform well in the experiments on non-convex tasks like training neural networks.

Acknowledgements

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

The work of A. Sadiev and E. Gorbunov (while affiliated with MIPT) was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730324P540002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: optimal rates without component convexity and large epoch requirements. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/cb8acb1dc9821bf74e6ca9068032d623-[]Abstract.html.

Dan Alistarh, Demjan Grubic, Jerry Z. Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1707–1718, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol.*, 13(4), 2022. ISSN 2157-6904. doi: 10.1145/3501813. URL https://doi.org/10.1145/3501813.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. volume 32, 2019.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, abs/2002.12410, 2020. URL https://arXiv.org/abs/2002.12410.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pages 2624–2633. Citeseer, 2009.

Konstantin Burlachenko, Samuel Horváth, and Peter Richtárik. Fl_pytorch: optimization research simulator for federated learning. In *Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning*, pages 1–7, 2021.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Margalit R. Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local SGD) and continuous perspective. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9050–9090. PMLR, 2022. URL https://proceedings.mlr.press/v151/glasgow22a.html.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3556–3564. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/gorbunov21a.html.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2350–2358. PMLR, 2021. URL http://proceedings.mlr.press/v130/haddadpour21a.html.

K. He et al. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, abs/1904.05115, 2019. URL https://arXiv.org/abs/1904.05115.

Samuel Horváth, Maziar Sanjabi, Lin Xiao, Peter Richtárik, and Michael Rabbat. FedShuffle: Recipes for better use of local work in federated learning. *arXiv preprint arXiv:2204.13169*, abs/2204.13169, 2022. URL https://arXiv.org/abs/2204.13169.

Kun Huang, Xiao Li, Andre Milzarek, Shi Pu, and Junwen Qiu. Distributed random reshuffling over networks. *arXiv preprint arXiv:2112.15287*, abs/2112.15287, 2021. URL https://arXiv.org/abs/2112.15287.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, abs/1912.04977, 2019. URL https://arXiv.org/abs/1912.04977.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 2020. URL http://proceedings.mlr.press/v119/karimireddy20a.html.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, abs/1806.06573, 2018. URL https://arXiv.org/abs/1806.06573.

Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Ming Liu, Stella Ho, Mengqi Wang, Longxiang Gao, Yuan Jin, and He Zhang. Federated learning meets natural language processing: a survey. *arXiv preprint arXiv:2107.12603*, abs/2107.12603, 2021. URL https://arXiv.org/abs/2107.12603.

Grigory Malinovsky and Peter Richtárik. Federated random reshuffling with compression and variance reduction. *arXiv preprint arXiv:2205.03914*, abs/2205.03914, 2022. URL https://arXiv.org/abs/2205.03914.

Grigory Malinovsky, Alibek Sailanbayev, and Peter Richtárik. Random reshuffling with variance reduction: New analysis and better rates. *arXiv preprint arXiv:2104.09342*, 2021.

Grigory Malinovsky, Konstantin Mishchenko, and Peter Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization. *arXiv preprint arXiv:2201.11066*, abs/2201.11066, 2022. URL https://arXiv.org/abs/2201.11066.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019a.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takác, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, abs/1901.09269, 2019b. URL https://arXiv.org/abs/1901.09269.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8cc6e90ccbff44c9cee23611711cdc4-[]Abstract.html.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. *arXiv preprint arXiv:2102.06704*, abs/2102.06704, 2021. URL https://arXiv.org/abs/2102.06704.

Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14606–14619. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7a6bda9ad6ffdac035c752743b7e9d0e-[]Paper.pdf.

Tomoya Murata and Taiji Suzuki. Bias-variance reduced local SGD for less heterogeneous federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7872–7881. PMLR, 2021. URL http://proceedings.mlr.press/v139/murata21a.html.

Jose Javier Gonzalez Ortiz, Jonathan Frankle, Mike Rabbat, Ari Morcos, and Nicolas Ballas. Trade-offs of local SGD at scale: an empirical study. *arXiv preprint arXiv:2110.08133*, abs/2110.08133, 2021. URL https://arXiv.org/abs/2110.08133.

Shashank Rajput, Anant Gupta, and Dimitris S. Papailiopoulos. Closing the convergence gap of SGD without replacement. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7964–7973. PMLR, 2020. URL http://proceedings.mlr.press/v119/rajput20a.html.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2021–2031. PMLR, 2020. URL http://proceedings.mlr.press/v108/reisizadeh20a.html.

Peter Richtárik, Elnur Gasanov, and Konstantin Burlachenko. Error feedback reloaded: From quadratic to arithmetic mean of smoothness constants. *arXiv preprint arXiv:2402.10774*, 2024.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Mher Safaryan, Filip Hanzely, and Peter Richtárik. Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25688–25702. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d79c6256b9bdac53a55801a066b70da3-[]Paper.pdf.

Itay Safran and Ohad Shamir. Random shuffling beats SGD only after many epochs on ill-conditioned problems. *arXiv preprint arXiv:2106.06880*, abs/2106.06880, 2021. URL https://arXiv.org/abs/2106.06880.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, abs/1907.04232, 2019. URL https://arXiv.org/abs/1907.04232.

Sebastian U. Stich. On communication compression for distributed optimization on heterogeneous data. *arXiv preprint arXiv:2009.02388*, abs/2009.02388, 2020. URL https://arXiv.org/abs/2009.02388.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4452–4463, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/b440509a0106086a67bc2ea9df0a1dab-[]Abstract.html.

Zhenheng Tang, Shaohuai Shi, Xiaowen Chu, Wei Wang, and Bo Li. Communication-efficient distributed deep learning: a comprehensive survey. *arXiv preprint arXiv:2003.06307*, abs/2003.06307, 2020. URL https://arXiv.org/abs/2003.06307.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/3328bdf9a4b9504b9398284244fe97c2-[]Paper.pdf.

Blake Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. *arXiv preprint arXiv:2102.01583*, abs/2102.01583, 2021. URL https://arXiv.org/abs/2102.01583.

Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/45713f6ff2041d3fdfae927b82488db8-[]Abstract.html.

Blake E. 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? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10334–10343. PMLR, 2020b. URL http://proceedings.mlr.press/v119/woodworth20a.html.

Zhaohui Yang, Mingzhe Chen, Kai-Kit Wong, H. Vincent Poor, and Shuguang Cui. Federated learning for 6g: Applications, challenges, and opportunities. *Engineering*, 8:33–41, 2022. ISSN 2095-8099. doi: https://doi.org/10.1016/j.eng.2021.12.002. URL https://www.sciencedirect.com/science/article/pii/S2095809921005245.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Chulhee Yun, Shashank Rajput, and Suvrit Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. *arXiv preprint arXiv:2110.10342*, abs/2110.10342, 2021. URL https://arXiv.org/abs/2110.10342.

Contents

A Extra Related Works

Federated optimization has been the subject of intense study, with many open questions even in the setting when all clients have identical data [Woodworth et al., 2020b,a, 2021]. The FedAvg algorithm (also known as Local SGD) has also been a subject of intense study, with tight bounds obtained only very recently by Glasgow et al. [2022]. It is now understood that using many local steps adds bias to distributed SGD, and hence several methods have been developed to mitigate it, e.g. [Karimireddy et al., 2020, Murata and Suzuki, 2021], see the work of Gorbunov et al. [2021] for a unifying lens on many variants of Local SGD. Note that despite the bias, even vanilla FedAvg/Local SGD still reduces the overall communication overhead in practice [Ortiz et al., 2021].

The success of RR in the single-machine setting has inspired several recent methods that use it as a local update method as part of distributed training: Mishchenko et al. [2021] developed a distributed variant of random reshuffling, FedRR. FedRR uses RR as a local client update method in lieu of SGD. They show that FedRR can improve upon the convergence of Local SGD when the number of local steps is fixed as the local dataset size, i.e. when $H = n$. Yun et al. [2021] study the same method under the name Local RR under a more restrictive assumption of bounded inter-machine gradient deviation and show that by varying H to be smaller than n better rates can be obtained in this setting than the rates of Mishchenko et al. [2021]. Other work has explored more such combinations between RR and distributed training algorithms [Huang et al., 2021, Malinovsky et al., 2022, Horváth et al., 2022].

There are several methods that combine compression or quantization and local steps: both Basu et al. [2019] and Reisizadeh et al. [2020] combined Local SGD with quantization and sparsification, and Haddadpour et al. [2021] later improved their results using a gradient tracking method, achieving linear convergence under strong convexity. In parallel, Mitra et al. [2021] also developed a variance-reduced method, FedLin, that achieves linear convergence under strong convexity despite using local steps and compression. The paper most related to our work is [Malinovsky and Richtárik, 2022] in which the authors combine *iterate* compression, random reshuffling, and local steps. We study *gradient* compression instead, which is a more common form of compression in both theory and practice [Kairouz et al., 2019]. We compare our results against [Malinovsky and Richtárik, 2022] and show we obtain better rates compared to their work.

B Experimental details

In this section, we provide missing details on the experimental setting from Section 3. The codes are provided in the following anonymous repository: https://anonymous.4open.science/r/diana_rr-[]B0A5.

B.1 Logistic Regression

To confirm our theoretical results we conducted several numerical experiments on binary classification problem with L2 regularized logistic regression of the form

$$\min_{x\in\mathbb{R}^d}\left[f(x) \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m} f_{m,i}\right], \tag{10}$$

where $f_{m,i} \overset{\text{def}}{=} \log\left(1+\exp(-y_{mi}a_{mi}^\top x)\right) + \lambda\|x\|_2^2$ $(a_{mi}, y_{mi}) \in \mathbb{R}^d \times \in \{-1,1\}, i = 1,\dots,n_m$ are the training data samples stored on machines $m = 1,\dots,M$, and $\lambda > 0$ is a regularization parameter. In all experiments, for each method, we used the largest stepsize allowed by its theory multiplied by some individually tuned constant multiplier. For better parallelism, each worker m uses mini-batches of size $\approx 0.1 n_m$. In all algorithms, as a compression operator $\mathcal{Q}$, we use Rand-k [Beznosikov et al., 2020] with fixed compression ratio $k/d \approx 0.02$, where d is the number of features in the dataset.

Hardware and Software. All algorithms were written in Python 3.8. We used three different CPU cluster node types:

1. AMD EPYC 7702 64-Core;

2. Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz;

3. Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz.

Datasets. The datasets were taken from open LibSVM library Chang and Lin [2011], sorted in ascending order of labels, and equally split among 20 machines \clients\workers. The remaining part of size $N - 20 \cdot \lfloor N/20 \rfloor$ was assigned to the last worker, where $N = \sum_{m=1}^{M} n_m$ is the total size of the dataset. A summary of the splitting and the data samples distribution between clients can be found in Tables 2, 3, 4, 5.

Table 2: Summary of the datasets and splitting of the data samples among clients.

Dataset	M	N (dataset size)	d (# of features)	n_m (# of datasamples per client)
`mushrooms`	20	8120	112	406
`w8a`	20	49749	300	2487
`a9a`	20	32560	123	1628

Table 3: Partition of the `mushrooms` dataset among clients.

Client's №	# of datasamples of class "-1"	# of datasamples of class "+1"
1 – 9	406	0
10	262	144
11 – 19	0	406
20	0	410

Table 4: Partition of the `w8a` dataset among clients.

Client's №	# of datasamples of class "-1"	# of datasamples of class "+1"
1 – 19	2487	0
20	1017	1479

Table 5: Partition of the `a9a` dataset among clients.

Client's №	# of datasamples of class "-1"	# of datasamples of class "+1"
1 – 14	1628	0
15	1328	300
16 – 19	0	1628
20	0	1629

Hyperparameters. Regularization parameter λ was chosen individually for each dataset to guarantee the condition number L/μ to be approximately 10^4, where L and μ are the smoothness and strong-convexity constants of function f. For the chosen logistic regression problem of the form (10), smoothness and strong convexity constants L, L_m, $L_{i,m}$, μ, $\widetilde{\mu}$ of functions f, f_m and f_m^i were computed explicitly as

$$
\begin{aligned}
L &= \lambda_{\max}\left(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{4n_m}\mathbf{A}_m^\top\mathbf{A}_m + 2\lambda\mathbf{I}\right)\\
L_m &= \lambda_{\max}\left(\frac{1}{4n_m}\mathbf{A}_m^\top\mathbf{A}_m + 2\lambda\mathbf{I}\right)\\
L_{i,m} &= \lambda_{\max}\left(\frac{1}{4}a_{mi}a_{mi}^\top + 2\lambda\mathbf{I}\right)\\
\mu &= 2\lambda\\
\widetilde{\mu} &= 2\lambda,
\end{aligned}
$$

where $\mathbf{A}_m$ is the dataset associated with client m, and a_{mi} is the i-th row of data matrix $\mathbf{A}_m$. In general, the fact that f is L-smooth with

$$
L \leq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m}L_{i,m}
$$

follows from the $L_{i,m}$-smoothness of f_m^i (see Assumption 3).

In all algorithms, as a compression operator $\mathcal{Q}$, we use Rand-k as a canonical example of unbiased compressor with relatively bounded variance, and fix the compression parameter $k = \lfloor 0.02d \rfloor$, where d is the number of features in the dataset.

In addition, in all algorithms, for all clients $m = 1, \ldots, M$, we set the batch size for the SGD estimator to be $b_m = \lfloor 0.1 n_m \rfloor$, where n_m is the size of the local dataset.

The summary of the values L, L_m, $L_{i,m}$ $L_{\max}$, μ, b_m and k for each dataset can be found in Table 6.

Table 6: Summary of the hyperparameters.

Dataset	L	$L_{\max}$	μ	λ	k	b_m (batchsize)
`mushrooms`	2.59	5.25	$2.58 \cdot 10^{-4}$	$1.29 \cdot 10^{-4}$	2	40
`w8a`	0.66	28.5	$6.6 \cdot 10^{-5}$	$3.3 \cdot 10^{-5}$	6	248
`a9a`	1.57	3.5	$1.57 \cdot 10^{-4}$	$7.85 \cdot 10^{-5}$	2	162

In all experiments, we follow constant stepsize strategy within the whole iteration procedure. For each method, we set the largest possible stepsize predicted by its theory multiplied by some individually tuned constant multiplier. For a more detailed explanation of the tuning routine, see Sections B.1.1 and B.1.2.

SGD implementation. We considered two approaches to minibatching: random reshuffling and with-replacement sampling. In the first, all clients $m = 1, \ldots, M$ independently permute their local datasets and pass through them within the next subsequent $\lfloor \frac{n_m}{b_m} \rfloor$ steps. In our implementations of Q-RR, Q-NASTYA and DIANA-NASTYA, all clients permuted their datasets in the beginning of every new epoch, whereas for the DIANA-RR method they do so only once in the beginning of the iteration procedure. Second approach of minibatching is called with-replacement sampling, and it requires every client to draw b_m data samples from the local dataset uniformly at random. We used this strategy in the baseline algorithms (QSGD, DIANA, FedCOM and FedPAQ) we compared our proposed methods to.

Experimental setup. To compare the performance of methods within the whole optimization process, we track the functional suboptimality metric $f(x_t) - f(x_\star)$ that was recomputed after each epoch. For each dataset, the value $f(x_\star)$ was computed once at the preprocessing stage with 10^{-16} tolerance via conjugate gradient method. We terminate our algorithms after performing 5000 epochs.

B.1.1 Experiment 1: Comparison of the Proposed Non-Local Methods with Existing Baselines

In our first experiment (see Figure 1), we compare Q-RR, DIANA-RR, and DIANA-RR-1S with classical baselines (QSGD [Alistarh et al., 2017], DIANA [Mishchenko et al., 2019b]) that use

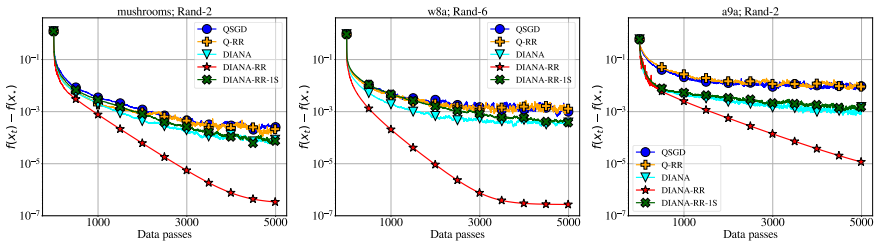

Figure 5: Non-local methods

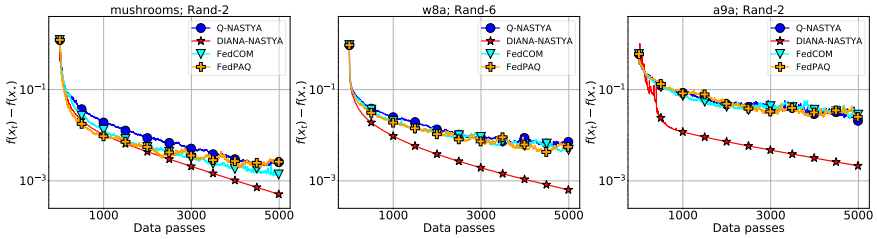

Figure 6: Local methods

Figure 7: The comparison of the proposed methods (Q-NASTYA, DIANA-NASTYA, Q-RR, DIANA-RR), DIANA-RR-1S (a modification of DIANA-RR), and existing baselines (QSGD, DIANA, FedCOM, FedPAQ). All methods use tuned stepsizes and the Rand-k compressor.

a with-replacement mini-batch SGD estimator. DIANA-RR-1S is a memory-friendly version of DIANA-RR that stores and uses a single shift $h_{t,m}$ on the worker side rather than n individual shifts $h_{t,m}^{\pi_m^i}$. Figure 1 illustrates that Q-RR exhibits similar behavior to QSGD, with both methods being slower than DIANA methods across all considered datasets. DIANA-RR-1S and DIANA show comparable convergence rates, indicating that random reshuffling alone, without introducing additional shifts, does not make a significant difference. Finally, DIANA-RR achieves the best rate among all considered non-local methods, efficiently reducing the variance and reaching the lowest functional sub-optimality tolerance. These experimental results align perfectly with our theoretical analysis. For each of the considered non-local methods, we take the stepsize as the largest one predicted by the theory premultiplied by the individually tuned constant factor from the set $\{0.000975, 0.00195, 0.0039, 0.0078, 0.0156, 0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096\}$.

Therefore, for each local method on every dataset, we performed 20 launches to find the stepsize multiplier showing the best convergence behavior (the fastest reaching the lowest possible level of functional suboptimality $f(x_t) - f(x_\star)$).

Theoretical stepsizes for methods Q-RR and DIANA-RR are provided by the Theorems 2.1 and 2.2, whereas stepsizes for QSGD and DIANA were taken from the paper Gorbunov et al. [2020].

B.1.2 Experiment 2: Comparison of the Proposed Local Methods with Existing Baselines

The second experiment shows that DIANA-based method can significantly outperform in practice when one applies it to local methods as well. In particular, whereas Q-NASTYA shows comparative behavior as existing methods FedCOM [Haddadpour et al., 2021], FedPAQ [Reisizadeh et al., 2020] in all considered datasets, DIANA-NASTYA noticeably outperforms other methods.

In this set of experiments, we tuned stepsizes similarly to the non-local methods. However, for algorithms Q-NASTYA, DIANA-NASTYA, and FedCOM we needed to independently adjust the client and server stepsizes, leading to a more extensive tunning routine.

As before, for each local method on every dataset, tuned client and server stepsizes are defined by the theoretical one and adjusted constant multiplier. Theoretical stepsizes for methods Q-NASTYA and DIANA-NASTYA are given by the Theorems E.1 and E.2, whereas FedCOM and FedPAQ stepsizes were taken from the papers by Haddadpour et al. [2021] and Reisizadeh et al. [2020] respectively.

We now list all the considered multipliers of client and server stepsizes for every method (i.e. γ and η respectively):

- Q-NASTYA:
  - Multipliers for γ : {0.000975, 0.00195, 0.0039, 0.0078, 0.0156, 0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64,128};
  - Multipliers for η : {0.0039, 0.0078, 0.0156, 0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128}.
- DIANA-NASTYA:
  - Multipliers for γ and η : {0.000975, 0.00195, 0.0039, 0.0078, 0.0156, 0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64,128};
- FedCOM:
  - Multipliers for γ : {0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512,1024, 2048, 4096, 8192, 16384, 32768 };
  - Multipliers for η : {0.000975, 0.00195, 0.0039, 0.0078, 0.0156, 0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128}.
- FedPAQ:
  - Multipliers for γ : {0.00195, 0.0039, 0.0078, 0.0156, 0.0312, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, 131072, 262144, 524288, 1048576 }.

For example, to find the best pair (γ, η) for FedCOM method on each dataset, we performed 378 launches. A similar subroutine was executed for all algorithms on all datasets independently.

B.1.3 Experiment 3: Comparison of DIANA-RR with EF21 and DIANA

In our third experiment (see Figure 8), we compared DIANA-RR with DIANA and EF21-SGD [Fatkhullin et al., 2021]. The EF21-SGD is the state-of-the-art algorithm for contractive compressors in distributed non-convex settings. All compared algorithms used a with-replacement mini-batch SGD estimator, consistent with the setup in Section B.1.1. However, in this experiment contrast with Section B.1.1, we employed reliable theoretical step sizes that ensure guaranteed convergence.

The EF21 algorithm family is designed for usage with contraction compressors in non-convex optimization for L-smooth objective functions in for of Equation 1. For scenarios involving unbiased compressors such as Rand-k, the EF21-SGD can be adapted through scaling [Fatkhullin et al., 2021]. More specifically, an unbiased compressor $\mathcal{C}(x) : \mathbb{R}^d \to \mathbb{R}^d$ which satisfied Assumption 1 via applying the transformation $C'(x) \overset{\text{def}}{=} (\omega+1)^{-1} \cdot \mathcal{C}(x)$ yields a contraction compressor $\mathbb{E}\left[\|C'(x) - x\|^2\right] \leq (1-\alpha)\|x\|^2, \forall x \in \mathbb{R}^d$ with $\alpha = 1/\omega+1$. In particular, this procedure makes Rand-k compatible with the EF21 algorithm family. In this experiment, we implemented EF21-SGD following the refined analysis from [Richtárik et al., 2024], which offers a stricter better convergence guarantee through improved bounds on the theoretical step size compared to [Fatkhullin et al., 2021].

As shown in Figure 8 EF21-SGD does not perform fast enough in scenarios involving using unbiased compressors for strongly-convex optimization problems compared to DIANA-RR and DIANA.

B.2 Training Deep Neural Network model: ResNet-18 on CIFAR-10

Since random reshuffling is a very popular technique in training neural networks, it is natural to test the proposed methods on such problems. Therefore, in the second set of experiments, we consider training `ResNet-18` [He et al., 2016] model on the `CIFAR10` dataset Krizhevsky and Hinton [2009]. To conduct these experiments we use `FL_PyTorch` simulator [Burlachenko et al., 2021].

The main goal of this experiment is to verify the phenomenon observed in Experiment 1 on the training of a deep neural network. That is, we tested Q-RR, QSGD, DIANA, and DIANA-RR in the distributed training of `ResNet-18` on `CIFAR10`, see Figure 9. As in the logistic regression experiments, we observe that (i) Q-RR and QSGD behave similarly and (ii) DIANA-RR outperforms DIANA.

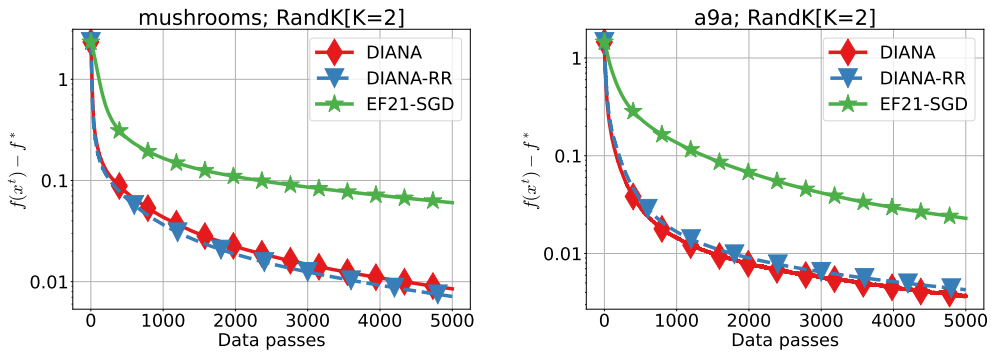

Figure 8: The comparison of the proposed variance-reduced DIANA-RR and baselines DIANA, EF21-SGD. All algorithms use theoretical step-sizes, Rand-k compressor, number of workers is 20.

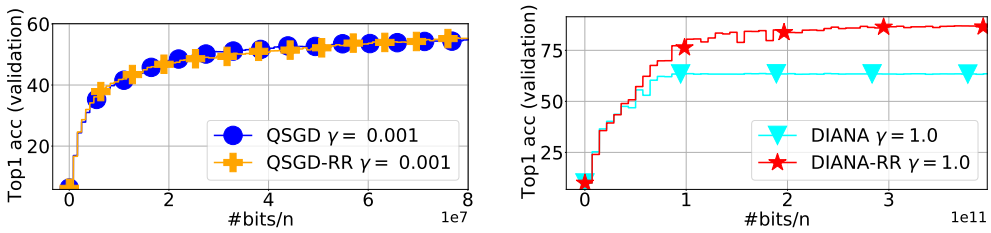

Figure 9: The comparison of Q-RR, QSGD, DIANA, and DIANA-RR on the task of training `ResNet-18` on `CIFAR-10` with $n = 10$ workers. Top-1 accuracy on test set is reported. Stepsizes were tuned and workers used Rand-k compressor with $k/d \approx 0.05$.

To illustrate the behavior of the proposed methods in training Deep Neural Networks (DNN), we consider the `ResNet-18` [He et al., 2016] model. This model is used for image classification, feature extraction for image segmentation, object detection, image embedding, and image captioning. We train all layers of `ResNet-18` model meaning that the dimension of the optimization problem equals $d = 11,173,962$. During the training, the `ResNet-18` model normalizes layer inputs via exploiting 20 Batch Normalization [Ioffe and Szegedy, 2015] layers that are applied directly before nonlinearity in the computation graph of this model. Batch normalization (BN) layers add 9600 trainable parameters to the model. Besides trainable parameters, a BN layer has its internal state that is used for computing the running mean and variance of inputs due to its own specific regime of working. We use *He* initialization [He et al., 2015].

B.2.1 Computing Environment

We performed numerical experiments on a server-grade machine running Ubuntu 18.04 and Linux Kernel v5.4.0, equipped with 16-cores (2 sockets by 16 cores per socket) 3.3 GHz Intel Xeon, and four NVIDIA A100 GPU with 40GB of GPU memory. The distributed environment is simulated in Python 3.9 via using the software suite `FL_PyTorch` [Burlachenko et al., 2021] that serves for carrying complex Federate Learning experiments. `FL_PyTorch` allowed us to simulate the distributed environment in the local machine. Besides storing trainable parameters per client, this simulator stores all not trainable parameters including BN statistics per client.

B.2.2 Loss Function

Training of `ResNet-18` can be formalized as problem (1) with the following choice of f_m^i

$$f_m(x) = \frac{1}{|n_m|} \sum_{j=1}^{|n_m|} CE(b^{(j)}, g(a^{(j)}, x)), \tag{11}$$

where $CE(p,q) \overset{\text{def}}{=} -\sum_{k=1}^{\#\text{classes}} p_i \cdot \log(q_i)$ with agreement $0 \cdot \log(0) = 0$ is a standard cross-entropy loss, function $g : \mathbb{R}^{28\times 28} \times \mathbb{R}^d \to [0,1]^{\#\text{classes}}$ is a neural network taking image $a^{(j)}$ and vector of parameters x as an input and returning a vector in probability simplex, and n_m is the size of the dataset on worker m.

B.2.3 Dataset and Metric

In our experiments, we used `CIFAR10` dataset Krizhevsky and Hinton [2009]. The dataset consists of input variables $a_i \in \mathbb{R}^{28\times 28\times 3}$, and response variables $b_i \in \{0,1\}^{10}$ and is used for training 10-way classification. The sizes of training and validation set are 5×10^4 and 10^4 respectively. The training set is partitioned heterogeneously across 10 clients. To measure the performance, we evaluate the loss function value $f(x)$, norm of the gradient $\|\nabla f(x)\|_2$ and the Top-1 accuracy of the obtained model as a function of passed epochs and the normalized number of bits sent from clients to the server.

B.2.4 Tuning Process

In this set of experiments, we tested QSGD [Alistarh et al., 2017], Q-RR (Algorithm 1), DIANA [Mishchenko et al., 2019a] and DIANA-RR (Algorithm 2) algorithms. For all algorithms, we tuned the strategy $\in \{A, B, C\}$ of decaying stepsize model via selecting the best in terms of the norm of the full gradient on the train set in the final iterate produced after 20000 rounds. The stepsize policies are described below.

A. Stepsizes decaying as inverse square root of the number epochs

$$\gamma_e = \begin{cases} \gamma_{init} \cdot \dfrac{1}{\sqrt{e-s+1}}, & \text{if } e \geq s, \\ \gamma_{init}, & \text{if } e < s, \end{cases}$$

where γ_e denotes the stepsize used during epoch $e+1$, s is a fixed shift.

B. Stepsizes decaying as inverse of number epochs

$$\gamma = \begin{cases} \gamma_{init} \cdot \dfrac{1}{e-s+1}, & \text{if } e \geq s, \\ \gamma_{init}, & \text{if } e < s. \end{cases}$$

C. Fixed stepsize

$$\gamma = \gamma_{init}.$$

We say that the algorithm passed e epochs if the total number of computed gradient oracles lies between $e\sum_{m=1}^M n_m$ and $(e+1)\sum_{m=1}^M n_m$. For each algorithm the used stepsize γ_{init} and shift parameter s were tuned via selecting from the following sets:

$$\gamma_{init} \in \gamma_{set} \overset{\text{def}}{=} \{4.0, 3.75, 3.00, 2.5, 2.00, 1.25, 1.0, 0.75, 0.5, 0.25, \\ 0.2, 0.1, 0.06, 0.03, 0.01, 0.003, 0.001, 0.0006\}.$$

$$s \in s_{set} \overset{\text{def}}{=} \{50, 100, 200, 500, 1000\}.$$

In all tested methods, clients independently apply Rand-k compression with carnality $k = \lfloor 0.05d \rfloor$. Computation for all gradient oracles is carried out in single precision float (FP64) arithmetic.

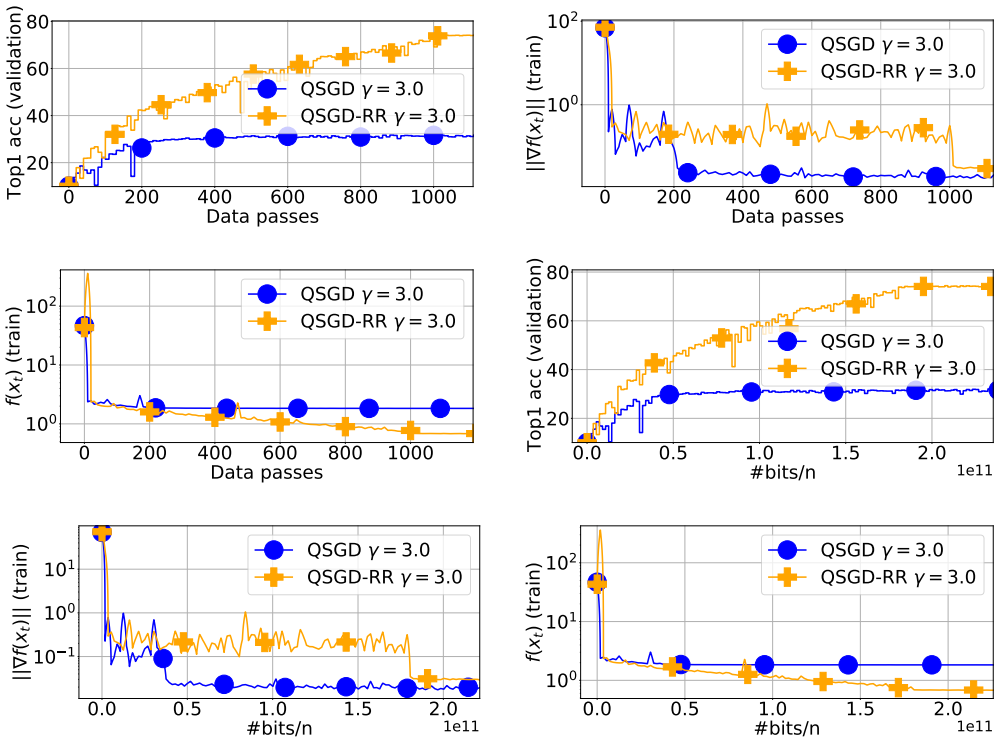

Figure 10

Figure 11: Comparison of QSGD and Q-RR in the training of `ResNet-18` on `CIFAR-10`, with $n = 10$ workers. Here (a) and (d) show Top-1 accuracy on test set, (b) and (e) – norm of full gradient on the train set, (c) and (f) – loss function value on the train set. Stepsizes and decay shift has been tuned from s_{set} and γ_{set} based on minimum achievable value of loss function on the train set.

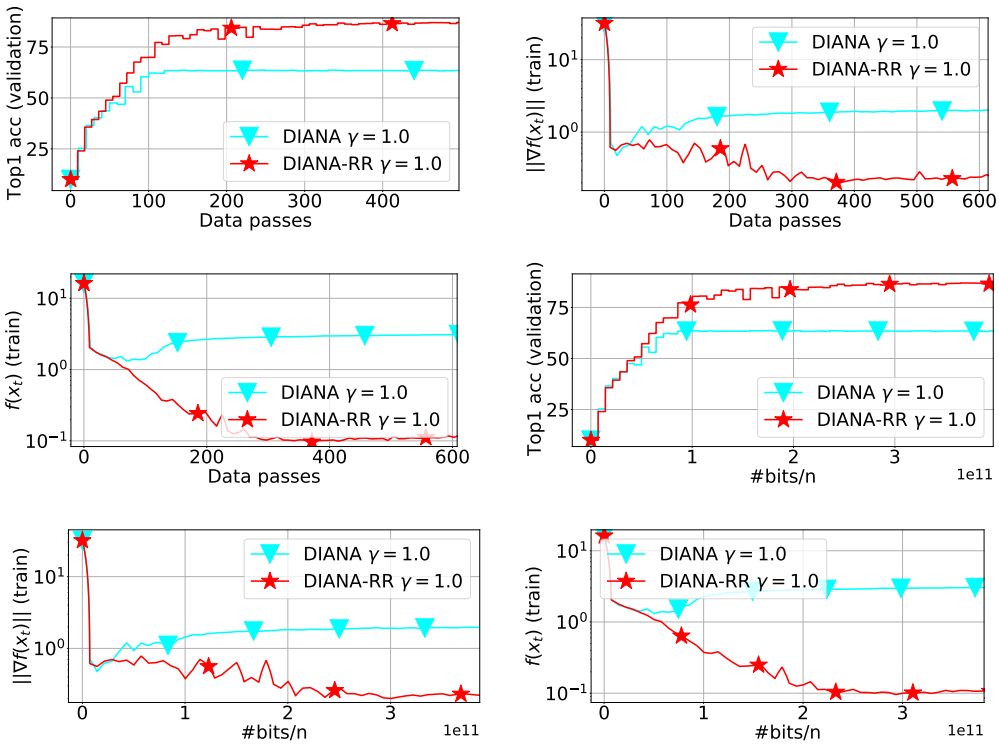

Figure 12: Comparison of DIANA and DIANA-RR in the training of `ResNet-18` on `CIFAR-10`, with $n = 10$ workers. Here (a) and (d) show Top-1 accuracy on test set, (b) and (e) – norm of full gradient on the train set, (c) and (f) – loss function value on the train set. Stepsizes and decay shift has been tuned from s_{set} and γ_{set} based on minimum achievable value of loss function on the train set. For both algorithms stepsize is fixed. For both algorithms stepsize is decaying according to srategy B.

B.2.5 Optimization-Based Fine-Tuning for Pretrained `ResNet-18`.

In this setting, we trained `ResNet-18` image classification in a distributed way across $n = 10$ clients. In this experiment, we have trained only the last linear layer.

Next, we have turned off batch normalization. Turning off batch normalization implies that the computation graph of NN $g(a, x)$ with weights of NN denoted as x is a deterministic function and does not include any internal state.

The loss function is a standard cross-entropy loss augmented with extra ℓ_2-regularization $\alpha\|x\|^2/2$ with $\alpha = 0.0001$. Initially used weights of NN are pretrained parameters after training the model on ImageNet.

The dataset distribution across clients has been set in a heterogeneous manner via presorting dataset D by label class and after this, it was split across 10 clients.

The comparison of stepsizes policies used in QSGD and Q-RR is presented in Figure 14. The behavior of the algorithms with best tuned step sizes is presented in Figure 13. These results demonstrate that in this setting there is no real benefit of using Q-RR in comparison to QSGD.

B.2.6 Experiments

The comparison of QSGD and Q-RR is presented in Figure 11. In particular, Figure 9 shows that in terms of the convergence to stationary points both algorithms exhibit similar behavior. However, Q-RR has better generalization and in fact, converges to the better loss function value. This experiment demonstrates that Q-RR with manually tuned stepsize can be better compared to QSGD in terms of the final quality of obtained Deep Learning model. For QSGD the tuned meta parameters are:

$\gamma_{init} = 3.0$,$s = 200$, strategy $= B$. For QSGD-RR tuned meta parameters are: $\gamma_{init} = 3.0$, $s = 1000$, strategy $= B$.

The results of comparison of DIANA and DIANA-RR are presented in Figure 12. For DIANA the tuned meta parameters are: $\gamma_{init} = 1.0$,$s = 0$, strategy $= C$ and for DIANA-RR tuned meta parameters are: $\gamma_{init} = 1.0$, $s = 0$, strategy $= C$. These results show that DIANA-RR outperforms DIANA in terms of all reported metrics.

B.3 Discussion

More about used arithmetics. We used FP64 (IEEE 754) due to its superior numerical stability compared to FP32, FP16, and BFloat16. While FP32 and FP16 are commonly used for inference tasks, the choice of precision for training depends on the specific requirements of the task. In certain cases, FP32 may be sufficient, but for others, FP64 is necessary to ensure stability.

The performance gain from switching from FP64 to FP32 can indeed vary based on the GPU model. For instance, the NVIDIA A100 40GB GPU used in our experiments offers approximately a two-fold increase in computational throughput with FP32 compared to FP64. The specific architecture of the GPU influences the choice of precision, and these characteristics can differ across various GPU models and updates.

The computational burden. The primary focus of the paper is to highlight the fundamental complexities and limits of algorithmic behavior. The experiments presented in our paper are intended for illustrative purposes.

The computational demands of our work are significant. Performing experiments beyond ResNet-18/CIFAR-10/FP64 with 10 clients is near the limit of what is feasible with our computational resources. In our simulation involving 10 clients sharing a common dataset, we ran 2000 rounds/epochs for fine-tuning. Based on an estimate of 2 minutes per epoch, the total computation time would be approximately 66 hours per run (2 minutes/epoch × 2000 epochs = 66 hours). Taking into account the grid search with 18 preset learning rates, 5 sets of decay parameters, and 4 algorithms, the total estimated computation time would be around 23760 hours (66 hours × 18 × 5 × 4). This represents a substantial amount of computation time. Therefore, conducting a comprehensive comparison involving four algorithms with an extensive grid of hyperparameters is already challenging for models larger than ResNet-18 on CIFAR10. To cover 23760 hours of training would indeed require approximately 40 GPUs running continuously for about 25 days. Nonetheless, we have conducted numerous experiments to ensure a thorough and fair comparison.

Training in overparameterized regime. During training image classification Convolution Neural Networks, we got two results for QSGD-RR as an improvement of QSGD. During training only the last layer (see Fig. 13) there are no benefits QSGD-RR, but QSGD does not behave worse.

When training the whole network (Fig. 11), the results suggest that Q-RR is much better than Q-SGD. Although we do not have formal proof explaining this phenomenon, we conjecture that this can be related to the significant overparameterization occurring during the training of a large model on a relatively small dataset. That is, the model can almost perfectly fit the training data on all clients, leading to a decrease in the heterogeneity parameter. In this case, there is no need for shifts since the variance coming from compression naturally goes to zero, and the complexities of QSGD and DIANA match (see Table 1). In this situation, Q-RR performs better than QSGD since the compression does not spoil the convergence of RR. Therefore, DIANA-type shifts are not always necessary to get improvements. Nevertheless, we conjecture that they are necessary when the datasets are larger and more complex.

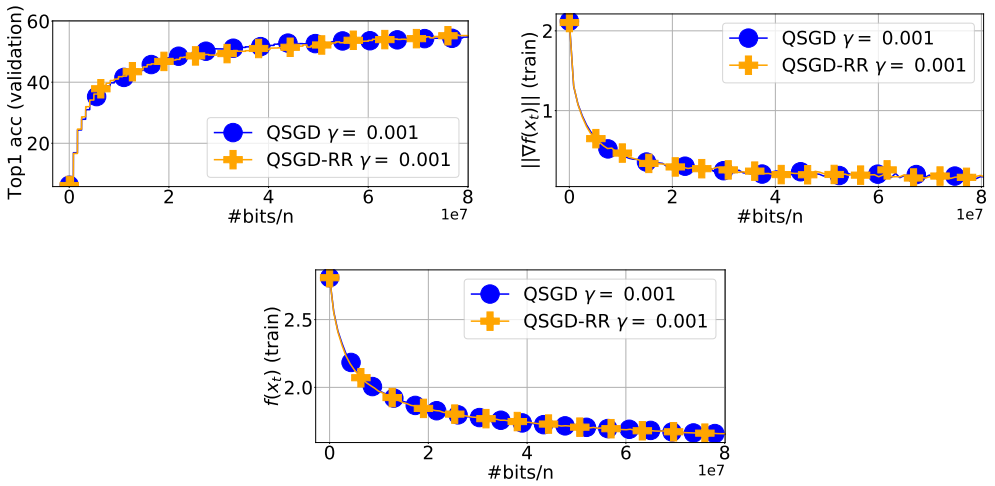

Figure 13: Comparison of QSGD and Q-RR in the training of the last linear layer of `ResNet-18` on `CIFAR-10`, with $n = 10$ workers. Here (a) shows Top-1 accuracy on test set, (b) – norm of full gradient on the train set, (c) – loss function value on the train set. Stepsizes and decay shift has been tuned from s_{set} and γ_{set} based on minimum achievable value of loss function on the train set. Both algorithms used fixed stepsize during training.

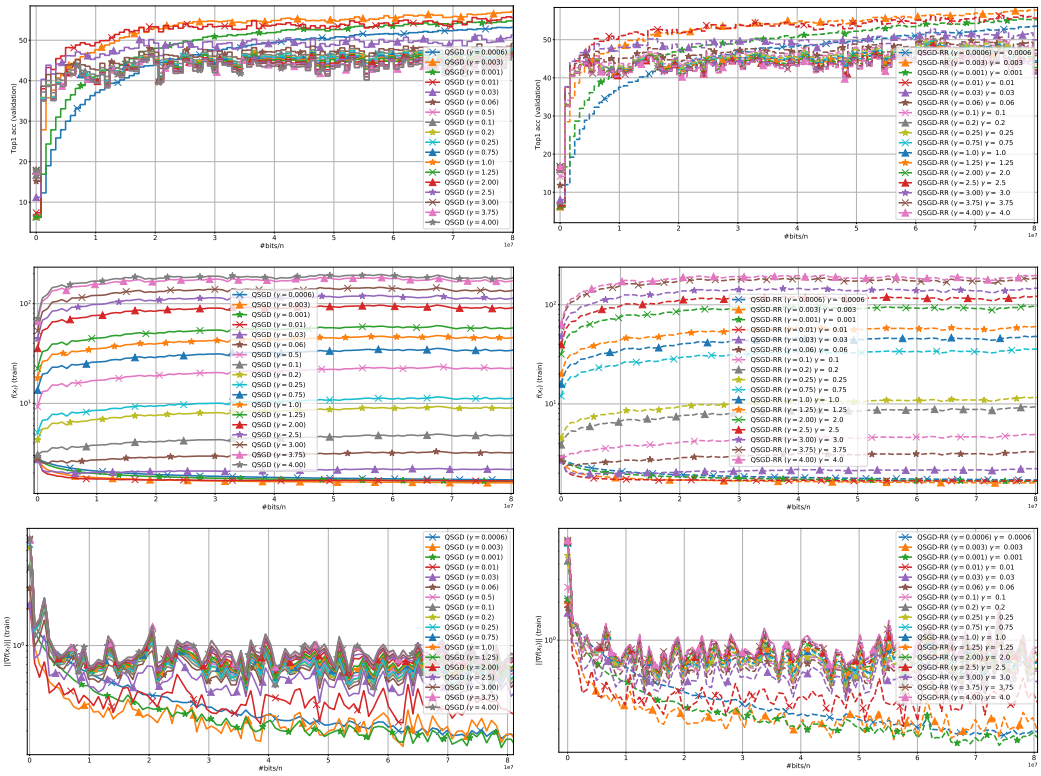

Figure 14: Comparison of QSGD and Q-RR in the training of the last linear layer of `ResNet-18` on `CIFAR-10`, with $n = 10$ workers. Here (a) and (b) show Top-1 accuracy on test set, (c) and (d) – loss function value on the train set, (e) and (f) – norm of full gradient on the train set. Stepsizes and decay shift has been tuned from s_{set} and γ_{set} based on minimum achievable value of loss function on the train set. During training stepsize was fixed. Batch Normalization was turned off.

C Missing Proofs for Q-RR

In the main part of the paper, we introduce Assumptions 3 and 4 for the analysis of Q-RR and DIANA-RR. These assumptions can be refined as follows.

Assumption 5. *Function* $f^{\pi^i} = \frac{1}{M}\sum_{i=1}^M f_m^{\pi_m^i} : \mathbb{R}^d \to \mathbb{R}$ *is* $\widetilde{L}$*-smooth for all sets of permutations* $\pi = (\pi_1, \ldots, \pi_m)$ *from* $[n]$ *and all* $i \in [n]$*, i.e.,*

$$\max_{i\in[n],\pi} \|\nabla f^{\pi^i}(x) - \nabla f^{\pi^i}(y)\| \leq \widetilde{L}\|x-y\| \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 6. *Function* $f^{\pi^i} = \frac{1}{M}\sum_{i=1}^M f_m^{\pi_m^i} : \mathbb{R}^d \to \mathbb{R}$ *is* $\widetilde{\mu}$*-strongly convex for all sets of permutations* $\pi = (\pi_1, \ldots, \pi_m)$ *from* $[n]$ *and all* $i \in [n]$*, i.e.,*

$$\min_{i\in[n],\pi} \left\{ f^{\pi^i}(x) - f^{\pi^i}(y) - \langle \nabla f^{i,\pi}(y), x-y\rangle \right\} \geq \frac{\widetilde{\mu}}{2}\|x-y\|^2 \quad \forall x, y \in \mathbb{R}^d.$$

Moreover, functions $f_1^i, f_2^i, \ldots, f_M^i : \mathbb{R}^d \to \mathbb{R}$ *are convex for all* $i = 1, \ldots, n$*.*

We notice that Assumptions 3 and 4 imply Assumptions 5 and 6. In the proofs of the results for Q-RR and DIANA-RR, we use Assumptions 5 in addition to Assumptions 3 and we use Assumption 6 instead of Assumption 4.

C.1 Shuffle Radius Clarification

Our results depend on the so-called shuffling radius proposed by Mishchenko et al. [2021]:

$$\sigma_{\text{rad}}^2 \overset{\text{def}}{=} \max_i \left\{ \frac{1}{\gamma^2 M} \sum_{m=1}^M \mathbb{E} D_{f_m^{\pi^i}}(x_\star^i, x_\star) \right\},$$

where $x_\star^{i+1} = x_\star^i - \frac{\gamma}{M}\sum_{m=1}^M \nabla f_m^{\pi_m^i}(x_\star)$.

One can think of the shuffling radius as a counterpart to the variance term in SGD. Both concepts measure how much the algorithm's performance can fluctuate near the optimal solution, but the cause of these fluctuations is different: in SGD, it is due to random sampling, and in RR, it is due to reshuffling. Additionally, Lemma 2.1 provides bounds for the shuffling radius — showing the maximum and minimum possible values — based on the variance at the optimum, reinforcing the shuffling radius as a useful way to understand how RR behaves. This relationship helps clarify how the reshuffling process influences the algorithm's path and its efficiency in reaching an optimal point.

C.2 Proof of Theorem 2.1

For convenience, we restate the theorem below.

Theorem C.1 (Theorem 2.1)**.** *Let Assumptions 1, 3, 5, 6 hold and* $0 < \gamma \leq \frac{1}{\widetilde{L}+2\frac{\omega}{M}L_{\max}}$*. Then, for all* $T \geq 0$ *the iterates produced by Q-RR satisfy*

$$\mathbb{E}\|x_T - x_\star\|^2 \leq (1-\gamma\widetilde{\mu})^{nT}\|x_0 - x_\star\|^2 + \frac{2\gamma^2\sigma_{rad}^2}{\widetilde{\mu}} + \frac{2\gamma\omega}{\widetilde{\mu}M}\left(\zeta_\star^2 + \sigma_\star^2\right),$$

where $\zeta_\star^2 = \frac{1}{M}\sum\limits_{m=1}^M \|\nabla f_m(x_\star)\|^2$*, and* $\sigma_\star^2 = \frac{1}{Mn}\sum\limits_{m=1}^M\sum\limits_{i=1}^n \|\nabla f_m^i(x^\star) - \nabla f_m(x^\star)\|^2$.

Proof. Using $x_\star^{i+1} = x_\star^i - \frac{\gamma}{M}\sum_{m=1}^M \nabla f_m^{\pi_m^i}(x_\star)$ and line 7 of Algorithm 1, we get

$$\begin{aligned}
\|x_t^{i+1} - x_\star^{i+1}\|^2 &= \left\| x_t^i - x_\star^i - \gamma\frac{1}{M}\sum_{m=1}^M \left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right) - \nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2 \\
&= \left\|x_t^i - x_\star^i\right\|^2 - 2\gamma\left\langle \frac{1}{M}\sum_{m=1}^M \left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right) - \nabla f_m^{\pi_m^i}(x_\star)\right), x_t^i - x_\star^i\right\rangle \\
&\quad + \gamma^2\left\|\frac{1}{M}\sum_{m=1}^M \left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right) - \nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2.
\end{aligned}$$

Taking the expectation w.r.t. $\mathcal{Q}$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1}-x_\star^{i+1}\|^2\right] &= \left\|x_t^i-x_\star^i\right\|^2-2\gamma\left\langle\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\right),x_t^i-x_\star^i\right\rangle\\ &\quad+\gamma^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{M}\sum_{m=1}^M\left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)-\nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\right].\end{aligned}$$

In view of Assumption 1 and $\mathbb{E}_\xi\|\xi-c\|^2=\mathbb{E}_\xi\|\xi-\mathbb{E}_\xi\xi\|^2+\|\mathbb{E}_\xi\xi-c\|^2$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1}-x_\star^{i+1}\|^2\right] &= \left\|x_t^i-x_\star^i\right\|^2-\frac{2\gamma}{M}\sum_{m=1}^M\left\langle\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star),x_t^i-x_\star^i\right\rangle\\ &\quad+\gamma^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{M}\sum_{m=1}^M\left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)-\nabla f_m^{\pi_m^i}(x_t^i)\right)\right\|^2\right]\\ &\quad+\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\\ &\le \left\|x_t^i-x_\star^i\right\|^2-\frac{2\gamma}{M}\sum_{m=1}^M\left\langle\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star),x_t^i-x_\star^i\right\rangle\\ &\quad+\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\\ &\quad+\frac{\gamma^2\omega}{M^2}\sum_{m=1}^M\left\|\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2,\end{aligned}$$

where in the last step we apply independence of $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)$ for $m\in[M]$. Next, we use three-point identity[4] and obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1}-x_\star^{i+1}\|^2\right] &\le \left\|x_t^i-x_\star^i\right\|^2\\ &\quad-\frac{2\gamma}{M}\sum_{m=1}^M\left(D_{f_m^{\pi_m^i}}(x_\star^i,x_t^i)+D_{f_m^{\pi_m^i}}(x_t^i,x_\star)-D_{f_m^{\pi_m^i}}(x_\star^i,x_\star)\right)\\ &\quad+\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\\ &\quad+\frac{\gamma^2\omega}{M^2}\sum_{m=1}^M\left\|\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2.\end{aligned}$$

[4]For any differentiable function $f:\mathbb{R}^d\to\mathbb{R}^d$ we have: $\langle\nabla f(x)-\nabla f(y),x-z\rangle=D_f(z,x)+D_f(x,y)-D_f(z,y)$.

Applying $\widetilde{L}$-smoothness and convexity of $\frac{1}{M}\sum_{m=1}^{m} f_m^{\pi_m^i}$, $\widetilde{\mu}$-strong convexity of $\frac{1}{M}\sum_{m=1}^{m} f_m^{\pi_m^i}$, and $L_{\max}$-smoothness and convexity of f_m^i, we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1}-x_\star^{i+1}\|^2\right] &\leq (1-\gamma\widetilde{\mu})\left\|x_t^i-x_\star^i\right\|^2 - 2\gamma\left(1-\widetilde{L}\gamma\right)\frac{1}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t^i,x_\star) \\
&\quad +2\gamma\frac{1}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_\star^i,x_\star) + \frac{\gamma^2\omega}{M^2}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2 \\
&\leq (1-\gamma\widetilde{\mu})\left\|x_t^i-x_\star^i\right\|^2 - 2\gamma\left(1-\widetilde{L}\gamma\right)\frac{1}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t^i,x_\star) \\
&\quad +2\gamma\frac{1}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_\star^i,x_\star) + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_\star)\right\|^2 \\
&\quad +\frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\right\|^2.
\end{aligned}$$

So, we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1}-x_\star^{i+1}\|^2\right] &\leq (1-\gamma\widetilde{\mu})\left\|x_t^i-x_\star^i\right\|^2 + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_\star)\right\|^2 \\
&\quad +\frac{2\gamma}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_\star^i,x_\star) \\
&\quad -2\gamma\left(1-\gamma\left(\widetilde{L}+\frac{2\omega L_{\max}}{M}\right)\right)\frac{1}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t^i,x_\star).
\end{aligned}$$

Taking the full expectation and using a definition of shuffle radius, $0<\gamma\leq\frac{1}{\left(\widetilde{L}+2\frac{\omega}{M}L_{\max}\right)}$, and $D_{f_m^{\pi_m^i}}(x_t^i,x_\star)\geq 0$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|x_t^{i+1}-x_\star^{i+1}\|^2\right] &\leq (1-\gamma\widetilde{\mu})\,\mathbb{E}\left[\left\|x_t^i-x_\star^i\right\|^2\right] + 2\gamma^3\sigma_{\text{rad}}^2 + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right] \\
&= (1-\gamma\widetilde{\mu})\,\mathbb{E}\left[\left\|x_t^i-x_\star^i\right\|^2\right] + 2\gamma^3\sigma_{\text{rad}}^2 + \frac{2\gamma^2\omega}{M^2 n}\sum_{m=1}^{M}\sum_{j=1}^{n}\left\|\nabla f_m^j(x_\star)\right\|^2 \\
&\leq (1-\gamma\widetilde{\mu})\,\mathbb{E}\left[\left\|x_t^i-x_\star^i\right\|^2\right] + 2\gamma^3\sigma_{\text{rad}}^2 + \frac{2\gamma^2\omega}{M}\left(\zeta_\star^2+\sigma_\star^2\right).
\end{aligned}$$

Unrolling the recurrence in i, we derive

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1}-x_\star\|^2\right] &\leq (1-\gamma\widetilde{\mu})^n\,\mathbb{E}\left[\|x_t-x_\star\|^2\right] + 2\gamma^3\sigma_{\text{rad}}^2\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \\
&\quad +\frac{2\gamma^2\omega}{M}\left(\zeta_\star^2+\sigma_\star^2\right)\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j.
\end{aligned}$$

Unrolling the recurrence in t, we derive

$$\begin{aligned}
\mathbb{E}\left[\|x_T-x_\star\|^2\right] &\leq (1-\gamma\widetilde{\mu})^{nT}\|x_0-x_\star\|^2 + 2\gamma^3\sigma_{\text{rad}}^2\sum_{t=0}^{T-1}(1-\gamma\widetilde{\mu})^{nt}\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \\
&\quad +\frac{2\gamma^2\omega}{M}\left(\zeta_\star^2+\sigma_\star^2\right)\sum_{j=0}^{nT-1}(1-\gamma\widetilde{\mu})^{nt}\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j.
\end{aligned}$$

Since $\sum_{j=0}^{nT-1}(1-\gamma\widetilde{\mu})^j\leq\frac{1}{\gamma\widetilde{\mu}}$, we get the result. □

Corollary 5. *Let the assumptions of Theorem C.1 hold and*

$$\gamma = \min\left\{\frac{1}{\widetilde{L} + 2\frac{\omega}{M}L_{\max}}, \sqrt{\frac{\varepsilon\widetilde{\mu}}{6\sigma_{rad}^2}}, \frac{\varepsilon\widetilde{\mu}M}{6\omega\left(\zeta_\star^2 + \sigma_\star^2\right)}\right\}. \tag{12}$$

Then, Q-RR finds a solution with accuracy $\varepsilon > 0$ after the following number of communication rounds:

$$\widetilde{\mathcal{O}}\left(\frac{\widetilde{L}}{\widetilde{\mu}} + \frac{\omega}{M}\frac{L_{\max}}{\widetilde{\mu}} + \frac{\omega}{M}\frac{\zeta_\star^2 + \sigma_\star^2}{\varepsilon\widetilde{\mu}^2} + \frac{\sigma_{rad}}{\sqrt{\varepsilon\widetilde{\mu}^3}}\right).$$

Proof. Theorem C.1 implies

$$\mathbb{E}\|x_T - x_\star\|^2 \leq (1-\gamma\widetilde{\mu})^{nT}\|x_0 - x_\star\|^2 + \frac{2\gamma^2\sigma_{\text{rad}}^2}{\widetilde{\mu}} + \frac{2\gamma\omega}{\widetilde{\mu}M}\left(\zeta_\star^2 + \sigma_\star^2\right). \tag{13}$$

To estimate the number of communication rounds required to find a solution with accuracy $\varepsilon > 0$, we need to upper-bound each term from the right-hand side by $\varepsilon/3$. Thus, we get additional conditions on γ:

$$\frac{2\gamma^2\sigma_{\text{rad}}^2}{\widetilde{\mu}} < \frac{\varepsilon}{3}, \quad \frac{2\gamma\omega}{\widetilde{\mu}M}\left(\zeta_\star^2 + \sigma_\star^2\right) < \frac{\varepsilon}{3}$$

and also the upper bound on the number of communication rounds nT

$$nT = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma\widetilde{\mu}}\right).$$

Substituting (12), we get a final result. □

C.3 Non-Strongly Convex Summands

In this section, we provide the analysis of Q-RR without using Assumptions 4, 6. Before we move one to the proofs, we would like to emphasize that

$$x_t^{i+1} = x_t^i - \gamma\frac{1}{M}\sum_{m=1}^{M}\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right).$$

Then we have

$$x_{t+1} = x_t - \gamma\sum_{i=0}^{n-1}\frac{1}{M}\sum_{m=1}^{M}\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right) = x_t - \tau\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right),$$

where $\tau = \gamma n$. For convenience, we denote

$$g_t = \frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)$$

allowing to write the update rule as $x_{t+1} = x_t - \tau g_t$.

Lemma C.1 (Lemma 1 from [Malinovsky et al., 2022]). *For any $k \in [n]$, let $\xi_{\pi_1}, \ldots, \xi_{\pi_k}$ be sampled uniformly without replacement from a set of vectors $\{\xi_1, \ldots, \xi_n\}$ and $\bar{\xi}_\pi$ be their average. Then, it holds*

$$\mathbb{E}\bar{\xi}_\pi = \bar{\xi}, \quad \mathbb{E}\left[\|\bar{\xi}_\pi - \bar{\xi}\|^2\right] = \frac{n-k}{k(n-1)}\sigma^2, \tag{14}$$

where $\bar{\xi} = \frac{1}{n}\sum_{i=1}^n \xi_i$, $\bar{\xi}_\pi = \frac{1}{k}\sum_{i=1}^k \xi_{\pi_i}$, $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \|\xi_i - \bar{\xi}\|^2$

Lemma C.2. *Under Assumptions 1, 2, 3, 5, the following inequality holds*

$$\mathbb{E}_{\mathcal{Q}}\left[-2\tau\langle g_t, x_t - x_\star\rangle\right] \leq -\frac{\tau\mu}{2}\|x_t - x_\star\|^2 - \tau(f(x_t) - f(x_\star)) + \frac{\tau\widetilde{L}}{n}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2.$$

Proof. Using that $\mathbb{E}_{\mathcal{Q}}\left[g_t\right] = \frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t^i)$ and definition of $h^\star$, we get

$$\begin{aligned}-2\tau\mathbb{E}_{\mathcal{Q}}\left[\langle g_t, x_t - x_\star\rangle\right] &= -\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\langle\nabla f_m^{\pi_m^i}(x_t^i), x_t - x_\star\right\rangle \\ &= -\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\langle\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_\star), x_t - x_\star\right\rangle.\end{aligned}$$

Using three-point identity, we obtain

$$\begin{aligned}-2\tau\mathbb{E}_{\mathcal{Q}}\left[\langle g_t, x_t - x_\star\rangle\right] &= -\frac{2\tau}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left(D_{f_m^{\pi_m^i}}(x_t, x_\star) + D_{f_m^{\pi_m^i}}(x_\star, x_t^i) - D_{f_m^{\pi_m^i}}(x_t, x_t^i)\right) \\ &= -2\tau D_f(x_t, x_\star) - \frac{2\tau}{n}\sum_{i=0}^{n-1}D_{f^{\pi^i}}(x_\star, x_t^i) + \frac{2\tau}{n}\sum_{i=0}^{n-1}D_{f^{\pi^i}}(x_t, x_t^i) \\ &\leq -2\tau D_f(x_t, x_\star) + \frac{\tau\widetilde{L}}{n}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2,\end{aligned}$$

where in the last inequality we apply $\widetilde{L}$-smoothness and convexity of each function f^{π^i}. Finally, using μ-strong convexity of f, we finish the proof of the lemma.

□

Lemma C.3. *Under Assumptions 1, 2, 3, 5, the following inequality holds*

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &\leq 2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] + \frac{4\omega}{Mn}\left(\zeta_\star^2 + \sigma_\star^2\right) \\ &\quad +8\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)(f(x_t) - f(x_\star)).\end{aligned}$$

Proof. Taking the expectation w.r.t. $\mathcal{Q}$ and using variance decomposition $\mathbb{E}\left[\|\xi\|^2\right] = \mathbb{E}\left[\|\xi - \mathbb{E}\left[\xi\right]\|^2\right] + \|\mathbb{E}\xi\|^2$, we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)\right\|^2\right] \\ &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right) - \nabla f_m^{\pi_m^i}(x_t^i)\right)\right\|^2\right] \\ &\quad + \left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2.\end{aligned}$$

Next, Assumption 1 and conditional independence of $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right)$ for $m = 1, \ldots, M, i = 0, \ldots, n-1$ imply

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &= \frac{1}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\left\|\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i)\right) - \nabla f_m^{\pi_m^i}(x_t^i)\right\|^2\right] \\
&\quad + \left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2 \\
&\leq \frac{\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2 + \left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t^i)\right\|^2 \\
&\leq \frac{2\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_t)\right\|^2 + \frac{2\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t)\right\|^2 \\
&\quad + 2\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left(\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_t)\right)\right\|^2 \\
&\quad + 2\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t)\right\|^2.
\end{aligned}$$

Using $L_{\max}$-smoothness and convexity of f_m^i and $\widetilde{L}$-smoothness and convexity of $f^{\pi^i} = \frac{1}{M}\sum_{m=1}^{M} f_m^{\pi_m^i}$, we derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &\leq \frac{4\omega}{M^2n^2}L_{\max}\sum_{i=0}^{n-1}\sum_{m=1}^{M}D_{f_m^{\pi_m^i}}(x_t^i, x_t) + \frac{2\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t)\right\|^2 \\
&\quad + 4\widetilde{L}\frac{1}{n}\sum_{i=0}^{n-1}D_{f^{\pi^i}}(x_t^i, x_t) + 2\left\|\nabla f(x_t)\right\|^2 \\
&\leq 4\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}D_{f^{\pi^i}}(x_t^i, x_t) + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_\star)\right\|^2 \\
&\quad + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t) - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2 + 2\left\|\nabla f(x_t) - \nabla f(x_\star)\right\|^2 \\
&\leq 2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\left\|x_t^i - x_t\right\|^2 + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_\star)\right\|^2 \\
&\quad + \frac{8\omega}{M^2n^2}L_{\max}\sum_{i=0}^{n-1}\sum_{m=1}^{M}D_{f_m^{\pi_m^i}}(x_t, x_\star) + 4\widetilde{L}\left(f(x_t) - f(x_\star)\right).
\end{aligned}$$

Taking the full expectation, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|g_t\|^2\right] &\leq 2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\left\|x_t^i - x_t\right\|^2\right] + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right] \\
&\quad + \left(4\widetilde{L} + \frac{8\omega}{Mn}L_{\max}\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&= 2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\left\|x_t^i - x_t\right\|^2\right] + \frac{4\omega}{Mn}\left(\zeta_\star^2 + \sigma_\star^2\right) \\
&\quad + \left(4\widetilde{L} + \frac{8\omega}{Mn}L_{\max}\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right].
\end{aligned}$$

□

Lemma C.4. *Let Assumptions 1, 2, 3, 5 hold and* $\tau \leq \frac{1}{2\sqrt{\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)}}$. *Then, the following inequality holds*

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 24\tau^2\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\mathbb{E}\left[f(x_t)-f(x_\star)\right] \\ &\quad +8\tau^2\frac{\omega}{Mn}\left(\zeta_\star^2+\sigma_\star^2\right)+8\tau^2\frac{\sigma_{\star,n}^2}{n},\end{aligned}$$

where $\sigma_{\star,n}^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f^i(x_\star)\|^2$, $f^i(x) = \frac{1}{M}\sum_{m=1}^M f_m^i(x)$, $i \in [n]$.

Proof. Since $x_t^i = x_t - \frac{\tau}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\mathcal{Q}\left(\nabla f_m^{\pi_m^j}(x_t^j)\right)$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^i - x_t\|^2\right] &= \tau^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\mathcal{Q}\left(\nabla f_m^{\pi_m^j}(x_t^j)\right)\right\|^2\right] \\ &= \tau^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\left(\mathcal{Q}\left(\nabla f_m^{\pi_m^j}(x_t^j)\right)-\nabla f_m^{\pi_m^j}(x_t^j)\right)\right\|^2\right] \\ &\quad +\tau^2\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2 \\ &\leq \frac{\tau^2}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{i-1}\mathbb{E}_{\mathcal{Q}}\left[\left\|\mathcal{Q}\left(\nabla f_m^{\pi_m^j}(x_t^j)\right)-\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2\right] \\ &\quad +\tau^2\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2.\end{aligned}$$

Using Assumption 1, $\widetilde{L}$-smoothness and convexity of $f^{\pi^i} = \frac{1}{M}\sum_{m=1}^M f_m^{\pi_m^i}$ and $L_{\max}$-smoothness and convexity of f_m^i, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|x_t^i - x_t\|^2\right] &\leq \frac{\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{i-1}\left\|\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2 + \tau^2\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2 \\
&\leq \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{i-1}\left\|\nabla f_m^{\pi_m^j}(x_t^j) - \nabla f_m^{\pi_m^j}(x_t)\right\|^2 + 2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2 \\
&\quad +2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\left(\nabla f^{\pi^j}(x_t^j) - \nabla f^{\pi^j}(x_t)\right)\right\|^2 \qquad (15) \\
&\quad +\frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{i-1}\left\|\nabla f_m^{\pi_m^j}(x_t)\right\|^2 \\
&\leq \frac{4\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}L_{\max}D_{f_m^{\pi_m^j}}(x_t^j, x_t) + 2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2 \\
&\quad +4\widetilde{L}\tau^2\frac{1}{n}\sum_{j=0}^{n-1}D_{f^{\pi^j}}(x_t^j, x_t) + \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}\left\|\nabla f_m^{\pi_m^j}(x_t)\right\|^2 \\
&= 4\tau^2\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{j=0}^{n-1}D_{f^{\pi^j}}(x_t^j, x_t) \\
&\quad +2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2 + \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}\left\|\nabla f_m^{\pi_m^j}(x_t)\right\|^2. \qquad (16)
\end{aligned}$$

Next, we need to estimate the second term from the previous inequality. Taking the full expectation and using Lemma C.1 and using new notation $\sigma_t^2 = \frac{1}{n}\sum_{j=1}^n \mathbb{E}[\|\nabla f^j(x_t) - \nabla f(x_t)\|^2]$, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2\right] &= \frac{i^2}{n^2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{i^2}{n^2}\mathbb{E}\left[\left\|\frac{1}{i}\sum_{j=0}^{i-1}\left(\nabla f^{\pi^j}(x_t) - \nabla f(x_t)\right)\right\|^2\right] \\
&\leq \frac{i^2}{n^2}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{i^2}{n^3}\frac{n-i}{i(n-1)}\sum_{j=1}^n\mathbb{E}\left[\|\nabla f^j(x_t) - \nabla f(x_t)\|^2\right] \\
&\leq \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{1}{n}\sigma_t^2. \qquad (17)
\end{aligned}$$

Taking the full expectation from (16) and using (17), we obtain

$$\begin{aligned}
\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 4\tau^2\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\sum_{j=0}^{n-1}\mathbb{E}\left[D_{f^{\pi^j}}(x_t^j, x_t)\right] \\
&\quad +2\tau^2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{2\tau^2}{n}\sigma_t^2 + \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_t)\right\|^2\right].
\end{aligned}$$

Using $\widetilde{L}$-smoothness of f^{π^j}, we get

$$\begin{aligned}
\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 2\widetilde{L}\tau^2\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\sum_{j=0}^{n-1}\mathbb{E}\left[\|x_t^j - x_t\|^2\right] \\
&\quad +2\tau^2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{2\tau^2}{n}\sigma_t^2 + \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_t)\right\|^2\right].
\end{aligned}$$

Since $\tau \leq \frac{1}{2\sqrt{\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)}}$, we have

$$\begin{aligned}
\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 2\left(1 - 2\widetilde{L}\tau^2\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\right)\sum_{j=0}^{n-1}\mathbb{E}\left[\|x_t^j - x_t\|^2\right] \\
&\leq 4\tau^2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{4\tau^2}{n}\sigma_t^2 + \frac{4\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_t)\right\|^2\right] \\
&\leq \frac{8\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_t) - \nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] \\
&\quad + \frac{8\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] + 4\tau^2\mathbb{E}\left[\|\nabla f(x_t) - \nabla f(x_\star)\|^2\right] \\
&\quad + \frac{4\tau^2}{n}\left(\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f^j(x_t)\|\right] - \mathbb{E}\left[\|\nabla f(x_t)\|^2\right]\right) \\
&\leq \frac{8\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_t) - \nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] \\
&\quad + \frac{8\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] + 8\tau^2\mathbb{E}\left[\|\nabla f(x_t) - \nabla f(x_\star)\|^2\right] \\
&\quad + \frac{8\tau^2}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f^j(x_t) - \nabla f^j(x_\star)\|^2\right] + \frac{8\tau^2}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f^j(x_\star)\|^2\right].
\end{aligned}$$

Summing from $i = 0$ to $n-1$ and using $\widetilde{L}$-smoothness of f^i and $L_{\max}$-smoothness of f_m^i, we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq \frac{16\tau^2\omega}{Mn}L_{\max}\mathbb{E}\left[f(x_t) - f(x_\star)\right] + \frac{16\tau^2}{n}\widetilde{L}\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad + \frac{8\tau^2\omega}{Mn}\left(\zeta_\star^2 + \sigma_\star^2\right) + \frac{8\tau^2}{n}\sigma_{\star,n}^2 + 8\tau^2\widetilde{L}\mathbb{E}\left[f(x_t) - f(x_\star)\right].
\end{aligned}$$

□

Theorem C.2. *Let Assumptions 1, 2, 3, 5 hold and stepsize γ satisfy*

$$0 < \gamma \leq \frac{1}{16n\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)}. \tag{18}$$

Then, for all $T \geq 0$ the iterates produced by Q-RR *satisfy*

$$\begin{aligned}
\mathbb{E}\left[\|x_T - x_\star\|^2\right] &\leq \left(1 - \frac{n\gamma\mu}{2}\right)^T\|x_0 - x_\star\|^2 + 18\frac{\gamma^2 n\widetilde{L}}{\mu}\left(\frac{\omega}{M}(\zeta_\star^2 + \sigma_\star^2) + \sigma_{\star,n}^2\right) \\
&\quad + 8\frac{\gamma\omega}{\mu M}(\zeta_\star^2 + \sigma_\star^2),
\end{aligned}$$

where

$$\sigma_{\star,n}^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f^i(x_\star)\|^2. \tag{19}$$

Proof. Taking expectation w.r.t. $\mathcal{Q}$ and using Lemma C.3, we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &= \|x_t-x_\star\|^2 - 2\tau\mathbb{E}_{\mathcal{Q}}\left[\langle g_t, x_t-x_\star\rangle\right] + \tau^2\mathbb{E}_{\mathcal{Q}}\left[\|g^t\|^2\right]\\ &\le \|x_t-x_\star\|^2 - 2\tau\mathbb{E}_{\mathcal{Q}}\left[\left\langle g^t, x_t-x_\star\right\rangle\right]\\ &\quad +2\tau^2\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i-x_t\|^2\right]\\ &\quad +8\tau^2\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)(f(x_t)-f(x_\star))+\frac{4\tau^2\omega}{Mn}(\zeta_\star^2+\sigma_\star^2).\end{aligned}$$

Using Lemma C.2, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &\le \|x_t-x_\star\|^2\\ &\quad -\frac{\tau\mu}{2}\|x_t-x_\star\|^2-\tau(f(x_t)-f(x_\star))+\frac{\tau\widetilde{L}}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2\\ &\quad +2\tau^2\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i-x_t\|^2\right]\\ &\quad +8\tau^2\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)(f(x_t)-f(x_\star))+\frac{4\tau^2\omega}{Mn}(\zeta_\star^2+\sigma_\star^2)\\ &\le \left(1-\frac{\tau\mu}{2}\right)\|x_t-x_\star\|^2\\ &\quad -\tau\left(1-8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)(f(x_t)-f(x_\star))\\ &\quad +\tau\widetilde{L}\left(1+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i-x_t\|^2\right]\\ &\quad +\frac{4\tau^2\omega}{Mn}(\zeta_\star^2+\sigma_\star^2).\end{aligned}$$

Next, we take the full expectation and apply Lemma C.4:

$$\begin{aligned}&\mathbb{E}\left[\|x_{t+1}-x_\star\|^2\right]\le\left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\|x_t-x_\star\|^2\right]\\ &-\tau\left(1-8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\mathbb{E}\left[f(x_t)-f(x_\star)\right]\\ &+24\tau^3\widetilde{L}\left(1+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)(f(x_t)-f(x_\star))\\ &+8\tau^3\widetilde{L}\left(1+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(\frac{\omega}{Mn}(\zeta_\star^2+\sigma_\star^2)+\frac{\sigma_{\star,n}^2}{n}\right)+\frac{4\tau^2\omega}{Mn}(\zeta_\star^2+\sigma_\star^2).\end{aligned}$$

Using (18), we derive

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1}-x_\star\|^2\right] &\le \left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\|x_t-x_\star\|^2\right]\\ &\quad +9\tau^3\widetilde{L}\left(\frac{\omega}{Mn}(\zeta_\star^2+\sigma_\star^2)+\frac{\sigma_{\star,n}^2}{n}\right)+\frac{4\tau^2\omega}{Mn}(\zeta_\star^2+\sigma_\star^2)\end{aligned}$$

Recursively unrolling the inequality, substituting $\tau=n\gamma$ and using $\sum\limits_{t=0}^{+\infty}\left(1-\frac{\tau\mu}{2}\right)^t\le\frac{2}{\mu\tau}$, we get the result. □

Corollary 6. *Let the assumptions of Theorem C.2 hold and*

$$\gamma=\min\left\{\frac{1}{16n\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)},\sqrt{\frac{\varepsilon\mu}{8^2n\widetilde{L}}}\left(\frac{\omega}{M}\Delta_\star^2+\sigma_{\star,n}^2\right)^{-\frac{1}{2}},\frac{\varepsilon\mu M}{24\omega\Delta_\star^2}\right\},\tag{20}$$

where $\Delta_\star^2 = \zeta_\star^2 + \sigma_\star^2$. *Then,* Q-RR *finds a solution with accuracy* $\varepsilon > 0$ *after the following number of communication rounds:*

$$\widetilde{\mathcal{O}}\left(\frac{n\widetilde{L}}{\mu} + \frac{\omega}{M}\frac{L_{\max}}{\mu} + \frac{\omega}{M}\frac{\zeta_\star^2 + \sigma_\star^2}{\varepsilon\mu^2} + \sqrt{\frac{n\widetilde{L}}{\varepsilon\mu^3}\sqrt{\frac{\omega}{M}\left(\zeta_\star^2 + \sigma_\star^2\right) + \sigma_{\star,n}^2}}\right).$$

Proof. Theorem C.2 implies

$$\begin{aligned}\mathbb{E}\left[\|x_T - x_\star\|^2\right] \quad &\leq \quad \left(1 - \frac{n\gamma\mu}{2}\right)^T \|x_0 - x_\star\|^2 + 18\frac{\gamma^2 n\widetilde{L}}{\mu}\left(\frac{\omega}{M}\left(\zeta_\star^2 + \sigma_\star^2\right) + \sigma_{\star,n}^2\right) \\ &\quad + 8\frac{\gamma\omega}{\mu M}\left(\zeta_\star^2 + \sigma_\star^2\right).\end{aligned}$$

To estimate the number of communication rounds required to find a solution with accuracy $\varepsilon > 0$, we need to upper bound each term from the right-hand side by $\varepsilon/3$. Thus, we get additional conditions on γ:

$$18\frac{\gamma^2 n\widetilde{L}}{\mu}\left(\frac{\omega}{M}\left(\zeta_\star^2 + \sigma_\star^2\right) + \sigma_{\star,n}^2\right) < \frac{\varepsilon}{3}, \quad 8\frac{\gamma\omega}{\mu M}\left(\zeta_\star^2 + \sigma_\star^2\right) < \frac{\varepsilon}{3},$$

and also the upper bound on the number of communication rounds nT

$$nT = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma\mu}\right).$$

Substituting (20) in the previous equation, we get the result. □

D Missing Proofs for DIANA-RR

D.1 Proof of Theorem 2.2

Lemma D.1. *Let Assumptions 1, 3, 5, 6 hold and $\alpha \leq \frac{1}{1+\omega}$. Then, the iterates of DIANA-RR satisfy*

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\right] \leq \frac{1-\alpha}{M}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2 + \frac{2\alpha L_{\max}}{M}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t^i,x_\star).$$

Proof. Taking expectation w.r.t. $\mathcal{Q}$, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\right] &= \mathbb{E}_{\mathcal{Q}}\left[\|h_{t,m}^{\pi_m^i}+\alpha\mathcal{Q}(\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i})-\nabla f_m^{\pi_m^i}(x_\star)\|^2\right]\\
&= \|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\quad +2\alpha\mathbb{E}_{\mathcal{Q}}\left[\left\langle\mathcal{Q}(\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i}),h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\right\rangle\right]\\
&\quad +\alpha^2\mathbb{E}_{\mathcal{Q}}\left[\|\mathcal{Q}(\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i})\|^2\right]\\
&= \|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\quad +2\alpha\left\langle\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i},h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\right\rangle\\
&\quad +\alpha^2\mathbb{E}_{\mathcal{Q}}\left[\|\mathcal{Q}(\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i})\|^2\right].
\end{aligned}$$

Assumption 1, $L_{\max}$-smoothness and convexity of f_m^i and $\alpha \leq 1/(1+\omega)$ imply

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\right] &\leq \|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\quad +2\alpha\left\langle\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i},h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\right\rangle\\
&\quad +\alpha^2(1+\omega)\|\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i}\|^2\\
&\leq \|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\quad +\alpha\left\langle\nabla f_m^{\pi_m^i}(x_t^i)-h_{t,m}^{\pi_m^i},h_{t,m}^{\pi_m^i}+\nabla f_m^{\pi_m^i}(x_t^i)-2\nabla f_m^{\pi_m^i}(x_\star)\right\rangle\\
&\leq \|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\quad +\alpha\|\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\|^2-\alpha\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\leq (1-\alpha)\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\
&\quad +\alpha\|\nabla f_m^{\pi_m^i}(x_t^i)-\nabla f_m^{\pi_m^i}(x_\star)\|^2 \qquad (21)\\
&\leq (1-\alpha)\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2+2\alpha L_{\max}D_{f_m^{\pi_m^i}}(x_t^i,x_\star).
\end{aligned}$$

Summing up the above inequality for $m=1,\ldots,M$, we get the result. □

Theorem D.1. *Let Assumptions 1, 3, 5, 6 hold and $0<\gamma\leq\min\left\{\frac{\alpha}{2n\widetilde{\mu}},\frac{1}{\widetilde{L}+\frac{6\omega}{M}L_{\max}}\right\}$, $\alpha\leq\frac{1}{1+\omega}$. Then, for all $T\geq 0$ the iterates produced by DIANA-RR satisfy*

$$\mathbb{E}\left[\Psi_T\right]\leq(1-\gamma\widetilde{\mu})^{nT}\Psi_0+\frac{2\gamma^2\sigma_{rad}^2}{\widetilde{\mu}},$$

where Ψ_t is defined in (6).

Proof. Using $x_\star^{i+1} = x_\star^i - \frac{\gamma}{M}\sum_{m=1}^M \nabla f_m^{\pi_m^i}(x_\star)$ and line 9 of Algorithm 2, we derive

$$\begin{aligned}\|x_t^{i+1} - x_\star^{i+1}\|^2 &= \left\|x_t^i - x_\star^i - \gamma\frac{1}{M}\sum_{m=1}^M\left(\hat{g}_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\\ &= \left\|x_t^i - x_\star^i\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^M\left\langle\left(\hat{g}_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right), x_t^i - x_\star^i\right\rangle\\ &\quad +\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\hat{g}_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2.\end{aligned}$$

Taking expectation w.r.t. $\mathcal{Q}$ and using $\mathbb{E}\|\xi - c\|^2 = \mathbb{E}\|\xi - \mathbb{E}\xi\|^2 + \|\mathbb{E}\xi - c\|^2$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1} - x_\star^{i+1}\|^2\right] &= \left\|x_t^i - x_\star^i\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^M\left\langle\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_\star), x_t^i - x_\star^i\right\rangle\\ &\quad +\gamma^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{M}\sum_{m=1}^M\left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right) + h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right)\right\|^2\right]\\ &\leq \left\|x_t^i - x_\star^i\right\|^2 - \frac{2\gamma}{M}\sum_{m=1}^M\left\langle\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_\star), x_t^i - x_\star^i\right\rangle\\ &\quad +\gamma^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{M}\sum_{m=1}^M\left(\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right) - \nabla f_m^{\pi_m^i}(x_t^i) + h_{t,m}^{\pi_m^i}\right)\right\|^2\right]\\ &\quad +\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_\star) - \nabla f_m^{\pi_m^i}(x_t^i)\right)\right\|^2.\end{aligned}$$

Independence of $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right)$, $m \in [M]$, assumption 1, and three-point identity imply

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1} - x_\star^{i+1}\|^2\right] &\leq \left\|x_t^i - x_\star^i\right\|^2\\ &\quad -\frac{2\gamma}{M}\sum_{m=1}^M\left(D_{f_m^{\pi_m^i}}(x_\star^i, x_t^i) + D_{f_m^{\pi_m^i}}(x_t^i, x_\star) - D_{f_m^{\pi_m^i}}(x_\star^i, x_\star)\right)\\ &\quad +\frac{\gamma^2\omega}{M^2}\sum_{m=1}^M\left\|\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right\|^2\\ &\quad +\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_\star) - \nabla f_m^{\pi_m^i}(x_t^i)\right)\right\|^2\\ &\leq \left\|x_t^i - x_\star^i\right\|^2\\ &\quad -\frac{2\gamma}{M}\sum_{m=1}^M\left(D_{f_m^{\pi_m^i}}(x_\star^i, x_t^i) + D_{f_m^{\pi_m^i}}(x_t^i, x_\star) - D_{f_m^{\pi_m^i}}(x_\star^i, x_\star)\right)\\ &\quad +\frac{2\gamma^2\omega}{M}\frac{1}{M}\sum_{m=1}^M\left\|\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2\\ &\quad +\gamma^2\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla f_m^{\pi_m^i}(x_\star) - \nabla f_m^{\pi_m^i}(x_t^i)\right)\right\|^2\\ &\quad +\frac{2\gamma^2\omega}{M^2}\sum_{m=1}^M\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2.\end{aligned}$$

Using $L_{\max}$-smoothness and μ-strong convexity of functions f_m^i and $\widetilde{L}$-smoothness and $\widetilde{\mu}$-strong convexity of $f^{\pi^i} = \frac{1}{M}\sum_{i=1}^M f_m^{\pi_m^i}$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^{i+1} - x_\star^{i+1}\|^2\right] &\leq (1-\gamma\widetilde{\mu})\left\|x_t^i - x_\star^i\right\|^2 \\ &\quad -2\gamma\left(1-\gamma\left(\widetilde{L}+\frac{2\omega}{M}L_{\max}\right)\right)\frac{1}{M}\sum_{m=1}^M D_{f_m^{\pi_m^i}}(x_t^i, x_\star) \\ &\quad +\frac{2\gamma}{M}\sum_{m=1}^M D_{f_m^{\pi_m^i}}(x_\star^i, x_\star) + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^M\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2.\end{aligned}$$

Taking the full expectation and using Defenition 2, we derive

$$\begin{aligned}\mathbb{E}\left[\|x_t^{i+1} - x_\star^{i+1}\|^2\right] &\leq (1-\gamma\widetilde{\mu})\mathbb{E}\left[\left\|x_t^i - x_\star^i\right\|^2\right] \\ &\quad -2\gamma\left(1-\gamma\left(\widetilde{L}+\frac{2\omega}{M}L_{\max}\right)\right)\frac{1}{M}\sum_{m=1}^M \mathbb{E}\left[D_{f_m^{\pi_m^i}}(x_t^i, x_\star)\right] \\ &\quad +2\gamma^3\sigma_{\text{rad}}^2 + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^M\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right].\end{aligned}$$

Recursively unrolling the inequality, we get

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_\star\|^2\right] &\leq (1-\gamma\widetilde{\mu})^n\mathbb{E}\left[\|x_t - x_\star\|^2\right] \\ &\quad +\frac{2\gamma^2\omega}{M^2}\sum_{m=1}^M\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right] \\ &\quad -2\gamma\left(1-\gamma\left(\widetilde{L}+\frac{2\omega}{M}L_{\max}\right)\right)\frac{1}{M}\sum_{m=1}^M\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j\mathbb{E}\left[D_{f_m^{\pi_m^i}}(x_t^i, x_\star)\right] \\ &\quad +2\gamma^3\sigma_{\text{rad}}^2\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j.\end{aligned}$$

Next, we apply (6) and Lemma D.1:

$$\begin{aligned}&\mathbb{E}\left[\Psi_{t+1}\right] \leq (1-\gamma\widetilde{\mu})^n\mathbb{E}\left[\|x_t - x_\star\|^2\right] + 2\gamma^3\sigma_{\text{rad}}^2\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \\ &+\left(c(1-\alpha)+\frac{2\omega}{M}\right)\frac{\gamma^2}{M}\sum_{m=1}^M\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right] \\ &-2\gamma\left(1-c\gamma\alpha L_{\max}-\gamma\left(\widetilde{L}+\frac{2\omega}{M}L_{\max}\right)\right)\frac{1}{M}\sum_{m=1}^M\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j\mathbb{E}\left[D_{f_m^{\pi_m^i}}(x_\star^i, x_\star)\right],\end{aligned}$$

where $c = \frac{4\omega}{\alpha M^2}$. Using $\alpha \leq \frac{1}{1+\omega}$ and $\gamma \leq \min\left\{\frac{\alpha}{2n\mu}, \frac{1}{\left(\widetilde{L}+6\omega/ML_{\max}\right)}\right\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}\right] &\leq (1-\gamma\widetilde{\mu})^n \mathbb{E}\left[\|x_t - x_\star\|^2\right] \\
&\quad + \left(1-\frac{\alpha}{2}\right)\frac{4\omega\gamma^2}{\alpha M^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \mathbb{E}\left[\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right] \\
&\quad + 2\gamma^2\sigma_{\text{rad}}^3 \sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \\
&\leq \max\left\{(1-\gamma\widetilde{\mu})^n, \left(1-\frac{\alpha}{2}\right)\right\}\mathbb{E}\left[\Psi_t\right] \\
&\quad + 2\gamma^2\sigma_{\text{rad}}^3 \sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \\
&\leq (1-\gamma\widetilde{\mu})^n \mathbb{E}\left[\Psi_t\right] + 2\gamma^3\sigma_{\text{rad}}^2 \sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j.
\end{aligned}$$

Recursively rewriting the inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Psi_T\right] &\leq (1-\gamma\widetilde{\mu})^{nT}\Psi_0 + 2\gamma^3\sigma_{\text{rad}}^2 \sum_{t=0}^{T-1}(1-\gamma\widetilde{\mu})^{tn}\sum_{j=0}^{n-1}(1-\gamma\widetilde{\mu})^j \\
&\leq (1-\gamma\widetilde{\mu})^{nT}\Psi_0 + 2\gamma^3\sigma_{\text{rad}}^2 \sum_{k=0}^{nT-1}(1-\gamma\widetilde{\mu})^k
\end{aligned}$$

Using that $\sum\limits_{k=0}^{+\infty}\left(1-\frac{\gamma\widetilde{\mu}}{2}\right)^k \leq \frac{2}{\widetilde{\mu}\gamma}$, we finish proof. □

Corollary 7. *Let the assumptions of Theorem D.1 hold, $\alpha = \frac{1}{1+\omega}$ and*

$$\gamma = \min\left\{\frac{\alpha}{2n\widetilde{\mu}}, \frac{1}{\widetilde{L}+\frac{6\omega}{M}L_{\max}}, \frac{\sqrt{\varepsilon\widetilde{\mu}}}{2\sigma_{rad}}\right\}. \tag{22}$$

Then DIANA-RR *finds a solution with accuracy $\varepsilon > 0$ after the following number of communication rounds:*

$$\widetilde{\mathcal{O}}\left(n(1+\omega) + \frac{\widetilde{L}}{\widetilde{\mu}} + \frac{\omega}{M}\frac{L_{\max}}{\widetilde{\mu}} + \frac{\sigma_{rad}}{\sqrt{\varepsilon\widetilde{\mu}^3}}\right).$$

Proof. Theorem D.1 implies

$$\mathbb{E}\left[\Psi_T\right] \leq (1-\gamma\widetilde{\mu})^{nT}\Psi_0 + \frac{2\gamma^2\sigma_{\text{rad}}^2}{\widetilde{\mu}}.$$

To estimate the number of communication rounds required to find a solution with accuracy $\varepsilon > 0$, we need to upper bound each term from the right-hand side by $\frac{\varepsilon}{2}$. Thus, we get an additional condition on γ:

$$\frac{2\gamma^2\sigma_{\text{rad}}^2}{\widetilde{\mu}} < \frac{\varepsilon}{2},$$

and also the upper bound on the number of communication rounds nT

$$nT = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma\mu}\right).$$

Substituting (22) in the previous equation, we get the result. □

D.2 Non-Strongly Convex Summands

In this section, we provide the analysis of DIANA-RR without using Assumptions 4, 6. We emphasize that $x_t^{i+1} = x_t^i - \gamma \frac{1}{M}\sum_{m=1}^M \hat{g}_{t,m}^{\pi_m^i}$. Then we have

$$x_{t+1} = x_t - \gamma \sum_{i=0}^{n-1} \frac{1}{M} \sum_{m=1}^{M} \hat{g}_{t,m}^{\pi_m^i} = x_t - \tau \frac{1}{Mn} \sum_{i=0}^{n-1} \sum_{m=1}^{M} \hat{g}_{t,m}^{\pi_m^i}.$$

We denote $\hat{g}_t = \frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M} \hat{g}_{t,m}^{\pi_m^i}$.

Lemma D.2. *Let Assumptions 1, 2, 3, 5 hold. Then, the following inequality holds*

$$-2\tau \mathbb{E}_{\mathcal{Q}}\left[\langle \hat{g}_t - h_\star, x_t - x_\star\rangle\right] \leq -\frac{\tau\mu}{2}\|x_t - x_\star\|^2 - \tau\left(f(x_t) - f(x_\star)\right) + \tau\widetilde{L}\frac{1}{n}\sum_{i=1}^{n-1}\|x_t - x_t^i\|^2,$$

where $h^\star = \nabla f(x_\star) = 0$.

Proof. Since $h^\star = \nabla f(x_\star) = 0$, the proof of Lemma D.2 is identical to the proof of Lemma C.2. □

Lemma D.3. *Let Assumptions 1, 2, 3, 5 hold. Then, the following inequality holds*

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t - h_\star\|^2\right] &\leq 2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2 + 8\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\left(f(x_t) - f(x_\star)\right)\\ &\quad + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2\end{aligned}$$

Proof. Taking expectation w.r.t. $\mathcal{Q}$, we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t - h_\star\|^2\right] &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\hat{g}_{t,m}^{\pi_m^i} - h_\star\right\|^2\right]\\ &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left(h_{t,m}^{\pi_m^i} + \mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right)\right) - h_\star\right\|^2\right]\\ &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left(h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_t^i) + \mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right)\right)\right\|^2\right]\\ &\quad + \left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t^i) - h_\star\right\|^2.\end{aligned}$$

Independence of $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right)$, $m \in [M]$ and Assumption 1 imply

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t - h_\star\|^2\right] &= \frac{1}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_t^i) + \mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right)\right\|^2\right]\\ &\quad + \left\|\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\nabla f_m^{\pi_m^i}(x_t^i) - h_\star\right\|^2\\ &\leq \frac{\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t^i) - h_{t,m}^{\pi_m^i}\right\|^2 + \left\|\frac{1}{n}\sum_{i=0}^{n-1}\nabla f^{\pi^i}(x_t^i) - h_\star\right\|^2\\ &\leq \frac{2\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_t)\right\|^2 + \frac{2}{n}\sum_{i=0}^{n-1}\left\|\nabla f^{\pi^i}(x_t^i) - \nabla f^{\pi^i}(x_t)\right\|^2\\ &\quad + \frac{2\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_t)\right\|^2 + 2\left\|\frac{1}{n}\sum_{i=0}^{n-1}\nabla f^{\pi^i}(x_t) - h_\star\right\|^2.\end{aligned}$$

Using $L_{\max}$-smoothness and convexity of f_m^i and $\widetilde{L}$-smoothness and convexity of f^{π^i}, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t - h_\star\|^2\right] &\leq \frac{4\omega L_{\max}}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t^i, x_t) + \frac{4\widetilde{L}}{n}\sum_{i=0}^{n-1} D_{f^{\pi^i}}(x_t^i, x_t)\\ &\quad + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2 + 4\widetilde{L}\left(f(x_t) - f(x_\star)\right)\\ &\quad + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|\nabla f_m^{\pi_m^i}(x_t) - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2\\ &\leq 2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2 + 4\widetilde{L}\left(f(x_t) - f(x_\star)\right)\\ &\quad + \frac{8\omega}{Mn}L_{\max}\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t, x_\star)\\ &\quad + \frac{4\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\left\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2.\end{aligned}$$

□

Lemma D.4. *Let $\alpha \leq \frac{1}{1+\omega}$ and Assumptions 1, 2, 3, 5 hold. Then, the iterates produced by* DIANA-RR *satisfy*

$$\begin{aligned}\frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2\right] &\leq \frac{1-\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\quad + \frac{2\alpha\widetilde{L}L_{\max}}{n}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2\\ &\quad + 4\alpha L_{\max}\left(f(x_t) - f(x^\star)\right).\end{aligned}$$

Proof. Fist of all, we introduce new notation: $\mathcal{H}_t = \frac{1}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2\right]$. Using (21) and summing it up for $i = 0, \ldots, n-1$, we obtain

$$\begin{aligned}\mathcal{H}_{t+1} &\leq \frac{1-\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2 + \frac{\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\leq \frac{1-\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2 + \frac{2\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|\nabla f_m^{\pi_m^i}(x_t^i) - \nabla f_m^{\pi_m^i}(x_t)\|^2\\ &\quad + \frac{2\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|\nabla f_m^{\pi_m^i}(x_t) - \nabla f_m^{\pi_m^i}(x_\star)\|^2.\end{aligned}$$

Next, we apply $L_{\max}$-smoothness and convexity of f_m^i and $\widetilde{L}$-smoothness and convexity of f^{π^i}:

$$\begin{aligned}\mathcal{H}_{t+1} &\leq \frac{1-\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2 + \frac{4\alpha}{Mn}L_{\max}\sum_{i=0}^{n-1}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t^i, x_t)\\ &\quad + \frac{4\alpha}{Mn}L_{\max}\sum_{i=0}^{n-1}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t, x_\star)\\ &\leq \frac{1-\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2 + \frac{2\alpha}{n}\widetilde{L}L_{\max}\sum_{i=0}^{n-1}\|x_t^i - x_t\|^2\\ &\quad + \frac{4\alpha}{Mn}L_{\max}\sum_{i=0}^{n-1}\sum_{m=1}^{M} D_{f_m^{\pi_m^i}}(x_t, x_\star).\end{aligned}$$

□

Lemma D.5. *Let Assumptions 1, 2, 3, 5 and* $\tau \leq \frac{1}{2\sqrt{\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)}}$. *Then, the following inequality holds*

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 24\tau^2\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\mathbb{E}\left[f(x_t)-f(x_\star)\right]+8\tau^2\frac{\sigma_{\star,n}^2}{n}\\ &\quad +8\frac{\tau^2\omega}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}\left[\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\right],\end{aligned}$$

where $\sigma_{\star,n}^2 = \frac{1}{n}\sum_{i=1}^n\|\nabla f^i(x_\star)\|^2$.

Proof. Since $x_t^i = x_t - \frac{\tau}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\left(h_{t,m}^{\pi_m^j}+\mathcal{Q}\left(\nabla f_m^{\pi_m^j}(x_t^j)-h_{t,m}^{\pi_m^j}\right)\right)$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^i - x_t\|^2\right] &= \tau^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\left(h_{t,m}^{\pi_m^j}+\mathcal{Q}\left(\nabla f_m^{\pi_m^j}(x_t^j)-h_{t,m}^{\pi_m^j}\right)\right)\right\|^2\right]\\ &= \tau^2\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\left(h_{t,m}^{\pi_m^j}-\nabla f_m^{\pi_m^j}(x_t^j)+\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^j)-h_{t,m}^{\pi_m^j}\right)\right)\right\|^2\right]\\ &\quad +\tau^2\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2.\end{aligned}$$

Independence of $\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^j)-h_{t,m}^{\pi_m^j}\right)$, $m\in[M]$ and Assumption 1 imply

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^i - x_t\|^2\right] &= \frac{\tau^2}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{i-1}\mathbb{E}_{\mathcal{Q}}\left[\left\|h_{t,m}^{\pi_m^j}-\nabla f_m^{\pi_m^j}(x_t^j)+\mathcal{Q}\left(\nabla f_m^{\pi_m^i}(x_t^j)-h_{t,m}^{\pi_m^j}\right)\right\|^2\right]\\ &\quad +\tau^2\left\|\frac{1}{Mn}\sum_{m=1}^M\sum_{j=0}^{i-1}\nabla f_m^{\pi_m^j}(x_t^j)\right\|^2\\ &\leq \frac{\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{i-1}\left\|\nabla f^{\pi_m^j}(x_t^j)-h_{t,m}^{\pi_m^j}\right\|^2+\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t^j)\right\|^2\\ &\leq \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}\left\|\nabla f_m^{\pi_m^j}(x_t^j)-\nabla f_m^{\pi_m^j}(x_t)\right\|^2+2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2\\ &\quad +\frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^M\sum_{j=0}^{n-1}\left\|h_{t,m}^{\pi_m^j}-\nabla f_m^{\pi_m^j}(x_t)\right\|^2+\frac{2\tau^2}{n}\sum_{j=0}^{n-1}\left\|\nabla f^{\pi^j}(x_t^j)-\nabla f^{\pi^j}(x_t)\right\|^2.\end{aligned}$$

Using $L_{\max}$-smoothness and convexity of f_m^i and $\widetilde{L}$-smoothness and convexity of f^{π^j}, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_t^i - x_t\|^2\right] &\leq \frac{4\tau^2\omega}{M^2n^2}L_{\max}\sum_{m=1}^{M}\sum_{j=0}^{n-1}D_{f_m^{\pi_m^j}}(x_t^j, x_t) + 2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2 \\ &\quad + \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\left\|h_{t,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_t)\right\|^2 + \frac{2\tau^2\widetilde{L}^2}{n}\sum_{j=0}^{n-1}\left\|x_t^j - x_t\right\|^2 \\ &\leq 2\tau^2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{j=0}^{n-1}\|x_t^j - x_t\|^2 + 2\tau^2\left\|\frac{1}{n}\sum_{j=0}^{i-1}\nabla f^{\pi^j}(x_t)\right\|^2 \\ &\quad + \frac{2\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\left\|h_{t,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_t)\right\|^2.\end{aligned}$$

Taking the full expectation and using (17), we derive

$$\begin{aligned}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 2\tau^2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{E}\left[\|x_t^j - x_t\|^2\right] + 2\tau^2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\ &\quad + \frac{4\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] + \frac{2\tau^2}{n}\mathbb{E}\left[\sigma_t^2\right] \\ &\quad + \frac{8\tau^2\omega}{M^2n^2}L_{\max}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[D_{f_m^{\pi_m^j}}(x_t, x_\star)\right].\end{aligned}$$

Using $L_{\max}$-smoothness and convexity of f_m^i and $\widetilde{L}$-smoothness and convexity of f^{π^j}, we obtain

$$\begin{aligned}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 2\tau^2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{E}\left[\|x_t^j - x_t\|^2\right] \\ &\quad + \frac{4\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] + \frac{2\tau^2}{n}\mathbb{E}\left[\sigma_t^2\right] \\ &\quad + 4\tau^2\left(\widetilde{L} + \frac{2\omega}{M^2n^2}L_{\max}\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right].\end{aligned}$$

Now we need to estimate $\frac{2\tau^2}{n}\mathbb{E}\left[\sigma_t^2\right]$. Due to $\mathbb{E}\left[\sigma_t^2\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f^i(x_t)\|^2\right]$, we get

$$\begin{aligned}\frac{2\tau^2}{n}\mathbb{E}\left[\sigma_t^2\right] &\leq \frac{2\tau^2}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f^j(x_t)\|^2\right] \\ &\leq \frac{4\tau^2}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f^j(x_t) - \nabla f^j(x_\star)\|^2\right] + \frac{4\tau^2}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f^j(x_\star)\|^2\right] \\ &\leq \frac{8\tau^2}{n^2}\widetilde{L}\sum_{j=1}^{n}\mathbb{E}\left[D_{f^j}(x_t, x_\star)\right] + \frac{4\tau^2}{n^2}\sum_{j=1}^{n}\sigma_{n,\star}^2.\end{aligned}$$

Combining two previous inequalities, we get

$$\begin{aligned}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 2\tau^2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{E}\left[\|x_t^j - x_t\|^2\right] \\ &\quad + \frac{4\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] \\ &\quad + 4\tau^2\left(\widetilde{L} + \frac{2\omega}{M^2n^2}L_{\max}\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\ &\quad + \frac{8\tau^2}{n}\widetilde{L}\mathbb{E}\left[f(x_t) - f(x_\star)\right] + \frac{4\tau^2}{n^2}\sum_{j=1}^{n}\sigma_{n,\star}^2.\end{aligned}$$

Summing from $i = 0$ to $n-1$ and using $\tau \leq \frac{1}{2\sqrt{\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)}}$, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] &\leq 2\left(1 - 2\tau^2\widetilde{L}\left(\widetilde{L} + \frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] \\ &\leq \frac{8\tau^2\omega}{M^2n^2}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|h_{t,m}^{\pi_m^j} - \nabla f_m^{\pi_m^j}(x_\star)\right\|^2\right] \\ &\quad + 8\tau^2\left(\widetilde{L} + \frac{2\omega}{M^2n^2}L_{\max}\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\ &\quad + \frac{16\tau^2}{n}\widetilde{L}\mathbb{E}\left[f(x_t) - f(x_\star)\right] + \frac{8\tau^2}{n^2}\sum_{j=1}^{n}\sigma_{n,\star}^2.\end{aligned}$$

□

We consider the following Lyapunov function:

$$\Psi_{t+1} = \|x_{t+1} - x_\star\|^2 + \frac{c\tau^2}{Mn}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\left\|h_{t+1,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\right\|^2. \tag{23}$$

Theorem D.2. *Let Assumptions 1, 2, 3, 5 hold and*

$$\gamma \leq \min\left\{\frac{\alpha}{n\mu}, \frac{1}{12n\left(\widetilde{L} + \frac{11\omega}{Mn}L_{\max}\right)}\right\}, \quad \alpha \leq \frac{1}{1+\omega}, \quad c = \frac{10\omega}{\alpha Mn}.$$

Then, for all $T \geq 0$ the iterates produced by DIANA-RR *satisfy*

$$\mathbb{E}\left[\Psi_T\right] \leq \left(1 - \frac{n\gamma\mu}{2}\right)^T\Psi_0 + 20\frac{\gamma^2 n\widetilde{L}}{\mu}\sigma_{\star,n}^2.$$

Proof. Taking expectation w.r.t. $\mathcal{Q}$ and using Lemma D.2, we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1} - x_\star\|^2\right] &= \|x_t - \tau\hat{g}_t - x_\star + \tau h^\star\|^2 \\ &= \|x_t - x_\star\|^2 - 2\tau\mathbb{E}_{\mathcal{Q}}\left[\langle\hat{g}_t - h^\star, x_t - x_\star\rangle\right] + \tau^2\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t - h^\star\|^2\right] \\ &\leq \|x_t - x_\star\|^2 - \frac{\tau\mu}{2}\|x_t - x_\star\|^2 + \tau^2\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t - h^\star\|^2\right] \\ &\quad - \tau\left(f(x_t) - f(x_\star)\right) + \tau\widetilde{L}\frac{1}{n}\sum_{i=1}^{n-1}\|x_t - x_t^i\|^2.\end{aligned}$$

Next, due to Lemma D.3 we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &\leq \left(1-\frac{\tau\mu}{2}\right)\|x_t-x_\star\|^2-\tau\left(f(x_t)-f(x_\star)\right)+\tau\widetilde{L}\frac{1}{n}\sum_{i=1}^{n-1}\|x_t-x_t^i\|^2\\ &\quad+2\tau^2\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2\\ &\quad+8\tau^2\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+\frac{4\omega\tau^2}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\leq \left(1-\frac{\tau\mu}{2}\right)\|x_t-x_\star\|^2+\frac{4\omega\tau^2}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\quad-\tau\left(1-8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+\tau\widetilde{L}\left(1+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2.\end{aligned}$$

Using (23), we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\tau\mu}{2}\right)\|x_t-x_\star\|^2+\frac{4\omega\tau^2}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\quad-\tau\left(1-8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+\tau\widetilde{L}\left(1+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2\\ &\quad+\frac{c\tau^2}{Mn}\sum_{m=1}^{M}\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|h_{t+1,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\right\|^2\right].\end{aligned}$$

To estimate the last term in the above inequality, we apply Lemma D.4:

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\tau\mu}{2}\right)\|x_t-x_\star\|^2+\frac{4\omega\tau^2}{M^2n^2}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\quad-\tau\left(1-8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+\tau\widetilde{L}\left(1+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2\\ &\quad+c\tau^2\frac{1-\alpha}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\quad+c\tau^2\frac{2\alpha\widetilde{L}L_{\max}}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2+4c\tau^2\alpha L_{\max}\left(f(x_t)-f(x^\star)\right)\\ &\leq \left(1-\frac{\tau\mu}{2}\right)\|x_t-x_\star\|^2+\left(1-\alpha+\frac{4\omega}{cMn}\right)\frac{c\tau^2}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\|h_{t,m}^{\pi_m^i}-\nabla f_m^{\pi_m^i}(x_\star)\|^2\\ &\quad-\tau\left(1-4c\tau\alpha L_{\max}-8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+\tau\widetilde{L}\left(1+2c\tau\alpha L_{\max}+2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\|x_t^i-x_t\|^2.\end{aligned}$$

Let $\mathcal{H}_t = \frac{c\tau^2}{Mn}\sum_{i=0}^{n-1}\sum_{m=1}^{M}\mathbb{E}\left[\|h_{t,m}^{\pi_m^i} - \nabla f_m^{\pi_m^i}(x_\star)\|^2\right]$. Taking the full expectation and using Lemma D.5, we get

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\|x_t - x_\star\|^2\right] + \left(1-\alpha+\frac{4\omega}{cMn}\right)\mathcal{H}_t \\
&\quad -\tau\left(1 - 4c\tau\alpha L_{\max} - 8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad +\tau\widetilde{L}\left(1 + 2c\tau\alpha L_{\max} + 2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_t^i - x_t\|^2\right] \\
&\leq \left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\|x_t - x_\star\|^2\right] + \left(1-\alpha+\frac{4\omega}{cMn}\right)\mathcal{H}_t \\
&\quad -\tau\left(1 - 4c\tau\alpha L_{\max} - 8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad +24\tau^3\widetilde{L}\left(1 + 2c\tau\alpha L_{\max} + 2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad +8\tau^3\widetilde{L}\left(1 + 2c\tau\alpha L_{\max} + 2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\frac{\sigma_{\star,n}^2}{n} \\
&\quad +\frac{8\tau\widetilde{L}\omega}{cMn}\left(1 + 2c\tau\alpha L_{\max} + 2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\right)\mathcal{H}_t.
\end{aligned}$$

Selecting $c = \frac{A\omega}{\alpha Mn}$, where A is a positive number to be specified later, we have

$$\begin{aligned}
1 + 2c\tau\alpha L_{\max} + 2\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right) &= 1 + 2\tau\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right), \\
1 - 4c\tau\alpha L_{\max} - 8\tau\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right) &\geq 1 - 8\tau\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\|x_t - x_\star\|^2\right] + \left(1-\alpha+\frac{4\alpha}{A}\right)\mathcal{H}_t \\
&\quad -\tau\left(1 - 8\tau\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right)\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad +24\tau^3\widetilde{L}\left(\widetilde{L}+\frac{\omega}{Mn}L_{\max}\right)\left(1 + 2\tau\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right)\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad +8\tau^3\widetilde{L}\left(1 + 2\tau\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right)\right)\frac{\sigma_{\star,n}^2}{n} \\
&\quad +\frac{8\alpha}{A}\tau\widetilde{L}\left(1 + 2\tau\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right)\right)\mathcal{H}_t.
\end{aligned}$$

Taking $\tau = \frac{1}{B\left(\widetilde{L}+\frac{(A+1)\omega}{Mn}L_{\max}\right)}$, where B is some positive constant, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\|x_t - x_\star\|^2\right] + \left(1-\alpha+\frac{4\alpha}{A}+\frac{8\alpha}{A}\tau\widetilde{L}\left(1+\frac{2}{B}\right)\right)\mathcal{H}_t \\
&\quad -\tau\left(1 - \frac{8}{B} - \frac{24}{B^2}\left(1+\frac{2}{B}\right)\right)\mathbb{E}\left[f(x_t) - f(x_\star)\right] \\
&\quad +8\tau^3\widetilde{L}\left(1+\frac{2}{B}\right)\frac{\sigma_{\star,n}^2}{n}.
\end{aligned}$$

Choosing $A = 10$, $B = 12$, $\tau \leq \frac{\alpha}{\mu}$, we have

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1 - \min\left\{\frac{\tau\mu}{2}, \frac{\alpha}{2}\right\}\right)\mathbb{E}\left[\Psi_t\right] + 10\tau^3\widetilde{L}\frac{\sigma_{\star,n}^2}{n} \\
&\leq \left(1-\frac{\tau\mu}{2}\right)\mathbb{E}\left[\Psi_t\right] + 10\tau^3\widetilde{L}\frac{\sigma_{\star,n}^2}{n}
\end{aligned}$$

Recursively unrolling the inequality, substituting $\tau = n\gamma$ and using $\sum\limits_{t=0}^{+\infty} \left(1 - \frac{\tau\mu}{2}\right)^t \leq \frac{2}{\mu\tau}$, we finish proof. □

Corollary 8. *Let the assumptions of Theorem D.2 hold,* $\alpha = \frac{1}{1+\omega}$*, and*

$$\gamma = \min\left\{\frac{\alpha}{2n\mu}, \frac{1}{12n\left(\widetilde{L} + \frac{11\omega}{Mn}L_{\max}\right)}, \sqrt{\frac{\varepsilon\mu}{40n\widetilde{L}\sigma_{\star,n}^2}}\right\}. \tag{24}$$

Then, DIANA-RR *finds a solution with accuracy* $\varepsilon > 0$ *after the following number of communication rounds:*

$$\widetilde{\mathcal{O}}\left(n(1+\omega) + \frac{n\widetilde{L}}{\mu} + \frac{\omega}{M}\frac{L_{\max}}{\mu} + \sqrt{\frac{n\widetilde{L}}{\varepsilon\mu^3}}\sigma_{\star,n}\right).$$

Proof. Theorem D.2 implies

$$\mathbb{E}\left[\Psi_T\right] \leq (1-\gamma\mu)^{nT}\Psi_0 + 20\frac{\gamma^2 n\widetilde{L}}{\mu}\sigma_{\star,n}^2.$$

To estimate the number of communication rounds required to find a solution with accuracy $\varepsilon > 0$, we need to upper bound each term from the right-hand side by $\frac{\varepsilon}{2}$. Thus, we get an additional condition on γ:

$$20\frac{\gamma^2 n\widetilde{L}}{\mu}\sigma_{\star,n}^2 < \frac{\varepsilon}{2},$$

and also the upper bound on the number of communication rounds nT

$$nT = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma\mu}\right).$$

Substituting (24) in the previous equation, we obtain the result. □

E Theoretical Results for Q-NASTYA and DIANA-NASTYA

Theorem E.1. *Let Assumptions 1, 2, 3 hold. Let the stepsizes γ, η satisfy $0 < \eta \leq \frac{1}{16L_{\max}\left(1+\frac{\omega}{M}\right)}$, $0 < \gamma \leq \frac{1}{5nL_{\max}}$. Then, for all $T \geq 0$ the iterates produced by Q-NASTYA (Algorithm 3) satisfy*

$$\mathbb{E}\left[\|x_T - x_\star\|^2\right] \leq \left(1 - \frac{\eta\mu}{2}\right)^T \|x_0 - x_\star\|^2 + 8\frac{\eta\omega}{\mu M}\zeta_\star^2 + \frac{9}{2}\frac{\gamma^2 n L_{\max}}{\mu}\left((n+1)\zeta_\star^2 + \sigma_\star^2\right).$$

Corollary 9. *Under the same conditions as Theorem E.1 and for Algorithm 3, there exist stepsizes $\gamma = \eta/n$ and $\eta > 0$ such that the number of communication rounds T to find a solution with accuracy $\varepsilon > 0$ is $\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right) + \frac{\omega}{M}\frac{\zeta_\star^2}{\varepsilon\mu^3} + \sqrt{\frac{L_{\max}}{\varepsilon\mu^3}}\sqrt{\zeta_\star^2 + \frac{\sigma_\star^2}{n}}\right)$. If $\gamma \to 0$, one can choose $\eta > 0$ such that the above complexity bound improves to $\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right) + \frac{\omega}{M}\frac{\zeta_\star^2}{\varepsilon\mu^3}\right)$.*

We emphasize several differences with the known theoretical results. First, the FedCOM method of Haddadpour et al. [2021] was analyzed in the homogeneous setting only, i.e., $f_m(x) = f(x)$ for all $m \in [M]$, which is an unrealistic assumption for FL applications. In contrast, our result holds in the fully heterogeneous case. Next, the analysis of FedPAQ of Reisizadeh et al. [2020] uses a bounded variance assumption, which is also known to be restrictive. Nevertheless, let us compare to their result. Reisizadeh et al. [2020] derive the following complexity for their method: $\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right) + \frac{\omega}{M}\frac{\sigma^2}{\mu^2\varepsilon} + \frac{\sigma^2}{M\mu^2\varepsilon}\right)$. This result is inferior to the one we show for Q-NASTYA: when ω is small, the main term in the complexity bound of FedPAQ is $\widetilde{\mathcal{O}}\left(1/\varepsilon\right)$, while for Q-NASTYA the dominating term is of the order $\widetilde{\mathcal{O}}\left(1/\sqrt{\varepsilon}\right)$ (when ω and ε are sufficiently small). We also highlight that FedCRR [Malinovsky and Richtárik, 2022] does not converge if $\omega > M^2\gamma\mu\varepsilon/\left(2\left\|x^n_{*,m}\right\|^2\right)$, while Q-NASTYA does for any $\omega \geq 0$. Finally, when $\omega = 0$ (no compression) we recover NASTYA as a special case, and using $\gamma = \eta/n$, we recover the rate of FedRR [Mishchenko et al., 2021].

Theorem E.2. *Let Assumptions 1, 2, 3 hold. Suppose the stepsizes γ, η, α satisfy $0 < \gamma \leq \frac{1}{16L_{\max}n}$, $0 < \eta \leq \min\left\{\frac{\alpha}{2\mu}, \frac{1}{16L_{\max}\left(1+\frac{9\omega}{M}\right)}\right\}$, and $\alpha \leq \frac{1}{1+\omega}$. Define the following Lyapunov function:*

$$\Psi_{t+1} \overset{\text{def}}{=} \|x_{t+1} - x_\star\|^2 + \frac{8\omega\eta^2}{\alpha M^2}\sum_{m=1}^M \|h_{t+1,m} - h_m^\star\|^2. \tag{25}$$

Then, for all $T \geq 0$ the iterates produced by DIANA-NASTYA (Algorithm 4) satisfy

$$\mathbb{E}\left[\Psi_T\right] \leq \left(1 - \frac{\eta\mu}{2}\right)^T \Psi_0 + \frac{9}{2}\frac{\gamma^2 nL}{\mu}\left((n+1)\zeta_\star^2 + \sigma_\star^2\right). \tag{26}$$

Corollary 10. *Under the same conditions as Theorem E.2 and for Algorithm 4, there exist stepsizes $\gamma = \eta/n$, $\eta > 0$, $\alpha > 0$ such that the number of communication rounds T to find a solution with accuracy $\varepsilon > 0$ is $\widetilde{\mathcal{O}}\left(\omega + \frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right) + \sqrt{\frac{L_{\max}}{\varepsilon\mu^3}}\sqrt{\zeta_\star^2 + \frac{\sigma_\star^2}{n}}\right)$. If $\gamma \to 0$, one can choose $\eta > 0$ such that the above complexity bound improves to $\widetilde{\mathcal{O}}\left(\omega + \frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right)\right)$.*

In contrast to Q-NASTYA, DIANA-NASTYA does not suffer from the $\widetilde{\mathcal{O}}(1/\varepsilon)$ term in the complexity bound. This shows the superiority of DIANA-NASTYA to Q-NASTYA. Next, FedCRR-VR [Malinovsky and Richtárik, 2022] has the rate $\widetilde{\mathcal{O}}\left(\frac{(\omega+1)\left(1-\frac{1}{\kappa}\right)^n}{\left(1-\left(1-\frac{1}{\kappa}\right)^n\right)^2} + \frac{\sqrt{\kappa}(\zeta_\star+\sigma_\star)}{\mu\sqrt{\varepsilon}}\right)$, which depends on $\widetilde{\mathcal{O}}\left(1/\sqrt{\varepsilon}\right)$. However, the first term is close to $\widetilde{\mathcal{O}}\left((\omega+1)\kappa^2\right)$ for a large condition number. FedCRR-VR-2 utilizes variance reduction technique from Malinovsky et al. [2021] and it allows to get rid of permutation variance. This method has $\widetilde{\mathcal{O}}\left(\frac{(\omega+1)\left(1-\frac{1}{\kappa\sqrt{\kappa n}}\right)^{\frac{n}{2}}}{\left(1-\left(1-\frac{1}{\kappa\sqrt{\kappa n}}\right)^{\frac{n}{2}}\right)^2} + \frac{\sqrt{\kappa}\zeta_\star}{\mu\sqrt{\varepsilon}}\right)$ complexity, but it requires additional assumption on number of functions n and thus not directly comparable with our result. Note that if we have no compression ($\omega = 0$), DIANA-NASTYA recovers rate of NASTYA.

F Missing Proofs for Q-NASTYA

We start with deriving a technical lemma along with stating several useful results from [Malinovsky et al., 2022]. For convenience, we also introduce the following notation:

$$g_{t,m} = \frac{1}{n}\sum_{i=0}^{n-1} \nabla f_m^{\pi_m^i}(x_{t,m}^i).$$

Lemma F.1. *Let Assumptions 1, 2, 3 hold. Then, for all $t \geq 0$ the iterates produced by Q-NASTYA satisfy*

$$\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] \leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 + 8L_{\max}\left(1+\frac{\omega}{M}\right)\left(f(x_t)-f(x_\star)\right) + \frac{4\omega}{M}\zeta_\star^2,$$

where $\mathbb{E}_{\mathcal{Q}}$ is expectation w.r.t. $\mathcal{Q}$, and $\zeta_\star^2 = \frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(x_\star)\|^2$.

Proof. Using the variance decomposition $\mathbb{E}\left[\|\xi\|^2\right] = \mathbb{E}\left[\|\xi - \mathbb{E}[\xi]\|^2\right] + \|\mathbb{E}\xi\|^2$, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\left\|\mathcal{Q}\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)\right) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)\right\|^2\right] \\
&\quad + \left\|\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)\right\|^2 \\
&\overset{\text{Asm.1}}{\leq} \frac{\omega}{M^2}\sum_{m=1}^{M}\left\|\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)\right\|^2 + \left\|\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)\right\|^2.
\end{aligned}$$

Next, we use $\nabla f_m(x_t) = \frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_t)$ and $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &\leq \frac{2\omega}{M^2}\sum_{m=1}^{M}\left\|\frac{1}{n}\sum_{i=0}^{n-1}\left(\nabla f_m^{\pi_m^i}(x_{t,m}^i) - \nabla f_m^{\pi_m^i}(x_t)\right)\right\|^2 + \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x_t)\|^2 \\
&\quad + 2\left\|\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left(\nabla f_m^{\pi_m^i}(x_{t,m}^i) - \nabla f_m^{\pi_m^i}(x_t)\right)\right\|^2 + 2\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(x_t)\right\|^2 \\
&\leq \frac{2\left(1+\frac{\omega}{M}\right)}{M}\sum_{m=1}^{M}\left\|\frac{1}{n}\sum_{i=0}^{n-1}\left(\nabla f_m^{\pi_m^i}(x_{t,m}^i) - \nabla f_m^{\pi_m^i}(x_t)\right)\right\|^2 \\
&\quad + \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x_t)\|^2 + 2\|\nabla f(x_t)\|^2.
\end{aligned}$$

Using $L_{i,m}$-smoothness of f_m^i and f and also convexity of f_m, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] &\leq \frac{2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|\nabla f_m^{\pi_m^i}(x_{t,m}^i) - \nabla f_m^{\pi_m^i}(x_t)\right\|^2 + \frac{4\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x_t) - \nabla f_m(x_\star)\|^2 \\
&\quad + \frac{4\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x_\star)\|^2 + 2\|\nabla f(x_t) - \nabla f(x_\star)\|^2 \\
&\leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 + \frac{8L_{\max}\left(1+\frac{\omega}{M}\right)}{M}\sum_{m=1}^{M}D_{f_m}(x_t, x_\star) + \frac{4\omega}{M}\zeta_\star^2.
\end{aligned}$$

□

Lemma F.2 (see [Malinovsky et al., 2022]). *Under Assumptions 1, 2, 3, it holds*

$$-\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\langle f_m^{\pi_m^i}(x_{t,m}^i), x_t - x_\star\right\rangle \leq -\frac{\mu}{4}\|x_t - x_\star\|^2 - \frac{1}{2}\left(f(x_t) - f(x_\star)\right) + \frac{L_{\max}}{2Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2.$$

Lemma F.3 (see [Malinovsky et al., 2022]). *Under Assumptions 1, 2, 3 and $\gamma \leq \frac{1}{2L_{\max}n}$, it holds*

$$\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 \leq 8\gamma^2 n^2 L_{\max}\left(f(x_t) - f(x_\star)\right) + 2\gamma^2 n\left(\sigma_\star^2 + (n+1)\zeta_\star^2\right).$$

Theorem F.1. *Let Assumptions 1, 2, 3 hold and stepsizes γ, η satisfy*

$$0 < \eta \leq \frac{1}{16L_{\max}\left(1 + \frac{\omega}{M}\right)}, \quad 0 < \gamma \leq \frac{1}{5nL_{\max}}. \tag{27}$$

Then, for all $T \geq 0$ the iterates produced by Q-NASTYA *satisfy*

$$\mathbb{E}\left[\|x_T - x_\star\|^2\right] \leq \left(1 - \frac{\eta\mu}{2}\right)^T\|x_0 - x_\star\|^2 + \frac{9}{2}\frac{\gamma^2 nL_{\max}}{\mu}\left(\sigma_\star^2 + (n+1)\zeta_\star^2\right) + 8\frac{\eta\omega}{\mu M}\zeta_\star^2.$$

Proof. Taking expectation w.r.t. $\mathcal{Q}$ and using Lemma F.1, we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1} - x_\star\|^2\right] &= \|x_t - x_\star\|^2 - 2\eta\mathbb{E}_{\mathcal{Q}}\left[\langle g_t, x_t - x_\star\rangle\right] + \eta^2\mathbb{E}_{\mathcal{Q}}\left[\|g^t\|^2\right] \\
&\leq \|x_t - x_\star\|^2 - 2\eta\mathbb{E}_{\mathcal{Q}}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}\mathcal{Q}\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)\right), x_t - x_\star\right\rangle\right] \\
&\quad + \frac{2\eta^2 L_{\max}^2\left(1 + \frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 \\
&\quad + 8\eta^2 L_{\max}\left(1 + \frac{\omega}{M}\right)\left(f(x_t) - f(x_\star)\right) + 4\eta^2\frac{\omega}{M}\zeta_\star^2 \\
&\leq \|x_t - x_\star\|^2 - 2\eta\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\langle \nabla f_m^{\pi_m^i}(x_{t,m}^i), x_t - x_\star\right\rangle \\
&\quad + \frac{2\eta^2 L_{\max}^2\left(1 + \frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 \\
&\quad + 8\eta^2 L_{\max}\left(1 + \frac{\omega}{M}\right)\left(f(x_t) - f(x_\star)\right) + 4\eta^2\frac{\omega}{M}\zeta_\star^2.
\end{aligned}$$

Next, Lemma F.2 implies

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1} - x_\star\|^2\right] &\leq \|x_t - x_\star\|^2 - \frac{\eta\mu}{2}\|x_t - x_\star\|^2 - \eta\left(f(x_t) - f(x_\star)\right) \\
&\quad + 8\eta^2 L_{\max}\left(1 + \frac{\omega}{M}\right)\left(f(x_t) - f(x_\star)\right) + \frac{\eta L_{\max}}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 \\
&\quad + \frac{2\eta^2 L_{\max}^2\left(1 + \frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 + 4\eta^2\frac{\omega}{M}\zeta_\star^2 \\
&\leq \left(1 - \frac{\eta\mu}{2}\right)\|x_t - x_\star\|^2 - \eta\left(1 - 8\eta L_{\max}\left(1 + \frac{\omega}{M}\right)\right)\left(f(x_t) - f(x_\star)\right) \\
&\quad + \frac{\eta L_{\max}\left(1 + 2\eta L_{\max}\left(1 + \frac{\omega}{M}\right)\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 + 4\eta^2\frac{\omega}{M}\zeta_\star^2.
\end{aligned}$$

Using Lemma F.3, we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &\leq \left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2-\eta\left(1-8\eta L\left(1+\frac{\omega}{M}\right)\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+\eta L_{\max}\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)\right)\cdot 8\gamma^2 n^2 L_{\max}\left(f(x_t)-f(x_\star)\right)\\ &\quad+\eta L_{\max}\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)\right)\cdot 2\gamma^2 n\left(\sigma_\star^2+(n+1)\zeta_\star^2\right)\\ &\quad+4\eta^2\frac{\omega}{M}\zeta_\star^2.\end{aligned}$$

In view of (27), we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &\leq \left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2+4\eta^2\frac{\omega}{M}\zeta_\star^2\\ &\quad-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{M}\right)-8\gamma^2 n^2 L_{\max}^2\left(1+2L_{\max}\eta\left(1+\frac{\omega}{M}\right)\right)\right)\left(f(x_t)-f(x_\star)\right)\\ &\quad+2\gamma^2 n\eta L_{\max}\left(1+2\eta L\left(1+\frac{\omega}{M}\right)\right)\left(\sigma_\star^2+n\zeta_\star^2\right)\\ &\leq \left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2+4\eta^2\frac{\omega}{M}\zeta_\star^2+\frac{9}{4}\eta L_{\max}\gamma^2 n\left(\sigma_\star^2+(n+1)\sigma_\star^2\right).\end{aligned}$$

Recursively unrolling the inequality and using $\sum\limits_{t=0}^{+\infty}\left(1-\frac{\eta\mu}{2}\right)^t\leq\frac{2}{\mu\eta}$, we get the result. □

Corollary 11. *Let the assumptions of Theorem E.1 hold, $\gamma=\eta/n$, and*

$$\eta=\min\left\{\frac{1}{16L_{\max}\left(1+\frac{\omega}{M}\right)},\sqrt{\frac{\varepsilon\mu n}{9L_{\max}}}\left((n+1)\zeta_\star^2+\sigma_\star^2\right)^{-1/2},\frac{\varepsilon\mu M}{24\omega\zeta_\star^2}\right\}. \tag{28}$$

Then, Q-NASTYA *finds a solution with accuracy $\varepsilon>0$ after the following number of communication rounds:*

$$\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right)+\frac{\omega}{M}\frac{\zeta_\star^2}{\varepsilon\mu^3}+\sqrt{\frac{L_{\max}}{\varepsilon\mu^3}}\sqrt{\zeta_\star^2+\sigma_\star^2/n}\right).$$

If $\gamma\to 0$, one can choose $\eta=\min\left\{\frac{1}{16L_{\max}\left(1+\frac{\omega}{M}\right)},\frac{\varepsilon\mu M}{24\omega\zeta_\star^2}\right\}$ such that the above complexity bound improves to

$$\widetilde{\mathcal{O}}\left(\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right)+\frac{\omega}{M}\frac{\zeta_\star^2}{\varepsilon\mu^3}\right).$$

Proof. Theorem E.1 implies

$$\mathbb{E}\left[\|x_T-x_\star\|^2\right]\leq\left(1-\frac{\eta\mu}{2}\right)^T\|x_0-x_\star\|^2+\frac{9}{2}\frac{\gamma^2 n L_{\max}}{\mu}\left((n+1)\zeta_\star^2+\sigma_\star^2\right)+8\frac{\eta\omega}{\mu M}\zeta_\star^2.$$

To estimate the number of communication rounds required to find a solution with accuracy $\varepsilon>0$, we need to upper bound each term from the right-hand side by $\varepsilon/3$. Thus, we get additional conditions on η:

$$\frac{9}{2}\frac{\eta^2 L_{\max}}{n\mu}\left((n+1)\zeta_\star^2+\sigma_\star^2\right)<\frac{\varepsilon}{3},\quad 8\frac{\eta\omega}{\mu M}\zeta_\star^2<\frac{\varepsilon}{3}$$

and also the upper bound on the number of communication rounds T

$$T=\widetilde{\mathcal{O}}\left(\frac{1}{\eta\mu}\right).$$

Substituting (31) in the previous equation, we get the first part of the result. When $\gamma\to 0$, the proof follows similar steps. □

G Missing Proofs for DIANA-NASTYA

Lemma G.1. *Under Assumptions 1, 2, 3, the iterates produced by* DIANA-NASTYA *satisfy*

$$-\mathbb{E}_{\mathcal{Q}}\left[\frac{1}{M}\sum_{m=1}^{M}\langle \hat{g}_{t,m}-h^\star, x_t-x_\star\rangle\right] \leq -\frac{\mu}{4}\|x_t-x_\star\|^2-\frac{1}{2}\left(f(x_t)-f(x_\star)\right) -\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i) +\frac{L_{\max}}{2Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_t-x_{t,m}^i\|^2,$$

where $h^\star=\nabla f(x_\star)$.

Proof. Using that $\mathbb{E}_{\mathcal{Q}}\left[\hat{g}_{t,m}\right]=g_{t,m}$ and definition of $h^\star$, we get

$$\begin{aligned}-\mathbb{E}_{\mathcal{Q}}\left[\frac{1}{M}\sum_{m=1}^{M}\langle \hat{g}_{t,m}-h^\star, x_t-x_\star\rangle\right] &= -\frac{1}{M}\sum_{m=1}^{M}\langle g_{t,m}-h^\star, x_t-x_\star\rangle\\ &= -\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\langle \nabla f_m^{\pi_m^i}(x_{t,m}^i)-\nabla f_m^{\pi_m^i}(x_\star), x_t-x_\star\right\rangle.\end{aligned}$$

Next, three-point identity and $L_{\max}$-smoothness of each function f_m^i imply

$$\begin{aligned}-\mathbb{E}_{\mathcal{Q}}\left[\frac{1}{M}\sum_{m=1}^{M}\langle \hat{g}_{t,m}-h^\star, x_t-x_\star\rangle\right] &= -\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left(D_{f_m^{\pi_m^i}}(x_t,x_\star)+D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i)-D_{f_m^{\pi_m^i}}(x_t,x_{t,m}^i)\right)\\ &\leq -D_f(x_t,x_\star)-\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i)\\ &\quad+\frac{L_{\max}}{2Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_t-x_{t,m}^i\|^2\end{aligned}$$

Finally, using μ-strong convexity of f, we finish the proof of lemma. □

Lemma G.2. *Under Assumptions 1, 2, 3, the iterates produced by* DIANA-NASTYA *satisfy*

$$\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t-h^\star\|^2\right] \leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+8L_{\max}\left(1+\frac{\omega}{M}\right)\left(f(x_t)-f(x_\star)\right) +\frac{4\omega}{M^2}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2.$$

Proof. Since $g_t=\frac{1}{M}\sum_{m=1}^{M}g_{t,m}$ and $\mathbb{E}\|\xi-c\|^2=\mathbb{E}\|\xi-\mathbb{E}\xi\|^2+\mathbb{E}\|\mathbb{E}\xi-c\|^2$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t-h^\star\|^2\right] &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\left(h_{t,m}+\mathcal{Q}\left(g_{t,m}-h_{t,m}\right)-h_m^\star\right)\right\|^2\right]\\ &= \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\left(h_{t,m}+\mathcal{Q}\left(g_{t,m}-h_{t,m}\right)\right)-g_t\right\|^2\right]+\|g_t-h^\star\|^2.\end{aligned}$$

Next, independence of $\mathcal{Q}\left(g_{t,m}-h_{t,m}\right)$, $m\in M$, Assumption 1, and $L_{\max}$-smoothness and convexity of each function f_m^i imply

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t-h^\star\|^2\right] &\leq \frac{\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{t,m}\|^2+\|g_t-h^\star\|^2\\
&\leq \frac{2\omega}{M^2}\sum_{m=1}^M\left\|\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)-\nabla f_m(x_t)\right\|^2+\frac{2\omega}{M^2}\sum_{m=1}^M\|\nabla f_m(x_t)-h_{t,m}\|^2\\
&\quad+2\|g_t-\nabla f(x_t)\|^2+2\|\nabla f(x_t)-h^\star\|^2\\
&\leq \frac{2\omega}{M^2}\sum_{m=1}^M\left\|\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)-\nabla f_m(x_t)\right\|^2+\frac{2\omega}{M^2}\sum_{m=1}^M\|\nabla f_m(x_t)-h_{t,m}\|^2\\
&\quad+2\left\|\frac{1}{M}\sum_{m=1}^M\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}(x_{t,m}^i)-\nabla f_m(x_t)\right)\right\|^2+2\|\nabla f(x_t)-h^\star\|^2\\
&\leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^M\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+\frac{2\omega}{M^2}\sum_{m=1}^M\|\nabla f_m(x_t)-h_{t,m}\|^2\\
&\quad+2\|\nabla f(x_t)-h^\star\|^2.
\end{aligned}$$

Using $L_{\max}$-smoothness and convexity of f_m, we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t-h^\star\|^2\right] &\leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^M\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+\frac{2\omega}{M^2}\sum_{m=1}^M\|\nabla f_m(x_t)-h_{t,m}\|^2\\
&\quad+4L_{\max}\left(f(x_t)-f(x_\star)\right)\\
&\leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^M\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+\frac{4\omega}{M^2}\sum_{m=1}^M\|\nabla f_m(x_t)-h_m^\star\|^2\\
&\quad+\frac{4\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_m^\star\|^2+4L_{\max}\left(f(x_t)-f(x_\star)\right)\\
&\leq \frac{2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^M\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+\frac{8L_{\max}\omega}{M^2}\sum_{m=1}^M D_{f_m}(x_t,x_\star)\\
&\quad+\frac{4\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_m^\star\|^2+4L_{\max}\left(f(x_t)-f(x_\star)\right).
\end{aligned}$$

□

Lemma G.3. *Under Assumptions 1, 2, 3, and $\alpha\leq\frac{1}{1+\omega}$, the iterates produced by* DIANA-NASTYA *satisfy*

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^M\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}-h_m^\star\|^2\right] &\leq \frac{1-\alpha}{M}\sum_{m=1}^M\|h_{t,m}-h_m^\star\|^2+\frac{2\alpha L_{\max}^2}{Mn}\sum_{m=1}^M\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2\\
&\quad+4\alpha L_{\max}\left(f(x_t)-f(x_\star)\right).
\end{aligned}$$

Proof. Taking expectation w.r.t. $\mathcal{Q}$ and using Assumption 1, we obtain

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}-h_m^\star\|^2\right] &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t,m}+\alpha\mathcal{Q}(g_{t,m}-h_{t,m})-h_m^\star\|^2\right]\\ &\le \frac{1}{M}\sum_{m=1}^{M}\left(\|h_{t,m}-h_m^\star\|^2+2\alpha\mathbb{E}_{\mathcal{Q}}\left[\langle\mathcal{Q}(g_{t,m}-h_{t,m}),h_{t,m}-h_m^\star\rangle\right]\right)\\ &\quad+\alpha^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|\mathcal{Q}(g_{t,m}-h_{t,m})\|^2\right]\\ &\le \frac{1}{M}\sum_{m=1}^{M}\left(\|h_{t,m}-h_m^\star\|^2+2\alpha\left\langle g_{t,m}-h_{t,m},h_{t,m}-h_m^\star\right\rangle\right)\\ &\quad+\alpha^2(1+\omega)\frac{1}{M}\sum_{m=1}^{M}\|g_{t,m}-h_{t,m}\|^2\end{aligned}$$

Using $\alpha\le\frac{1}{1+\omega}$, we get

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}-h_m^\star\|^2\right] &\le \frac{1}{M}\sum_{m=1}^{M}\left(\|h_{t,m}-h_m^\star\|^2+\alpha\left\langle g_{t,m}-h_{t,m},h_{t,m}+g_{t,m}-2h_m^\star\right\rangle\right)\\ &\le \frac{1}{M}\sum_{m=1}^{M}\left(\|h_{t,m}-h_m^\star\|^2+\alpha\|g_{t,m}-h_m^\star\|^2-\alpha\|h_{t,m}-h_m^\star\|^2\right)\\ &\le \frac{1-\alpha}{M}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2+\frac{\alpha}{M}\sum_{m=1}^{M}\|g_{t,m}-h_m^\star\|^2.\end{aligned}$$

Finally, $L_{\max}$-smoothness and convexity of f_m imply

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\left[\|h_{t+1,m}-h_m^\star\|^2\right] &\le \frac{1-\alpha}{M}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2\\ &\quad+\frac{2\alpha}{M}\sum_{m=1}^{M}\left(\|g_{t,m}-\nabla f_m(x_t)\|^2+\|\nabla f_m(x_t)-h_m^\star\|^2\right)\\ &\le \frac{1-\alpha}{M}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2+\frac{4L_{\max}\alpha}{M}\sum_{m=1}^{M}D_{f_m}(x_t,x_\star)\\ &\quad+\frac{2\alpha}{M}\sum_{m=1}^{M}\left\|\frac{1}{n}\sum_{i=0}^{n-1}(\nabla f_m^{\pi_m^i}(x_{t,m}^i)-\nabla f_m^i(x_t))\right\|^2\\ &\le \frac{1-\alpha}{M}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2+4L_{\max}\alpha\left(f(x_t)-f(x_\star)\right)\\ &\quad+\frac{2L_{\max}^2\alpha}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2.\end{aligned}$$

□

Theorem G.1. *Let Assumptions 1, 2, 3 hold and stepsizes γ, η, α satisfy*

$$0<\gamma\le\frac{1}{16L_{\max}n},\quad 0<\eta\le\min\left\{\frac{\alpha}{2\mu},\frac{1}{16L_{\max}\left(1+\frac{9\omega}{M}\right)}\right\},\quad \alpha\le\frac{1}{1+\omega}. \tag{29}$$

Then, for all $T\ge 0$ the iterates produced by DIANA-NASTYA *satisfy*

$$\mathbb{E}\left[\Psi_T\right]\le\left(1-\frac{\eta\mu}{2}\right)^T\Psi_0+\frac{9}{2}\frac{\gamma^2nL}{\mu}\left(\sigma_\star^2+(n+1)\zeta_\star^2\right). \tag{30}$$

Proof. We have

$$\begin{aligned}\|x_{t+1}-x_\star\|^2 &= \|x_t-\eta\hat{g}_t-x_\star+\eta h^\star\|^2\\ &= \|x_t-x_\star\|^2-2\eta\langle\hat{g}_t-h^\star,x_t-x_\star\rangle+\eta^2\|\hat{g}_t-h^\star\|^2.\end{aligned}$$

Taking expectation w.r.t. $\mathcal{Q}$ and using Lemma G.1, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &= \|x_t-x_\star\|^2-2\eta\mathbb{E}_{\mathcal{Q}}\left[\langle\hat{g}_t-h^\star,x_t-x_\star\rangle\right]+\eta^2\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t-h^\star\|^2\right]\\ &\le \left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2-\eta(f(x_t)-f(x_\star))-\frac{2\eta}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i)\\ &\quad+\frac{L_{\max}\eta}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+\eta^2\mathbb{E}_{\mathcal{Q}}\left[\|\hat{g}_t-h^\star\|^2\right].\end{aligned}$$

Next, Lemma G.2 implies

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|x_{t+1}-x_\star\|^2\right] &\le \left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2-\eta(f(x_t)-f(x_\star))-\frac{2\eta}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i)\\ &\quad+\frac{L_{\max}\eta}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+\frac{2\eta^2L_{\max}^2\left(1+\frac{\omega}{M}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2\\ &\quad+\eta^2\left(8L_{\max}\left(1+\frac{\omega}{M}\right)(f(x_t)-f(x_\star))+\frac{4\omega}{M^2}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2\right)\\ &\le \left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{M}\right)\right)(f(x_t)-f(x_\star))\\ &\quad+L_{\max}\eta\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)\right)\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2\\ &\quad-\frac{2\eta}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i)+\frac{4\eta^2\omega}{M^2}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2.\end{aligned}$$

Using (6) and Lemma G.3, we get

$$\begin{aligned}&\mathbb{E}_{\mathcal{Q}}\left[\Psi_{t+1}\right]\\ \le\ &\left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{M}\right)\right)(f(x_t)-f(x_\star))\\ &+L_{\max}\eta\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)\right)\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2\\ &-\frac{2\eta}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}D_{f_m^{\pi_m^i}}(x_\star,x_{t,m}^i)+\frac{4\eta^2\omega}{M^2}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2\\ &+c\eta^2\left(\frac{1-\alpha}{M}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2+\frac{2\alpha L_{\max}^2}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2+4\alpha L_{\max}(f(x_t)-f(x_\star))\right)\\ \le\ &\left(1-\frac{\eta\mu}{2}\right)\|x_t-x_\star\|^2+\eta^2\left(c(1-\alpha)+\frac{4\omega}{M}\right)\frac{1}{M}\sum_{m=1}^{M}\|h_{t,m}-h_m^\star\|^2\\ &-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{M}\right)-4\alpha\eta cL_{\max}\right)(f(x_t)-f(x_\star))\\ &+L\eta\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)+2\alpha\eta cL_{\max}\right)\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\|x_{t,m}^i-x_t\|^2.\end{aligned}$$

Taking the full expectation, we derive

$$\begin{aligned}\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\left[\|x_t-x_\star\|^2\right]+\eta^2\left(c(1-\alpha)+\frac{4\omega}{M}\right)\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|h_{t,m}-h_m^\star\|^2\right]\\ &\quad-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{M}\right)-4\alpha\eta cL\right)\mathbb{E}\left[f(x_t)-f(x_\star)\right]\\ &\quad+L_{\max}\eta\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)+2\alpha\eta cL_{\max}\right)\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_{t,m}^i-x_t\|^2\right].\end{aligned}$$

Using Lemma F.3, we get

$$\begin{aligned}\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\left[\|x_t-x_\star\|^2\right]+\eta^2\left(c(1-\alpha)+\frac{4\omega}{M}\right)\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|h_{t,m}-h_m^\star\|^2\right]\\ &\quad-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{M}\right)-4\alpha\eta cL_{\max}\right)\mathbb{E}\left[f(x_t)-f(x_\star)\right]\\ &\quad+8\gamma^2n^2L_{\max}^2\eta\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)+2\alpha\eta cL_{\max}\right)\mathbb{E}\left[f(x_t)-f(x_\star)\right]\\ &\quad+2\gamma^2nL_{\max}\eta\left(1+2\eta L_{\max}\left(1+\frac{\omega}{M}\right)+2\alpha\eta cL_{\max}\right)\left(\sigma_\star^2+(n+1)\zeta_\star^2\right).\end{aligned}$$

In view of (29), we have

$$\begin{aligned}\mathbb{E}\left[\Psi_{t+1}\right] &\leq \left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\left[\|x_t-x_\star\|^2\right]+\left(1-\frac{\alpha}{2}\right)\frac{c\eta^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|h_{t,m}-h_m^\star\|^2\right]\\ &\quad+\frac{9}{4}\gamma^2nL_{\max}\eta\left(\sigma_\star^2+(n+1)\zeta_\star^2\right)\end{aligned}$$

Using definition of Lyapunov function and using $\sum\limits_{t=0}^{+\infty}\left(1-\frac{\eta\mu}{2}\right)^t\leq\frac{2}{\mu\eta}$, we get the result. □

Corollary 12. *Let the assumptions of Theorem E.2 hold, $\gamma=\eta/n$, $\alpha=\frac{1}{1+\omega}$, and*

$$\eta=\min\left\{\frac{\alpha}{2\mu},\frac{1}{16L_{\max}\left(1+\frac{9\omega}{M}\right)},\sqrt{\frac{\varepsilon\mu n}{9L_{\max}}}\left((n+1)\zeta_\star^2+\sigma_\star^2\right)^{-1/2}\right\}. \tag{31}$$

Then, DIANA-NASTYA finds a solution with accuracy $\varepsilon>0$ after the following number of communication rounds:

$$\widetilde{\mathcal{O}}\left(\omega+\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right)+\sqrt{\frac{L_{\max}}{\varepsilon\mu^3}}\sqrt{\zeta_\star^2+\sigma_\star^2/n}\right).$$

If $\gamma\to0$, one can choose $\eta=\min\left\{\frac{\alpha}{2\mu},\frac{1}{16L_{\max}\left(1+\frac{9\omega}{M}\right)}\right\}$ such that the number of communication rounds T to find solution with accuracy $\varepsilon>0$ is

$$\widetilde{\mathcal{O}}\left(\omega+\frac{L_{\max}}{\mu}\left(1+\frac{\omega}{M}\right)\right).$$

Proof. Theorem E.2 implies

$$\mathbb{E}\left[\Psi_T\right]\leq\left(1-\frac{\eta\mu}{2}\right)^T\Psi_0+\frac{9}{2}\frac{\gamma^2nL_{\max}}{\mu}\left((n+1)\zeta_\star^2+\sigma_\star^2\right).$$

To estimate the number of communication rounds required to find a solution with accuracy $\varepsilon>0$, we need to upper bound each term from the right-hand side by $\frac{\varepsilon}{2}$. Thus, we get an additional restriction on η:

$$\frac{9}{2}\frac{\eta^2L_{\max}}{n\mu}\left((n+1)\zeta_\star^2+\sigma_\star^2\right)<\frac{\varepsilon}{2},$$

and also the upper bound on the number of communication rounds T

$$T=\widetilde{\mathcal{O}}\left(\frac{1}{\eta\mu}\right).$$

Substituting (31) in the previous equation, we get the first part of the result. When $\gamma\to0$, the proof follows similar steps. □

H Alternative Analysis of Q-NASTYA

In this analysis, we will use additional sequence:

$$x^i_{\star,m} = x_\star - \gamma \sum_{j=0}^{i-1} \nabla f_m(x_\star). \tag{32}$$

Theorem H.1. *Let Assumptions 1, 3, 4 hold. Moreover, we assume that* $(1-\gamma\mu)^n \leq \frac{9/10-1/C}{1+1/C} = \widehat{C} < 1$ *for some numerical constant* $C > 1$*. Also let* $\beta = \frac{\eta}{\gamma n} \leq \frac{1}{3C\frac{\omega}{M}+1}$ *and* $\gamma \leq \frac{1}{L_{\max}}$*. Then, for all* $T \geq 0$ *the iterates produced by Q-NASTYA satisfy*

$$\leq \max\left(1-\frac{\beta}{10}, 1-\frac{\alpha}{2}\right)\Psi_t + \frac{2}{\mu}\beta\gamma^2\hat{\sigma}^2_{rad} \tag{33}$$

$$\mathbb{E}\left[\|x_T - x_\star\|^2\right] \leq \left(1-\frac{\beta}{10}\right)\|x_t - x_*\|^2 + \frac{4}{\mu}\beta\gamma^2\hat{\sigma}^2_{\text{rad}} + 3\beta^2\frac{\omega}{M}\frac{1}{M}\hat{\Delta}_\star,$$

where $\hat{\Delta}_\star = \frac{1}{M}\sum_{m=1}^M \|x^n_{\star,m} - x_\star\|^2$ and $\hat{\sigma}^2_{\text{rad}} \leq L_{\max}\left(\zeta^2_\star + {}^{n\sigma^2_\star}/_4\right)$.

Proof. The update rule for one epoch can be rewritten as

$$x_{t+1} = x_t - \eta\frac{1}{M}\sum_{M}^{m=1} Q\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right).$$

Using this, we derive

$$\begin{aligned}
\|x_{t+1} - x_*\|^2 &= \left\|x_t - \eta\frac{1}{M}\sum_{m=1}^M Q\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right) - x_*\right\|^2 \\
&= \|x_t - x_*\|^2 - 2\eta\left\langle x_t - x_*, \frac{1}{M}\sum_{m=1}^M Q\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right)\right\rangle \\
&+ \eta^2\left\|\frac{1}{M}\sum_{m=1}^M Q\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right)\right\|^2.
\end{aligned}$$

Taking conditional expectation w.r.t. the randomness comming from compression, we get

$$\begin{aligned}
\mathbb{E}_Q\|x_{t+1} - x_*\|^2 &= \|x_t - x_*\|^2 - 2\eta\left\langle x_t - x_*, \frac{1}{M}\sum_{m=1}^M\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right)\right\rangle \\
&+ \eta^2\mathbb{E}_Q\left\|\frac{1}{M}\sum_{m=1}^M Q\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right)\right\|^2.
\end{aligned}$$

Next, we use the definition of quantization operator and independence of $Q\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right)$, $m \in [M]$:

$$\begin{aligned}
\mathbb{E}_Q\|x_{t+1} - x_*\|^2 &\leq \|x_t - x_*\|^2 - 2\eta\left\langle x_t - x_*, \frac{1}{M}\sum_{m=1}^M\left(\frac{x_t - x^n_{t,m}}{\gamma n}\right)\right\rangle \\
&+ \eta^2\left(\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^M\left\|\frac{x_t - x^n_{t,m}}{\gamma n}\right\|^2 + \left\|\frac{1}{M}\sum_{m=1}^M\frac{x_t - x^n_{t,m}}{\gamma n}\right\|^2\right).
\end{aligned}$$

Since $\beta = \frac{\eta}{\gamma n}$, we obtain

$$\begin{aligned}
\mathbb{E}_Q\|x_{t+1} - x_*\|^2 &\leq \|x_t - x_*\|^2 - 2\beta \left\langle x_t - x_*, \frac{1}{M}\sum_{m=1}^{M}\left(x_t - x_{t,m}^n\right)\right\rangle \\
&+ \beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|x_t - x_{t,m}^n\right\|^2 + \beta^2\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_t - x_{t,m}^n\right)\right\|^2 \\
&= \|x_t - x_*\|^2 + 2\beta \left\langle x_t - x_*, \frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n - x_t\right)\right\rangle \\
&+ \beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|x_t - x_{t,m}^n\right\|^2 + \beta^2\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n - x_t\right)\right\|^2 \\
&= \left\|x_t - x_* + \beta\left(\frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n - x_t\right)\right)\right\|^2 + \beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|x_t - x_{t,m}^n\right\|^2 \\
&= \left\|(1-\beta)(x_t - x_*) + \beta\left(\frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n\right) - x_*\right)\right\|^2 \\
&+ \beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|x_t - x_{t,m}^n\right\|^2.
\end{aligned}$$

Using the condition that $x_* = \frac{1}{M}\sum_{m=1}^{M} x_{*,m}^n$ we have:

$$\mathbb{E}_Q\|x_{t+1} - x_*\|^2 \leq \left\|(1-\beta)(x_t - x_*) + \beta\left(\frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n - x_{*,m}^n\right)\right)\right\|^2 + \beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|x_t - x_{t,m}^n\right\|^2.$$

Convexity of squared norm and Jensen's inequality imply

$$\mathbb{E}_Q\|x_{t+1} - x_*\|^2 \leq (1-\beta)\|x_t - x_*\|^2 + \beta\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n - x_{*,m}^n\right)\right\|^2 + \beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|x_t - x_{t,m}^n\right\|^2.$$

Next, from Young's inequality we get

$$\begin{aligned}
\mathbb{E}_Q\|x_{t+1} - x_*\|^2 &\leq (1-\beta)\|x_t - x_*\|^2 + \beta\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t,m}^n - x_{*,m}^n\right)\right\|^2 + 3\beta^2\frac{\omega}{M}\|x_t - x_*\|^2 \\
&+ 3\beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\|x_{t,m}^n - x_{*,m}^n\|^2 + 3\beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\|x_{*,m}^n - x_*\|^2.
\end{aligned}$$

Theorem 4 from [Mishchenko et al., 2021] gives

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\|x_{t,m}^n - x_{*,m}^n\|^2\right] &\leq (1-\gamma\mu)^n\left[\|x_t - x_\star\|^2\right] + 2\gamma^3\hat{\sigma}_{\text{rad}}^2\left(\sum_{j=0}^{n-1}(1-\gamma\mu)^j\right) \\
&= (1-\gamma\mu)^n\left[\|x_t - x_\star\|^2\right] + 2\gamma^2\hat{\sigma}_{\text{rad}}^2\frac{1}{\gamma\mu}.
\end{aligned}$$

It leads to

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_*\|^2 &\leq (1-\beta)\|x_t - x_*\|^2 + \beta\left((1-\gamma\mu)^n\left[\|x_t - x_\star\|^2\right] + 2\gamma^3\hat{\sigma}_{\mathrm{rad}}^2\frac{1}{\gamma\mu}\right) \\
&\quad + 3\beta^2\frac{\omega}{M}\|x_t - x_*\|^2 + 3\beta^2\frac{\omega}{M}\left((1-\gamma\mu)^n\left[\|x_t - x_\star\|^2\right] + 2\gamma^3\hat{\sigma}_{\mathrm{rad}}^2\frac{1}{\gamma\mu}\right) \\
&\quad + 3\beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\|x_{*,m}^n - x_*\|^2 \\
&\leq \left(1 - \beta + \beta(1-\gamma\mu)^n + 3\beta^2\frac{\omega}{M} + 3\beta^2\frac{\omega}{M}(1-\gamma\mu)^n\right)\|x_t - x_*\|^2 \\
&\quad + 2\beta\gamma^3\hat{\sigma}_{\mathrm{rad}}^2\frac{1}{\gamma\mu}\left(1 + 3\beta\frac{\omega}{M}\right) + 3\beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\|x_{*,m}^n - x_*\|^2.
\end{aligned}$$

Using $(1-\gamma\mu)^n \leq \frac{9/10 - 1/C}{1 + 1/C}$, we have

$$\begin{aligned}
&(1-\gamma\mu)^n \leq \frac{9/10 - 1/C}{1 + 1/C} \\
&(1-\gamma\mu)^n\left(1 + \frac{1}{C}\right) \leq \frac{9}{10} - \frac{1}{C} \\
&-\frac{9}{10}\beta + \beta(1-\gamma\mu)^n + \frac{\beta}{C} + \frac{\beta}{C}(1-\gamma\mu)^n \leq 0 \\
&1 - \beta + \beta(1-\gamma\mu)^n + \frac{\beta}{C} + \frac{\beta}{C}(1-\gamma\mu)^n \leq 1 - \frac{\beta}{10}.
\end{aligned}$$

Next, applying $\beta \leq \frac{1}{1+3C\frac{\omega}{M}}$, we derive

$$1 - \beta + \beta(1-\gamma\mu)^n + 3\beta^2\frac{\omega}{M} + 3\beta^2\frac{\omega}{M}(1-\gamma\mu)^n \leq 1 - \frac{\beta}{10}.$$

Finally, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_\star\|^2 &\leq \left(1 - \frac{\beta}{10}\right)\|x_t - x_*\|^2 + 2\beta\gamma^2\hat{\sigma}_{\mathrm{rad}}^2\frac{1}{\mu}\left(1 + \frac{1}{C}\right) \\
&\quad + 3\beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\|x_{*,m}^n - x_*\|^2 \\
&\leq \left(1 - \frac{\beta}{10}\right)\|x_t - x_*\|^2 + \frac{4}{\mu}\beta\gamma^2\hat{\sigma}_{\mathrm{rad}}^2 \\
&\quad + 3\beta^2\frac{\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\|x_{*,m}^n - x_*\|^2.
\end{aligned}$$

□

I Alternative Analysis of DIANA-NASTYA

Theorem I.1. *Let Assumptions 1, 3, 4 hold. Moreover, we assume that* $(1-\gamma\mu)^n \leq \frac{9/10-1/B}{1+1/B} = \widehat{B} < 1$ *for some numerical constant* $B > 1$. *Also let* $\beta = \frac{\eta}{\gamma n} \leq \frac{1}{12B\frac{\omega}{M}+1}$ *and* $\gamma \leq \frac{1}{L_{\max}}$ *and also* $\alpha \leq \frac{1}{\omega+1}$. *Then, for all* $T \geq 0$ *the iterates produced by* DIANA-NASTYA *satisfy*

$$\mathbb{E}\Psi_T \leq \max\left(1-\frac{\beta}{10}, 1-\frac{\alpha}{2}\right)^T \Psi_0 + \frac{2}{\mu\min(\frac{\beta}{10},\frac{\alpha}{2})}\beta\gamma^2\hat{\sigma}_{rad}^2. \tag{34}$$

Proof. We start with expanding the square:

$$\begin{aligned}
\|x_{t+1}-x_*\|^2 &= \|x_t-\eta\hat{g}_t-x_*\|^2 \\
&= \left\|x_t-\eta\frac{1}{M}\sum_{m=1}^M\left(h_{t,m}+Q(g_{t,m}-h_{t,m})\right)-x_*\right\|^2 \\
&= \|x_t-x_*\|^2 - 2\eta\left\langle\frac{1}{M}\sum_{m=1}^M\left(h_{t,m}+Q(g_{t,m}-h_{t,m})\right), x_t-x_*\right\rangle \\
&\quad + \eta^2\left\|\frac{1}{M}\sum_{m=1}^M\left(h_{t,m}+Q(g_{t,m}-h_{t,m})\right)\right\|^2.
\end{aligned}$$

Taking the expectation w.r.t. $\mathcal{Q}$, we get

$$\begin{aligned}
\mathbb{E}_Q\|x_{t+1}-x_*\|^2 &= \|x_t-x_*\|^2 - 2\eta\left\langle\frac{1}{M}\sum_{m=1}^M g_{t,m}, x_t-x_*\right\rangle \\
&\quad + \eta^2\mathbb{E}_Q\left\|\frac{1}{M}\sum_{m=1}^M\left(h_{t,m}+Q(g_{t,m}-h_{t,m})\right)\right\|^2 \\
&= \|x_t-x_*\|^2 - 2\eta\left\langle\frac{1}{M}\sum_{m=1}^M g_{t,m}, x_t-x_*\right\rangle \\
&\quad + \eta^2\mathbb{E}_Q\left\|\frac{1}{M}\sum_{m=1}^M\left(h_{t,m}+Q(g_{t,m}-h_{t,m})-g_{t,m}\right)\right\|^2 + \eta^2\left\|\frac{1}{M}\sum_{m=1}^M g_{t,m}\right\|^2 \\
&\leq \|x_t-x_*\|^2 - 2\eta\left\langle\frac{1}{M}\sum_{m=1}^M g_{t,m}, x_t-x_*\right\rangle \\
&\quad + \eta^2\frac{\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{t,m}\|^2 + \eta^2\left\|\frac{1}{M}\sum_{m=1}^M g_{t,m}\right\|^2 \\
&\leq \|x_t-x_*\|^2 - 2\eta\left\langle\frac{1}{M}\sum_{m=1}^M g_{t,m}, x_t-x_*\right\rangle \\
&\quad + \eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{*,m}\|^2 + \eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_{*,m}\|^2 + \eta^2\left\|\frac{1}{M}\sum_{m=1}^M g_{t,m}\right\|^2.
\end{aligned}$$

Next, using definition of $g_{t,m}$, we obtain

$$\begin{aligned}
\mathbb{E}\|x_{t+1}-x_*\|^2 &\le \|x_t-x_*\|^2 - 2\eta\left\langle \frac{1}{M}\sum_{m=1}^M \frac{x_t-x_{t,m}^n}{\gamma n}, x_t-x_*\right\rangle + \eta^2\left\|\frac{1}{M}\sum_{m=1}^M \frac{x_t-x_{t,m}^n}{\gamma n}\right\|^2\\
&+\eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{*,m}\|^2 + \eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_{*,m}\|^2\\
&= \|x_t-x_*\|^2 + 2\alpha\left\langle \frac{1}{M}\sum_{m=1}^M \left(x_{t,m}^n-x_t\right), x_t-x_*\right\rangle + \alpha^2\left\|\frac{1}{M}\sum_{m=1}^M \left(x_{t,m}^n-x_t\right)\right\|^2\\
&+\eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{*,m}\|^2 + \eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_{*,m}\|^2\\
&= \left\|x_t-x_*+\alpha\frac{1}{M}\sum_{m=1}^M \left(x_{t,m}^n-x_t\right)\right\|^2\\
&+\eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{*,m}\|^2 + \eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_{*,m}\|^2\\
&= \left\|(1-\beta)(x_t-x_*)+\beta\left(\frac{1}{M}\sum_{m=1}^M \left(x_{t,m}^n-x_{*,m}^n\right)\right)\right\|^2\\
&\le (1-\beta)\|x_t-x_*\|^2 + \beta\frac{1}{M}\sum_{m=1}^M\|x_{t,m}^n-x_{*,m}^n\|^2\\
&+\eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|g_{t,m}-h_{*,m}\|^2 + \eta^2\frac{2\omega}{M^2}\sum_{m=1}^M\|h_{t,m}-h_{*,m}\|^2.
\end{aligned}$$

Let us consider recursion for control variable:

$$\begin{aligned}
\|h_{t+1,m}-h_{*,m}\|^2 &= \|h_{t,m}+\alpha Q(g_{t,m}-h_{t,m})-h_{*,m}\|^2\\
&= \|h_{t,m}-h_{*,m}\|^2 + \alpha\left\langle Q(g_{t,m}-h_{t,m}), h_{t,m}-h_{*,m}\right\rangle + \alpha^2\|Q(g_{t,m}-h_{t,m})\|^2.
\end{aligned}$$

Taking the expectation w.r.t. $\mathcal{Q}$, we have

$$\mathbb{E}_{\mathcal{Q}}\|h_{t+1,m}-h_{*,m}\|^2 \le \|h_{t,m}-h_{*,m}\|^2 + 2\alpha\left\langle g_{t,m}-h_{t,m}, h_{t,m}-h_{*,m}\right\rangle + \alpha^2(\omega+1)\|g_{t,m}-h_{t,m}\|^2.$$

Using $\alpha \le \frac{1}{\omega+1}$ we have

$$\begin{aligned}
\mathbb{E}\|h_{t+1,m}-h_{*,m}\|^2 &\le \|h_{t,m}-h_{*,m}\|^2\\
&+2\alpha\left\langle g_{t,m}-h_{t,m}, h_{t,m}-h_{*,m}\right\rangle + \alpha\|g_{t,m}-h_{t,m}\|^2\\
&= \|h_{t,m}-h_{*,m}\|^2\\
&+2\alpha\left\langle g_{t,m}-h_{t,m}, h_{t,m}-h_{*,m}\right\rangle + \alpha\left\langle g_{t,m}-h_{t,m}, g_{t,m}-h_{t,m}\right\rangle\\
&= \|h_{t,m}-h_{*,m}\|^2\\
&+\alpha\left\langle g_{t,m}-h_{t,m}, g_{t,m}-h_{t,m}+2h_{t,m}-2h_{*,m}\right\rangle\\
&= \|h_{t,m}-h_{*,m}\|^2\\
&+\alpha\left\langle g_{t,m}-h_{t,m}, g_{t,m}+h_{t,m}-2h_{*,m}\right\rangle\\
&= \|h_{t,m}-h_{*,m}\|^2\\
&+\alpha\left\langle g_{t,m}-h_{t,m}-h_{*,m}+h_{*,m}, g_{t,m}+h_{t,m}-2h_{*,m}\right\rangle\\
&= \|h_{t,m}-h_{*,m}\|^2\\
&+\alpha\left\langle g_{t,m}-h_{*,m}-(h_{t,m}-h_{*,m}), (g_{t,m}-h_{*,m})+(h_{t,m}-h_{*,m})\right\rangle\\
&= \|h_{t,m}-h_{*,m}\|^2 + \alpha\|g_{t,m}-h_{*,m}\|^2 - \alpha\|h_{t,m}-h_{*,m}\|^2\\
&= (1-\alpha)\|h_{t,m}-h_{*,m}\|^2 + \alpha\|g_{t,m}-h_{*,m}\|^2.
\end{aligned}$$

Using this bound we get that

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{Q}}\|h_{t+1,m}-h_{*,m}\|^2 \leq (1-\alpha)\frac{1}{M}\sum_{m=1}^{M}\|h_{t,m}-h_{*,m}\|^2+\alpha\frac{1}{M}\sum_{m=1}^{M}\|g_{t,m}-h_{*,m}\|^2.$$

Let us consider Lyapunov function:

$$\Psi_t = \|x_t-x_*\|^2+\frac{4\omega\eta^2}{\alpha M}\frac{1}{M}\sum_{m=1}^{M}\|h_{t,m}-h_{*,m}\|^2.$$

Using previous bounds and Theorem 4 from [Mishchenko et al., 2021] we have

$$\begin{aligned}
\mathbb{E}\Psi_{t+1} &\leq (1-\beta)\|x_t-x_*\|^2+\beta\left((1-\gamma\mu)^n\mathbb{E}\|x_t-x_*\|^2+\gamma^3\frac{1}{\gamma\mu}\hat{\sigma}_{rad}^2\right)\\
&\quad+\eta^2\frac{2\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|g_{t,m}-h_{*,m}\|^2+\eta^2\frac{2\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|h_{t,m}-h_{*,m}\|^2\\
&\quad+(1-\alpha)\frac{4\omega\eta^2}{\alpha M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|h_{t,m}-h_{*,m}\|^2+\alpha\frac{4\omega\eta^2}{\alpha M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|g_{t,m}-h_{*,m}\|^2\\
&\leq \left(1-\frac{\alpha}{2}\right)\frac{4\omega\eta^2}{\alpha M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|h_{t,m}-h_{*,m}\|^2+\eta^2\frac{6\omega}{M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|g_{t,m}-h_{*,m}\|^2\\
&\quad+(1-\beta)\mathbb{E}\|x_t-x_*\|^2+\beta\left((1-\gamma\mu)^n\mathbb{E}\|x_t-x_*\|^2+\gamma^3\frac{1}{\gamma\mu}\hat{\sigma}_{rad}^2.\right)
\end{aligned}$$

Let us consider

$$\begin{aligned}
\eta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|g_{t,m}-h_{*,m}\|^2 &= \eta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\frac{x_t-x_{t,m}^n}{\gamma n}-\frac{x_*-x_{*,m}^n}{\gamma n}\right\|^2\\
&\leq 2\eta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\frac{x_t-x_*}{\gamma n}\right\|^2+2\eta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\frac{x_{t,m}^n-x_{*,m}^n}{\gamma n}\right\|^2\\
&\leq 2\beta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|x_t-x_*\right\|^2+2\beta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|x_{t,m}^n-x_{*,m}^n\right\|^2\\
&\leq 2\beta^2\mathbb{E}\left\|x_t-x_*\right\|^2+2\beta^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|x_{t,m}^n-x_{*,m}^n\right\|^2.
\end{aligned}$$

Putting all the terms together and using $(1-\gamma\mu)^n \leq \frac{9/10-1/B}{1+1/B} = \widehat{B} < 1$, $\beta \leq \frac{1}{12B\frac{\omega}{M}+1}$ we have

$$\begin{aligned}
\mathbb{E}\Psi_{t+1} &\leq \left(1-\beta+12\frac{\omega}{M}\beta^2+12\frac{\omega}{M}\beta^2(1-\gamma\mu)^n+\beta(1-\gamma\mu)^n\right)\mathbb{E}\|x_t-x_*\|^2+\beta\gamma^3\frac{1}{\gamma\mu}\hat{\sigma}_{rad}^2\\
&\quad+2\beta^2\frac{6\omega}{M}\gamma^3\frac{1}{\gamma\mu}\hat{\sigma}_{rad}^2+\left(1-\frac{\alpha}{2}\right)\frac{4\omega\eta^2}{\alpha M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|h_{t,m}-h_{*,m}\|^2\\
&\leq \left(1-\frac{\beta}{10}\right)\mathbb{E}\|x_t-x_\star\|^2+\frac{2}{\mu}\beta\gamma^2\hat{\sigma}_{\text{rad}}^2+\left(1-\frac{\alpha}{2}\right)\frac{4\omega\eta^2}{\alpha M}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|h_{t,m}-h_{*,m}\|^2\\
&\leq \max\left(1-\frac{\beta}{10},1-\frac{\alpha}{2}\right)\Psi_t+\frac{2}{\mu}\beta\gamma^2\hat{\sigma}_{\text{rad}}^2.
\end{aligned}$$

Unrolling this recursion we get the final result. □

Algorithm 5 Q-NASTYA-PP

Input: x_0 – starting point, $\gamma > 0$ – local stepsize, $\eta > 0$ – global stepsize

for $t = 0, 1, \ldots, T-1$ **do**
  Sample a cohort S_t with cardinality C uniformly
  for $m \in S_t$ in parallel **do**
    Receive x_t from the server and set $x_{t,m}^0 = x_t$
    Sample random permutation of $[n]$: $\pi_m = (\pi_m^0, \ldots, \pi_m^{n-1})$
    for $i = 0, 1, \ldots, n-1$ **do**
      Set $x_{t,m}^{i+1} = x_{t,m}^i - \gamma \nabla f_m^{\pi_m^i}(x_{t,m}^i)$
    end for
    Compute $g_{t,m} = \frac{1}{\gamma n}\left(x_t - x_{t,m}^n\right)$ and send $\mathcal{Q}_t(g_{t,m})$ to the server
  end for
  Compute $g_t = \frac{1}{C}\sum_{m \in S_t} \mathcal{Q}_t(g_{t,m})$
  Compute $x_{t+1} = x_t - \eta g_t$ and send x_{t+1} to the workers
end for

Output: x_T

J Partial Participation for Method with Local Steps

J.1 Analysis of Q-NASTYA with Partial Participation

Lemma J.1. *Let Assumptions 1, 2, 3 hold. Then, for all $t \geq 0$ the iterates produced by Q-NASTYA-PP (Algorithm 5) satisfy*

$$\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|g_t\right\|^2\right] \leq \frac{2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i - x_t\right\|^2 + 8L_{\max}\left(1+\frac{\omega}{C}\right)\left(f\left(x_t\right) - f\left(x^\star\right)\right) + 4\left(\frac{\omega}{C} + \frac{M-C}{C\max M-1,1}\right)\sigma_\star^2,$$

where $\mathbb{E}_{\mathcal{Q},S_t}$ is expectation w.r.t. $\mathcal{Q}, S_t$ and $\sigma_\star^2 = \frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x^\star\right)\right\|^2$.

Proof. $\mathbb{E}\left[\|\xi\|^2\right] = \mathbb{E}\left[\|\xi - \mathbb{E}[\xi]\|^2\right] + \|\mathbb{E}\xi\|^2$, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Q}}\left[\left\|g_t\right\|^2\right]\\
=&\mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{C}\sum_{m\in S_t}\left(\mathcal{Q}\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right) + \frac{1}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2\right]\\
=&\frac{1}{C^2}\mathbb{E}_{\mathcal{Q}}\|\sum_{m\in S_t}(\underbrace{\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)}_{=\xi_m})\|^2]\\
&+\left\|\frac{1}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2\\
=&\frac{1}{C^2}\mathbb{E}_{\mathcal{Q}}\left[\sum_{m\in S_t}\left\|\xi_m\right\|^2 + \sum_{m,l\in S_t:m\neq l}2\left\langle\xi_m,\xi_l\right\rangle\right] + \left\|\frac{1}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2.
\end{aligned}$$

Using independence between ξ_m and ξ_l for different m, l and Using (2), (3), we get

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] =& \frac{1}{C^2}\sum_{m\in S_t}\mathbb{E}_{\mathcal{Q}}\left[\left\|\mathcal{Q}\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right)-\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2\right]\\ &+\left\|\frac{1}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2\\ \leq& \frac{\omega}{C^2}\sum_{m\in S_t}\left\|\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2+\left\|\frac{1}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right\|^2.\end{aligned}$$

Rewriting previous inequality and using $\nabla f_m(x)=\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_t\right)$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] \leq& \frac{2\omega}{C^2}\sum_{m\in S_t}\left\|\frac{1}{n}\sum_{i=0}^{n-1}\left(\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)-\nabla f_m^{\pi_m^i}\left(x_t\right)\right)\right\|^2+\frac{2\omega}{C^2}\sum_{m\in S_t}\left\|\nabla f_m\left(x_t\right)\right\|^2\\ &+2\left\|\frac{1}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\left(\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)-\nabla f_m^{\pi_m^i}\left(x_t\right)\right)\right\|^2+2\left\|\frac{1}{C}\sum_{m\in S_t}\nabla f_m\left(x_t\right)\right\|^2\\ \leq& \frac{2\left(1+\frac{\omega}{C}\right)}{C}\sum_{m\in S_t}\left\|\frac{1}{n}\sum_{i=0}^{n-1}\left(\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)-\nabla f_m^{\pi_m^i}\left(x_t\right)\right)\right\|^2\\ &+\frac{2\omega}{C^2}\sum_{m\in S_t}\left\|\nabla f_m\left(x_t\right)\right\|^2+2\left\|\frac{1}{C}\sum_{m\in S_t}\nabla f_m\left(x_t\right)\right\|^2\end{aligned}$$

Using L-smoothness of f_m^i and f and also convexity of f_m, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}\left[\|g_t\|^2\right] \leq& \frac{2\left(1+\frac{\omega}{C}\right)}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\left\|\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)-\nabla f_m^{\pi_m^i}\left(x_t\right)\right\|^2\\ &+\frac{4\omega}{C^2}\sum_{m\in S_t}\left\|\nabla f_m\left(x_t\right)-\nabla f_m\left(x^\star\right)\right\|^2\\ &+\frac{4\omega}{C^2}\sum_{m\in S_t}\left\|\nabla f_m\left(x^\star\right)\right\|^2+4\left\|\frac{1}{C}\sum_{m\in S_t}\left(\nabla f_m\left(x_t\right)-\nabla f_m\left(x^\star\right)\right)\right\|^2\\ &+4\left\|\frac{1}{C}\sum_{m\in S_t}\nabla f_m\left(x^\star\right)\right\|^2\\ \leq& \frac{2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{Cn}\sum_{m\in S_t}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+\frac{8L_{\max}\left(1+\frac{\omega}{C}\right)}{C}\sum_{m\in S_t}D_{f_m}\left(x_t,x^\star\right)\\ &+\frac{4\omega}{C^2}\sum_{m\in S_t}\left\|\nabla f_m\left(x^\star\right)\right\|^2+4\left\|\frac{1}{C}\sum_{m\in S_t}\nabla f_m\left(x^\star\right)\right\|^2.\end{aligned}$$

Taking expectation w.r.t. S_t and using uniform sampling, we receive

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|g_t\right\|^2\right] \leq & \frac{2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{n}\mathbb{E}_{S_t}\left[\frac{1}{C}\sum_{m\in S_t}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right|^2\right] \\
& + 8L_{\max}\left(1+\frac{\omega}{C}\right)\mathbb{E}_{S_t}\left[\frac{1}{C}\sum_{m\in S_t}D_{f_m}\left(x_t,x^\star\right)\right] \\
& + \frac{4\omega}{C}\mathbb{E}_{S_t}\left[\frac{1}{C}\sum_{m\in S_t}\left\|\nabla f_m\left(x^\star\right)\right\|^2\right] + 4\mathbb{E}_{S_t}\left[\left\|\frac{1}{C}\sum_{m\in S_t}\nabla f_m\left(x^\star\right)\right\|^2\right] \\
\leq & \frac{2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2 + \frac{8L_{\max}\left(1+\frac{\omega}{C}\right)}{M}\sum_{m=1}^{M}D_{f_m}\left(x_t,x^\star\right) \\
& + \frac{4\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x^\star\right)\right\|^2 + 4\frac{M-C}{MC\max M-1,1}\sum_{m=1}^{M}\left\|\nabla f_m\left(x^\star\right)\right\|^2.
\end{aligned}
$$

□

Theorem J.1. *Let step sizes η,γ satisfy the following equations*

$$
\eta = \frac{1}{16L_{\max}\left(1+\frac{\omega}{C}\right)}, \quad \gamma = \frac{1}{5nL_{\max}}
$$

Under the Assumptions 1, 2, 3 iterates of Q-NASTYA-PP *(Algorithm 5) satisfy*

$$
\begin{aligned}
\mathbb{E}\left[\left\|x_T-x^\star\right\|^2\right] \leq & \left(1-\frac{\eta\mu}{2}\right)^T\left\|x_0-x^\star\right\|^2 + \frac{9}{2}\frac{\gamma^2 nL_{\max}}{\mu}\left(\frac{1}{M}\sum_{m=1}^{M}\sigma_{\star,m}^2+n\sigma_\star^2\right) \\
& + 8\frac{\eta}{\mu}\left(\frac{\omega}{C}\sigma_\star^2+\frac{M-C}{C\max(M-1,1)}\sigma_\star^2\right),
\end{aligned}
$$

where

$$
\sigma_\star^2 = \frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x^\star\right)\right\|^2, \quad \sigma_{\star,m}^2 = \frac{1}{n}\left\|\nabla f_m^i\left(x^\star\right)\right\|^2
$$

As we can see, there is an additional error term proportional to $\frac{M-C}{C\max(M-1,1)}$ that arises due to client sampling in the partial participation setting. Note that when $C=M$ (all clients are participating), this error term vanishes, allowing us to recover the previous result for the full participation case. This shows the consistency of our theoretical framework across different participation scenarios.

Proof.

Taking expectation w.r.t. $\mathcal{Q}, S_t$ and using Lemma J.1 updated, we get

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|x_{t+1}-x^\star\right|^2\right] \\
=&\left\|x_t-x^\star\right|^2-2\eta\mathbb{E}_{\mathcal{Q},S_t}\left[\langle g_t,x_t-x^\star\rangle\right]+\eta^2\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|g^t\right\|^2\right] \\
\leq&\left\|x_t-x^\star\right\|^2-2\eta\mathbb{E}_{\mathcal{Q},S_t}\left[\left\langle\frac{1}{C}\sum_{m\in S_t}\mathcal{Q}\left(\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_m^{\pi_m^i}\left(x_{t,m}^i\right)\right),x_t-x^\star\right\rangle\right] \\
&+\frac{2\eta^2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+8\eta^2L_{\max}\left(1+\frac{\omega}{C}\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&+4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2 \\
\leq&\left\|x_t-x^\star\right\|^2-2\eta\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\langle\nabla f_m^{\pi i}\left(x_{t,m}^i\right),x_t-x^\star\right\rangle \\
&+\frac{2\eta^2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+8\eta^2L_{\max}\left(1+\frac{\omega}{C}\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&+4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2.
\end{aligned}
$$

Using Lemma F.2, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|x_{t+1}-x^\star\right\|^2\right] \\
\leq&\left\|x_t-x^\star\right\|^2-\frac{\eta\mu}{2}\left\|x_t-x^\star\right\|^2-\eta\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&+8\eta^2L_{\max}\left(1+\frac{\omega}{C}\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right)+\frac{\eta L_{\max}}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2 \\
&+\frac{2\eta^2L_{\max}^2\left(1+\frac{\omega}{C}\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2 \\
\leq&\left(1-\frac{\eta\mu}{2}\right)\left\|x_t-x^\star\right\|^2-\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{C}\right)\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&+\frac{\eta L_{\max}\left(1+2\eta L_{\max}\left(1+\frac{\omega}{C}\right)\right)}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+ \\
&4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2.
\end{aligned}
$$

Using Lemma F.3, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|x_{t+1}-x^\star\right\|^2\right]\leq&\left(1-\frac{\eta\mu}{2}\right)\left\|x_t-x^\star\right\|^2-\eta\left(1-8\eta L\left(1+\frac{\omega}{C}\right)\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&+\eta L_{\max}\left(1+2\eta L_{\max}\left(1+\frac{\omega}{C}\right)\right)\cdot 8\gamma^2n^2L_{\max}\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&+\eta L_{\max}\left(1+2\eta L_{\max}\left(1+\frac{\omega}{C}\right)\right)\cdot 2\gamma^2n\left(\frac{1}{M}\sum_{m=1}^{M}\sigma_{\star,m}^2+n\sigma_\star^2\right) \\
&+4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2.
\end{aligned}
$$

Finally, we receive

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|x_{t+1}-x^\star\right\|^2\right] \\
&\leq \left(1-\frac{\eta\mu}{2}\right)\left\|x_t-x^\star\right\|^2 + 4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2 \\
&\quad -\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{C}\right)-8\gamma^2n^2L_{\max}^2\left(1+2L_{\max}\eta\left(1+\frac{\omega}{C}\right)\right)\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
&\quad +2\gamma^2n\eta L_{\max}\left(1+2\eta L\left(1+\frac{\omega}{C}\right)\right)\left(\frac{1}{M}\sum_{m=1}^M\sigma_{\star,m}^2+n\sigma_\star^2\right) \\
&\leq \left(1-\frac{\eta\mu}{2}\right)\left\|x_t-x^\star\right\|^2 + 4\eta^2\left(\frac{\omega}{C}+\frac{M-C}{C\max M-1,1}\right)\sigma_\star^2 \\
&\quad +\frac{9}{4}\eta L_{\max}\gamma^2n\left(\frac{1}{M}\sum_{m=1}^M\sigma_{\star,m}^2+n\sigma_\star^2\right)
\end{aligned}
$$

Recursively rewriting the inequality and using $\sum_{t=0}^{+\infty}\left(1-\frac{\eta\mu}{2}\right)^t \leq \frac{2}{\mu\eta}$, we finish proof.

□

Algorithm 6 DIANA-NASTYA-PP

Input: x_0 – starting point, $\{h_{0,m}\}_{m=1}^M$ – initial shift-vectors, $\gamma > 0$ – local stepsize, $\eta > 0$ – global stepsize, $\alpha > 0$ – stepsize for learning the shifts

for $t = 0, 1, \ldots, T-1$ **do**
  Sample a cohort S_t with cardinality C uniformly
  for $m \in S_t$ in parallel **do**
    Receive x_t from the server and set $x_{t,m}^0 = x_t$
    Sample random permutation of $[n]$: $\pi_m = (\pi_m^0, \ldots, \pi_m^{n-1})$
    for $i = 0, 1, \ldots, n-1$ **do**
      Set $x_{t,m}^{i+1} = x_{t,m}^i - \gamma \nabla f_m^{\pi_m^i}(x_{t,m}^i)$
    end for
    Compute $g_{t,m} = \frac{1}{\gamma n}\left(x_t - x_{t,m}^n\right)$ and send $\mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$ to the server
    Set $h_{t+1,m} = h_{t,m} + \alpha \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
    Set $\hat{g}_{t,m} = h_{t,m} + \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
  end for
  $h_{t+1} = \frac{1}{C}\sum_{m \in S_t} h_{t+1,m} = h_t + \frac{\alpha}{C}\sum_{m \in S_t} \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
  $\hat{g}_t = \frac{1}{C}\sum_{m \in S_t} \hat{g}_{t,m} = h_t + \frac{1}{C}\sum_{m \in S_t} \mathcal{Q}_t\left(g_{t,m} - h_{t,m}\right)$
  $x_{t+1} = x_t - \eta \hat{g}_t$
end for

Output: x_T

J.2 Analysis of DIANA-NASTYA with Partial Participation

Theorem J.2. *Let step sizes η, γ satisfy the following equations*

$$\eta = \min\left(\frac{1}{80L_{\max}\left(1 + \frac{\omega}{C}\right)}, \frac{C}{\mu(1+\omega)M}\right), \quad \gamma = \frac{1}{5nL_{\max}}$$

Define the Lyapunov function:

$$\Psi_t = \|x_t - x^\star\|^2 + \frac{A}{M}\sum_{m=1}^M \|h_{t,m} - h_m^\star\|^2,$$

where $A = \lambda\eta^2$. Selecting parameters $\alpha = \frac{1}{1+\omega}; \lambda = \frac{8\omega}{\alpha M}, \gamma = \frac{1}{5nL_{\max}}$, also using $\eta \leq \min\left[\frac{C}{\mu(1+\omega)M}, \frac{1}{80L_{\max}\left(1+\frac{\omega}{C}\right)}\right]$ Under the Assumptions 1, 2, 3 iterates of DIANA-NASTYA-PP (Algorithm 6) satisfy

$$\mathbb{E}\left[\Psi_T\right] \leq \left(1 - \frac{\eta\mu}{2}\right)^T \mathbb{E}\left[\Psi_0\right] + \frac{3\gamma^2 n^2 L_{\max}^2}{\mu}\left(\frac{1}{M}\sum_{m=1}^M \sigma_{\star,m}^2 + n\sigma_\star^2\right) + \frac{2\eta(M-C)}{\mu C \max(1, M-1)}\sigma_\star^2.$$

Note that we eliminate the variance term proportional to ω : $8\frac{\eta}{\mu}\frac{\omega}{C}\sigma_\star^2$. In the Partial Participation regime, we have a variance term proportional to $\frac{(M-C)}{C\max(1,M-1)}$, which equals zero if $C = M$. This term decreases as $\mathcal{O}\left(\frac{1}{C}\right)$, so we achieve the expected linear speedup.

Proof. STEP 1: we need to estimate inner product. By $\hat{g}_t = \frac{1}{C}\sum_{m\in S_t}\hat{g}_{t,m}$, we have

$$\begin{aligned}
-\mathbb{E}_t\left[\left\langle \frac{1}{C}\sum_{m\in S_t}\hat{g}_{t,m}, x_t - x^\star\right\rangle\right] =& -\left\langle \frac{1}{C}\mathbb{E}_t\left[\sum_{m\in S_t}\hat{g}_{t,m}\right], x_t - x^\star\right\rangle \\
=& -\left\langle \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_t\left[\hat{g}_{t,m}\right], x_t - x_\star\right\rangle \\
=& -\frac{1}{M}\sum_{m=1}^{M}\left\langle g_{t,m}, x_t - x^\star\right\rangle \\
=& -\frac{1}{M}\sum_{m=1}^{M}\left\langle g_{t,m} - h_m^\star, x_t - x^\star\right\rangle \\
\leq& -\frac{\mu}{4}\left\|x_t - x^\star\right\|^2 - \frac{1}{2}\left(f\left(x_t\right) - f\left(x^\star\right)\right) \\
& -\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1} D_{f_m^{\pi_m^{im}}}\left(x^\star, x_{t,m}^i\right) \\
& +\frac{L_{\max}}{2Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_t - x_{t,m}^i\right\|^2.
\end{aligned}$$

STEP 2: We need to bound $\mathbb{E}\left\|\hat{g}_t\right\|^2$. By $\hat{g}_t = \dfrac{1}{C}\sum_{m\in S_t}\hat{g}_{t,m}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}\left[\left\|\hat{g}_t\right\|^2\right] =& \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{C}\sum_{m\in S_t}\left(h_{t,m} + \mathcal{Q}\left(g_{t,m} - h_{t,m}\right) - g_{t,m} + g_{t,m}\right)\right\|^2\right] \\
=& \mathbb{E}_{\mathcal{Q}}\left[\left\|\frac{1}{C}\sum_{m\in S_t}\left(h_{t,m} + \mathcal{Q}\left(g_{t,m} - h_{t,m}\right) - g_{t,m}\right)\right\|^2\right] + \left\|\frac{1}{C}\sum_{m\in S_t} g_{t,m}\right\|^2 \\
=& \frac{1}{C^2}\sum_{m\in S_t}\mathbb{E}_{\mathcal{Q}}\left[\left\|h_{t,m} + \mathcal{Q}\left(g_{t,m} - h_{t,m}\right)\right\|^2\right] + \left\|\frac{1}{C}\sum_{m\in S_t} g_{t,m}\right\|^2 \\
\leq& \frac{\omega}{C^2}\sum_{m\in S_t}\left\|g_{t,m} - h_{t,m}\right\|^2 + \left\|\frac{1}{C}\sum_{m\in S_t} g_{t,m}\right\|^2 \\
\leq& \frac{2\omega}{C^2}\sum_{m\in S_t}\left\|g_{t,m} - \nabla f_m\left(x_t\right)\right\|^2 + \frac{2\omega}{C^2}\sum_{m\in S_t}\left\|\nabla f_m\left(x_t\right) - h_{t,m}\right\|^2 \\
& + 2\left\|\frac{1}{C}\sum_{m\in S_t} g_{t,m} - h_m^\star\right\|^2 + 2\left\|\frac{1}{C}\sum_{m\in S_t} h_m^\star\right\|^2
\end{aligned}$$

Taking expectation by subsamling, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|\hat{g}_t\right\|^2\right] \leq & \frac{2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|g_{t,m}-\nabla f_m\left(x_t\right)\right\|^2+\frac{2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_{t,m}\right\|^2 \\
& +\frac{2}{M}\sum_{m=1}^{M}\left\|g_{t,m}-h_m^\star\right\|^2+\frac{2(M-C)}{C(M-1)M}\sum_{m=1}^{M}\left\|h_m^\star\right\|^2 \\
\leq & \frac{2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|g_{t,m}-\nabla f_m\left(x_t\right)\right\|^2+\frac{2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_{t,m}\right\|^2 \\
& +\frac{4}{M}\sum_{m=1}^{M}\left\|g_{t,m}-\nabla f_m\left(x_t\right)\right\|^2+\frac{4}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_m^\star\right\|^2 \\
& +\frac{2(M-C)}{C(M-1)M}\sum_{m=1}^{M}\left\|h_m^\star\right\|^2 \\
\leq & 4\left(1+\frac{\omega}{C}\right)\frac{L_{\max}^2}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+\frac{2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_{t,m}\right\|^2 \\
& +\frac{8L_{\max}}{M}\sum_{m=1}^{M}D_{f_m}\left(x_t,x^\star\right)+\frac{2(M-C)}{C(M-1)M}\sum_{m=1}^{M}\left\|h_m^\star\right\|^2 \\
= & \left(1+\frac{\omega}{C}\right)\frac{L_{\max}^2}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+\frac{2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_{t,m}\right\|^2 \\
& +8L_{\max}\left(f\left(x_t\right)-f\left(x_\star\right)\right)+\frac{2(M-C)}{C(M-1)M}\sum_{m=1}^{M}\left\|h_m^\star\right\|^2
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Q},S_t}\left[\left\|x_{t+1}-x^\star\right\|^2\right] \leq & \left(1-\frac{\eta\mu}{2}\right)\left\|x_t-x^\star\right\|^2-\eta\left(1-4L_{\max}\eta\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\
& +\eta L_{\max}\left(1+4\left(1+\frac{\omega}{C}\right)L_{\max}\eta\right)\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2 \\
& +\frac{2\eta^2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_{t,m}\right\|^2+\frac{2\eta^2(M-C)}{C(M-1)M}\sum_{m=1}^{M}\left\|h_m^\star\right\|^2.
\end{aligned}
$$

STEP 3: Note that

$$
\frac{1}{M}\sum_{m=1}^{M}\left\|h_{t+1,m}-h_m^\star\right\|^2=\frac{C}{M}\frac{1}{C}\sum_{m\in S_t}\left\|h_{t+1,m}-h_m^\star\right\|^2+\frac{M-C}{M}\frac{1}{M-C}\sum_{m\notin S_t}\left\|h_{t+1,m}-h_m^\star\right\|^2.
$$

Taking expectation by compression, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q}}\left[\frac{1}{C}\sum_{m\in S_t}\left\|h_{t+1,m}-h_m^\star\right\|^2\right]=\mathbb{E}_{\mathcal{Q}}\left[\frac{1}{C}\sum_{m\in S_t}\left\|h_{t,m}+\alpha\mathcal{Q}\left(g_{t,m}-h_{t,m}\right)-h_m^\star\right\|^2\right]\\
&=\frac{1}{C}\sum_{m\in S_t}\left(\left\|h_{t,m}-h_m^\star\right\|^2+2\alpha\left\langle g_{t,m}-h_{t,m},h_{t,m}-h_m^\star\right\rangle+\alpha^2(1+\omega)\left\|g_{t,m}-h_{t,m}\right\|^2\right)\\
&\overset{\alpha\le 1/1+\omega}{\le}\frac{1}{C}\sum_{m\in S_t}\left(\left\|h_{t,m}-h_m^\star\right\|^2+2\alpha\left\langle g_{t,m}-h_{t,m},h_{t,m}-h_m^\star\right\rangle+\alpha\left\|g_{t,m}-h_{t,m}\right\|^2\right)\\
&=\frac{1-\alpha}{C}\sum_{m\in S_t}\left\|h_{t,m}-h_m^\star\right\|^2+\frac{\alpha}{C}\sum_{m\in S_t}\left\|g_{t,m}-h_{t,m}\right\|^2.
\end{aligned}
$$

Taking expectation by subsampling, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Q},S_t}\left[\frac{1}{C}\sum_{m\in S_t}\left\|h_{t+1,m}-h_m^\star\right\|^2\right]&\le\mathbb{E}_{S_t}\left[\frac{1-\alpha}{C}\sum_{m\in S_t}\left\|h_{t,m}-h_m^\star\right\|^2+\frac{\alpha}{C}\sum_{m\in S_t}\left\|g_{t,m}-h_m^\star\right\|^2\right]\\
&=\frac{1-\alpha}{M}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2+\frac{\alpha}{M}\sum_{m=1}^{M}\left\|g_{t,m}-h_m^\star\right\|^2.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathbb{E}_{S_t,Q_t}\left[\frac{1}{M}\sum_{m=1}^{M}\left\|h_{t+1,m}-h_m^\star\right\|^2\right]=&\frac{(1-\alpha)C}{M^2}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2+\frac{\alpha C}{M^2}\sum_{m=1}^{M}\left\|g_{t,m}-h_m^\star\right\|^2\\
&+\frac{M-C}{M}\mathbb{E}_{S_t,Q_t}\left[\frac{1}{M-C}\sum_{m\notin S_t}\left\|h_{t,m}-h_m^\star\right\|^2\right]\\
=&\frac{(1-\alpha)C}{M^2}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2+\frac{\alpha C}{M^2}\sum_{m=1}^{M}\left\|g_{t,m}-h_m^\star\right\|^2\\
&+\frac{M-C}{M}\frac{1}{M}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2\\
\le&\left(1-\frac{\alpha C}{M}\right)\frac{1}{M}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2\\
&+\frac{2\alpha L_{\max}^2C}{M^2n}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2\\
&+\frac{4L_{\max}\alpha C}{M^2}\sum_{m=1}^{M}D_{f_m}\left(x_t,x^\star\right).
\end{aligned}
$$

STEP 4: Defining Lyapunov function as follows

$$
\Psi_t=\left\|x_t-x^\star\right\|^2+\frac{A}{M}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2,
$$

we have

$$\begin{aligned}\mathbb{E}_{\mathcal{Q},S_t}\left[\Psi_{t+1}\right] \leq & \left(1-\frac{\eta\mu}{2}\right)\left\|x_t-x^\star\right\|^2-\eta\left(1-4L_{\max}\eta\right)\left(f\left(x_t\right)-f\left(x^\star\right)\right) \\ & +\eta L_{\max}\left(1+4\left(1+\frac{\omega}{C}\right)L_{\max}\eta\right)\frac{1}{Mn}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2 \\ & +\frac{2\eta^2\omega}{C}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla f_m\left(x_t\right)-h_{t,m}\right\|^2+\frac{2\eta^2(M-C)}{C(M-1)M}\sum_{m=1}^{M}\left\|h_m^\star\right\|^2 \\ & +\left(1-\frac{\alpha C}{M}\right)\frac{A}{M}\sum_{m=1}^{M}\left\|h_{t,m}-h_m^\star\right\|^2 \\ & +\frac{2\alpha L_{\max}^2 AC}{M^2n}\sum_{m=1}^{M}\sum_{i=0}^{n-1}\left\|x_{t,m}^i-x_t\right\|^2+\frac{4L_{\max}\alpha AC}{M}\left(f\left(x_t\right)-f\left(x^\star\right)\right).\end{aligned}$$

Setting $A=\lambda\eta^2$ and using Lemma F.3, we have

$$\begin{aligned}\mathbb{E}\left[\Psi_{t+1}\right] \leq & \left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\left[\left\|x_t-x^\star\right\|^2\right]+\left(1-\frac{\alpha C}{M}+\frac{4\omega}{\lambda C}\right)\frac{\lambda\eta^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|h_{t,m}-h_m^\star\right\|^2\right] \\ & -\eta\left(1-8\eta L_{\max}\left(1+\frac{\omega}{C}\right)-4\eta L_{\max}\alpha\lambda\frac{C}{M}\right)\mathbb{E}\left[f\left(x_t\right)-f\left(x^\star\right)\right] \\ & +8\gamma^2n^2L_{\max}^2\eta\left(1+4\eta L_{\max}\left(1+\frac{\omega}{C}\right)+2\eta L_{\max}\alpha\lambda\frac{C}{M}\right)\mathbb{E}\left[f\left(x_t\right)-f\left(x^\star\right)\right] \\ & +2\gamma^2n^2L_{\max}^2\eta\left(1+4\eta L_{\max}\left(1+\frac{\omega}{C}\right)+2\eta L_{\max}\alpha\lambda\frac{C}{M}\right)\left(\frac{1}{M}\sum_{m=1}^{M}\sigma_{\star,m}^2+n\sigma_\star^2\right) \\ & +\frac{2\eta^2(M-C)}{C(M-1)}\sigma_\star^2.\end{aligned}$$

Selecting $\alpha=\frac{1}{1+\omega}$; $\lambda=\frac{8\omega}{\alpha M}$; $\eta\leq\frac{C}{\mu(1+\omega)M}$, also using $\eta=\frac{1}{80L_{\max}\left(1+\frac{\omega}{C}\right)}$, $\gamma=\frac{1}{5nL_{\max}}$ and applying previous steps we obtain

$$\begin{aligned}\mathbb{E}\left[\Psi_{t+1}\right] & \leq\left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\left[\Psi_k\right]+3\gamma^2n^2L_{\max}^2\eta\left(\frac{1}{M}\sum_{m=1}^{M}\sigma_{\star,m}^2+n\sigma_\star^2\right)+\frac{2\eta^2(M-C)}{C(M-1)}\sigma_\star^2 \\ & \quad-\eta\left(\frac{1}{2}-10\gamma^2n^2L_{\max}^2\right)\mathbb{E}\left[f\left(x_t\right)-f\left(x^\star\right)\right] \\ & \leq\left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\left[\Psi_k\right]+3\gamma^2n^2L_{\max}^2\eta\left(\frac{1}{M}\sum_{m=1}^{M}\sigma_{\star,m}^2+n\sigma_\star^2\right)+\frac{2\eta^2(M-C)}{C(M-1)}\sigma_\star^2,\end{aligned}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly outline our contributions in the abstract and introduction, and we also include a dedicated Contributions section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly highlight all assumptions and limitations in the text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The main contribution of the paper is its theoretical analysis. We support the paper with necessary definitions, assumptions, and lemmas.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper is supported by reproducible experiments, with all stochastic elements from pseudo-random generators fixed in advance. For details, see the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We aim to make the paper and all source code for experiments open-sourced to accelerate scientific findings in the field of Federated Learning and Machine Learning in general. For details, see the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide detailed guidelines for experiments setup in Appendix and in the folder with experiment source code. For details, see the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide detailed guidelines for experiments setup in Appendix and in the folder with experiment source code. For details, see the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides information on the computer resources in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research focuses on mathematical objects and does not involve human subjects or participants. The data used for our experiments consists of publicly available datasets. Our work does not explicitly address or examine the social implications of applying this research in practice.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work operates on mathematical objects, and the essence of our work provides a new optimization algorithm. Because of theoretical nature of our work the impact discussion is not applied.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper provides an optimization algorithm with the required theory. The data or models are not output assets of our work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide references for used datasets and deep learning models used in experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The output assets of our paper is Algorithm and Source code for experiments.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.