We want to thank the reviewers for their comments! **For all reviewers**, we want to answer the common questions

Why is this problem a challenge?

In (SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator, Fang et al. 2018) was presented a new method SPIDER that improved the oracle complexity the classical SGD method from ${}^{\sigma^2}\!/_{\varepsilon^2}$ to ${}^{\sigma}\!/_{\varepsilon^{3/2}}$. Numerous papers that we present in Table 1, tried to understand how to apply this idea to distributed optimization. As the reviewers can see, that is not a very trivial task, and all previous methods have significant drawbacks. We made a new step that fixed all previously known drawbacks. *The fact that at least 7 papers from different research groups could not completely solve the problem is a challenge.*

What is challenging in theoretical analysis, and what is the difference from previous methods?

We discuss it in Section 5 and explain the difference between the new method and the DASHA method. But the proofs are the central part of our paper. As in all mathematical papers, all insights are hidden there. For instance, one of the challenges we mention in the main part:"while in DASHA, the randomness from compressors is independent of the randomness from stochastic gradients, in DASHA-PP, these two randomnesses are coupled by the randomness from the partial participation." This is the best we can do in the main part of the paper. But we hear your comments, and we will try to elaborate more.

Experiments

This is purely a theoretical paper that solves a popular optimization problem. In this field, numerous excellent papers provide the same "amount" of experiments. So it is not apparent why our paper is treated differently. In our paper, we added partial participation to DASHA. So in the experiments, we present how it affects the convergence. It is not clear what else we should add.

Experiments with Neural Network

We, practitioners and theoreticians, know very little about neural networks. Still, nobody — even people who focus on this problem — understands how the vanilla GD and SGD methods work with neural networks (See recent works on "The Edge of Stability" phenomenon (Understanding the Unstable Convergence of Gradient Descent, Kwangjun Ahn (ICML 2022))). Therefore, such experiments are not well motivated.

Reviewer 2:

2.a. "There are eight assumptions used in the analysis which might be too restrictive."

In Tables 1 and 2, in column "Limitations", we compare how the assumptions between the papers are different. The main purpose of "Limitations" is to show that our assumptions are not stronger than the assumptions of any papers from Tables 1 and 2, so we do not agree that they are restrictive. Some papers present these eight assumptions as two-three assumptions. For instance, we define 3 different smoothness constants, while other papers can define only 1 constant, which is just the maximum of our 3 smoothness constants.

2.b. "In particular, could you please provide more comments on Assumption 6?."

This is the mean-squared smoothness property that is used in **all papers with variance reduction**. See (Lower Bounds for Non-Convex Stochastic Optimization Arjevani et al.; SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator Fang et al.)

3. This part is indeed can be confusing. We mean that the compressors are *statistically* independent. In other words, $\mathcal{C}_1(x_1), \dots, \mathcal{C}_n(x_n)$ are independent random vectors for all $x_1, \dots, x_n \in \mathbb{R}^d$.

4. In order to get the computational and communication complexities, we have to substitute an explicit formula of ω from Definition 1. It may differ for different compressors (see On Biased Compression for Distributed Learning Beznosikov et al). So it is not possible to provide a nice corollary that will work for any compressor. The good news is that RandK is the simplest compressor. By showing an improvement for RandK, we can expect that more advanced compressors will have even better theoretical and practical guarantees. **5.** We also discuss it after Corollaries 1, 3, and 4.

Reviewer 4:

2. If you read papers from Tables 1 and 2, you will find that virtually all of them extend the theory from some previous papers. The research on the optimization methods is incremental. It is not very obvious why it is a weakness of our paper and not others. Our assumptions are general and not stronger than the assumptions of methods from Tables 1 and 2. Our analysis is independent and requires additional mathematical techniques that we provide in our proofs. Can the review kindly explain what do he/she mean by "fundamental analysis"?

Reviewer 5:

2. We are aware of the work by A. Defazio L. Bottou. But it does not mean that VR methods are hopeless for neural network optimization. See, for instance, a recent work (Momentum-Based Variance Reduction in Non-Convex SGD Ashok Cutkosky, Francesco Orabona), where they provided theoretical and practical improvement. The development and understanding of VR methods are still going.