We thank the reviewers for their feedback and time! We are encouraged they found our work "powerful" and "adding a significant value to the field" of distributed optimization (R1; score 8), "novel" and "rich in content" (R2; score 7), "technically interesting" and "can be useful in analyzing a wide range of error-compensated methods" (R3, score 4), "a great contribution to the field" (R4; score 9) and emphasized that our paper is well-written (R1, R2, R4).

R1: ***The paper is too large to sufficiently describe in the main body of the paper. The main novelty is that it accumulates a lot of methods in the literature to present a unified analysis whereas it does not seem to introduce many new ideas.*** Indeed, our work contains many new results. However, we do not merely "accumulate" existing methods in the literature: our new general approach allows us to design many new and innovative EC methods, with SOTA convergence guarantees (e.g., we design first EC methods with linear rates - an open problem since Seide et al in 2014; see Sec 2 for more)! In order to achieve this, we had to innovate substantially. So, our approach is both very general, as evidenced by the many concrete methods that it supports, and leads to SOTA theory.

R2: ***1) It is too rich in content probably given the page limitations: there is a lot (too much) material that is just relegated to the appendix...an appendix of 90 pages is excessive..maybe I would suggest to submit the paper to a journal rather than a conference...*** The size of the appendix is substantial because we did not want to hide important details of the proofs/convergence results for the different *special cases* and tried to be as precise as possible. As the result, we have many theorems requiring many proofs. These extra details will be also useful for other researchers who can build on our work. Moreover, there are no strict limits on the appendix enforced by the rules of NeurIPS. The key insights and results are explained in the main body of the paper. ***2) I find the title not so in line with the main content of the paper. The title should probably be focused on the uniform analysis of distributed SGD-based methods which is the main topic of the paper.*** The working version of our title was different and highlighted the unified analysis, as you suggest. We later decided it was better to change it by emphasizing the *new algorithms and SOTA convergence theory* presented in the paper. ***3) The authors are not discussing the following methods...*** These works are broadly relevant, but not particularly so: while they address distributed learning, they are not about compressed communication, nor about EC; both being the key aspects of our work. However, we are happy to mention papers in a larger neigborhood of our work in final version of the paper as an extra page is allowed there. ***4) It is a bit disappointing that for the experiments the distributed setting is only 'simulated' on a single machine with a for loop...*** One of the main goals of the paper is to develop a unifying theory for distributed methods with EC. The simulated experiments showing the number of bits sent by workers and the number of oracle calls are enough to justify the theoretical results, independently from a particular implementation. They correlate well with the performance of the methods in real distributed systems. We can easily add a few genuine distributed experiments, but this will not affect any of the findings of our work. ***5) Improve the readability of the plots by increasing their size (they are currently a bit too small).*** Thanks for the suggestion. We will increase the size of the plots in the final version of the paper.

R3: ***1) The proposed framework is too general so that the authors didn't find a good way to present it in an eight-page paper.*** We politely disagree. Our framework is general, but we reflected the key parts/results and insights in the main body: we have *6 pages* devoted to the presentation of our results including *a high-level explanation* of our framework and new methods, *formal statement of the main convergence result* and *12 plots* in the numerical part with a detailed description of our experimental setup. Furthermore, the appendix consists mostly of proofs and other technical details related to special cases that we believe should be in the appendix. ***2) I doubt there is something missing in the main theorem: I cannot find constant δ in the main theorem or its proceeding assumptions. This is not normal: δ should play a critical role in the convergence.*** All is OK with the result. First, the def. of δ is given at the end of page 2. Second, δ of course plays a critical role in the convergence results for *EC* methods. However, our framework and, in particular, Theorem 3.1 works even for the methods *without error-compensation*, see Sections E and G in the appendix. Third, the explicit dependencies on δ of the convergence results for EC methods are presented in Table 1, Thm F.1, and Section I (see also Table 5 for the dependency of the parameters from Assumption 3.4 on δ). ***3) I think the writing of this paper is awful.*** It is against reviewer guidelines and common courtesy to be impolite in a review. We respectfully ask the AC to ignore this comment. We also kindly ask the other reviewers who found our writing clear to defend us; thanks!! We did our best to present our many results in as simple and understandable way as possible, and we believe we achieved a very good result. No suggestions for improvement were proposed by the reviewer. ***4) Authors try to combine too many things in this paper (16 new methods!), and used more nearly 3 pages to demonstrate the contributions, leaving very few spaces for the technical part. Too many things are defined in the appendix, which makes the main paper really hard to read. Though NIPS allows supplementary material, the paper itself still needs to be as self-contained as possible.*** We politely disagree with the reviewer's criticism. In fact, we used *less than 2 pages* to demonstrate the contributions since Tables 1 and 2 contain a lot of technical details.

R4: ***In numerical experiments, it would be nice to include a comparison between "non-EC SGD" and "EC-SGD" to show effectiveness of error compensation.*** You are right: one can add this kind of comparison to justify the needing for EC for biased compressors like TopK. However, this question was theoretically addressed in [7], where authors proposed an example of distributed optimization problems for which "non-EC SGD" with TopK compression diverges exponentially fast. However, we will have extra page if our paper gets accepted, and can add a couple plots of this type.