Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-325-AC2, 2021

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Lossy Checkpoint Compression in Full Waveform Inversion" *by* Navjot Kukreja et al.

Navjot Kukreja et al.

nkukreja@imperial.ac.uk

Received and published: 14 April 2021

We thank the reviewer for the insightful comments. Every one of these points is helping make this paper much stronger.

Major points:

- *The idea of combining checkpointing and compression itself is not new and has been discussed by the same authors already in a previous publication. Some parts of the manuscript read more like a lab report and some parts are written quite sloppily (see examples in the minor comments below). I am also missing a discussion section that provides a bit of context for the validity of the results*

for other media. For instance, to what extent does the achievable compression factor depend on the medium, physics, aperture, misfit function, etc. We can rephrase the contributions section to not claim that the idea itself is new. This is an empirical study and we can only carry out a limited number of experiments. While we would have liked to have studied multiple media, physics, aperture and misfit functions, we have left this as future work and mentioned it as such. Happy to add a few lines to make these limitations of our study clearer.

- *I would have assumed that application-tailored compression algorithms would outperform black-box compression tools. Such a comparison would be very interesting and I find it currently missing. Examples of such techniques can be found, for instance, in Boehm et al. (Geophysics, 2016) or Weiser Gotschel (SISC 2012).* The two studies the reviewer mentions are both very relevant. Boehm et al is already cited as such, and Weiser and Gotschel is on the list of planned improvements for the next revision after comments from Reviewer 1. We would be happy to add a discussion about how different compression algorithms might affect the trade-off between accuracy, runtime and peak memory.

- *I am surprised that there is no adaptivity in steering the compression settings throughout the inversion. From an optimization perspective with inexact derivatives, it is well-known that more accurate gradients would be required when approaching a stationary point. I can't find this mentioned or analyzed in the manuscript.* Thanks for pointing this out. Our method takes atol as an input and not mentioning where that atol comes from is a clear oversight. We will add a couple of lines about this kind of adaptivity that we are enabling with our method. We would also like to point out that there is a level of adaptivity already present that we have left out from the paper for simplicity. The whole inversion problem presented in our paper is one instance of a multigrid method that starts off by inverting on the coarsest possible grid with the lower frequencies, and then slowly adding higher frequency components and interpolating onto finer grids over multiple inversions. We will also discuss this level of adaptivity in the next revision of the paper.

- *The manuscript would really benefit from a 3D visco-elastic example as this is where the memory bottleneck becomes a lot more prominent. Of course, I am not asking for a 3D FWI example, but I would strongly encourage the authors to provide an error analysis for a single visco-elastic gradient computation in 3D.* The memory bottleneck becomes more prominent as we move to more complex physics requiring multiple fields. There is also more computation to do, of course. This can be seen as another data point on the "arithmetic intensity" axis. We would be happy to add an example of how the tradeoff changes with different physics - potentially a plot along this arithmetic intensity axis. However, as we show in Table 1, the memory bottleneck is already significant for the isotropic acoustic equation. Also, as far as we are aware, isotropic-acoustic FWI is more common in practice than visco-elastic FWI as of today. Hence we don't think that isotropic-acoustic is irrelevant or too small an example to be interesting.

Minor comments:

- *You reference eq. (1) at least twice already in the introduction, way before the equation is actually introduced. Similar premature references exist for eq. (4) and symbols like Phi(m).* We would be happy to rewrite these sections to fix this in the next revision.
- *I also don't think this should be focusing on the non-dissipative acoustic wave equation. The compression approach applies to all time-dependent wave equations and it is a lot more relevant for visco-elastic media. You already hint that eq. (1) is not really the equation of interest on p.3, line 48.* As discussed above, we chose to focus on the isotropic acoustic case because that is the more commonly used case in practice. We agree that any memory-saving technique becomes

more relevant when the equation of interest has more fields. The change previously discussed should increase the applicability of this work beyond isotropic acoustic.

- *I think the statements on p.3 lines 50-56 are misleading and / or incorrect. The communication overhead of MPI-parallelized simulations can well be hidden behind computations by computing the halo first and performing asynchronous communication. But even on distributed compute architectures reducing the memory footprint is highly desired. Furthermore, many frequency-domain methods indeed have a HUGE memory footprint when factorizing the Helmholtz operator. At least, you would need to be more specific what you mean by frequency-domain here.* About the frequency-domain methods, we are happy to rephrase this to be clearer. About MPI, we are already accounting for computation-communication overlap in what we say. The point we are trying to make is that there is a lower limit to the amount of computation required to hide the communication behind. The communication time depends on the interconnect - which is not always very fast on cloud platforms, so the smallest per-rank subdomain that would be able to hide the communication behind its in-domain computation might be too big on a system without a high-speed interconnect. Happy to expand this discussion in the next revision of the paper.
- *p.5, line 106. Typo: There is an additional ")" in (2016)).* Will be fixed in the next revision.
- *Section 2. The notation mixes bold and italic symbols, for instance, for the model m and d_obs / d_sim, phi. The operator A is not introduced properly. It is the discretization of the PDE operator, and not of the equation.* Will be fixed in the next revision.
- *I would consider merging sections 3 and 4, because section 3 merely contains extended headings for the subsections of part 4.* We think Section 3 provides

value as a reference to understand what each experiment does, in a quick read that typically goes back and forth through the paper. Happy to remove if the reviewers think this is creating confusion.

- *Many figures contain similar quantities and could be merged. For instance, Fig 7/8, Fig 12/13, Fig 14/15, Fig 20/21 could all be merged into a single row each.* Happy to fix
- *The maximum buffer size used in Fig 7 and 8 is fairly small and I would consider extending the line to at least twice the number of time steps.* We understand the reviewer to mean that we should run this experiment for more timesteps than 300. Happy to update this plot to twice as many timesteps in the next revision.