Review of: "Annealed Stein Variational Gradient Descent for Improved Uncertainty Estimation in Full-Waveform Inversion"

Xuebin Zhao[1]

1 University of Edinburgh, United Kingdom

Potential competing interests: No potential competing interests to declare.

This paper performs uncertainty quantification in seismic full waveform inversion (FWI) problems using a specific variational inference algorithm - Stein variational gradient descent (SVGD). Considering that SVGD often suffers from the variance collapse problem since the number of samples (particles) used is far fewer than the dimensionality of the inverse problem (the number of unknown model parameters), the authors introduced an annealed version of SVGD by changing the relative weight of the two terms (i.e., the driving force and the repulsive force) in the SVGD objective function. Numerical examples show that the annealed version outperforms the vanilla SVGD in terms of inversion accuracy and uncertainty estimates. In addition, the authors performed uncertainty analysis using the inversion results, including PCA analysis and clustering. I have the following comments:

1. In the annealed version of the SVGD update (equation 8), the authors assign a weight $\alpha \in [0, 1]$ to down-weight the contribution of the driving force. Therefore, the algorithm tends to push samples away from each other such that they can explore a broader parameter space and hopefully overcome the variance collapse problem, especially at the beginning of the optimisation when α value is small. However, this looks heuristic to me. Mathematically, the original update rule in equation 5 represents the steepest minimisation direction of the KL divergence (see details in Liu and Wang 2016). By adding the annealing factor α, does equation 8 still provide the steepest direction? Does the ASVGD algorithm still minimise the KL divergence between the variational distribution and the posterior distribution? The answer might be no. Then the next questions would be, what do we actually obtain? Can we still interpret the obtained particles as approximate posterior samples? Is the final probability distribution a good approximation to the posterior pdf? I admit that this is an application work rather than a purely theoretical work, and that the presented inversion results are improved, but I think the authors still need to consider these theoretical questions carefully and discuss these aspects in the paper.
2. Following 1, Wang et al., 2023 (Reweighted variational full-waveform inversions - Geophysics) introduced a similar approach to this work, where the two terms in the objective function of variational inference are reweighted. It would be good to see a comparison between these two studies.
3. If the purpose of adding the annealing factor is to ensure that particles can facilitate broad coverage and exploration of the parameter space, heuristically, an easier approach might be to initialise particles by sampling from a probability distribution with large variances, such as a uniform distribution. This will provide initial samples that are broadly distributed within the parameter space. Then, during the optimisation process, we use the vanilla SVGD to update the particles. This seems to be similar to the annealing approach. Can you put some comments/discussion on this?

4. Equation 8: to make it consistent with equation 5, I suggest changing p(m) into the posterior pdf p(m|d).
5. Section 2.5 (Clustering) looks quite descriptive and slightly difficult to follow. Can the authors add some details/explanations on this?
6. What is the prior distribution used in the numerical examples? Please define the prior and likelihood terms explicitly.
7. Figure 7: Since the initial particles have relatively lower variances, it is difficult for the vanilla SVGD to push these particles away, as displayed in Figure 7a, whereas ASVGD results show higher variance, which is good. However, I am curious about the performance of the two algorithms if we initialise particles with larger variance (such as sampling from a uniform distribution with a large interval). If similar results as displayed in Figure 7 are obtained, this will demonstrate the difference between ASVGD and initialising SVGD with, say, a uniform distribution, as pointed out in 3.
8. All standard deviation maps are difficult to compare. Can you use a smaller color scale or other more informative colormaps?
9. For the clustering results, it would be interesting to compare data misfit values calculated from samples in cluster -1 with those from clusters 0 and 1. This will prove whether samples from cluster -1 are truly model solutions that fit the observed data but present geologically meaningless features or they are not good solutions (possibly due to the introduction of the annealing factor such that the algorithm doesn't minimise the KL divergence). This might help to answer the first comment.