Paper ID: 6909 **Title: Learning Hawkes Processes from a handful of events**

We would like to thank the reviewers for their detailed feedback and insightful comments, we will incorporate the suggested clarifications in the paper.

Remarks for Reviewer 1. We will add the reference of the SVI approach of Linderman and Adams in the introduction.

Remarks for Reviewer 2.

- *On the normalization coefficient and prior p_α in (4)*: It is known that the regularized-MLE objective is equivalent to MAP objective up to the constant normalization coefficient. The regularization term $R(\mu, W)/\alpha$ in (4) can be seen as the negative of the logarithm of the unnormalized prior. So to derive the prior $p_\alpha(\mu, W)$ we only need to compute the normalization term, which is the integral of $\exp\left(-R(\mu, W)/\alpha\right)$ over μ and W, and which therefore cannot be a function of the integration variables μ and W.
- *On Optimizing α in (4) and (8) directly*: In line 142 we actually mean that the MAP estimator cannot be optimized over α. Indeed, as demonstrated in the example of Appendix B, the MAP objective is an unbounded function (from above) of α.
- *On line 147*: Indeed it should be "maximum likelihood", we apologize for the confusion caused by the typo.
- *On performance if only 1 dimension has few observations*: In our setting, data scarcity comes from the short length of the observation window and not from missing data. So, if one dimension i has much fewer timestamps than others, it means that overall, it has smaller intensity (i.e. small μ_i and incoming W_{ij}). Therefore the likelihood function naturally enforces small values in these parameters to explain the observed intervals with no timestamps. The setting where the data scarcity comes from both short observation window and missing data requires extending our probabilistic model and is an interesting direction for future work.

Remarks for Reviewer 3.

- *Computational complexity of the algorithm*: Our gradient-based method is computationally efficient and scales well to large data regimes. For small data-regimes, the state-of-the-art methods empirically seem to converge faster as they need fewer iterations (even if there is no proof of convergence rate in the papers). However, when the number of nodes gets large and the number of observations increases, the per-iteration cost of the state-of-the-art methods grows faster than our gradient-based approach, which we expect to converge faster for such settings. Indeed, the complexity of MLE-ADM4 is $O(N_{iter} n^3 d^2)$ (see Table 1, Achab et al. 2016, [1]), whereas the complexity of our approach can be reduced to $O(N_{iter} n d^2 + d^2 n^2)$ where the reduced cost comes from efficiently pre-computing some constant terms in the log-likelihood function (at the cost of memory), which is a one shot cost of $O(d^2 n^2)$.
- To evaluate the scalability of our approach, we ran additional simulations on increasingly large-dimensional problems. As shown in Figure 1, the per-iteration running time of our approach VI-EXP (implemented in python) scales better than the one of MLE-ADM4 (implemented in C++). In addition, even if our gradient descent algorithm requires more iterations to converge, we show in Figure 2 that VI-EXP reaches the same F1-score as MLE-ADM4 faster. Empirically, we expect similar results for the non-parametric setting. We performed simulations only for the parametric setting due to the time constraint of the rebuttal.
- *Optimizing the decay parameters*: It is possible to optimize the decay parameters but we chose to use this particular form of exponential kernel as an example designed to match the efficient C++ implementation of MLE-ADM4, which takes advantage of the convexity of the problem. In addition, considering a fixed decay enables the use of caching as discussed above.
- *Visualizing the estimated causal-networks*: In high-dimensions, the resulting networks are not easy to visualize. We tried to draw the learned networks on top of a map of the Ebola dataset, but the figure needs to be rendered too large to be clear. Given space limitation, we did not plot any.

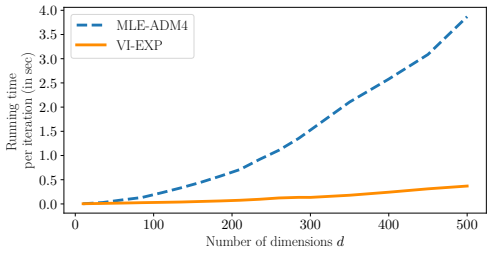

Figure 1: Comparison of per-iteration running time.

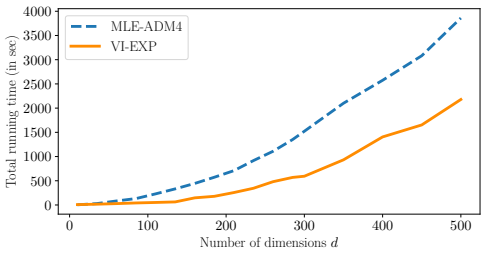

Figure 2: Running time required for our approach VI-EXP to reach the same F1-Score as MLE-ADM4.