Uncertainty on Asynchronous Event Prediction: Author Response

Compound distribution. We would like to emphasize that the uncertainty is not modeled through a compound distribution in our models. Indeed, the compound distribution would be $\text{Cat}(\bar{\boldsymbol{p}}_i(\tau))$ where $\bar{\boldsymbol{p}}_i(\tau) = \mathbb{E}_{\boldsymbol{p}\sim P_i(\theta(\tau))}[\boldsymbol{p}] = \int \boldsymbol{p} P_i(\theta(\tau))(\boldsymbol{p}) d\boldsymbol{p}$, and (see lines 167-169) the CE loss would only use this distribution. In contrast, the UCE does not use the compound distribution but considers the expected cross-entropy (note the order of $\int$ and $\log$).

Intuition on UCE. To give more intuition about the UCE loss, we propose the following example where we have two distributions on the simplex $P_i^{(1)}(\theta(\tau_i^*))$ and $P_i^{(2)}(\theta(\tau_i^*))$ such that $\bar{\boldsymbol{p}}_i(\tau_i^*) = \mathbb{E}_{\boldsymbol{p}\sim P_i^{(1)}(\theta(\tau_i^*))}[\boldsymbol{p}] = \mathbb{E}_{\boldsymbol{p}\sim P_i^{(2)}(\theta(\tau_i^*))}[\boldsymbol{p}]$. In this case the CE will be the same for both distributions, $\mathcal{L}_i^{(1)\text{ CE}} = \mathcal{L}_i^{(2)\text{ CE}}$. Now assume that all the probability mass is concentrated around the mean $\bar{\boldsymbol{p}}_i(\tau_i^*)$ for $P_i^{(1)}(\theta(\tau_i^*))$ but not for $P_i^{(2)}(\theta(\tau_i^*))$. Hence, $P_i^{(1)}(\theta(\tau_i^*))$ is very certain on the mean prediction. In contrast to CE, the UCE can distinguish the two distributions and especially $\mathcal{L}_i^{(1)\text{ UCE}} < \mathcal{L}_i^{(2)\text{ UCE}}$. Hence, an important property of the UCE is that the variance of the distribution on the simplex plays a substantial role in its value. In particular, high variance is penalized by the UCE which is particularly important during training. Indeed, the UCE will reduce the uncertainty for the categorical distributions predicted for the observed data. In combination with a prior value for the variance (which is done by the regularization term, lines 186-199), we keep the variance high for non-observed data while being more certain on the data we observed, as desired. Note that the regularization applied with CE would only set the variance of all (observed and non-observed) data/time points to the same prior. The CE would not reduce the variance on observed data and only adjust the mean prediction.

Objective criteria for loss selection. We propose the anomaly detection experiment with the distribution uncertainty (lines 305-322) as an objective criteria. The comparison of the different losses (CE, CE + reg, UCE + reg) for the FD-Dir model are shown in Fig. 1. The loss UCE + reg consistently improves the anomaly detection based on the distribution uncertainty. Furthermore, in the appendix, we proposed a visual representation of the benefit of UCE compared to CE on a simple classification task.

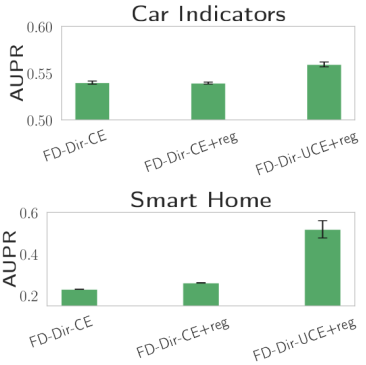

Figure 1: Loss comparison

Number of pseudo points. In our initial hyperparameter search we tuned the number of points but for the final results we kept it fixed across datasets (see lines 485-486 in the supp. mat.). Figure 2 shows that changing the number of points does not significantly affect the accuracy (same for the other datasets). Additionally, Figure 5 in the paper shows that both models learn to give lower weights to unnecessary points, essentially discarding them if we have too many.

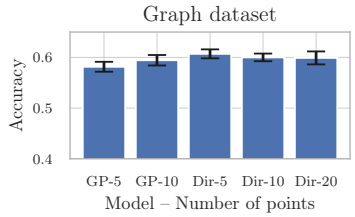

Figure 2: Number of pseudo points

Training time w.r.t. M. If the size of the RNN's hidden state is D, and we have M pseudo points, adding one more point leads to D more parameters. In the case of GP, we have to take into account the increase in computation time due to the inverse. Since the number of points is always lower than D and often $M < 10$, the increase is negligible. We found that the number of epochs until the early stopping is similar for different M. Therefore, neither the accuracy (see above) nor the training time are strongly affected when varying M.

Sampling. The Neural Hawkes Process [13] needs sampling to evaluate the integral and does so by passing time points through the RNN-based model which is expensive. In our case, sampling is (i) only required if we wish to use regularization or a point process version (note that obtaining the M pseudo points does *not* require sampling), and (ii) very cheap. The reason is that the evolution of the distributions over time is represented by pseudo points, which after computing the RNN's hidden state are given. That is, for the Dirichlet model, sampling only requires to evaluate the Gaussian function; and for the GP model to evaluate the kernel function. The computation of the hidden state, the inverse of the covariance etc. can all be reused across multiple samples. We will add these discussions to the paper.

Related Work. We will extend the related work section based on your feedback. In particular, we will mention the ability of Neural Hawkes Process to model multi-modal distributions and the possibility of RMTPP to model decaying intensities (like many classic point processes, e.g. Cox, Hawkes).