We ran **new experiments** to address your concerns: you'll like the results! These results (and related discussion) could be easily included in camera-ready, using supplementary material + the extra page that NeurIPS 2020 allows.

New experiment A (R3, R4): comparison with Guo et al. & LSE. We ran experiments to compare with Guo et al. and least-squares estimation (LSE) on multivariate point processes. Figure 1 shows the learning curves of MLE (red), our NCE (blue), Guo et al. NCE (green) and LSE (orange). Both Guo et al. and LSE converged (eventually) to much worse log-likelihood than our method and did so more slowly. We promise to use larger figures and fonts for the final version.

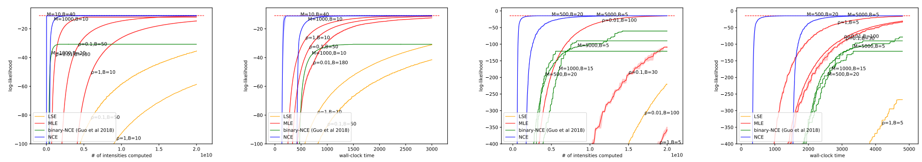

Figure 1: Learning curves on Synthetic-2 and BitcoinOTC datasets. (x-axis is truncated.) Similar patterns hold on all our other datasets, with the best curve always being NCE.

New experiment B (R2): prediction accuracy. Models that achieved comparable log-likelihood—whether they were trained by MLE or NCE—achieved comparable prediction accuracies (measured by RMSE for time and Error Rate for type). Therefore, NCE still beats MLE at converging quickly to the highest prediction accuracy.

New experiment C (R3, R5): simple baseline models. We checked the classical Hawkes process on our datasets, with MLE training. Classical Hawkes performed far worse on both training and test data than the deep models in our paper (similar to the experiments of Mei & Eisner 2017). The deep models have more flexibility to fit the data.

Now we clarify our **contributions**: new method, new theorems, sampling speedup, analysis of runtime, etc.

Mild assumption (R2). No, our theorems only require the intensity functions to be Riemann integrable, *not* continuous! Indeed, in our experiments, they are typically *discontinuous* at events, though continuous between events (line 55). This setting is Riemann integrable, as are the other widely-used point processes that R2 mentioned—so our results apply.

New theorems (R3). Did we merely inherit the theoretical properties from the discrete case? No, we needed non-trivial additional work. Lemma 1 in Appendix A showed that if θ and θ^* are meaningfully different in that they predict different intensities at time t for *some history*, then they actually do so for a *set of histories of non-zero measure*, making this difference visible in the objective function. (Note to R2: This lemma does require Riemann integrability, not Lebesgue integrability.) Previous work did not encounter this since they worked on non-sequential data (e.g., Gutmann & Hyvarinen 2010 + 2012) or discrete-time sequential data (e.g., Ma & Collins 2018). We'll highlight this difference.

New method (R3, R4). There are two approaches to NCE—BINARY and RANKING. We chose RANKING because we are working with conditional intensity functions. Our key idea of how to apply this to continuous time (line 111) is new, and required new analysis. Guo et al. used the older BINARY version, which is *not* well-suited to conditional distributions (see Ma & Collins 2018). This complicates their method since they needed to build a parametric model of the local normalizing constant, giving them weaker theoretical guarantees (see lines 243–247) and worse performance (Figure 1 above). So our contribution was *not* (merely) to extend to multivariate point processes as R3 implies.

Sampling speedup + analysis of runtime (R3). Sure, NCE requires sampling. But so does MLE (lines 73–74). We compared them **analytically** (§3.2) and showed **experimentally** (§5) that NCE evaluates on *fewer* samples and is practically faster (often by a factor of 5–10), in part because of our efficient method for sampling NCE noise (§3.1).

Least-squares estimator (R3). For classical or neural Hawkes processes, LSE is not faster than MLE—it requires sampling to estimate an integral, just like MLE. (R3 notes that it is faster for linear Hawkes processes, but those are extremely simple and of limited use.) In practice, LSE underperforms MLE in our experimental settings: see Figure 1.

Other methods (R5). We *did* use a Monte Carlo estimate for the log-likelihood. MCMC and variational methods aren't needed in our experiments, because the complete-data likelihood is defined by the simple equation (2). It's simple because our point process models are autoregressive; there's no need to fit a tractable lower bound (ELBO) as in globally normalized models. It contains no difficult log-sum-exp, only an integral over a summation. This can be estimated directly and without bias by sampling, which is what we do. (So isn't MLE always "feasible"?)

More evaluation (R3). Our evaluation was actually rather wide-ranging for an 8-page paper that also includes theorems. We evaluated on 6 (dataset, model) pairs: 2 synthetic and 4 real, spanning both NHP and NDTT models. Our baselines included both MLE and ablated versions of our full NCE method. We had so many results that many of them (figures & discussion) went to an appendix—maybe they were missed? We established that NCE is often more efficient than MLE and thus is worth trying in practice. We don't think this conclusion would change by testing the method on other similar parametric point process models. An 8-page paper needn't test a method on every conceivable (dataset, model) pair.

Other (R2, R3, R5). We'll take other suggestions, including correcting the venue of Xu et al. (R2), using larger fonts in figures (R2, R5), directly comparing convergence for diff. hyperparams (R5), and broader impact on society (R3).