We thank all the reviewers for their efforts and constructive comments, which help improve the quality of our paper.

To Reviewer #1

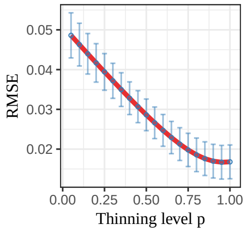
Q1. Lack of the evaluation of the uncertainty of the estimate. We actually derived an error bound of the empirical risk. Based on the analysis of the first and second moment of the estimator presented in Theorems 5.1 and 5.2, a Chebyshev's type error bound can be easily obtained:

$$\mathbb{P}\Big(\big|\nabla\tilde{R}^{\mathcal{S}} - \mathbb{E}\nabla\tilde{R}^{\mathcal{S}}(\theta)\big| > \epsilon\Big) \leqslant \frac{\mathbb{V}\Big[\nabla\tilde{R}^{\mathcal{S}}(\theta)\Big]}{\epsilon^2} \leqslant \frac{\mathbb{V}\Big[\nabla R_{\ell}(\theta)\Big]}{\epsilon^2} = \frac{\frac{1-p}{p}\mathbb{E}\Big[\nabla R(\theta)\Big]^2 + \mathbb{V}\big[\nabla R(\theta)\big]}{\epsilon^2},$$

for any $\epsilon > 0$. As we reckon this error bound of the gradient estimation is a direct deduction of Theorems 5.1 and 5.2, we didn't include this result due to page limit. We will present it as a corollary in the final version.

Q2. Thinning bootstrap procedure. Bootstrap is not suitable for point processes as it is a resampling method with replacement, that is, an event could be sampled more than once. This violates the fundamental assumption of point processes that events do not arrive simultaneously. A point cannot be sampled twice, otherwise the log-likelihood of the point process will be infinite. Therefore, we propose thinning as a downsampling method, rather than bootstrap as a resampling method with replacement. Alternatively, Jackknife resampling seems feasible for point processes, and we plan to explore it in the future. We will add this discussion to the final version.

Q3. Better figure representing MSE vs p. Thanks for the suggestion and we will revise our paper accordingly. We have conducted an experiment on this with the same setting in the paper and will add a new figure (shown below) in the final version. It shows that RMSE decreases as p increases.



To Reviewer #2

Q1. The definition of R. R is the likelihood/squared error loss, aka the residue. We will make this clear in the final version.

Q2. Make the transition to multi-variate processes a bit smoother. This is indeed a helpful suggestion, we will add a discussion to make it better.

Q3. Discuss the gradient for stochastic intensities calculated on sub-samples potentially being unbiased. The gradient estimation is unbiased if and only if $\mathbb{E}[\mathcal{H}(t;\theta)\lambda(t;\theta)] = \mathbb{E}\mathcal{H}(t;\theta)\mathbb{E}\lambda(t;\theta)$, as we shown in the proof of Theorem 5.1. However, this result is somehow inaccessible, therefore we particularly indicate, in the Theorem 5.1, that the gradient estimation for NHPPs are unbiased, but not for all the stochastic intensities. We also illustrated this result by the experiment shown in Fig. 3 and discussed in Line 266-269, in order to provide an intuitive understanding. We will add a remark after the theorem to clarify this result.

To Reviewer #3

We want to highlight that the aim of this paper is NOT to propose a specific optimization/learning algorithm, but to answer one of the most fundamental questions in point processes: what is the best sampling method for point processes? We propose thinning as a solution, and derive important theoretical results of thinning for parameter estimation, gradient estimation, and stochastic optimization, for point processes with decouplable intensities. In our empirical study, we applied thinning to learn various state-of-the-art models (MMEL, Granger Causality for Hawkes, and Sparse Low-rank Hawkes); to estimate the gradient in different types of point processes (NHPP and Hawkes) with different estimators (MLE and LSE); and also to perform stochastic optimization under different algorithms (SGD and Adam). Through this study, we show that thinning is a general downsampling solution for point processes with decouplable intensities, and is not restricted to a specific learning/optimization algorithm or estimator.

Q1. The applicability of the proposed model. Hawkes is a weak baseline. The thinning method is applicable to most, if not all, state-of-the-art models related to parametric point processes, including MMEL, GC, and sparse low-rank Hawkes, as we shown in the synthetic experiment in Table 1. In the experiment with real datasets for stochastic optimization, the aim is to test the applicability of thinning to typical and popular optimization methods such as SGD and Adam. We picked Hawkes as the model to learn as it is the basis of many derivative models. Note that, **Hawkes is not used as a baseline here**; our baselines are other sampling/optimization methods. The main purpose here is NOT to show that Hawkes is good, but rather to show that thinning is effective when coupled with various optimization algorithms (SGD and Adam).

Q2. The focus on decouplable intensities is a bit limited. As we mentioned in the paper, most state-of-the-art models of parametric point processes are decouplable (we would be grateful if the reviewer can point out if we have missed any). Besides, Netcodec (Long Tran, et al., 2015), parametric Hawkes (Liangda Li, et al., 2014), Hawkes with Stochastic Excitations (Young Lee, et al., 2016) and SLANT (Abir De, et al., 2016) are also decouplable. This general category covers all NHPPs, compound Poisson processes, renewal processes, marked point processes with independent markers, and a certain part of doubly stochastic Poisson processes (Cox processes). We believe this is a general assumption with many interesting properties. We choose this specific class of point processes in this first attempt of the problem as it facilitates the mathematical preciseness and rigorous theoretical proofs. We agree that a broader assumption will serve more scenarios, and we will investigate more general classes.