We thank the reviewers for their insightful comments and detailed analysis of our work. We provide clarifications below to address their comments.

Reviewer 1 We thank this reviewer for a thoughtful discussion of our work, and we hope that our comments to other reviewers below will be helpful in clarifying our contributions.

Reviewer 2

- *Comparison to signed DPPs:* There are several important differences between our work and the work on learning signed DPPs [1]: **(1)** Signed DPPs require $K_{ij} = \pm K_{ji}$, where $\boldsymbol{K}$ is the marginal DPP kernel (described in Sec. 2.1 of our paper). In contrast, our nonsymmetric DPP model is much more general, since it does not require that $|K_{ij}| = |K_{ji}|$. Since the off-diagonal elements of $\boldsymbol{K}$ determine the correlations between pairs of items, $\mathrm{cov}(\mathbb{1}_{i \in Y}, \mathbb{1}_{j \in Y}) = -K_{ij}K_{ji}$, this gives our model more flexibility. **(2)** The learning algorithm for signed DPPs presented in [1] assumes that the unknown kernel $\boldsymbol{K}$ is dense, i.e., all its entries are nonzero. In practice, this may not be a realistic assumption, because it implies that all pairs of items are correlated. **(3)** Moreover, our approach allows us to leverage a low-rank assumption on $\boldsymbol{L}$ (or, equivalently, $\boldsymbol{K}$), whereas the approach in [1] is not compatible with a low-rank assumption. **(4)** The learning algorithm in [1] has computational complexity of $O(M^6)$, where M is the size of the ground set (e.g., item catalog), making it computationally infeasible for most scenarios. In contrast, our learning algorithm has substantially lower time complexity, which allows our approach to be used on many real-world datasets. It is true that [1] inspired and informed our work. We will add some text to the camera-ready version of our paper to provide a comparison with this work.
- *Comparison to other baselines:* Since the focus of our work is on improving DPP modeling power and comparing nonsymmetric and symmetric DPPs, to keep the message of our paper clear we use the standard symmetric low-rank DPP as the baseline model for our experiments. We plan to perform an experimental comparison to other competing models for subset selection as part of future work. Regarding the comparison with the theory presented in [2], we emphasize, in our work, that the problem becomes significantly harder when we deal with nonsymmetric kernels, which shows that going from symmetric to nonsymmetric kernels is not a straightforward extension of previous work.

Reviewer 3

- *Low-rank representation of nonsymmetric DPP kernel:* The first term on the right side of Eq. 12 will be singular whenever $|Y_i| > D$, where Y_i is an observed subset. Therefore, to address this in practice we set D to the size of the largest subset observed in the data, as explained in [3]. Furthermore, the first term on the right side of Eq. 12 may be singular even when $|Y_i| \leq D$. In this case, we know that we are not at a maximum, since the value of the function becomes $-\infty$. Numerically, to prevent such singularities, in our implementation we add a small $\epsilon \boldsymbol{I}$ correction to each $\boldsymbol{L}_{Y_i}$ when optimizing Eq. 12 (we set $\epsilon = 10^{-5}$ in our experiments).

  Regarding the significance of our low-rank decomposition of $\boldsymbol{L}$ for nonsymmetric DPPs (described in lines 177 - 181 of our paper), this is indeed an extension of an idea developed in the symmetric case, and we do use well known decompositions for symmetric and skew symmetric matrices. We do not claim that we prove new matrix decompositions, but we rather propose a simple low-rank representation of a **subclass** of P_0-matrices. Please note that the claim, in the review, that a P_0-matrix can be decomposed as the sum of a PSD matrix and a skew-symmetric matrix is incorrect, and is not a consequence of Lemma 1 in our paper. Lemma 1 only states that if the symmetric component of a matrix is PSD, then that matrix is P_0, but the converse is not true (e.g., take the P_0-matrix $\boldsymbol{L} = ((1, -1), (5, 1))$, whose symmetric component, $(\boldsymbol{L} + \boldsymbol{L}^T)/2$, is the non-PSD matrix $((1, 2), (2, 1))$). Therefore, dealing with the class of all P_0-matrices seems very challenging, but leaves an exciting research topic open.

  Regarding the time complexity of the low-rank representation, we see from Eq. 12 that the time complexity required to compute the matrix multiplications associated with the gradient of the first and second terms of the log-likelihood will be $O(n\kappa^2 D + n\kappa^2 D' + DM^2 + D'M^2)$, where n is the number of observed subsets, κ is the size of the largest observed subset in the training data, and M is the size of the ground set (item catalog). We typically set $D \ll M$ and $D' \ll M$ in the low-rank representation; the associated matrix multiplications become much more expensive if we set $D = M$ (and presumably $D' = M$). In particular, the matrix multiplications for the second term of the log-likelihood will become $O(M^3)$ operations, instead of $O(DM^2 + D'M^2)$ operations. Therefore, we see that our low-rank representation still affords improvements in time complexity compared to the full-rank representation. We will add some text to the camera-ready version of our paper to make this point clear. We are confident that it is possible to approximate the DPP normalization constant, $\log\det(\boldsymbol{L} + \boldsymbol{I})$, using contrastive estimation for DPPs [4], and therefore address the remaining $O(M^3)$ bottleneck, but we leave this for future work.

References

[1] Victor-Emmanuel Brunel. Learning signed determinantal point processes through the principal minor assignment problem. In *NeurIPS*, pages 7365–7374, 2018.

[2] Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Rates of estimation for determinantal point processes. In *Conference on Learning Theory*, pages 343–345, 2017.

[3] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Low-rank factorization of Determinantal Point Processes. In *AAAI*, 2017.

[4] Zelda Mariet, Mike Gartrell, and Suvrit Sra. Learning determinantal point processes by corrective negative sampling. In *AISTATS*, pages 2251–2260, 2019.