We thank the reviewers for their insightful feedback. We are encouraged that they found our motivation and idea to be novel (R2, R4), practical (R1), interesting (R3) and scalable (R2, R4). Moreover, we are grateful that reviewers identified our contributions beyond just performance gains on different tasks (R1, R2, R4), and that R2 appreciated our theoretical contributions. Reviewers also found that our empirical studies are sound and convincing (R2, R4).

Below we first provide a recap on the goal of our work, and then give a point-by-point response to the comments.

[Recap] What is our goal? We investigated the role of sparsity and DAG constraints for learning linear DAGs, and established the corresponding *asymptotic guarantees*. Inspired by that, we formulate a likelihood-based method (GOLEM), and showed that one only has to apply soft sparsity and DAG constraints for structure learning.

@R1, R4 – Performance metrics: Thank you for raising this issue. Following your suggestions, we have computed and will report the normalized SHD and Structural Intervention Distance (SID) in the revision. We found that these new metrics yield consistent observations with the existing metrics reported in the paper.

@R2, R4 – Real data experiment: Thank you for the suggestion. We agree that these methods may not learn much on the Sachs dataset; in fact, even nonparametric extensions of NOTEARS, e.g., GraN-DAG (Lachapelle et al., 2020), do not perform well on this dataset, either. Given your comment, we have done and will include an additional experiment based on SynTReN generator (Van den Bulcke, 2006), which simulates gene expression data with ground truth. For each method, the SHDs, normalized SHDs and SIDs are: GOLEM-NV (35.9, 1.8, 128.4), GOLEM-EV (42.8, 2.1, 139.4) and NOTEARS-L1 (48.7, 2.4, 162.6), respectively. Note that lower is better for all three metrics reported here.

@R1 – Assumption of linearity: Great point. Learning DAGs is challenging, and some assumptions (e.g., linearity) are needed for developing estimation methods and corresponding theoretical guarantees, as in LiNGAM and NOTEARS. We plan to extend our method to nonlinear cases as future work (L372-373).

@R1 – "Regularization" vs. "soft constraint": Thank you for pointing this out. In order to avoid any possible confusion, we will follow your suggestion and replace the term "regularization" with "soft constraint" in the final paper.

@R1 – Assumption of DAGs: We agree that it is a huge assumption and will make it explicit in all relevant sections.

@R1 – Using a more effective variant of PC: We experimented with the original PC and its variant CPC (Conservative PC), and picked the latter which gave better results. We will include the detail in the revised paper. Thank you.

@R2 – Choice of regularization coefficients: Thank you for the insightful comment. For NOTEARS-L1, we experimented with several choices, and picked the one which yielded the best results (L570-573). For GOLEM, our focus is not to attain the best possible accuracy with the optimal hyperparameters, but rather to empirically validate our theoretical results. We thus did not perform a thorough hyperparameter search, and found that small coefficients suffice to work well (L588-590). We will include more details on this.

@R2 – Proof of Lemma 1: We will include the proof in the final version. Thank you.

@R2 – Regularization coefficients in asymptotic proof: Thanks a lot for this suggestion, which helps improve our presentation. We chose the sparsity term such that it scales with the order of $\mathcal{O}(\log n/n)$, as in the consistency result of BIC score. We will clarify this in the revised paper before Theorem 1 (L153).

@R3 – "puts much emphasis on beating NOTEARS" and "comes with no little surprises for more optimal ML loss is used rather than square loss from NOTEARS": Thank you for the comment. We would like to respectfully remind R3 that our main contributions include establishing the *asymptotic guarantees* and showing that soft DAG constraint is, under mild assumption, sufficient to recover the underlying DAGs (Section 3.1). The superior performance of GOLEM came as a *by-product* of our theoretical results.

@R3 – "transforming DAG constraint into a penalty is such an advantage w.r.t. NOTEARS is questionable" and "Is there a more efficient way for implementing constraint minimization than augmented Lagrangian": We would like to make it clear that our method *does not* involve augmented Lagrangian, since we proposed an easier *unconstrained* problem with a soft DAG constraint term (L213-216). This is entirely different from the constrained formulation of NOTEARS which increases the penalty to very large values, leading to optimization difficulties (L79-81). We also provided asymptotic guarantees and extensive empirical results to justify our motivation.

@R3 – Impact of sparsity on NOTEARS: It already includes a sparsity term, denoted as NOTEARS-L1 in Section 5.

@R4 – Causality and broader impact section: Thanks for pointing this out. We will remove the claim of "causality".

@R4 – Use of SHD: We use SHD to measure the discrepancy between DAGs, and SHD-CPDAG for equivalence classes (L624-628), which is what R4 nicely suggested. We will make these terminologies clear in the revised version.

We hope that the above discussion has addressed the reviewers' concern, and will incorporate all suggestions in the final version. We once again appreciate the reviewers' time dedicated to reviewing our paper.