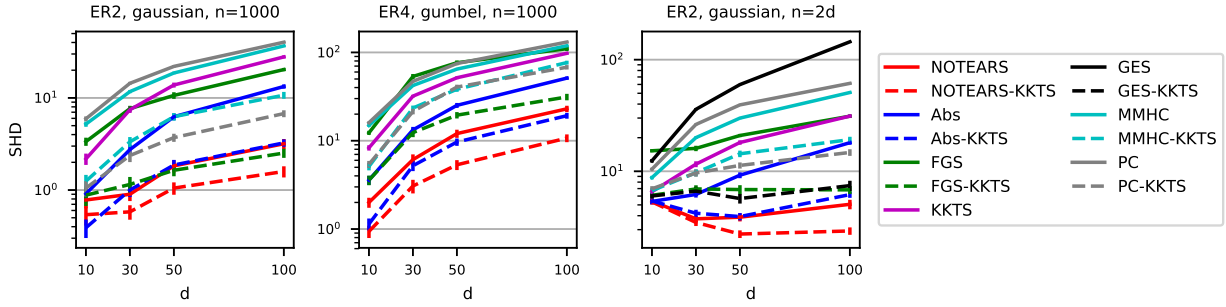
We thank the reviewers for their efforts. We are glad that R1, R2, R3 appreciated the improved understanding of NOTEARS ("useful negative results"). Thanks to R2 for describing our theoretical insights in general as "a potentially useful side effect for others" and "surprisingly readable for a paper with lots of theoretical results." R1, R2, R4 recognize our empirical results to be "successful" and "a significant improvement." Below we respond to reviewer comments.

R3: Questionable that NOTEARS, FGS outperform earlier methods, [2, Table 1] shows MMHC, PC perform much better than GES (here FGS) for moderate d. Thank you for pointing this out. We agree that the statement on line 276 is too broad and will remove it. To address R3's concern, we first compared with MMHC and PC in the experimental setting of Sec. 5. The significance level α was chosen from the range considered in [2] to minimize SHD. The two left panels below show that while MMHC and PC do not perform better than FGS, they are also significantly improved by KKTS (we will report full results in the paper). We then performed a second experiment with $n = 2d$ to be closer to the setting of [2, Table 1], also adding the GES implementation used in [2]. The right panel shows that FGS is actually an improvement over GES, remaining better than MMHC and PC (except $d = 10$) while GES is worse.



R3: "Paper is fairly incremental, developing a single heuristic local search method (namely NOTEARS that enjoys no non-trivial performance guarantees)." We naturally disagree about incrementality and are not sure what is meant by non-trivial. Prop. 3 provides a negative guarantee for NOTEARS (which is *not* our method), whereas Thms 9 and 7 provide positive guarantees for our KKTS method to yield KKT points and local minima. We would thus not call KKTS heuristic or lacking guarantees. To get from Prop. 3 to KKTS requires several more contributions: reformulating the problem and proving that KKT conditions are necessary (Thm 6), and relating the KKT conditions to edge absence constraints (Lem. 6, Thm 8). Sec. 2 makes additional contributions in generalizing acyclicity constraints from [32,30].

R4: Apparently only applicable to continuous case, no mention of categorical. More needed on limitations and position in literature. We thank R4 for prompting us to elaborate upon the problem setting and what remains for future work, which we will do in the paper. The theory and methods apply straightforwardly to binary variables (although we have not experimented with them) but not non-binary categorical variables. As Sec. 3, para. 1 states, the key assumption is that each edge is associated with a single parameter W_{ij}. In a (generalized) linear structural equation, a single parameter can account for the effect of a binary or continuous input, while a binary output can be handled by a suitable loss function (e.g. logistic). However, a single parameter is likely insufficient for a non-binary categorical input (typically encoded into multiple binary variables) or output (e.g. [14] proposes multi-logit regression with parameters for each level). Therefore the extension to multiple parameters per edge (Sec. 6) is desirable to address categorical variables as well as nonlinear models. **Abstract:** We will add a sentence on the one-parameter-per-edge assumption. **Title:** We find it difficult to capture this assumption in a few readily understood words, but perhaps R4 has a suggestion.

R1: "What leads to better or worse SHD...F (always squared error...danger of overfitting?), thresholding, acyclicity constraint and even centering." Thanks for the thoughtful questions. Below we summarize what we know/have reported. We think a proper exploration would best be left to a journal extension of this paper. **Score function F:** By keeping this as least squares, we have somewhat avoided overfitting to the noise type (Gaussian, Gumbel, etc., usually unknown) as opposed to using the log-likelihood. [32, Sec. 5.3] shows that NOTEARS can achieve scores close to those of the exact optimizer GOBNILP (especially before thresholding), i.e. it is fitting well but not over-fitting, but we have not done a similar comparison for NOTEARS-KKTS. **Acyclicity constraint:** The NOTEARS vs. Abs comparison shows that setting $A = W \circ W$ is empirically superior to $A = |W|$. As for the function h, Appendix C.1 states that the polynomial $h(A)$ from [30] performs slightly better and is slightly faster than the exponential $h(A)$ from [32] (the authors of [32] seem to agree as their code now uses polynomial h). Section 6 mentions that other h in the class of eq. (1) could be explored in future work. **Thresholding:** We followed [32] and fixed the threshold $\omega = 0.3$ for NOTEARS and for our NOTEARS-inspired algorithms (Abs, KKTS), in part to demonstrate success without too much parameter tuning. But we agree that the role of thresholding could be further explored. **Centering:** We were also surprised by the effect this had, as reported in Appendix C.3.

R1: "Fully continuous" is overstatement. We will remove "fully". Note that NOTEARS [32] also uses thresholding.