We thank all three reviewers for their time and feedback. Below we have done our best to respond to the major concerns.

A common concern was the surprising simplicity of the method. It is crucial to note that prior to our work, there was *no provably poly-time algorithm with sample complexity guarantees* for this problem in a nonparametric setting. It is difficult to overstate this point: Our analysis is completely model-free and comes with explicit guarantees. Existing analyses fail on nonparametric models (see **(A1)**), and existing nonparametric methods do not come with poly-time guarantees. This is the main contribution of our work, to resolve this open problem and prove that such an algorithm exists (L31-33; see also L1-4; L24-29; L36-37; L71-73; L181-182; L198-L202). Despite the similarity to existing work, our algorithm is *not* the same, and indeed several subtle but crucial changes were made to eliminate reliance on linearity, additivity, and independence of noise (see **(A1)**).

(A1) [R1, R4] Incremental. The analysis from existing work fails on nonparametric models, and our analysis is *completely* different—we analyze a *different* algorithm with a *different* technical approach. We must contrast Alg 1 and Alg 2: Alg 1 is a direct translation of existing algorithms (refs [5,11,12,32]) for which existing proofs *fail* when applied to nonparametric models, whereas Alg 2 contains crucial changes to adapt to the nonparametric setting, namely the use of the layer decomposition and sample splitting. We can show with an explicit counterexample that existing proofs would not generalize to Alg 1: If Alg 1 is applied to the null DAG model (no edges), then one needs to bound the estimation error for all $d2^{d-1}$ possible residual variances (this example is not unique or pathological; any DAG with more than one sort will have similar issues). This is subtle: Essentially, Alg 1 randomly chooses $O(d^2)$ parameters based on the data (note the estimated $\widehat{A}_j$ in Alg 2). This is precisely the reason for modifying Alg 1 into Alg 2: By learning the DAG layer-by-layer, this combinatorial explosion is avoided. By contrast, existing work crucially relies on linearity to write the residual variances in terms of the covariance matrix Σ (L167-171) in order to bound all $d2^{d-1}$ choices uniformly. For nonparametric models, there is no such representation via Σ, and each residual variance must be estimated separately. Regrettably this discussion was missing and we will add it to the camera ready version.

Furthermore, our analysis is nontrivial in several aspects: Our main results do not depend on any specific regression estimator, which uses several interesting tools (e.g. log-convexity and interpolation of L^p norms; see Appendix C.4). This makes our results more practical. The proof of Theorem 4.1 is also completely different from related work on equal variances, and informs the modifications made in Alg 2. See **(A2)** for a discussion of the novelty of Theorem 3.1.

(A2) [R1, R4] Relation to prior work. Prior work on equal variances crucially relies on linearity and independent, additive noise; see L102-105. Linearity is not crucial for identifiability, but is leveraged extensively to obtain statistical guarantees; please see **(A1)** for more details. One of our contributions is to show that these assumptions can be *completely* removed, without qualification: Our results apply to arbitrary nonlinear models with correlated, non-additive noise. For example, although the proof of Theorem 3.1 is straightforward, it is not quite a "simple generalization" of existing work: Our proof is completely different, and proves something much stronger using only the Markov property of BNs. We emphasize that existing results *completely miss this*, arguably because they rely on independence and additivity of noise in a *crucial* way (L96-101), though it is not needed. Faithfulness is not required by previous work on equal variances, but is commonly assumed in other work on BNs. We are happy to add this discussion to the paper.

(A3) [R2, R3] Sparsity and sample complexity. As pointed out at L250-257, there are several ways to improve the sample complexity. The most direct approach is to use a more sophisticated estimator of $\mathbb{E}\,\mathrm{var}(X_\ell \mid X_A)$, for which faster (root-n) rates are available (ref [9]; see also L308-311). Another approach, as suggested by R2, is to assume some kind of sparsity: By using adaptive estimators such as RODEO [21] or GRID [13], the sample complexity will depend only on the sparsity of $f_{\ell j}(X_{A_j})$, i.e. $d^* = \max_j \max_{\ell \notin A_j} |\{k \in A_j : \partial_k f_{\ell j} \neq 0\}|$ (∂_k is the kth partial derivative). Here is another way that does not require adaptive estimation: Suppose $|L_j| \leq w$ and define $r^* := \sup\{|i-j| : e = (e_1, e_2) \in E, e_1 \in L_i, e_2 \in L_j\}$. Then $\delta^2 \asymp n^{-2/(2+wr^*)}$, and the resulting sample complexity depends on wr^* instead of d. For a Markov chain with $w = r^* = 1$ this leads to a substantial improvement.

(A4) [R3] Assumptions. Certainly our main assumption may not hold in some practical situations. We have gone to great lengths to address this: (a) Our results hold in *far greater generality than existing work*, not requiring linearity, additivity, independent noise, or faithfulness (L3-4, L24-25, L181-182); (b) Our algorithm provably recovers models that state-of-the-art algorithms fail to recover (Sec 3.3, Ex 5); (c) The main assumption can be substantially weakened to unequal variances (L151-157, App B.1); (d) We have additional experiments on misspecified models in App B.1.

(A5) [R4] Causality / title. Here we are following convention in the literature, which refers to this problem as "causal discovery", "causal DAG learning", etc. We are happy to add more details on this point, e.g. by including a discussion of causal assumptions such as minimality and noting that under our assumptions there is a unique causal DAG.

(A6) [R4] Guarantees. A notable drawback of existing work is the failure to accomplish *both* nonasymptotic statistical and algorithmic guarantees in nonparametric settings. We have devoted an entire section (Sec 3.3) to highlight this point (see also L198-202). Most estimators for nonparametric DAG models come with one or the other: Finite-sample statistical guarantees that lack efficiency guarantees, or poly-time guarantees without explicit sample complexities. We are happy to add additional comparisons to emphasize this point.