- We would like to start by thanking the authors for their comments and suggestions.

Reviewer # 1:

- We also believe that these two elements are the most important points that should be augmented. We will add an extended discussion about the condition C1 and its relation to the conditional Fisher information matrix and a detailed discussion with respect to previous work as a conclusion. This will be possible using the 9th page in the camera-ready version of the paper

Reviewer # 2:

- In this paper, we have chosen to focus solely on the theoretical aspects of this algorithm as we believe that a serious and exhaustive numerical study of the algorithm's performance deserves a dedicated study. For pairwise and binary variables, [12] shows that RISE beats conditional likelihood and is fast to run (although in [12], the optimization is performed using a second-order method which is much less efficient than what we propose). We expect GRISE to be a very efficient algorithm in practice as it contains RISE as a special case.

- The efficient way to verify the identifiability condition is currently in the Appendix D1 and not in the main text. If space permits, we will put it back in the main text.

- Thank you for the list of typos, we will amend them in the final submission.

Reviewer # 3:

- It may look like our paper is a straightforward generalization of [17], but only superficially. This is the result of a deliberate choice on our part to find a compromise to facilitate the comprehension of this study for readers who are both familiar and unfamiliar with the field. We understand that this choice may have the unfortunate consequence of downplaying the innovations presented in this paper. For instance, the local learnability condition looks like it is similar to the restricted strong convexity condition from [17] but this is not the case. Strong convexity (restricted or not) is related to the notion of curvature of a function and is therefore linked to the standard Euclidian ℓ_2-norm. It turns out that the GRISE loss function is not strongly convex independently of the dimension of the problem (this implies in turns that the number of samples cannot be $\log p$ according to the standard theory). With the local learnability condition, we are showing that by looking at the problem using different types of geometry (in our case it is mostly the ℓ_∞ geometry) we can bypass this issue. This approach has been demonstrated using GRISE, but it is in fact an important result in itself for it is applicable to the analysis of any estimator that minimizes an empirical risk. We also provide a recipe to efficiently verify this new concept of convexity that is easily applicable to other cost functions. With respect to other innovations presented in this paper, we would also mention Algorithm 1 which uses a recursive call to GRISE. A straightforward generalization of [17] will not work for reconstructing graphical models precisely because such a loss function is not strongly convex independently of the dimension.

- The parameter $\hat{\gamma}$ is a prior on on the strength of the parameters of the problem. This implies that it should be bigger than the actual γ of the system. So unlike the λ from [17], $\hat{\gamma}$ does not depend on the size of the system or on the number of samples. Ideally one would like to take $\hat{\gamma} = \gamma$, but γ is an unknown quantity a priori. This implies that in practice we would chose $\hat{\gamma}$ much larger than the expected γ. Happily we show that this has only a limited impact on the sample complexity and the runtime of the algorithm as they both scale polynomially with $\hat{\gamma}$ while it is exponential in γ (which is the best one can hope for). Regardless of the size of the problem, in practice one would not expect γ to be bigger than 15 as it means that the intrinsic randomness of the variables is less than 1ppm. There are several reasons for switching from a penalty (the λ form) to a constraint (the γ form). The main reason is that the penalty is designed to enforce sparsity and is adapted to graphical models with bounded degrees. By proposing a constrained version, we can reconstruct graphical models irrespective of their degree while in [17] dense graphs can require up to $\mathcal{O}(p^4 \log(p))$ samples. Another reason is that the very nice property of λ in [17] being independent of some of the problem parameters is lost for more general models. In fact λ becomes dependant of the parameter γ for graphical models that are not pairwise and binary.

- The graphical models considered here can have a more general form (arbitrary order, arbitrary alphabet, arbitrary parameterization, arbitrary graph structure) while in [17] it is only applicable to pairwise models with bounded degrees with binary variables using monomial basis functions. The tools for the analysis may looks superficially similar to [17] at first glance but are in fact based on a radically new concept that does not use convexity.

Reviewer # 4:

- In our study we focus essentially on the theoretical aspect of GM learning (we believe that the current paper is already very dense). We plan to perform an exhaustive numerical study of this algorithm in a future work.

- In algorithm 1, L is not exactly an input parameter for it is determined by the family of models that is under consideration and is given by the highest order of the basis functions in the family.