We thank all reviewers for their time and valuable comments.

Reviewer #1

We thank this reviewer for the positive feedback!

"The theoretical sample complexity is not significantly improved over previously-known methods."

The main contribution of our paper is to show that an existing and popular algorithm (i.e., group-sparse regularized logistic regression) actually gives the state-of-the-art performance (in a setting where alternative algorithms are being proposed). We view the sample complexity improvement over the dependence on k as a side benefit of our analysis.

"It would be interesting to see a more thorough empirical evaluation, to compare with the interaction screening method and in more settings."

The main contribution of our paper is theoretical. A thorough empirical evaluation of different algorithms is definitely an interesting direction for future research, and we believe is beyond the scope of our current paper. Nevertheless, we did an experiment comparing the performance of the following algorithms: ℓ_1-constrained logistic regression, RISE (regularized interaction screening estimator) and its variant logRISE [LVMC18], and the Sparsitron algorithm [KM17]. Our graph has diamond shape (Figure 1 of our paper), 10 variables and edge weight 0.2. We focus on Ising models, because RISE and logRISE *cannot* be used to learn graphical models with general alphabet. With 1500 samples, the fraction of successful runs out of 100 runs is: 92 (logistic regression), 90 (RISE), 93 (logRISE), and 53 (Sparsitron).

"Extend the method to higher-order MRFs." Intuitively, it should not be difficult to prove that ℓ_1-constrained logistic regression can recover the structure of binary t-wise MRFs. One can prove it by combining results from Section 7 of [KM17] and the following fact: the Sparsitron algorithm can be viewed as an online mirror descent algorithm that approximately solves an ℓ_1-constrained logistic regression. This observation is actually the starting point of our paper. For higher-order MRFs with non-binary alphabet, we conjecture that similar result can be proved for group-sparse regularized logistic regression. Extending the current proof/method to higher-order MRFs is definitely an interesting direction for future research. We will include this discussion in our paper.

Reviewer #2

"The presentation is quite technical...the Ising case seems to be enough to introduce the main idea...but a lot of space is devoted to the generalization to larger alphabet..."

In this paper we consider the general alphabet setting for two reasons:

- This shows that our proof technique is actually quite general and can be easily extended to the setting with non-binary alphabet. In fact, there is a one-to-one correspondence between the lemmas used in learning Ising models (Lemma 8, 1, 5) and the non-binary graphical models (Lemma 11, 2, 6).
- For learning non-binary graphical models, we see a benefit of using the group-sparse (i.e., the $\ell_{2,1}$-norm) constraint instead of the ℓ_1-norm constraint used in [KM17]: the sample complexity improves from k^5 to k^4. A more general statement holds (by following a proof similar to ours): for any $1 \leq p \leq 2$, the $\ell_{p,1}$-constrained logistic regression gives a $k^{3+2/p}$ dependence. The case of $p > 2$ requires a proof different from ours and it is interesting to see if one can get a better dependence on k in that case.

"Experiments are only presented for rather small examples (up to 14 variables, up to $k = 6$)."

The main contribution of our paper is to theoretically prove the state-of-the-art performance of an existing and popular algorithm (i.e., group-sparse regularized logistic regression), in a setting where alternative algorithms are being proposed. Large-scale empirical evaluation is an interesting direction, and we think is beyond the scope of our current paper.

The biggest problem with large-scale simulation is that efficiently sampling from large graphical models is difficult. In our experiments, the samples are generated as follows: 1) We first *exactly* compute the probability distribution defined by a graphical model with n variables and alphabet size k; 2) We then sample from this probability distribution. Because the distribution contains k^n probabilities, the above sampling procedure is only possible for small n and k. When n is large (e.g., $n \sim 100$), exactly computing the probability distribution is impossible, and Gibbs sampling needs to be used. The mixing time for Gibbs sampling can be very large [BM09]. Because of this reason, we believe that large-scale empirical evaluation of different learning algorithms is itself a contribution to this area of research.

Reviewer #3: We thank this reviewer for all the positive comments!

References

[KM17] Klivans, Adam and Meka, Raghu. Learning graphical models using multiplicative weights. *FOCS*, 2017.
[BM09] Montanari, Andrea and Bento, Jose. Which graphical models are difficult to learn? *NeurIPS*, 2009.
[LVMC18] Lokhov, A. Y. et al. Optimal structure and parameter learning of Ising models. *Science advances*, 2018.