Author Response: From Boltzmann Machines to Neural Networks and Back Again We thank the reviewers for their input. First, we answer a few high-level questions asked by the reviewers:

Relation to practical algorithms for training RBMs We view this as a first step in importing tools from graphical models to combine with neural net methods. These new tools should be useful to improve over classic heuristics like contrastive divergence training, which have sometimes disappointing performance in practice.

Experimental results The purpose of our experiment was to compare our method with others with similar complexity and limitations, like classical CD training of RBMs, to which this method compares reasonably well. For reasons of comparison, this model has some obvious handicaps, e.g. it only uses binary units so it has to view grey as a probability and thus cannot really understand textures, it has no convolutional structure, etc; this is a proof-of-concept experiment.

Motivation for Supervised RBM Learning As discussed in the related work section, the distributional assumptions in the literature under which we have good theoretical neural net learning results are unfortunately very narrow. Many results either depend poorly on key parameters, or rely very strongly on Gaussianity of the input which is unlikely to hold in practice. In comparison, assuming data comes from some natural family of graphical models may be a more reasonable assumption, since a lot of data in practice is clearly structured and energy-based methods have seen a lot of success in modern image processing and machine learning. Our approach can be viewed as theoretically understanding how learning about the input distribution could play a role in supervised learning tasks.

Context for Theorem 4 (Structure Learning of RBMs) There is a huge literature on structure learning in the context of graphical models with no latent variables (e.g. references 3-8 and many more). However, for latent variable models like RBMs there is much less theory. For RBMs, the main previous works here are the cited results of Bresler, Koehler, and Moitra and the work of Goel; these results both require ferromagneticity (non-negative interactions). By viewing the distribution on observed variables as an MRF, it is possible to use general MRF-learning results (as in [5] and [8]), but the runtime of these methods is fairly poor: it is $n^{O(d)}$ where d is the max degree of a hidden unit.

Reviewer 4 asked for more context as to when the ℓ_1-norm is smaller than the degree. Specifically in the context of RBMs, Hinton's guide [1] says on page 9 "Care should be taken to ensure that the initial weight values do not allow typical visible vectors to drive the hidden unit probabilities very close to 1 or 0 as this significantly slows the learning" and suggests very small edge weights for initialization – standard deviation 0.01 in his example. In the context of a d-sparse RBM, as $d \to \infty$ we need the edge weights to scale down if we want the typical input to a hidden unit to be size $O(1)$ (e.g. if the visible units behave like they are independent, we need the edges to scale like $1/\sqrt{d}$). So the ℓ_1 norm will be much smaller than the degree.

Why are the complexity parameters in Definition 2 natural? ℓ_1-norm is perhaps the most popular complexity measure in the literature on learning graphical models (see e.g. [8]) because ℓ_1 regularization encourages sparsity, and sparse graphical models allow for drawing more powerful inferences (about conditional independence, etc.) than dense ones do. Mathematically, it's also the most natural because the spins X and H live in the ℓ_∞ unit ball and ℓ_1 is the dual norm. Finally, ℓ_1-norm bounds are the main assumption studied in the sampling literature (see lines 209-220 and Remark 4).

Next, we answer the remaining technical questions asked by the reviewers:

- Reviewer 4 asked for justifications of the following statements about RBMs: 1) they can represent arbitrary distributions (line 191), and 2) their parameters are not identifiable (i.e. impossible to estimate even with an infinite amount of data). Note that these are both statements about the observed distribution X, since H is unobserved; i.e. the joint distribution of (X, H) is not arbitrary (it's an Ising model), however the marginal on X is arbitrary as long as we have enough hidden units. In prior work, [18] showed that any order r Markov Random Field can be represented as the distribution on the observed units of an RBM with hidden units of degree r. By the Hammersley-Clifford Theorem, every distribution $p(x)$ on $\{\pm 1\}^n$ with $p(x) \neq 0$ for all x is an order r MRF for some $r \leq n$, so such a distribution can be exactly represented as the marginal over observable units in the RBM. They also gave several examples of RBMs which have different parameters but represent the same distribution (e.g. by having hidden units which cancel out each others effects on the observed units), which proves non-identifiability.
- Reviewer 1 asked for references for the statement on line 57, that learning full-observed Ising models is well understood. Some recent references with results for learning Ising models under weak assumptions are [5-8].
- Answers to the other questions asked by reviewer 1: on lines 209-220, here d stands for the maximum degree of nodes in the RBM. In Theorem 2, ℓ is indeed the logistic loss. In Examples 1 and 2, η indeed represents the maximum η such that all 2-hop neighbors are η-nondegenerate.
- Reviewer 3 questions: On line 129, the feature map is all monomials of degree up to D, d is a typo. Line 100: n_V here is the number of visible units, n_H is the number of hidden units.