On advancing MCMC-based methods for Markovian data structures with applications to deep learning, simulation, and resampling

Andee Kaplan

October 25, 2016

An exposition on the propriety of restricted Boltzmann machines

What is this?

A Restricted Boltzman Machine (RBM) is an undirected probabilistic graphical model (for discrete or continuous random variables) with two layers, one hidden and one visible, with conditional independence within a layer [Smolensky, 1986].

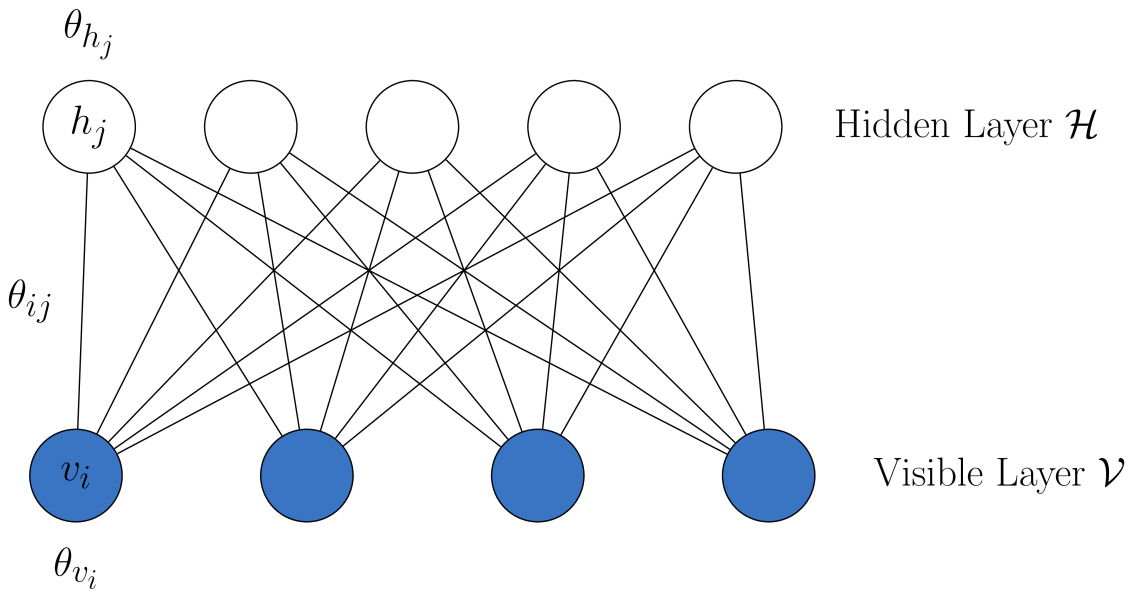

Figure 1: An example RBM, which consists of two layers. Hidden nodes are indicated by white circles and the visible nodes are indicated by blue circles

How is it used?

Typically used for image classification. Each image pixel is a node in the visible layer. The output creates features, which are passed to a supervised learning algorithm.

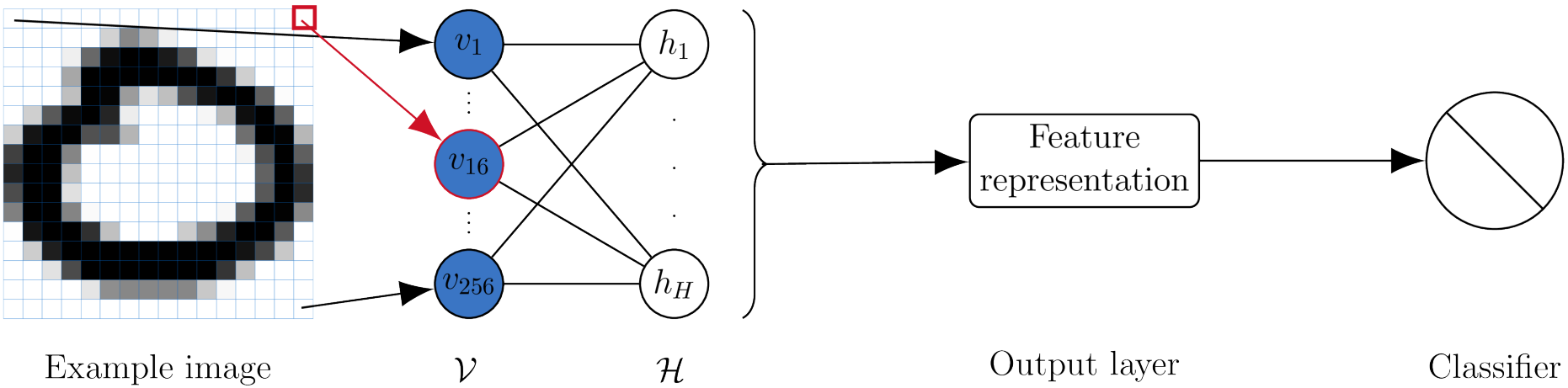

Figure 2: Image classification using a RBM. On the left, each image pixel comprises a node in the visible layer, $\mathcal{V}$. On the right, the output of the RBM is used to create features which are then passed to a supervised learning algorithm.

Joint distribution

Let $\boldsymbol{x} = h_1, \ldots, h_H, v_1, \ldots, v_V$ represent the states of the visible and hidden nodes in an RBM. We will consider both $\boldsymbol{x} \in \{0, 1\}^{H+V}$ and $\boldsymbol{x} \in \{-1, 1\}^{H+V}$.

A parametric form for probabilities corresponding to a potential vector of states of each node taking the value of 1.

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}{\gamma(\boldsymbol{\theta})} \tag{1}$$

Where

$$\gamma(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)$$

.

Deep learning

- By stacking layers of RBMs in a deep architecture, proponents of the models claim the ability to learn "internal representations that become increasingly complex, which is considered to be a promising way of solving object and speech recognition problems" [Salakhutdinov and Hinton, 2009, pp. 450].
- Stacking is achieved by treating a hidden layer of one RBM as the visible layer in a second RBM, etc.

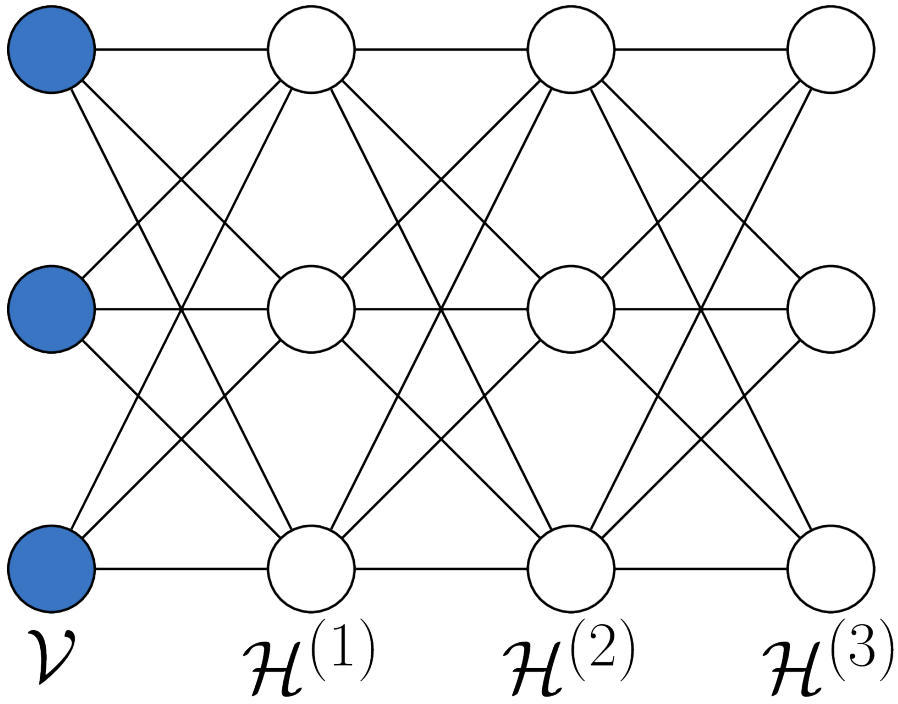

Figure 3: Three layer deep Boltzmann machine, with visible-to-hidden and hidden-to-hidden connections but no within-layer connections.

Why do I care?

- The model properties are largely unexplored in the literature and the commonly cited fitting methodology remains heuristic-based and abstruse.
- We want to
  1. provide steps toward a thorough understanding of the model and its behavior from the perspective of statistical model theory, and
  2. explore the possibility of a rigorous fitting methodology.

Degeneracy, instability, and uninterpretability. Oh my!

The highly flexible nature of a RBM (having as it does $H + V + HV$ parameters) makes three kinds of potential model impropriety of concern, *degeneracy*, *instability*, and *uninterpretability*.

> *A model should "provide an explanation of the mechanism underlying the observed phenomena" [Lehmann, 1990, Box and Hill [1967]].*

RBMs often fail to generate data that resemble realistic data and thus an unsatisfactory conceptualization of the data generation process. Additionally, we find that RBMs easily exhibit a kind of instability in the parameter space. In practice, this is seen when a single pixel change in an image results in a wildly different classification. Such model impropriety issues have been documented in other deep architectures recently [Szegedy et al., 2013, Nguyen et al. [2014]].

Near-degeneracy

Definition (Model Degeneracy)

There is a disproportionate amount of probability placed on only a few elements of the sample space, $\mathcal{X}$, by the model.

The RBM class of models exhibits *near-degeneracy* when random variables in the neg-potential function

$$Q(\boldsymbol{x}) = \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j,$$

have a mean vector $\mu(\boldsymbol{\theta})$ close to the boundary of the convex hull of $\mathcal{T}$ [Handcock et al., 2003], where
$t(\boldsymbol{x}) = \{v_1, \ldots, v_V, h_1, \ldots, h_H, v_1h_1, \ldots, v_Vh_H\} \in \mathcal{T}$ and the mean parameterization on the model parameters, $\mu(\boldsymbol{\theta})$ is

$$
\begin{aligned}
\mu(\boldsymbol{\theta}) &= \mathsf{E}_{\boldsymbol{\theta}}\, t(\boldsymbol{X}) \\
&= \sum_{x \in \mathcal{X}} \{ t(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \} \\
&= \sum_{x \in \mathcal{X}} \left\{ t(\boldsymbol{x}) \frac{\exp\left(\sum_{i=1}^{V} \sum_{j=1}^{H} \theta_{ij} v_i h_j + \sum_{i=1}^{V} \theta_{v_i} v_i + \sum_{j=1}^{H} \theta_{h_j} h_j \right)}{\sum_{\boldsymbol{x} \in \mathcal{X}} \exp\left(\sum_{i=1}^{V} \sum_{j=1}^{H} \theta_{ij} v_i h_j + \sum_{i=1}^{V} \theta_{v_i} v_i + \sum_{j=1}^{H} \theta_{h_j} h_j \right)} \right\}.
\end{aligned}
$$

Instability

Definition (Instability)

Characterized by excessive sensitivity in the model, where small changes in the components of potential data outcomes, $\boldsymbol{x}$, may lead to substantial changes in the probability mass function.

Schweinberger [2011] introduced a concept of model defficiency related to *instability* considering only a class of exponential families of distributions.

To quantify *instability* in the RBM, it is useful to imagine how a data model might be expanded to incorporate more and more observations. To increase the size of a RBM model to handle more "visible variables", it becomes necessary to expand the number of model parameters (and in the process of increasing the model size in visibles, one may also arbitrarily expand the number of hidden variables to be included).

Unstable RBMs

Definition (Unstable RBM)

A RBM model formulation is *unstable* if, as the number of visible variables increase ($V \to \infty$), it holds that

$$\lim_{V \to \infty} \frac{1}{V} ELPR_V(\boldsymbol{\theta}) = \infty.$$

where

$$ELPR_V(\boldsymbol{\theta}) = \log \left[\frac{\max_{(v_1,\ldots,v_V) \in \{-1,1\}^V} P_{\boldsymbol{\theta}}(v_1, \ldots, v_V)}{\min_{(v_1,\ldots,v_V) \in \{-1,1\}^V} P_{\boldsymbol{\theta}}(v_1, \ldots, v_V)} \right] \tag{2}$$

Unstable RBM models are undesirable for several reasons. One is that, as mentioned above, small changes in data can lead to overly-sensitive changes in probability.

One-pixel change

Consider, for example, the biggest log-probability ratio for a one-pixel (one component) change in data outcomes (visibles).

$$\Delta_V(\boldsymbol{\theta}) \equiv \max \left\{ \log \frac{P_{\boldsymbol{\theta}}(v_1, \ldots, v_V)}{P_{\boldsymbol{\theta}}(v_1^*, \ldots, v_V^*)} \right\},$$

where $(v_1, \ldots, v_V) \& (v_1^*, \ldots, v_V^*) \in \{-1, 1\}^V$ differ by exactly one component

Result

If $\frac{1}{V} ELPR_V(\boldsymbol{\theta}) > C$, *then* $\Delta_V(\boldsymbol{\theta}) > C$.

In other words, if the quantity (2) is too large, then the RBM model will exhibit large probability shifts for very small changes in the data configuration (i.e. instability mentioned earlier).

Tie to degeneracy

Unstable RBM models are connected to degenerate models (placing all probability on a small pieace of the sample space). Define a model set

$$M_{\epsilon,\boldsymbol{\theta}} \equiv \left\{ \boldsymbol{v} \in \{-1,1\}^V : \log P_{\boldsymbol{\theta}}(\boldsymbol{v}) > (1-\epsilon) \max_{\boldsymbol{v}^*} P_{\boldsymbol{\theta}}(\boldsymbol{v}^*) + \epsilon \min_{\boldsymbol{v}^*} P_{\boldsymbol{\theta}}(\boldsymbol{v}^*) \right\}$$

of possible outcomes, for a given $0 < \epsilon < 1$.

Result

For an unstable RBM model, and for any given $0 < \epsilon < 1$,
$P_{\boldsymbol{\theta}}\left((v_1, \ldots, v_V) \in M_{\epsilon,\boldsymbol{\theta}}\right) \to 1$ *holds as* $V \to \infty$.

In other words, as a consequence of unstable models, all probability will stack up on mode sets or potentially those few outcomes with the highest probability. Proofs of results 1-2 are provided in the appendix.

Computationally tractable lower bound

To numerically investigate unstable RBM models, it is useful to define a computationally tractable lower bound on (2) given by

$$\frac{1}{V}R(\boldsymbol{\theta}) \text{ for } R(\boldsymbol{\theta}) \equiv \max_{\boldsymbol{v}} \max_{\boldsymbol{h}} Q(\boldsymbol{x}) - \min_{\boldsymbol{v}} \max_{\boldsymbol{h}} Q(\boldsymbol{x}) - H \log 2.$$

That is, it holds that, in (2),

$$\frac{1}{V}\text{ELPR}_V(\boldsymbol{\theta}) \geq \frac{1}{V}R(\boldsymbol{\theta}). \tag{3}$$

(A proof of (3) appears in the appendix.)

Uninterpretability

Definition (Uninterpretability)

Characterized by marginal mean-structure not being maintained in the model due to dependence Kaiser [2007].

A measure of this is the magnitude of the difference between model expectations, $\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}]$, and expectations given independence, $\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}^*]$, where $\boldsymbol{\theta}^*$ is defined to equal $\boldsymbol{\theta}$ with all $\theta_{ij} = 0$ for $i = 1, \ldots, V, j = 1, \ldots, H$.

Using this concept it is possible to investigate what conditions lead to uninterpretability in a model versus those that guarantee interpretable models. A model is considered without uninterpretability if its expected value is not very different from the same model with assumed independence.

RBM quantities to compare

$$
\begin{aligned}
\mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}\right] &= \sum_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x} \frac{\exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}{\sum_{\boldsymbol{x}\in\mathcal{X}}\exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)} \\
&= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} x_k \exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}{\sum_{\boldsymbol{x}\in\mathcal{X}}\exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_i h_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}
\end{aligned}
$$

for $k = 1, \ldots, H + V$.

RBM quantities to compare

$$\mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}^*\right] = \sum_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x} \frac{\exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}{\sum_{\boldsymbol{x}\in\mathcal{X}}\exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}$$

$$= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} x_k \exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}{\sum_{\boldsymbol{x}\in\mathcal{X}}\exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right)}$$

for $k = 1, \ldots, H + V$.

If $|\mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}\right] - \mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}^*\right]|$ is large then the RBM with parameter vector $\boldsymbol{\theta}$ is *uninterpetable*.

Data coding to mitigate degeneracy

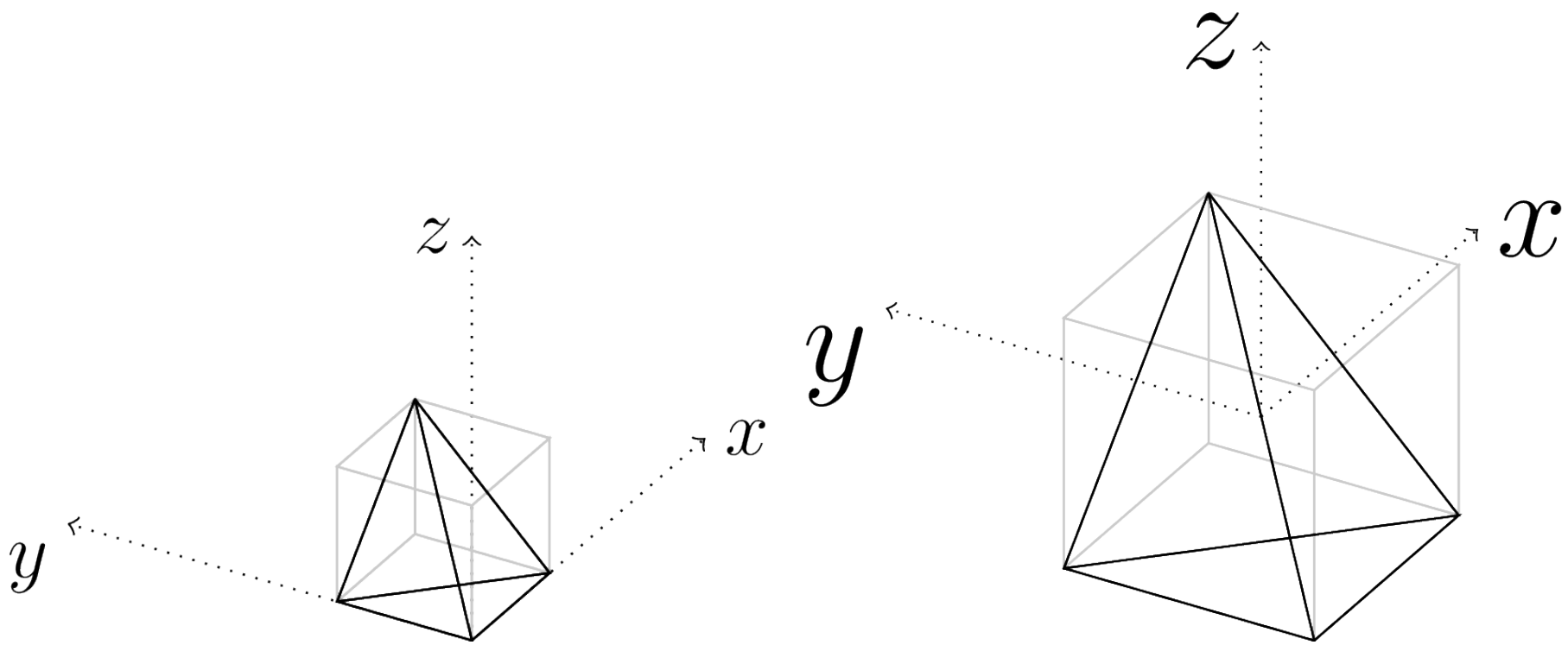

Figure 4: The convex hull of the "statistic space" in three dimensions for the toy RBM with one visible and one hidden node for $\{0, 1\}$ (left) and $\{-1, 1\}$ (right) data encoding. The convex hull of $\mathcal{T}$ does not fill the unit cube because of the relationship between statistics.

The center of the universe

- For the $\{-1, 1\}$ encoding of $\mathcal{V}$ and $\mathcal{H}$, the origin is the center of the parameter space
- At $\theta = 0$, RBM is equivalent to elements of X being distributed as iid Bernoulli$\left(\frac{1}{2}\right)$ $\Rightarrow$ No *near-degeneracy*, *instability*, or *uninterpretability*!

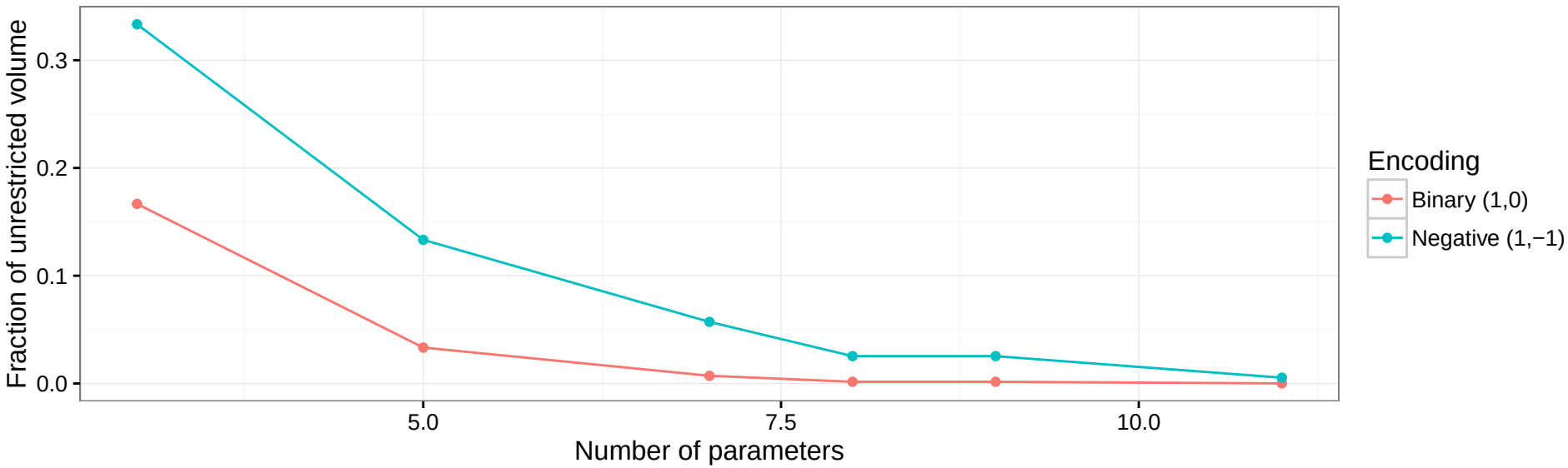

Figure 5: The relationship between volume of the convex hull of $\mathcal{T}$ and the cube containing it for different configurations of nodes.

Manageable (a.k.a. small) examples

- To explore the behavior of the RBM parameters $\boldsymbol{\theta}$ as it relates to near-degeneracy, instability, and uninterpretability, we consider models of small size. For $H, V \in \{1, \ldots, 4\}$, we sample 100 values of $\boldsymbol{\theta}$.
  1. Split $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_{interaction}$ and $\boldsymbol{\theta}_{main}$, in reference to which sufficient statistics the parameters correspond to.
  2. Allow the two types of terms to have varying average magnitudes, $||\boldsymbol{\theta}_{main}||/(H+V)$ and $||\boldsymbol{\theta}_{interaction}||/(H*V)$.
  3. Average magnitudes vary on a grid between 0.001 and 3 with 24 breaks, yielding 576 grid points.
- Calculate the three metrics of model impropriety, $\boldsymbol{\mu}(\boldsymbol{\theta})$, $R(\boldsymbol{x})/V$, and the coordinates of $|\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}] - \mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}^*]|$.
- In the case of *near-degeneracy*, we can go further and classify each model as near-degenerate or "viable" based on the distance of $\boldsymbol{\mu}(\boldsymbol{\theta})$ from the boundary of the convex hull of $\mathcal{T}$

Simulation results

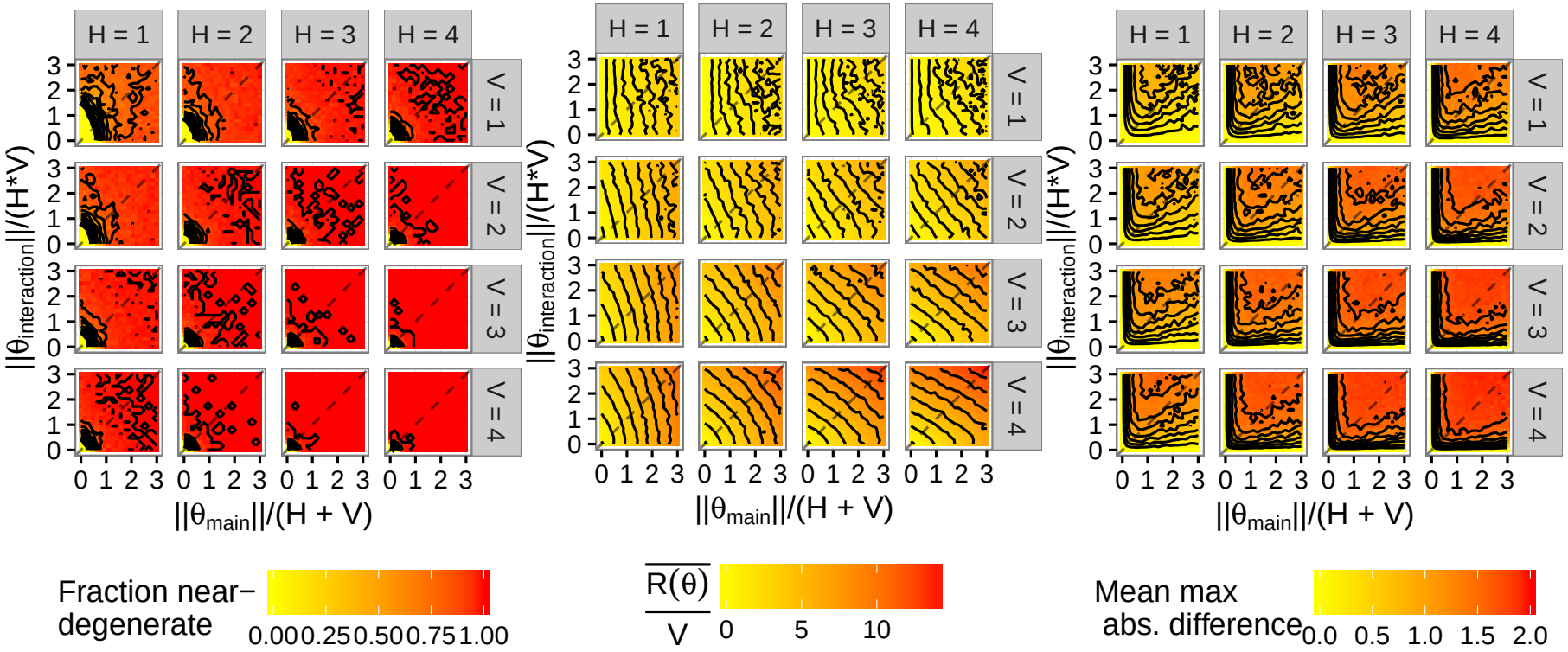

Figure 6: Results from the numerical experiment, here looking at the fraction of models that were near-degenerate (left), the sample mean value of $R(\boldsymbol{\theta})/V$ (middle), and the sample mean of the maximum component of the absolute difference between the model expectation vector, $\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}]$, and the expectation vector given independence, $\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}^*]$ (right).

Model fitting

- Typically, fitting a RBM via maximum likelihood (ML) methods will be infeasible mainly due to the intractibility of the normalizing term $\gamma(\boldsymbol{\theta})$ in a model of any realistic size
  - Ad hoc methods are used instead, which aim to avoid this problem by using stochastic ML that employ a small number of MCMC draws.
- Computational concerns are not the only issues with fitting an RBM using ML, the RBM model, has the potential to re-create any distribution for the data.
  - Based on a random sample of visible variables, the model for the cell probabilities that has the highest likelihood over *all possible model classes* is the empirical distribution, and the parametrization of the RBM model itself ensures that this empirical distribution can be arbitrarily well approximated.
  - Whenever the empirical distribution contains empty cells, fitting steps for the RBM model will aim to chase parameters that necessarily diverge in magnitude in order to zero out the corresponding RBM cell probabilities.

Bayesian methods

- We consider what might be done in a principled manner, testing on a $V = H = 4$ case that already stretched the limits of what is computable - in particular we consider Bayes methods.
- To avoid model impropriety for a fitted RBM, we want to avoid parts of the parameter space $\mathbb{R}^{V+H+VH}$ that lead to *near-degeneracy*, *instability*, and *uninterpretability*.
    - Shrink $\boldsymbol{\theta}$ toward $\mathbf{0}$ by specifying priors that place low probability on large values of $||\boldsymbol{\theta}||$, shrinking $\boldsymbol{\theta}_{interaction}$ more than $\boldsymbol{\theta}_{main}$.
- We considered a test case with $V = H = 4$ and parameters given in in appendix. This parameter vector was chosen as a sampled value of $\boldsymbol{\theta}$ that was not near the convex hull of the sufficient statistics for a grid point in figure 6 with $< 5\%$ near-degeneracy. We simulated $n = 5,000$ as a training set and fit the RBM using three Bayes methodologies.

Fitting methodologies

1. *A trick prior.* Here we cancel out normalizing term in the likelihood so that resulting full conditionals of $\boldsymbol{\theta}$ are multivariate Normal. The h_j are carried along as latent variables.

$$\pi(\boldsymbol{\theta}) \propto \gamma(\boldsymbol{\theta})^n \exp\left(-\frac{1}{2C_1}\boldsymbol{\theta}'_{main}\boldsymbol{\theta}_{main} - \frac{1}{2C_2}\boldsymbol{\theta}'_{interaction}\boldsymbol{\theta}_{interaction}\right),$$

   where $C_2 < C_1$. This is the method of Li [2014].
2. *A truncated Normal prior.* Use independent truncated spherical normal distributions as priors for $\boldsymbol{\theta}_{main}$ and $\boldsymbol{\theta}_{interaction}$ with $\sigma_{interaction} < \sigma_{main}$. Full conditionals are not conjugate, and simulaion from the posterior was accomplished using a geometric adaptive Metropolis Hastings step [Zhou, 2014] and calculation of likelihood normalizing constant. Here the h_j are carried along as latent variables.
3. *A truncated Normal prior and marginalized likelihood.* Marginalize out $\boldsymbol{h}$ in $f_\theta(\boldsymbol{x})$, and use the truncated Normal prior.

Hyperparameters

Table 1: The values used for the hyperparameters for all three fitting methods. A rule of thumb is imposed which decreases prior variances for the model parameters as the size of the model increases and also shrinks $\theta_{interaction}$ more than θ_{main}.

Method	Hyperparameter	Value
Trick Prior	C_1	$\frac{C}{N}\frac{1}{H+V}$
	C_2	$\frac{C}{N}\frac{1}{H*V}$
Truncated Normal	σ^2_{main}	$\frac{1}{H+V}$
	$\sigma^2_{interaction}$	$\frac{1}{H*V}$
Marginalized Likelihood	σ^2_{main}	$\frac{1}{H+V}$
	$\sigma^2_{interaction}$	$\frac{1}{H*V}$

Mixing

The truncated Normal method (2) and the marginalized likelihood method (3) are drawing from the same stationary posterior distribution for images.

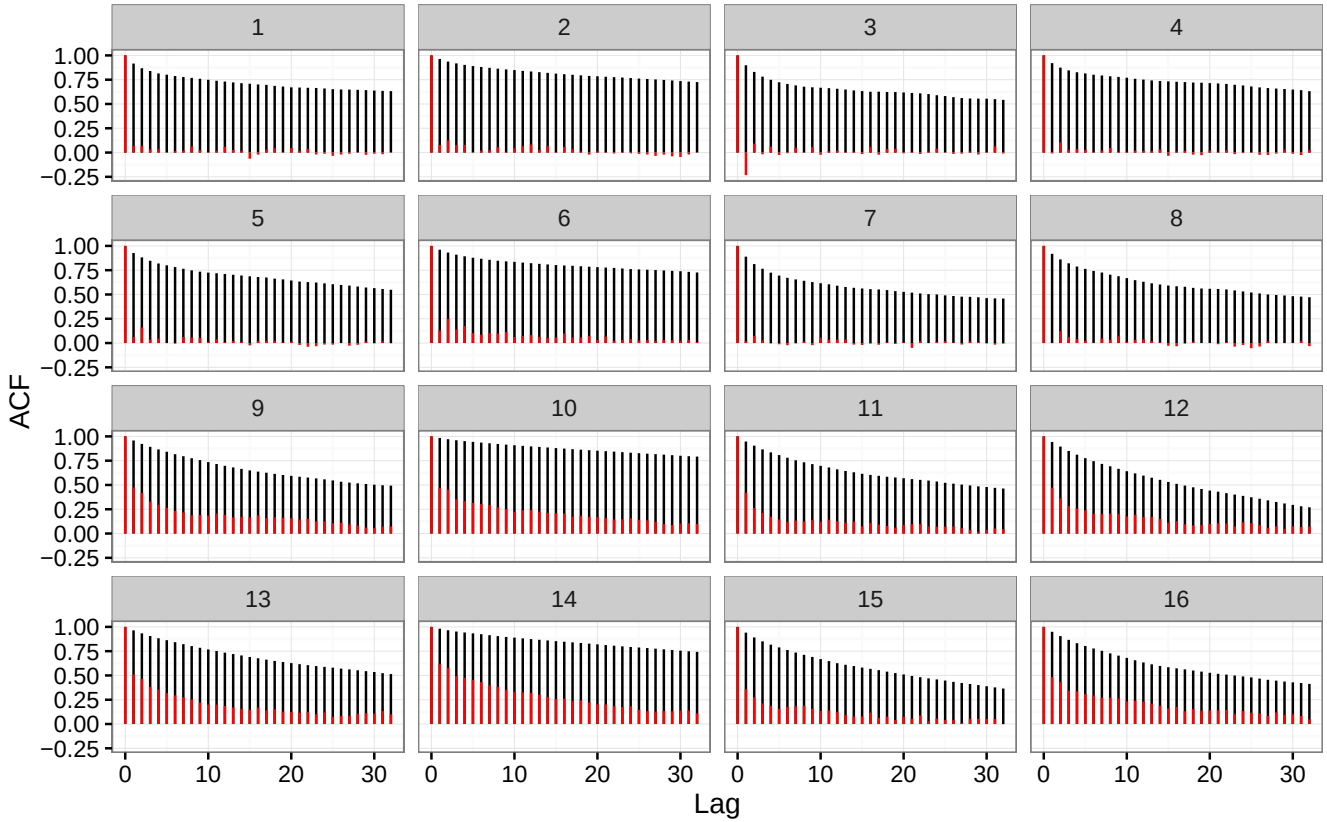

Figure 7: The autocorrelation functions (ACF) for each image with the truncated Normal method in black and the marginalized likelihood method in red.

Effective sample size

We can use an overlapping blockmeans approach to get a crude estimate for the aysmptotic variance of the probability of each image and compare it to an estimate of the asymptotic variance assuming IID draws from the target distribution.

Table 2: The effective sample sizes for a chain of length $M = 1000$ of all 16 images.

Image	Marginal Likelihood	Truncated Normal	Image	Marginal Likelihood	Truncated Normal
1	509.43	73.00	9	394.90	83.47
2	472.51	65.05	10	327.35	95.39
3	1229.39	87.10	11	356.56	70.74
4	577.73	72.64	12	338.30	81.40
5	452.01	71.67	13	373.59	105.98
6	389.78	66.49	14	306.91	132.61
7	660.37	84.30	15	365.30	82.15
8	515.09	75.46	16	304.57	98.05

Posterior distributions of images

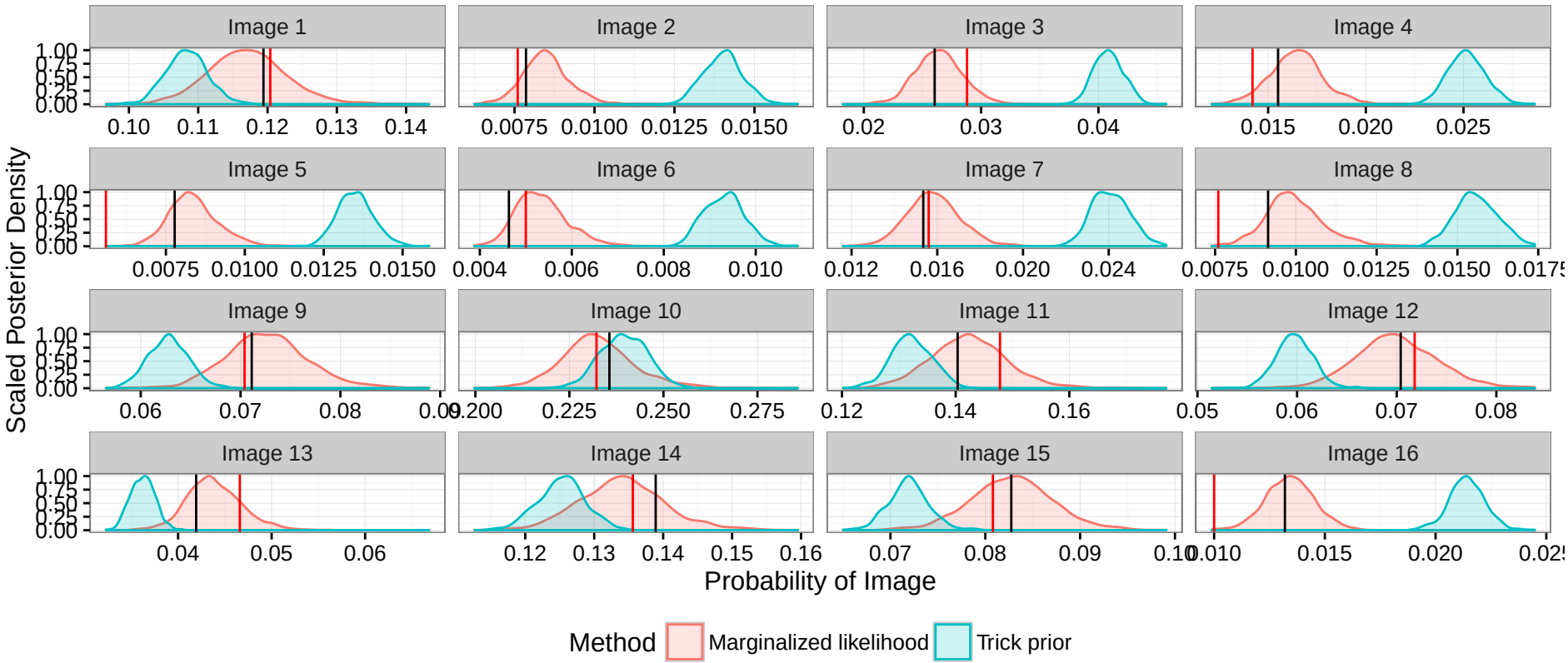

Figure 8: Posterior probability of each possible 4-pixel image using two of the three Bayesian fitting techniques, trick prior and marginalized likelihood. The black vertical lines show the true probabilities of each image.

Wrapping up

- While RBMs can be useful for classification, in the context of a statistical model as a representation of data, RBMs are a poor fit due to
  1. *near-degeneracy*,
  2. *instability*, and
  3. *uninterpretability*.
- Rigorous fitting methodology is possible but slow and as the size of the model grows, becomes intractible.
- There is no "smoothing" achieved with a RBM fit using a rigorous method, because any fully principled fitting method will reproduce the empirical distribution of most realistic training sets

$$\text{Concliques} + \text{Gibbs} = \text{Cool}^3$$

Future work

Ideas

- Using AdaBoost to create an ensemble learner for combining classifiers
- Goodness-of-fit test for network data
  - The model-based method of resampling re-frames network into a collection of neighborhoods by using covariate information
  - Creates concliques on a graph structure.
  - Use a conditionally specified network distribution to sample network data in a blockwise conclique-based Gibbs sampler.
- Non-parametric methods of resampling network data
  - Blockwise bootstrap, analogous to the blockwise bootstrap in time series, where network data blocks are created from the network by selecting a node and all nodes connected to it by an edge.
  - These blocks are then sampled with replacement and sewn together to re-create a total network by using iid Bernoulli draws (with an appropriate probability of success).
  - This method does not impose a neighborhood structure, nor particular distributional assumptions.

Appendix: Proof of Result 1

We prove the contrapositive. Suppose that $\Delta_V(\theta) \leq C$ holds for some $C > 0$. Under the RBM model for visibles, $P_\theta(\boldsymbol{v}) > 0$ holds for each outcome $\boldsymbol{v} \in \{-1, 1\}^V$. Let $\boldsymbol{v}_{min} \equiv \underset{\boldsymbol{v} \in \{-1,1\}^V}{\arg\min} \; P_\theta(\boldsymbol{v})$ and $\boldsymbol{v}_{max} \equiv \underset{\boldsymbol{v} \in \{-1,1\}^V}{\arg\max} \; P_\theta(\boldsymbol{v})$. Note there exists a sequence $\boldsymbol{v}_{min} \equiv \boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \equiv \boldsymbol{v}_{max}$ in $\{-1, 1\}^V$ of component-wise switches to move from $\boldsymbol{v}_{min}$ to $\boldsymbol{v}_{max}$ in the sample space (i.e. $\boldsymbol{v}_i, \boldsymbol{v}_{i+1} \in \{-1, 1\}^V$ differ by exactly 1 component for $i = 0, \ldots, k$) for some integer $k \in \{0, 1, \ldots, V\}$. Then

$$\begin{aligned}
\log\left[\frac{P_\theta(\boldsymbol{v}_{max})}{P_\theta(\boldsymbol{v}_{min})}\right] &= \left|\sum_{i=1}^{k} \log\left(\frac{P_\theta(\boldsymbol{v}_i)}{P_\theta(\boldsymbol{v}_{i-1})}\right)\right| \\
&\leq \sum_{i=1}^{k} \left|\log\left(\frac{P_\theta(\boldsymbol{v}_i)}{P_\theta(\boldsymbol{v}_{i-1})}\right)\right| \\
&\leq k\Delta_V(\theta) \leq VC
\end{aligned}$$

using $k \leq V$ and $\Delta_V(\theta) \leq C$.$\square$

Appendix: Proof of Result 2

Define $\boldsymbol{v}_{max}$ and $\boldsymbol{v}_{min}$ as in the proof of Proposition 1. Fix $0 < \epsilon < 1$ Then, $\boldsymbol{v}_{max} \in M_{\epsilon,\boldsymbol{\theta}}$, so $P_{\boldsymbol{\theta}}(M_{\epsilon,\boldsymbol{\theta}}) \geq P_{\boldsymbol{\theta}}(\boldsymbol{v}_{max})$. If $\boldsymbol{v} \in \{-1,1\}^V \setminus M_{\epsilon,\boldsymbol{\theta}}$, then by definition $P_{\boldsymbol{\theta}}(\boldsymbol{v}) \leq [P_{\boldsymbol{\theta}}(\boldsymbol{v}_{max})]^{1-\epsilon}[P_{\boldsymbol{\theta}}(\boldsymbol{v}_{min})]^{\epsilon}$ holds so that

$$\begin{aligned} 1 - P_{\boldsymbol{\theta}}(M_{\epsilon,\boldsymbol{\theta}}) &= P_{\boldsymbol{\theta}}(M_{\epsilon,\boldsymbol{\theta}}^C) \\ &= \sum_{\boldsymbol{v} \in \{-1,1\}^V \setminus M_{\epsilon,\boldsymbol{\theta}}} P_{\boldsymbol{\theta}}(\boldsymbol{v}) \\ &\leq (2^V)[P_{\boldsymbol{\theta}}(\boldsymbol{v}_{max})]^{1-\epsilon}[P_{\boldsymbol{\theta}}(\boldsymbol{v}_{min})]^{\epsilon} \end{aligned}$$

Then,

$$\begin{aligned} \frac{1}{V} \log \left[\frac{P_{\boldsymbol{\theta}}(M_{\epsilon,\boldsymbol{\theta}})}{1 - P_{\boldsymbol{\theta}}(M_{\epsilon,\boldsymbol{\theta}})} \right] &\geq \frac{1}{V} \log \left[\frac{P_{\boldsymbol{\theta}}(\boldsymbol{v}_{max})}{(2^V)[P_{\boldsymbol{\theta}}(\boldsymbol{v}_{max})]^{1-\epsilon}[P_{\boldsymbol{\theta}}(\boldsymbol{v}_{min})]^{\epsilon}} \right] \\ &= \frac{\epsilon}{V} \log \left[\frac{P_{\boldsymbol{\theta}}(\boldsymbol{v}_{max})}{P_{\boldsymbol{\theta}}(\boldsymbol{v}_{min})} \right] - \log 2 \to \infty \end{aligned}$$

as $V \to \infty$ by the definition of an unstable RBM model.□

Appendix: Proof of Equation (3)

$$\log\left[\frac{\max_{\boldsymbol{v}\in\{-1,1\}^V} P_{\boldsymbol{\theta}}\boldsymbol{v}}{\min_{\boldsymbol{v}\in\{-1,1\}^V} P_{\boldsymbol{\theta}}\boldsymbol{v}}\right] = \log\left[\frac{\max_{\boldsymbol{v}} \sum_{\boldsymbol{h}\in\{-1,1\}^H} \exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j\right)}{\min_{\boldsymbol{v}} \sum_{\boldsymbol{h}\in\{-1,1\}^H} \exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j\right)}\right]$$

$$\geq \log \frac{\max_{\boldsymbol{v}} \max_{\boldsymbol{h}\in\{-1,1\}^H} \exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j\right)}{\min_{\boldsymbol{v}} 2^H \max_{\boldsymbol{h}\in\{-1,1\}^H} \exp\left(\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j\right)}$$

$$= \max_{\boldsymbol{v}} \max_{\boldsymbol{h}\in\{-1,1\}^H} \left\{\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j\right\} -$$

$$\min_{\boldsymbol{v}} \max_{\boldsymbol{h}\in\{-1,1\}^H} \left\{\sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j + \sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j\right\} - H\log 2$$

$$\equiv R(\boldsymbol{\theta})$$

using above the monotonicity of $\log(\cdot)$.

Appendix: Parameters used

Table 3: Parameters used to fit a test case with $V = H = 4$. This parameter vector was chosen as a sampled value of $\boldsymbol{\theta}$ that was not near the convex hull of the sufficient statistics for a grid point in figure 6 with $< 5\%$ near-degeneracy.

Parameter	Value	Parameter	Value	Parameter	Value
θ_{v1}	-1.1043760	θ_{11}	-0.0006334	θ_{31}	-0.0038301
θ_{v2}	-0.2630044	θ_{12}	-0.0021401	θ_{32}	0.0032237
θ_{v3}	0.3411915	θ_{13}	0.0047799	θ_{33}	0.0020681
θ_{v4}	-0.2583769	θ_{14}	0.0025282	θ_{34}	0.0041429
θ_{h1}	-0.1939302	θ_{21}	0.0012975	θ_{41}	0.0089533
θ_{h2}	-0.0572858	θ_{22}	0.0000253	θ_{42}	-0.0042403
θ_{h3}	-0.2101802	θ_{23}	-0.0004352	θ_{43}	-0.0000480
θ_{h4}	0.2402456	θ_{24}	-0.0086621	θ_{44}	0.0004767

References I

G. E. P. Box and W. J. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

Mark S Handcock, Garry Robins, Tom AB Snijders, Jim Moody, and Julian Besag. Assessing degeneracy in statistical models of social networks. Technical report, Working paper, 2003.

Mark S Kaiser. Statistical dependence in markov random field models. *Statistics Preprints*, Paper 57, 2007. URL `http://lib.dr.iastate.edu/stat_las_preprints/57/`.

E. L. Lehmann. Model specification: The views of fisher and neyman, and later developments. *Statistical Science*, 5(2):160–168, 1990.

Jing Li. *Biclustering methods and a Bayesian approach to fitting Boltzmann machines in statistical learning*. PhD thesis, Iowa State University, 2014. URL `http://lib.dr.iastate.edu/etd/14173/`.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897*, 2014. URL `http://arxiv.org/abs/1412.1897`.

Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

Michael Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. URL `http://arxiv.org/abs/1312.6199`.

Wen Zhou. *Some Bayesian and multivariate analysis methods in statistical machine learning and applications*. PhD thesis, Iowa State University, 2014. URL `http://lib.dr.iastate.edu/etd/13816/`.