An exposition on the propriety of restricted Boltzmann machines

Andee Kaplan, Daniel Nordman, Stephen Vardeman
Department of Statistics, Iowa State University

Deep learning

Three layer deep Boltzmann machine, with visible-to-hidden and hidden-to-hidden connections but no within-layer connections. This can be considered as multiple single layer restricted Boltzmann machines with the lower stack hidden layer acting as the visible layer for the higher stacked model. Claimed ability to learn "internal representations that become increasingly complex" [5], used in classification problems.

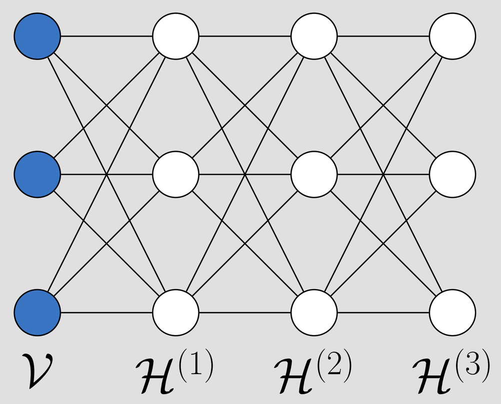

Figure 3: Deep RBM example.

Degeneracy, instability, and uninterpretability... Oh my!

The highly flexible nature of the RBM ($H+V+HV$ parameters) makes the following characteristics of model impropriety of particular concern.

Characteristic	Detection
Near-degeneracy. Occurs when there is a disproportionate amount of probability placed on only a few elements of the sample space by the model [2].	If the mean parametrization on the model parameters, $\mu(\boldsymbol{\theta})$, is close to the boundary of the convex hull of the set of statistics in the neg-potential function $Q(\boldsymbol{x})$.
Instability. Small changes in natural parameters result in large changes of the pmf, excessive sensitivity [6].	If for any $C > 0$ there exists $N_C > 0$ such that $\max_{\boldsymbol{x}_N \in \mathcal{X}_N} [Q(\boldsymbol{x}_N)] > CN$ for all $N > N_C$. Where $Q(\cdot)$ is the neg-potential function of the model.
Uninterpretability. Due to the existence of dependence, marginal mean-structure no longer maintained [3].	If the magnitude of the difference between model expectations and expectations under independence, $\lvert \mathrm{E}(\boldsymbol{X}\mid\boldsymbol{\theta}) - \mathrm{E}(\boldsymbol{X}\mid\emptyset)\rvert$, is large.

Table 1: Table of "improper model" characteristics.

Avoiding degeneracy

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

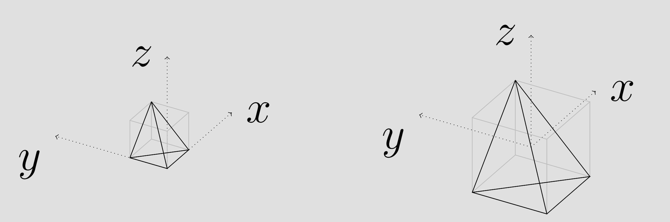

Figure 4: The convex hulls of the statistic space in three dimensions for a toy RBM with $|\mathcal{V}| = |\mathcal{H}| = 1$ for $\{0,1\}$-encoding (left) and $\{-1,1\}$-encoding (right) enclosed by an unrestricted hull of 3-space.

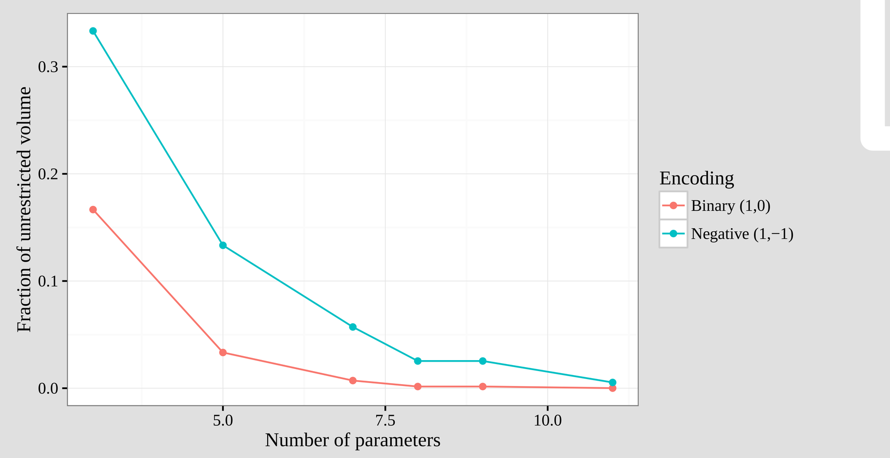

Figure 5: Volume relationship for the convex hulls of statistics in $Q(\cdot)$ vs. unrestricted space.

andeekaplan.com
ajkaplan@iastate.edu
github.com/andeek

Restricted Boltzmann machine (RBM)

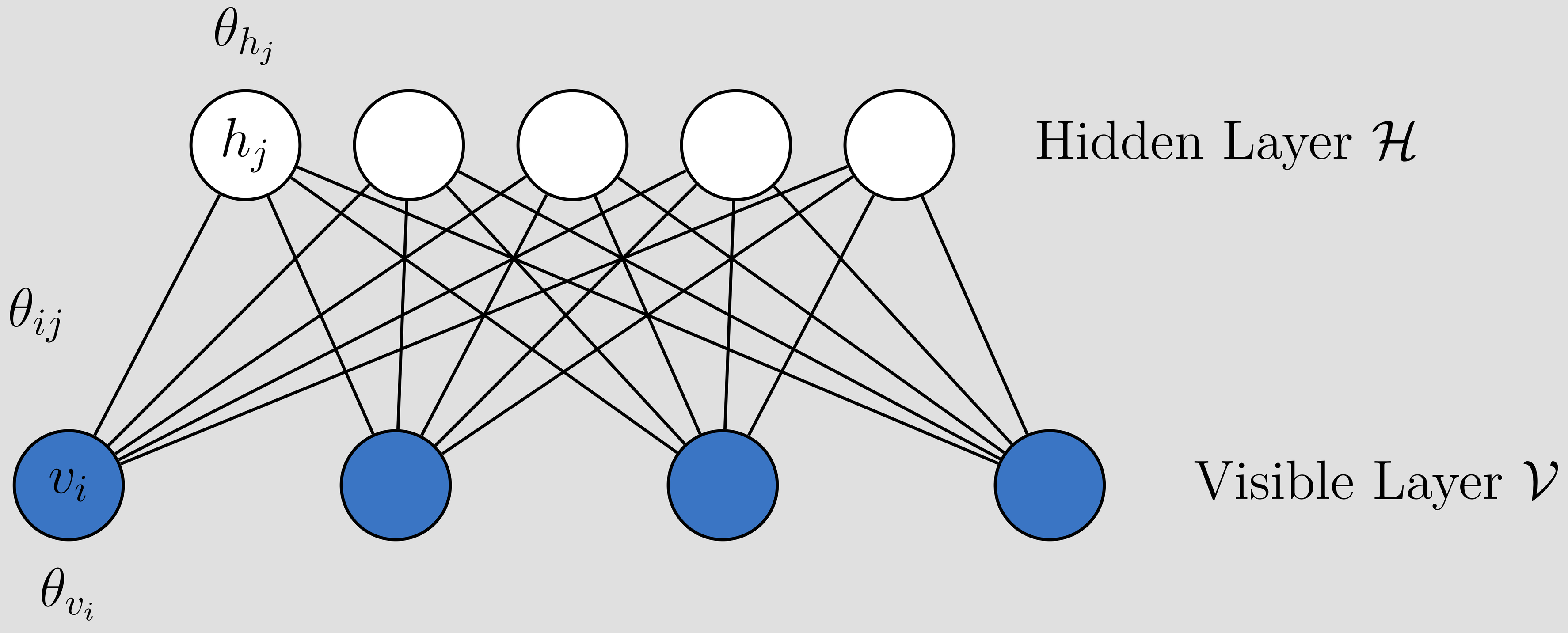

Figure 1: An example restricted Boltzmann machine (RBM), which consists of two layers, a hidden ($\mathcal{H}$) and a visible layer ($\mathcal{V}$), with no connections within a layer. Hidden nodes indicated by white circles and the visible nodes indicated by blue circles [1].

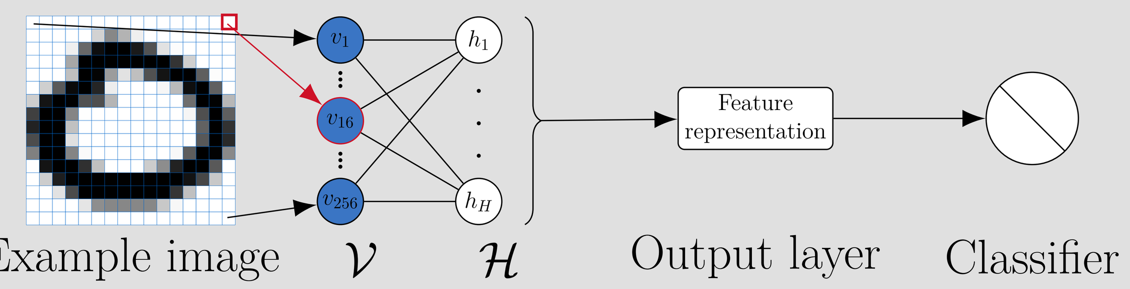

Figure 2: Image classification using a RBM. On the left, each image pixel comprises a node in the visible layer, $\mathcal{V}$. On the right, the output of the RBM is used to create features which are then passed to a supervised learning algorithm.

Joint distribution

Let $\boldsymbol{x} = \{h_1, \dots, h_H, v_1, \dots, v_V\}$ represent the states of the visible and hidden nodes in an RBM. Then the probability each node taking the the value corresponding to $\boldsymbol{x}$ is:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(\sum\limits_{i=1}^{V}\sum\limits_{j=1}^{H}\theta_{ij}v_ih_j + \sum\limits_{i=1}^{V}\theta_{v_i}v_i + \sum\limits_{j=1}^{H}\theta_{h_j}h_j\right)}{\sum\limits_{\boldsymbol{x}\in\mathcal{X}}\exp\left(\sum\limits_{i=1}^{V}\sum\limits_{j=1}^{H}\theta_{ij}v_ih_j + \sum\limits_{i=1}^{V}\theta_{v_i}v_i + \sum\limits_{j=1}^{H}\theta_{h_j}h_j\right)} \tag{1}$$

References

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[2] Mark S Handcock et al. *Assessing degeneracy in statistical models of social networks*. Tech. rep. Working paper, 2003.

[3] Mark S Kaiser. "Statistical Dependence in Markov Random Field Models". In: *Statistics Preprints* Paper 57 (2007). URL: http://lib.dr.iastate.edu/stat_las_preprints/57/.

[4] 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/.

[5] Ruslan Salakhutdinov and Geoffrey E Hinton. "Deep boltzmann machines". In: *International Conference on Artificial Intelligence and Statistics*. 2009, pp. 448–455.

[6] Michael Schweinberger. "Instability, sensitivity, and degeneracy of discrete exponential families". In: *Journal of the American Statistical Association* 106.496 (2011), pp. 1361–1370.

[7] 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/.

Manageable (a.k.a. small) examples

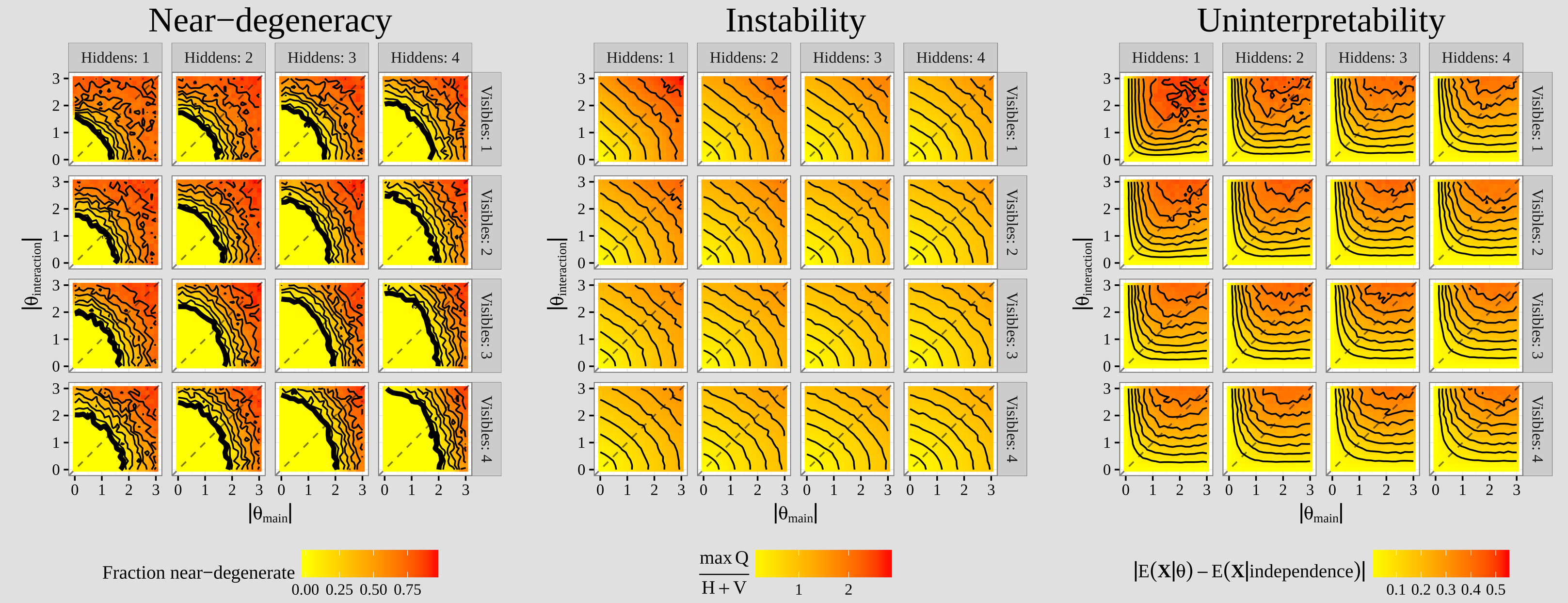

Figure 6: As the magnitude of $\boldsymbol{\theta}$ grows, so does the occurence of near-degeneracy, instability, and uninterpretability for RBMs of varying sizes.

Bayesian model fitting

Idea: To avoid model impropriety, avoid parts of the parameter space that lead to *near-degeneracy*, *instability*, and *uninterpretability* (i.e., shrink $\boldsymbol{\theta}$).

Simulated $n = 5{,}000$ images (4 pixels) from RBM model with 4 hiddens then fit using Bayesian methods,

- **Trick prior.** Cancel out the normalizing term, resulting full conditionals are normally distributed. **Conclusion:** Scalable solution, but requires tuning.

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

where $\gamma(\boldsymbol{\theta}) = \sum\limits_{\boldsymbol{x}\in\mathcal{X}}\exp\left(\sum\limits_{i=1}^{V}\sum\limits_{j=1}^{H}\theta_{ij}v_ih_j + \sum\limits_{i=1}^{V}\theta_{v_i}v_i + \sum\limits_{j=1}^{H}\theta_{h_j}h_j\right)$ and $C_2 < C_1$ [4].

- **Truncated Normal prior.** Use two independent truncated spherical normal distributions as priors for θ_{main} and $\theta_{interaction}$ with $\sigma_{interaction} < \sigma_{main}$. Full conditional distributions are not conjugate, requires a geometric adaptive MH step [7] and calculation of likelihood normalizing constant. **Conclusion:** Computationally intensive and convergence issues.

- **Marginalized likelihood.** Marginalize out $\boldsymbol{h}$ in $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and use the truncated Normal prior. **Conclusion:** Least scalable, but removes need to gain MCMC convergence for Hn sampled hidden nodes.

$$g_{\boldsymbol{\theta}}(\boldsymbol{v}) = \sum_{\boldsymbol{h}\in\{-1,1\}^H}\exp\left(\sum_{i=1}^{V}\sum_{j=1}^{H}\theta_{ij}v_ih_j + \sum_{i=1}^{V}\theta_{v_i}v_i + \sum_{j=1}^{H}\theta_{h_j}h_j\right).$$

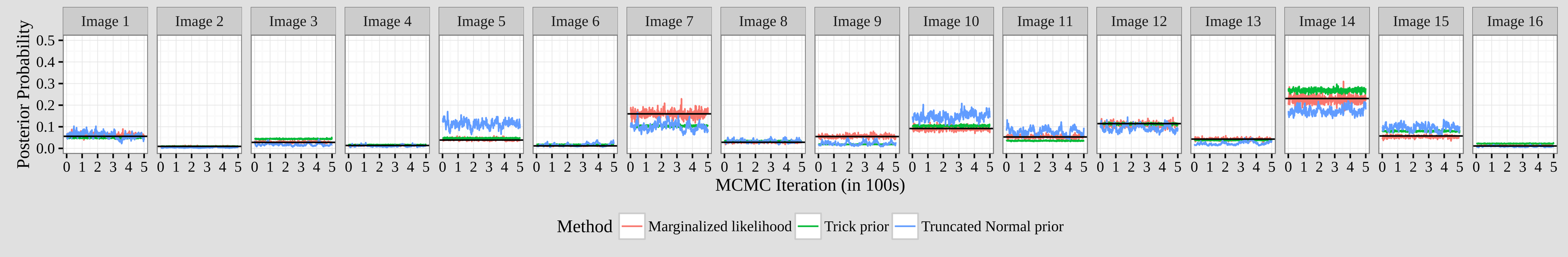

Figure 7: Posterior probability of each possible 4-pixel image using priors above.

Big takeaway: RBMs very easily are degenerate, unstable, and uninterpretable. To further complicate things, a rigorous fitting method for these models is not scalable and replicates the nonparametric solution (empirical distribution).