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

Andee Kaplan

Octover 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].

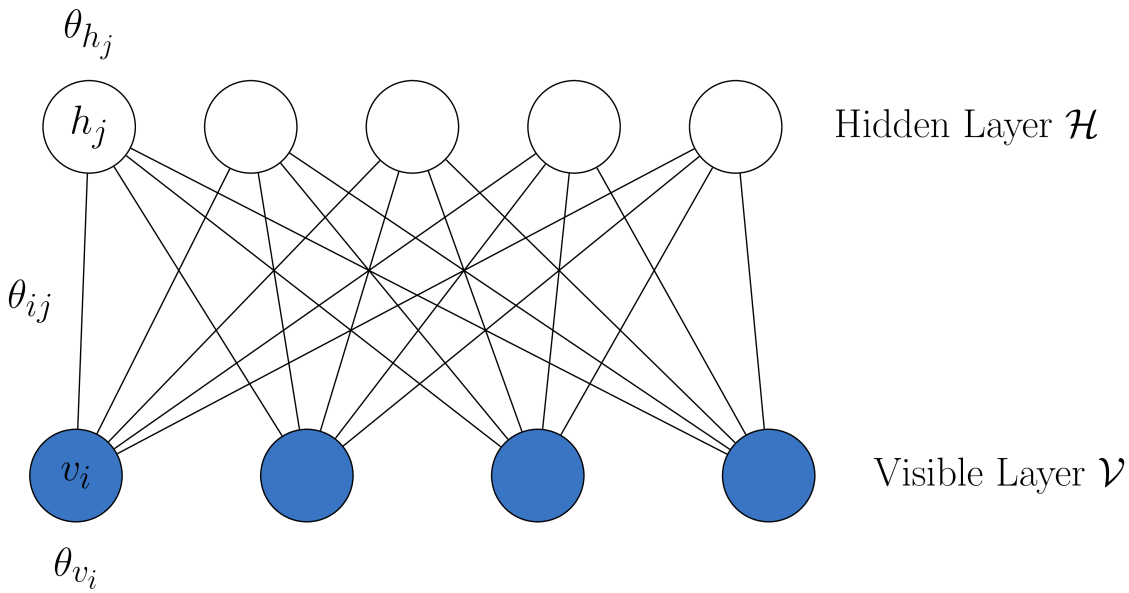

Figure 1: An example RBM, which consists of two layers. Hidden nodes are indicated by white circles and the visible nodes are indicated by blue circles

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.

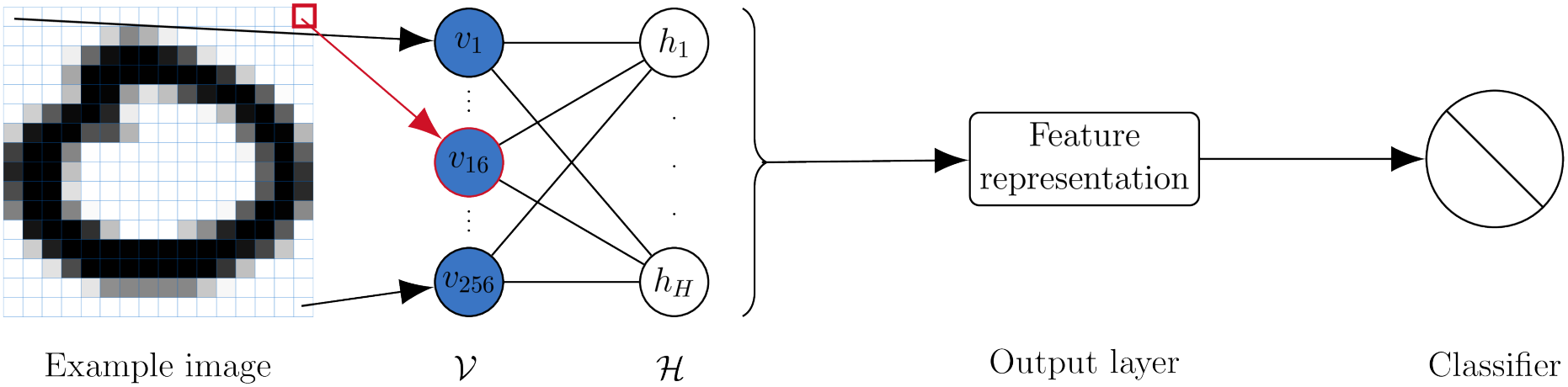

Figure 2: Image classification using a RBM. On the left, each image pixel comprises a node in the visible layer, $\mathcal{V}$. On the right, the output of the RBM is used to create features which are then passed to a supervised learning algorithm.

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.

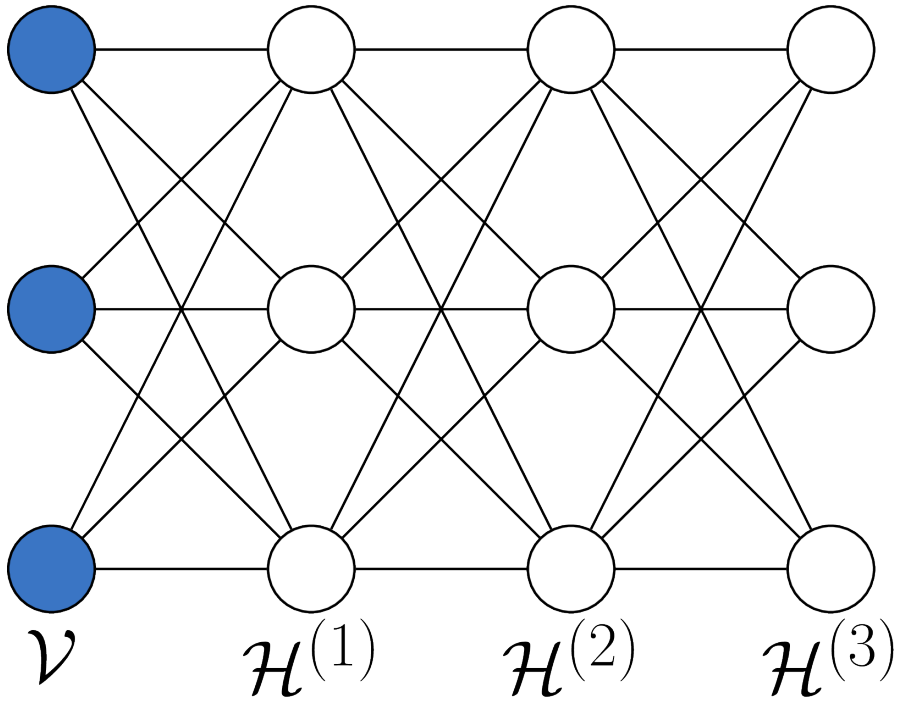

Figure 3: Three layer deep Boltzmann machine, with visible-to-hidden and hidden-to-hidden connections but no within-layer connections.

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, G. E. P. Box [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

Manageable (a.k.a. small) examples

Model fitting

Bayesian methods

Posterior distributions of images

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

References I

W. J. Hill G. E. P. Box. 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.

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`.