Restricted Boltzmann Machines

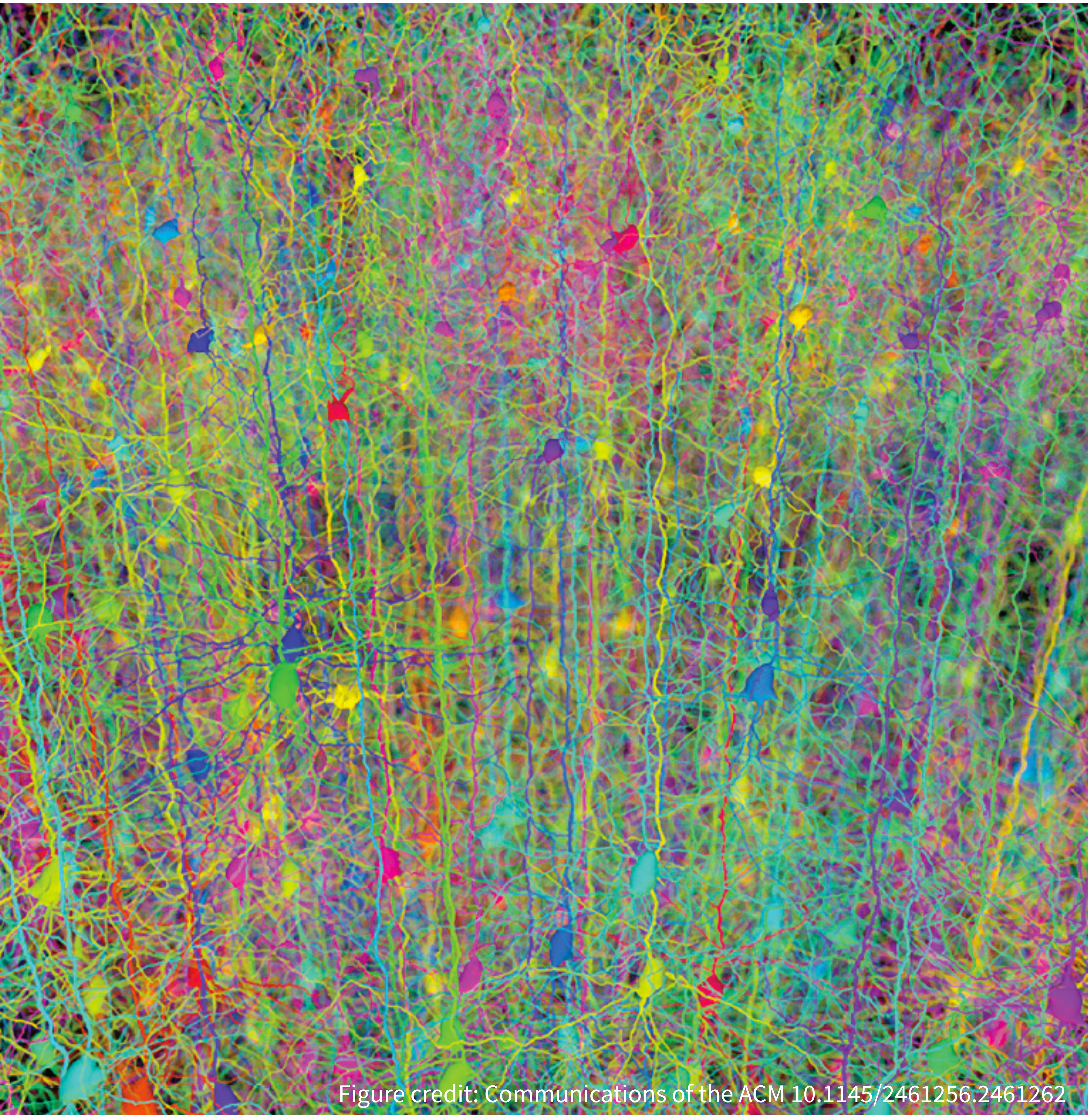


Figure credit: Communications of the ACM 10.1145/2461256.2461262

GRAHAM TAYLOR

VECTOR INSTITUTE

SCHOOL OF ENGINEERING
UNIVERSITY OF GUELPH

CANADIAN INSTITUTE
FOR ADVANCED RESEARCH

UNIVERSITY of GUELPH
CHANGING LIVES IMPROVING LIFE

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

Scaling ML to the Challenges of AI

Classification algorithms take an input from a high-dimensional distribution and summarize it with a category label

- During this process, the classifier discards most of the information in the input and produces a single output

It is possible to ask our ML models to do many other tasks

- Some of which require them to produce multiple outputs
- Most require a complete understanding of the entire structure of the input, with no option to ignore sections of it

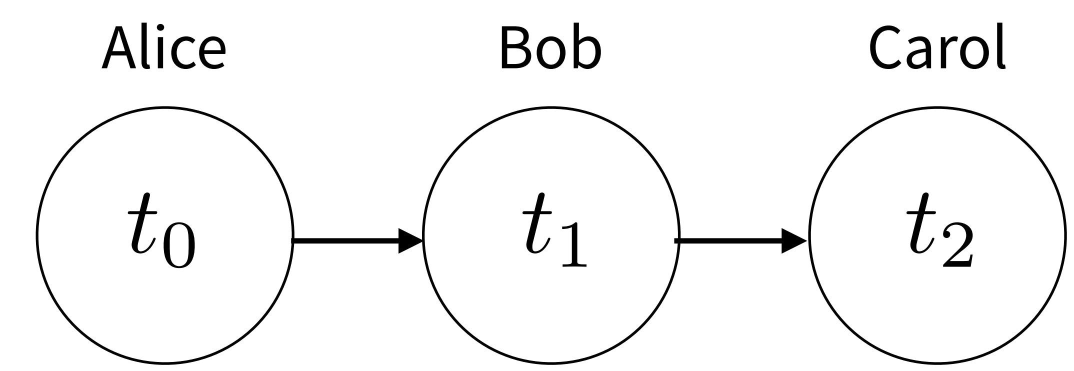
Structured Probabilistic Models for Machine Learning

Modeling a rich distribution over random variables is a challenging task, both computationally and statistically

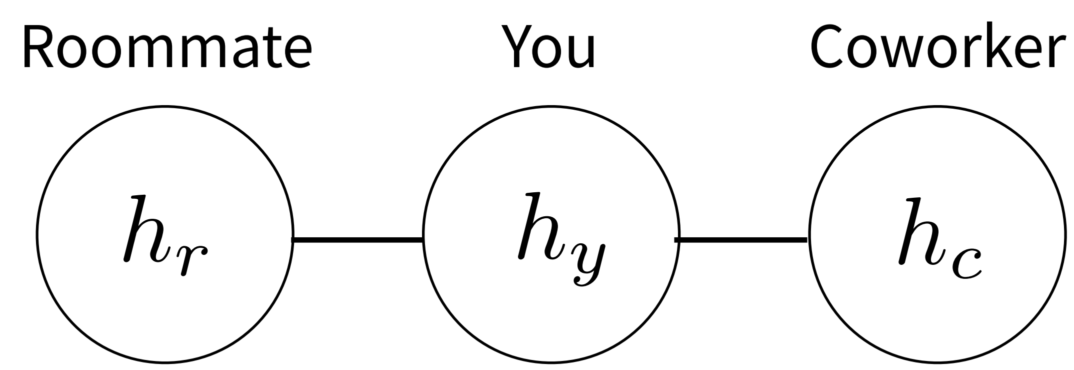
Structured probabilistic models (graphical models) provide a formal framework for modeling only direct interactions between random variables

- significantly fewer parameters
- estimated reliably from less data

directed graphical model



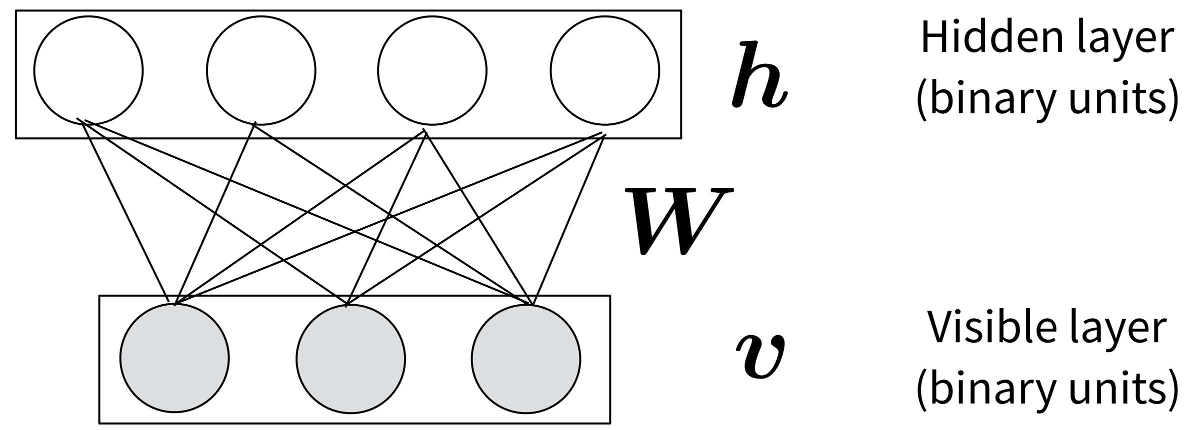
undirected graphical model



Reproduced from Goodfellow et al.

Restricted Boltzmann Machine

- Undirected graphical model with bipartite structure and restricted connectivity to make inference and learning easier



- Energy function:

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{b}^\top \boldsymbol{v} - \boldsymbol{c}^\top \boldsymbol{h} - \boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{h}$$

- Distribution:

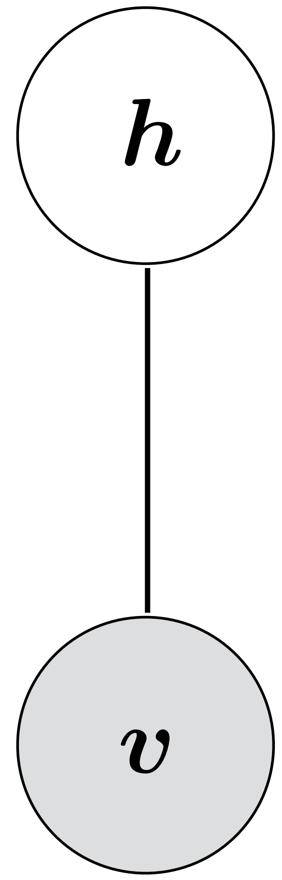
$$P(\mathbf{v} = \boldsymbol{v}, \mathbf{h} = \boldsymbol{h}) = \frac{1}{Z} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))$$

$$Z = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp\{-E(\boldsymbol{v}, \boldsymbol{h})\}$$

partition function
(intractable)

Markov Network View

Markov network with vector nodes:



$$
\begin{aligned}
P(\mathbf{v}=\boldsymbol{v},\mathbf{h}=\boldsymbol{h}) &= \exp(-E(\boldsymbol{v},\boldsymbol{h}))/Z \\
&= \exp(\boldsymbol{b}^\top\boldsymbol{v}+\boldsymbol{c}^\top\boldsymbol{h}+\boldsymbol{v}^\top\boldsymbol{W}\boldsymbol{h})/Z \\
&= \underbrace{\exp(\boldsymbol{b}^\top\boldsymbol{v})\exp(\boldsymbol{c}^\top\boldsymbol{h})\exp(\boldsymbol{v}^\top\boldsymbol{W}\boldsymbol{h})}_{\text{factors}}/Z
\end{aligned}
$$

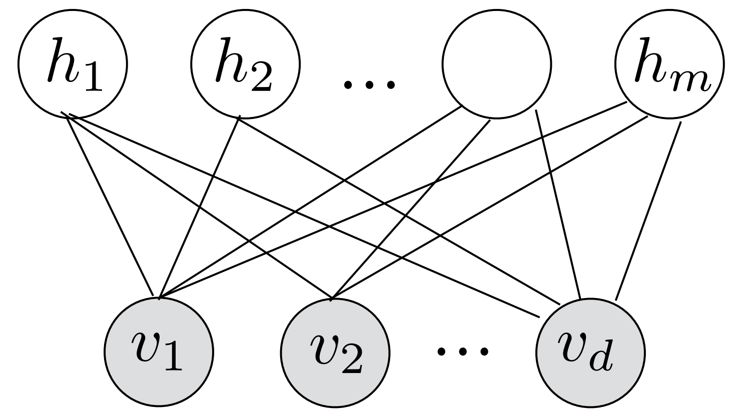
The notation based on an energy function is simply an alternative to the representation as the product of factors

Reproduced from Hugo Larochelle's slides

Markov Network View (2)

Markov network with scalar nodes:

$$P(\mathbf{v}=\boldsymbol{v},\mathbf{h}=\boldsymbol{h}) = \frac{1}{Z}\overbrace{\prod_j\prod_k \exp(W_{j,k}h_j v_k)}^{\text{pair-wise factors}} \left.\begin{array}{l}\displaystyle\prod_k \exp(b_k v_k)\\ \displaystyle\prod_j \exp(c_j h_j)\end{array}\right\}\text{unary factors}$$

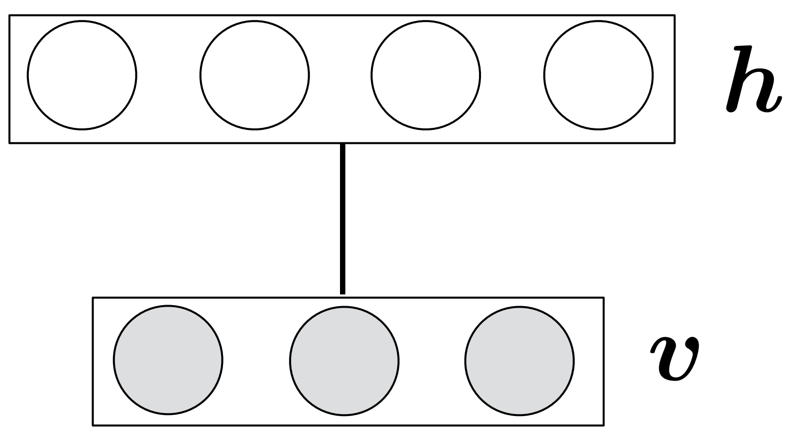


This scalar visualization is more informative of the structure within the vectors

Reproduced from Hugo Larochelle's slides

Inference in RBM

Though $P(\boldsymbol{v})$ is intractable, the bipartite graph structure of the RBM has the special property of its conditional distributions $P(\boldsymbol{h}|\boldsymbol{v})$ and $P(\boldsymbol{v}|\boldsymbol{h})$ being factorial and relatively simple to compute and sample:

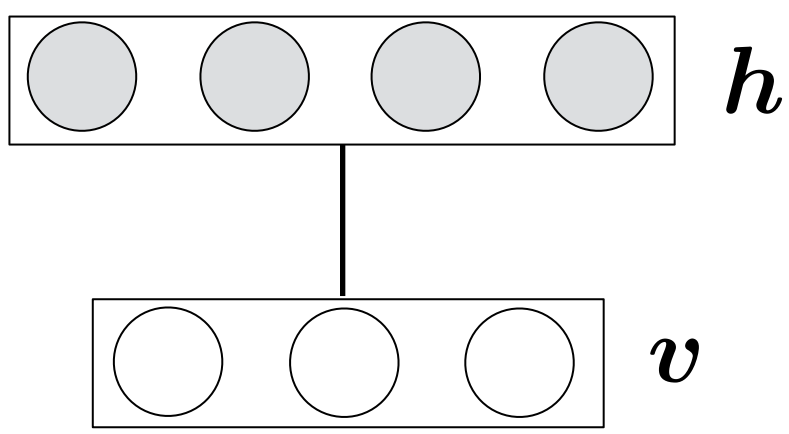


$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_j P(h_j = 1|\boldsymbol{v})$$

$$P(h_j = 1|\boldsymbol{v}) = \frac{1}{1 + \exp(-(c_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j}))}$$

$$= \sigma(c_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j})$$

jth column of $\boldsymbol{W}$



$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_k P(v_k = 1|\boldsymbol{h})$$

$$P(v_k = 1|\boldsymbol{h}) = \frac{1}{1 + \exp(-(b_k + \boldsymbol{W}_{k,:}\boldsymbol{h}))}$$

$$= \sigma(b_k + \boldsymbol{W}_{k,:}\boldsymbol{h})$$

kth row of $\boldsymbol{W}$

Reproduced from Hugo Larochelle's slides

Free Energy

- Many algorithms that operate on probabilistic models need to compute not $p_{\text{model}}(\boldsymbol{x})$ but only $\log \tilde{p}_{\text{model}}(\boldsymbol{x})$

- For energy-based models with latent variables $\boldsymbol{h}$, these algorithms are sometimes phrased in terms of the negative of this quantity, called the free energy:

$$\mathcal{F}(\boldsymbol{x}) = -\log \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{x}, \boldsymbol{h}))$$

- For the RBM:

$$\begin{aligned} P(\boldsymbol{v}) &= \sum_{\boldsymbol{h} \in \{0,1\}^m} P(\boldsymbol{v}, \boldsymbol{h}) = \sum_{\boldsymbol{h} \in \{0,1\}^m} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))/Z \\ &= \exp\left(\boldsymbol{b}^\top \boldsymbol{v} + \sum_{j=1}^{m} \log\left(1 + \exp\left(c_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j}\right)\right)\right)/Z \\ &= \exp\left(-\mathcal{F}(\boldsymbol{v})\right)/Z \end{aligned}$$