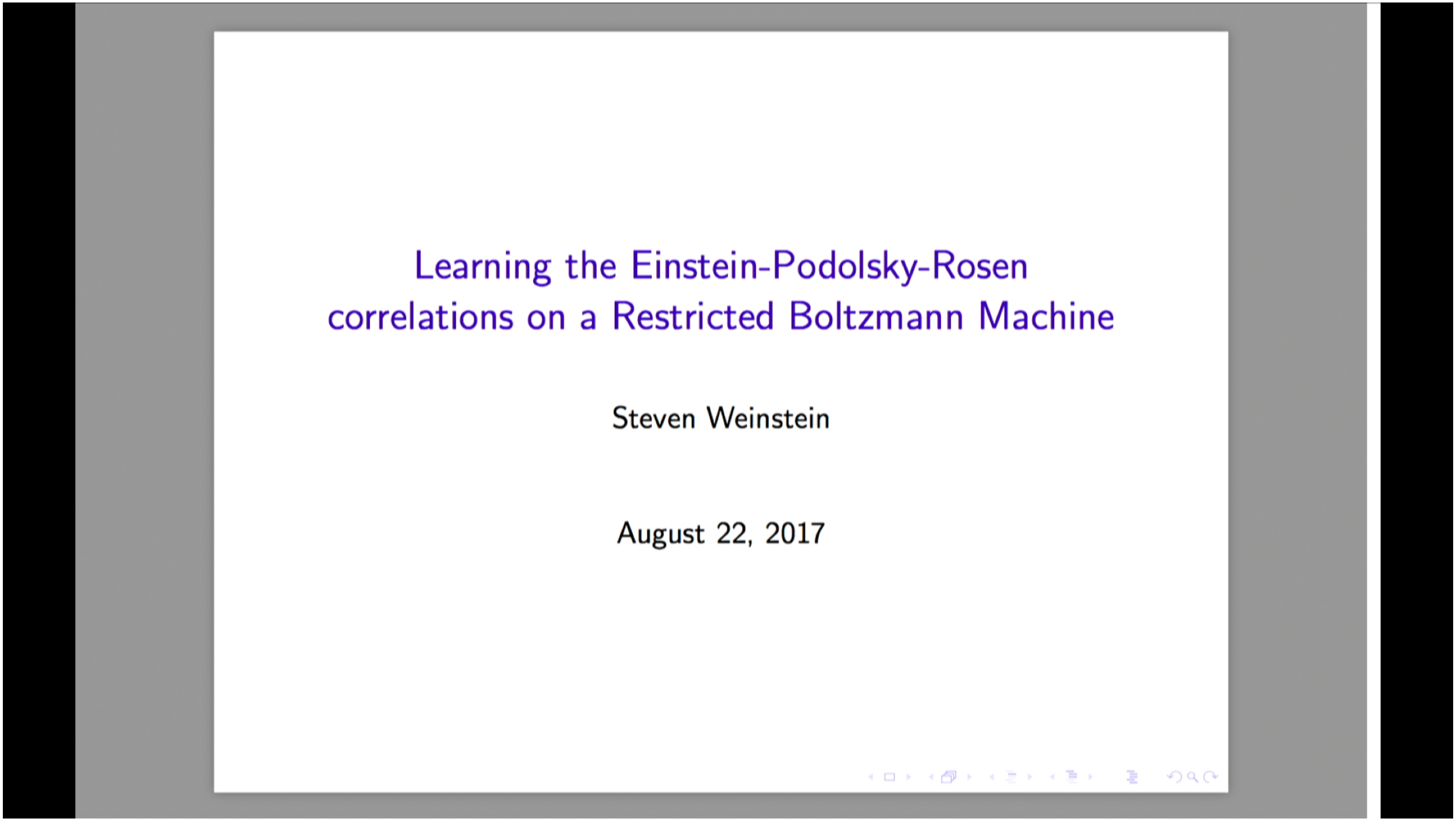
Title: Learning the Einstein-Podolsky-Rosen correlations on a Restricted Boltzmann Machine

Date: Aug 22, 2017 03:30 PM

URL: http://pirsa.org/17080074

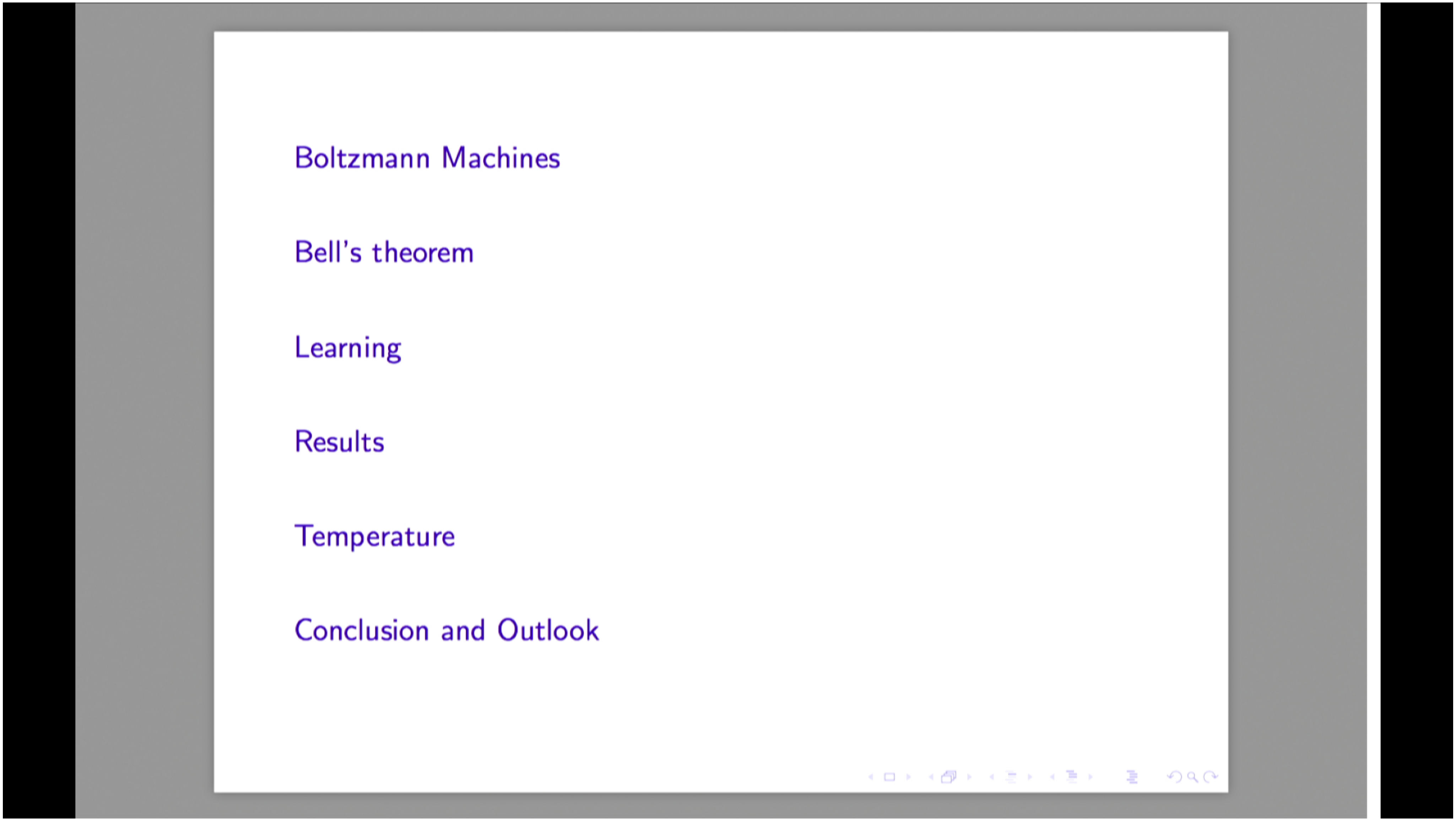
Abstract: <p>We construct a hidden variable model for the EPR correlations using a Restricted Boltzmann Machine. The model reproduces the expected correlations and thus&nbsp;violates the Bell inequality, as required by Bell's theorem. Unlike most hidden-variable models, this model does not violate the&nbsp;<em>locality</em>&nbsp;assumption in Bell's&nbsp;argument. Rather, it violates&nbsp;<em>measurement&nbsp;independence</em>, albeit in a decidedly non-conspiratorial way.</p>



Learning the Einstein-Podolsky-Rosen correlations on a Restricted Boltzmann Machine

Steven Weinstein

August 22, 2017



Boltzmann Machines

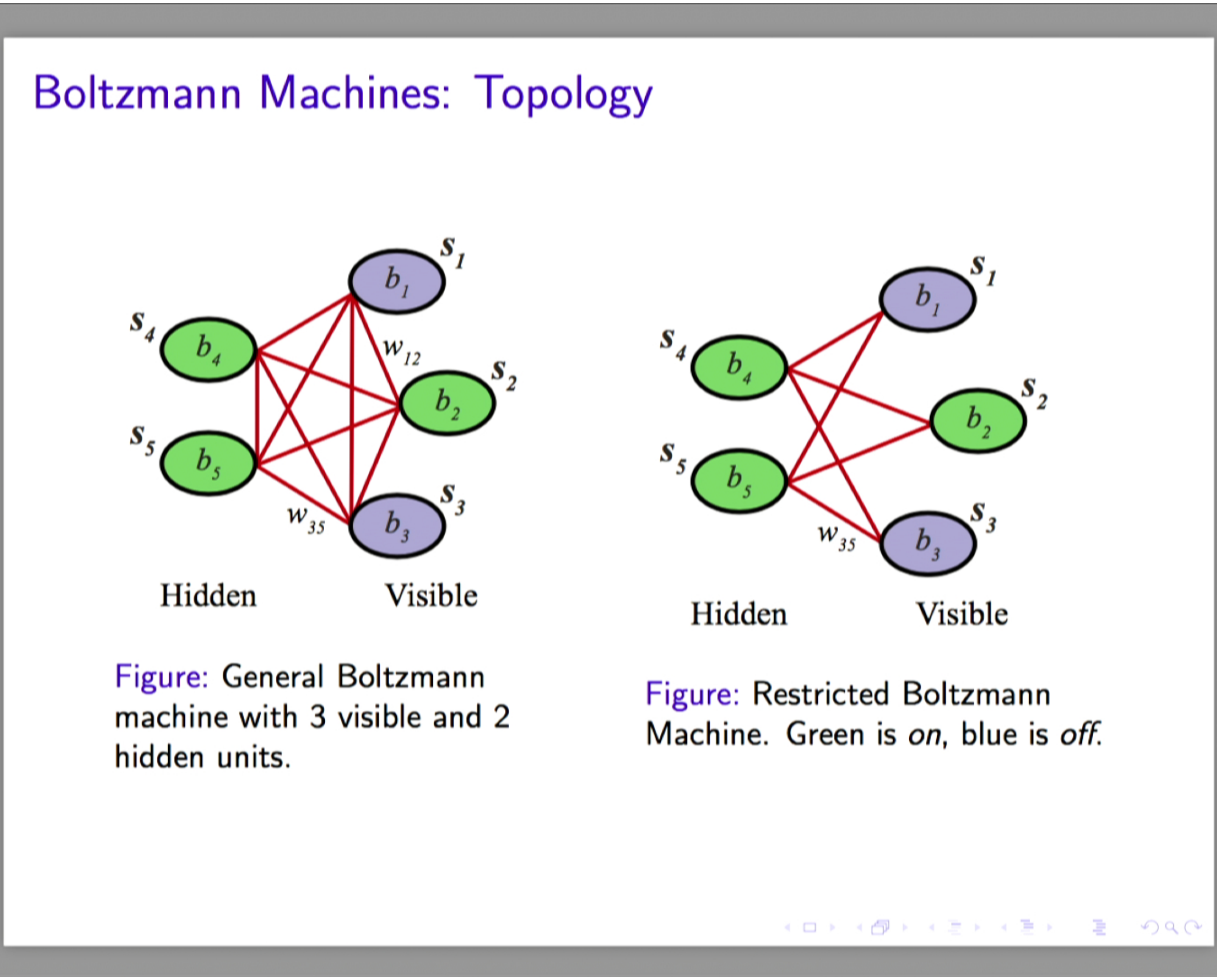
Bell's theorem

Learning

Results

Temperature

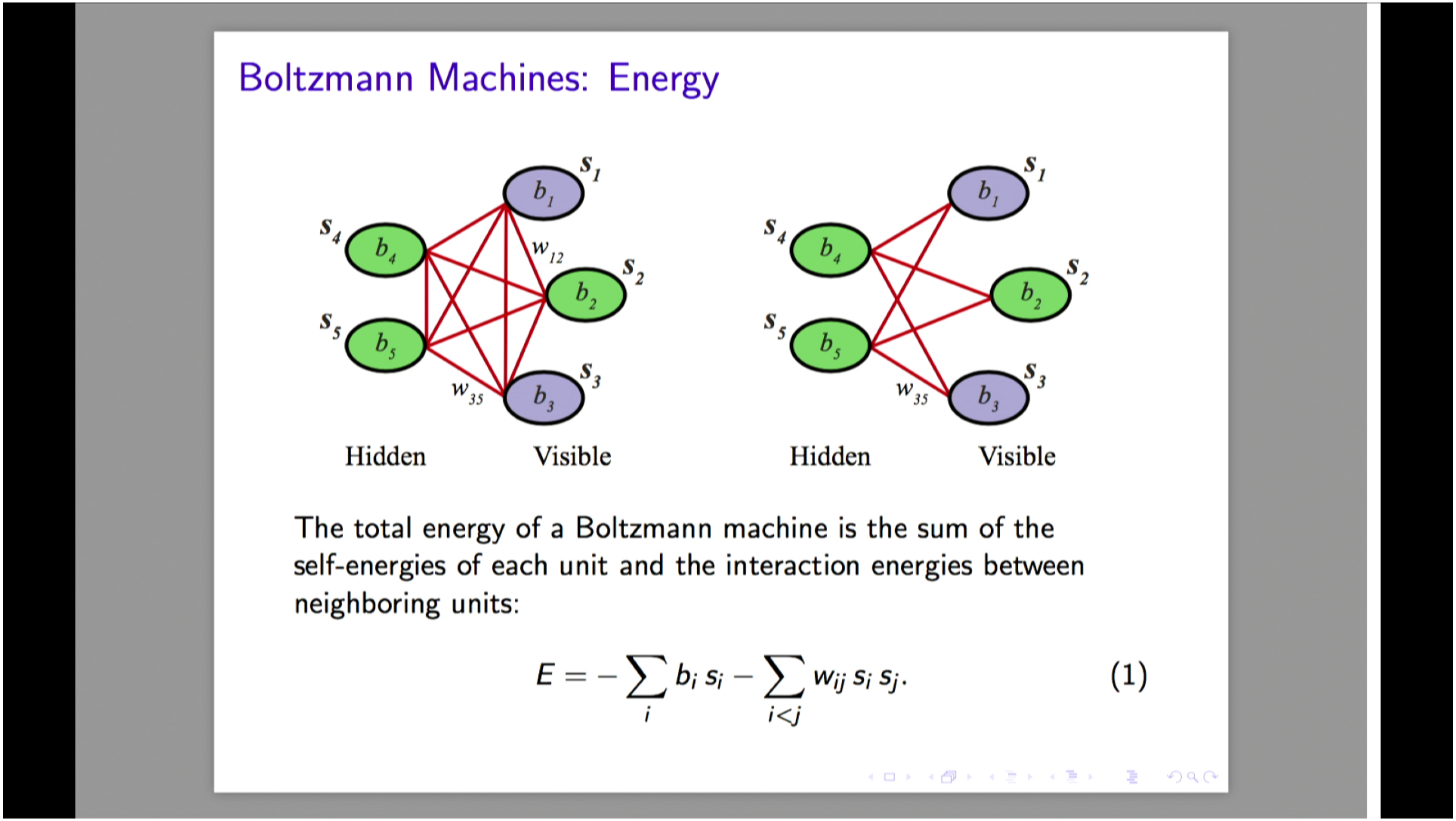
Conclusion and Outlook



Boltzmann Machines: Topology

Figure: General Boltzmann machine with 3 visible and 2 hidden units.

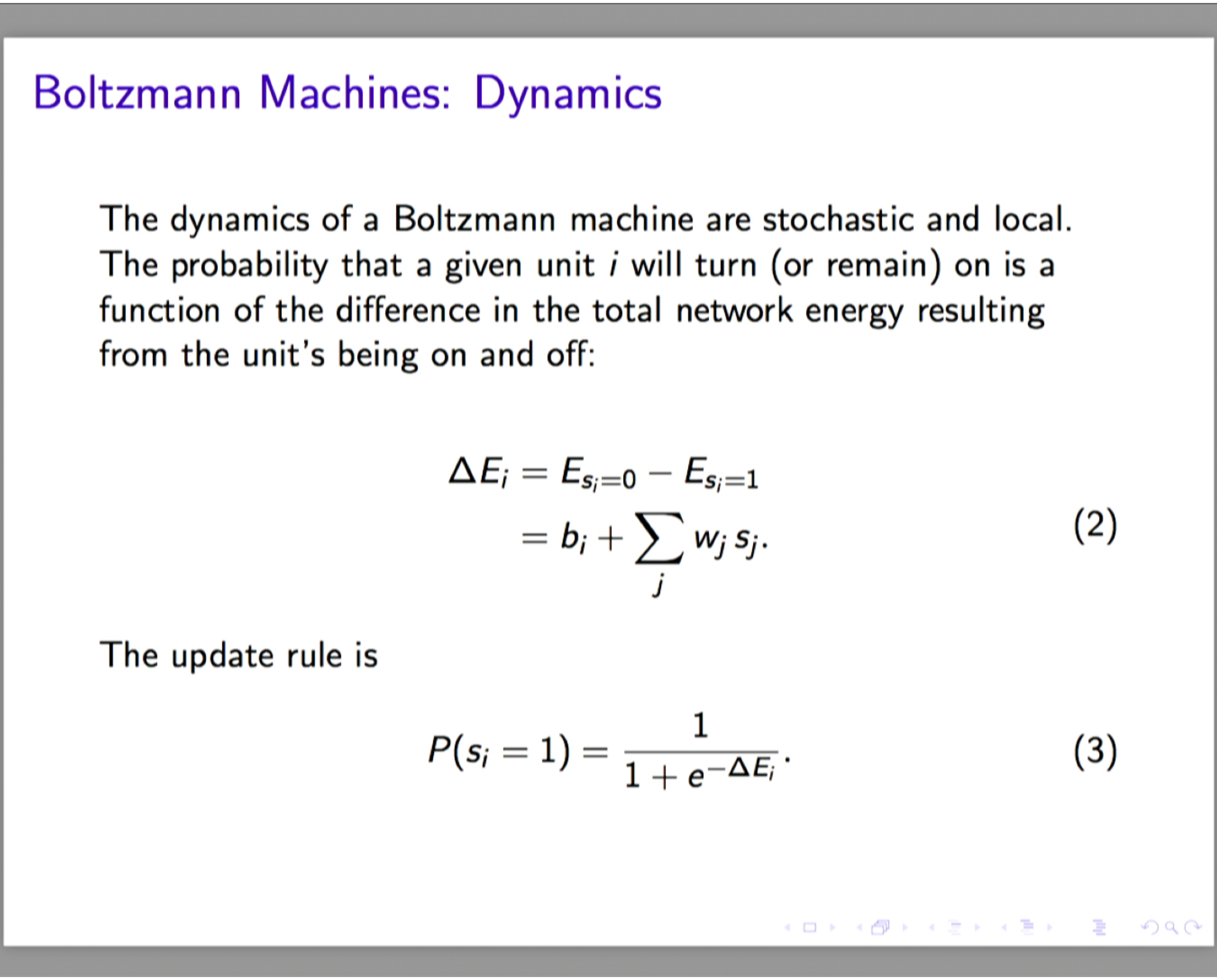
Figure: Restricted Boltzmann Machine. Green is *on*, blue is *off*.



Boltzmann Machines: Energy

The total energy of a Boltzmann machine is the sum of the self-energies of each unit and the interaction energies between neighboring units:

$$E = -\sum_{i} b_i \, s_i - \sum_{i<j} w_{ij} \, s_i \, s_j . \qquad (1)$$



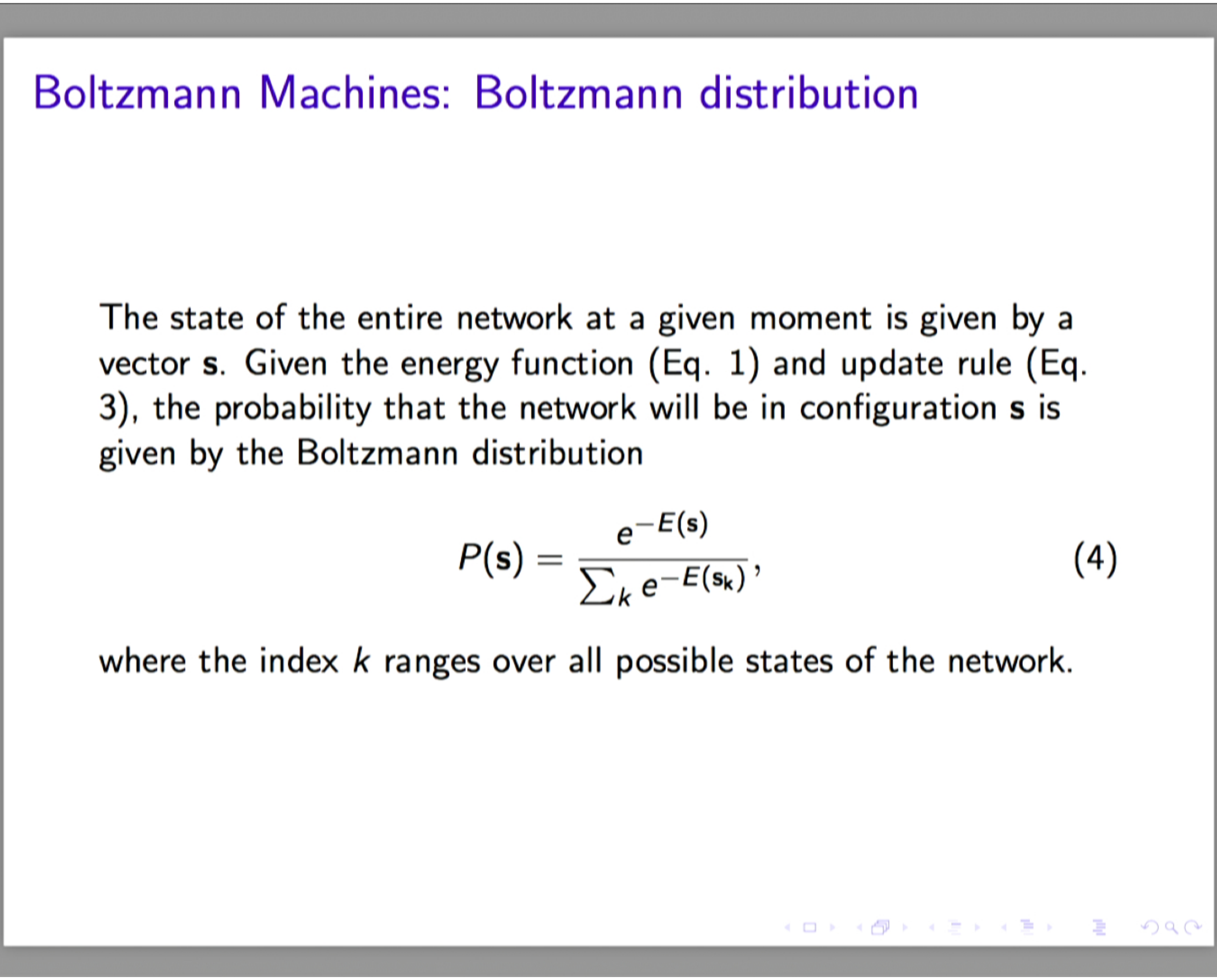
Boltzmann Machines: Dynamics

The dynamics of a Boltzmann machine are stochastic and local. The probability that a given unit *i* will turn (or remain) on is a function of the difference in the total network energy resulting from the unit's being on and off:

$$\begin{aligned} \Delta E_i &= E_{s_i=0} - E_{s_i=1} \\ &= b_i + \sum_j w_j\, s_j . \end{aligned} \tag{2}$$

The update rule is

$$P(s_i = 1) = \frac{1}{1 + e^{-\Delta E_i}} . \tag{3}$$

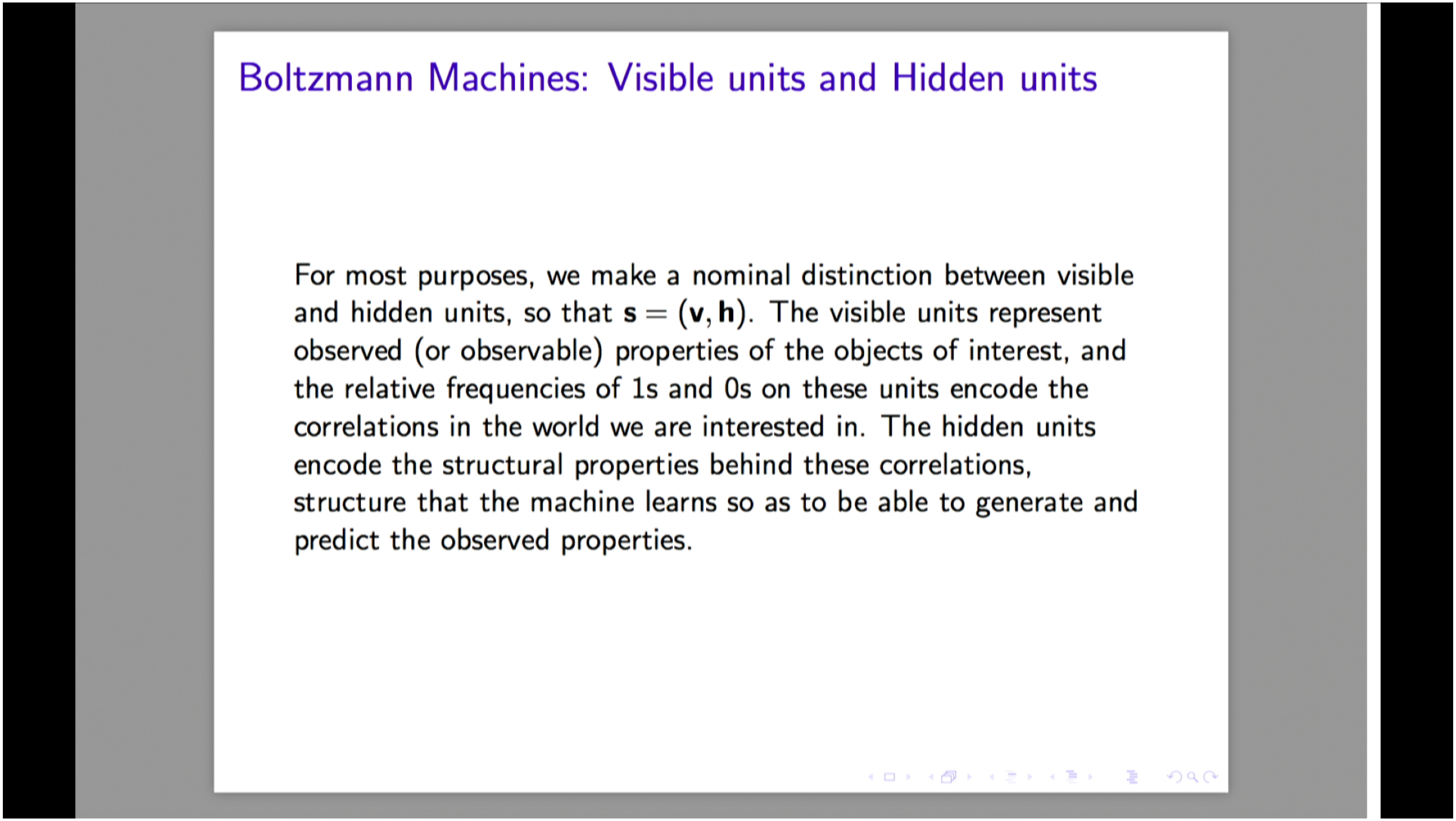


Boltzmann Machines: Boltzmann distribution

The state of the entire network at a given moment is given by a vector $\mathbf{s}$. Given the energy function (Eq. 1) and update rule (Eq. 3), the probability that the network will be in configuration $\mathbf{s}$ is given by the Boltzmann distribution

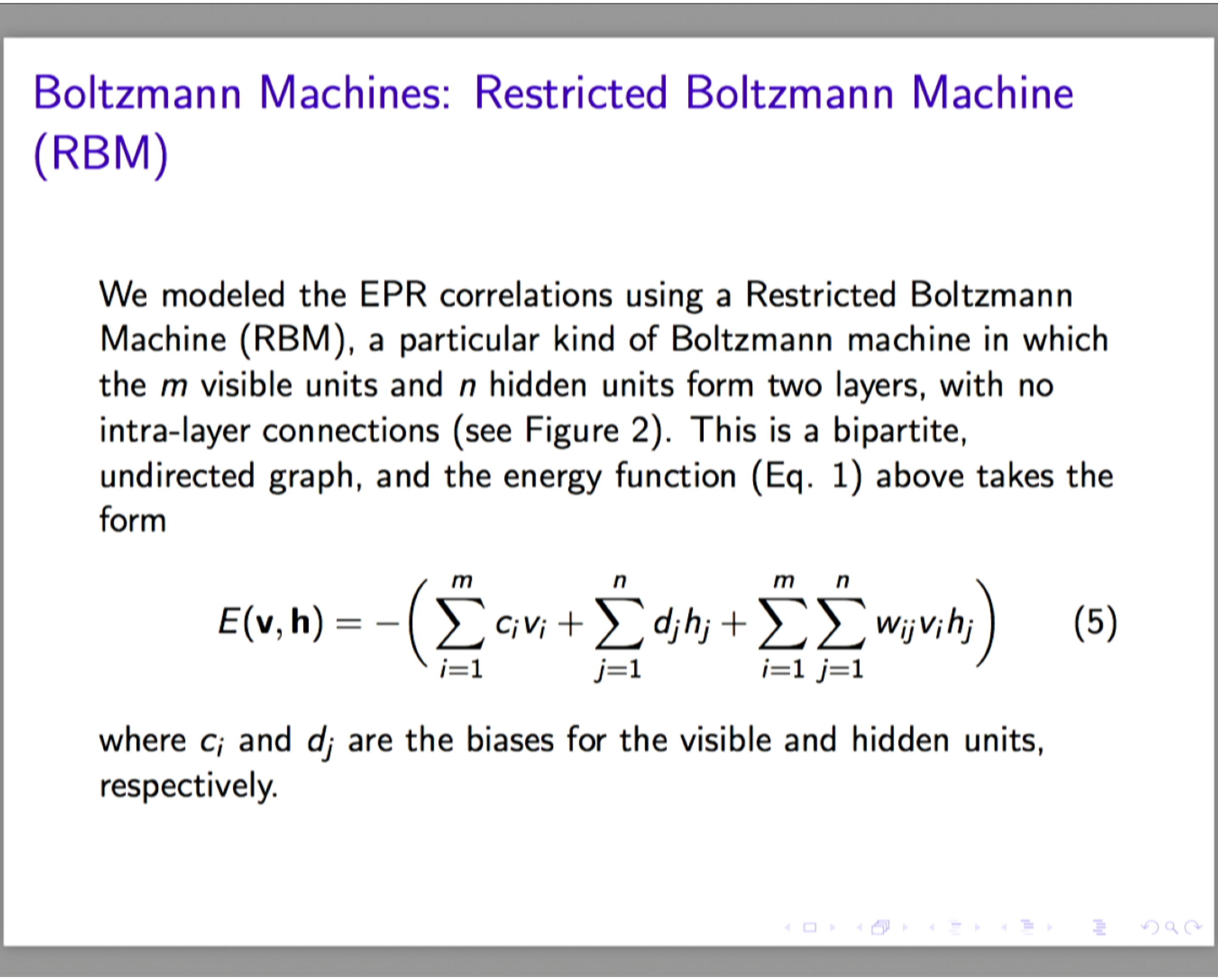
$$P(\mathbf{s}) = \frac{e^{-E(\mathbf{s})}}{\sum_k e^{-E(\mathbf{s_k})}}, \tag{4}$$

where the index k ranges over all possible states of the network.



Boltzmann Machines: Visible units and Hidden units

For most purposes, we make a nominal distinction between visible and hidden units, so that $\mathbf{s} = (\mathbf{v}, \mathbf{h})$. The visible units represent observed (or observable) properties of the objects of interest, and the relative frequencies of 1s and 0s on these units encode the correlations in the world we are interested in. The hidden units encode the structural properties behind these correlations, structure that the machine learns so as to be able to generate and predict the observed properties.

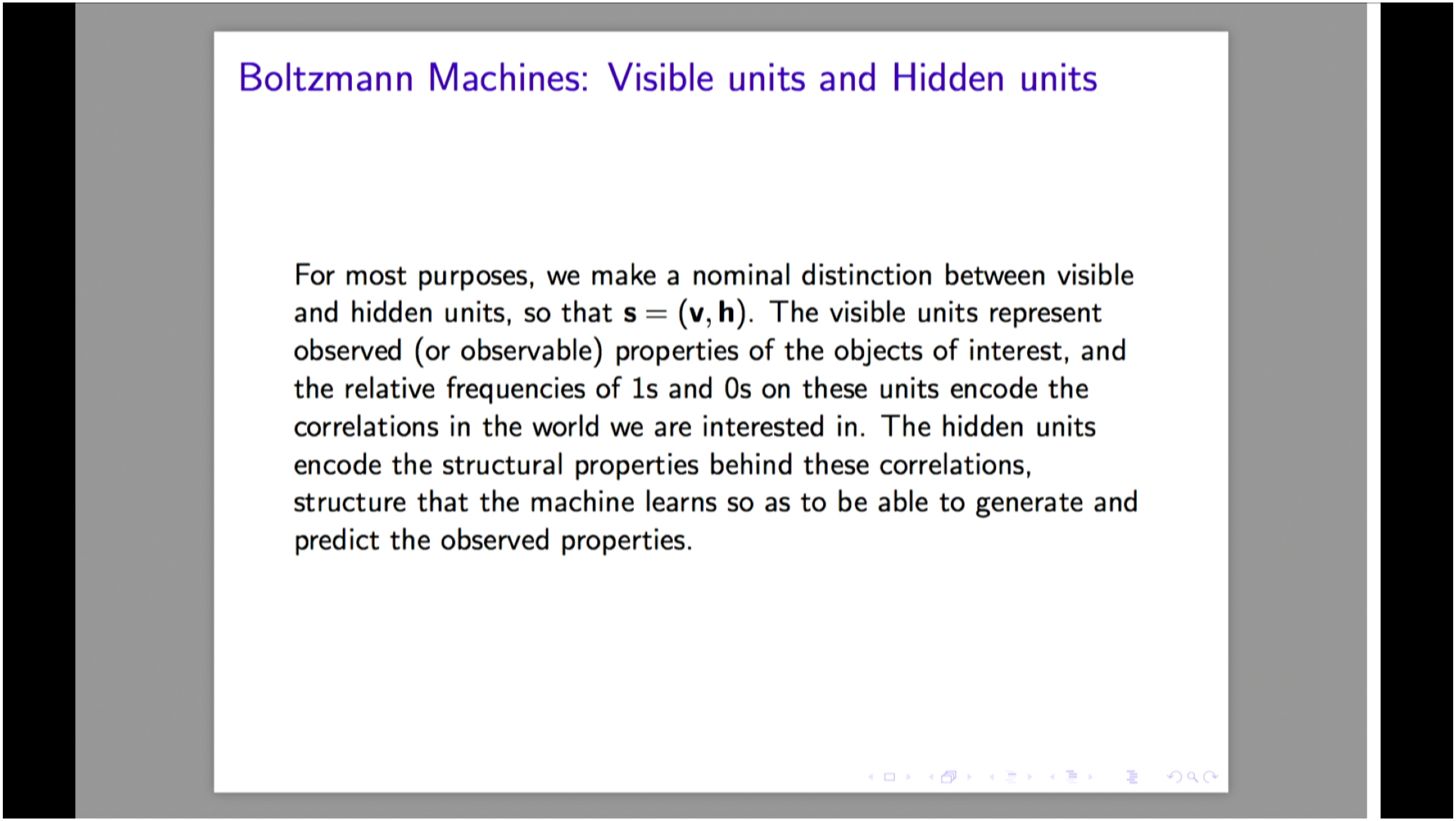


Boltzmann Machines: Restricted Boltzmann Machine (RBM)

We modeled the EPR correlations using a Restricted Boltzmann Machine (RBM), a particular kind of Boltzmann machine in which the m visible units and n hidden units form two layers, with no intra-layer connections (see Figure 2). This is a bipartite, undirected graph, and the energy function (Eq. 1) above takes the form

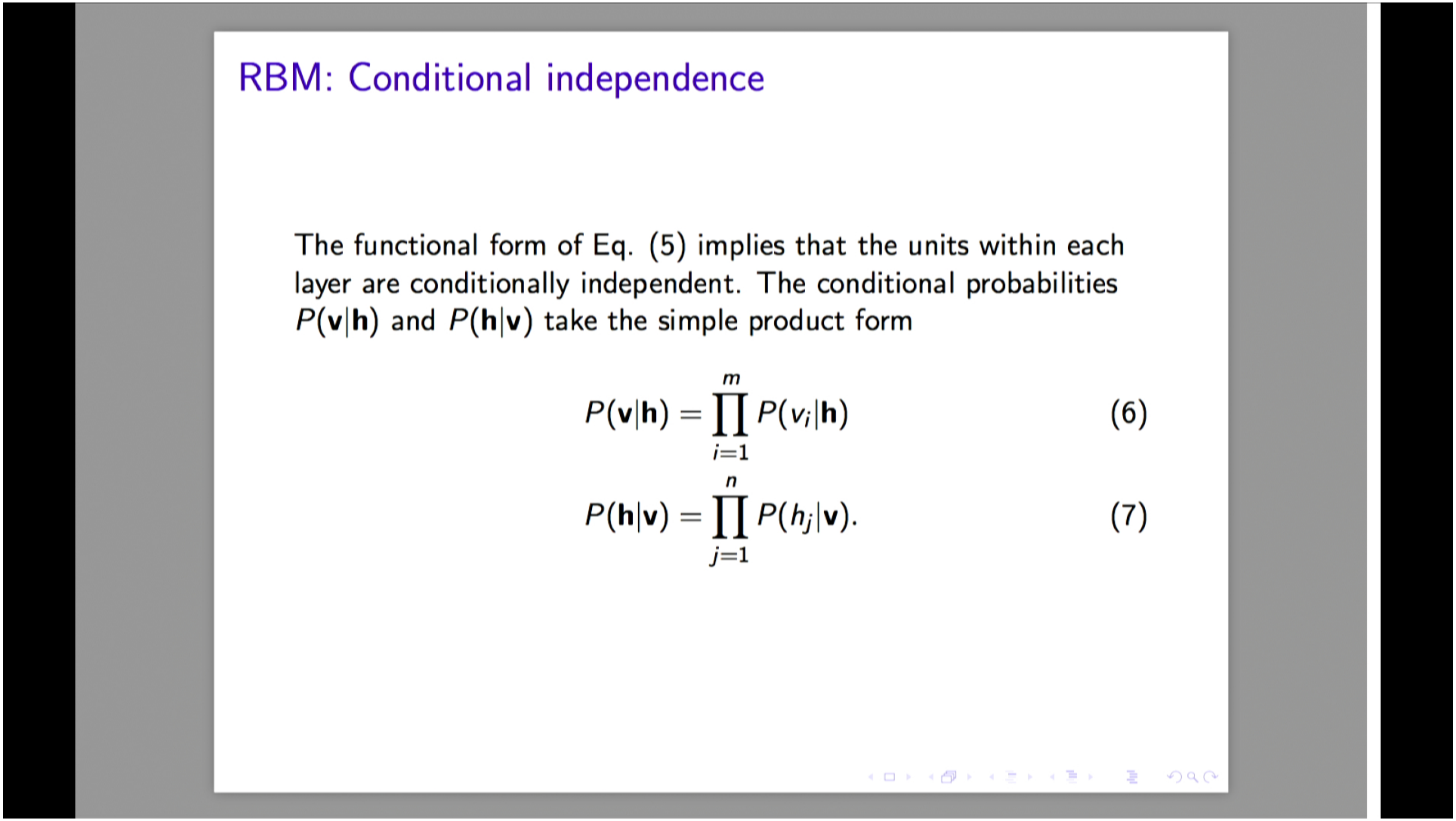
$$E(\mathbf{v},\mathbf{h}) = -\left(\sum_{i=1}^{m} c_i v_i + \sum_{j=1}^{n} d_j h_j + \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} v_i h_j\right) \tag{5}$$

where c_i and d_j are the biases for the visible and hidden units, respectively.



Boltzmann Machines: Visible units and Hidden units

For most purposes, we make a nominal distinction between visible and hidden units, so that $\mathbf{s} = (\mathbf{v}, \mathbf{h})$. The visible units represent observed (or observable) properties of the objects of interest, and the relative frequencies of 1s and 0s on these units encode the correlations in the world we are interested in. The hidden units encode the structural properties behind these correlations, structure that the machine learns so as to be able to generate and predict the observed properties.

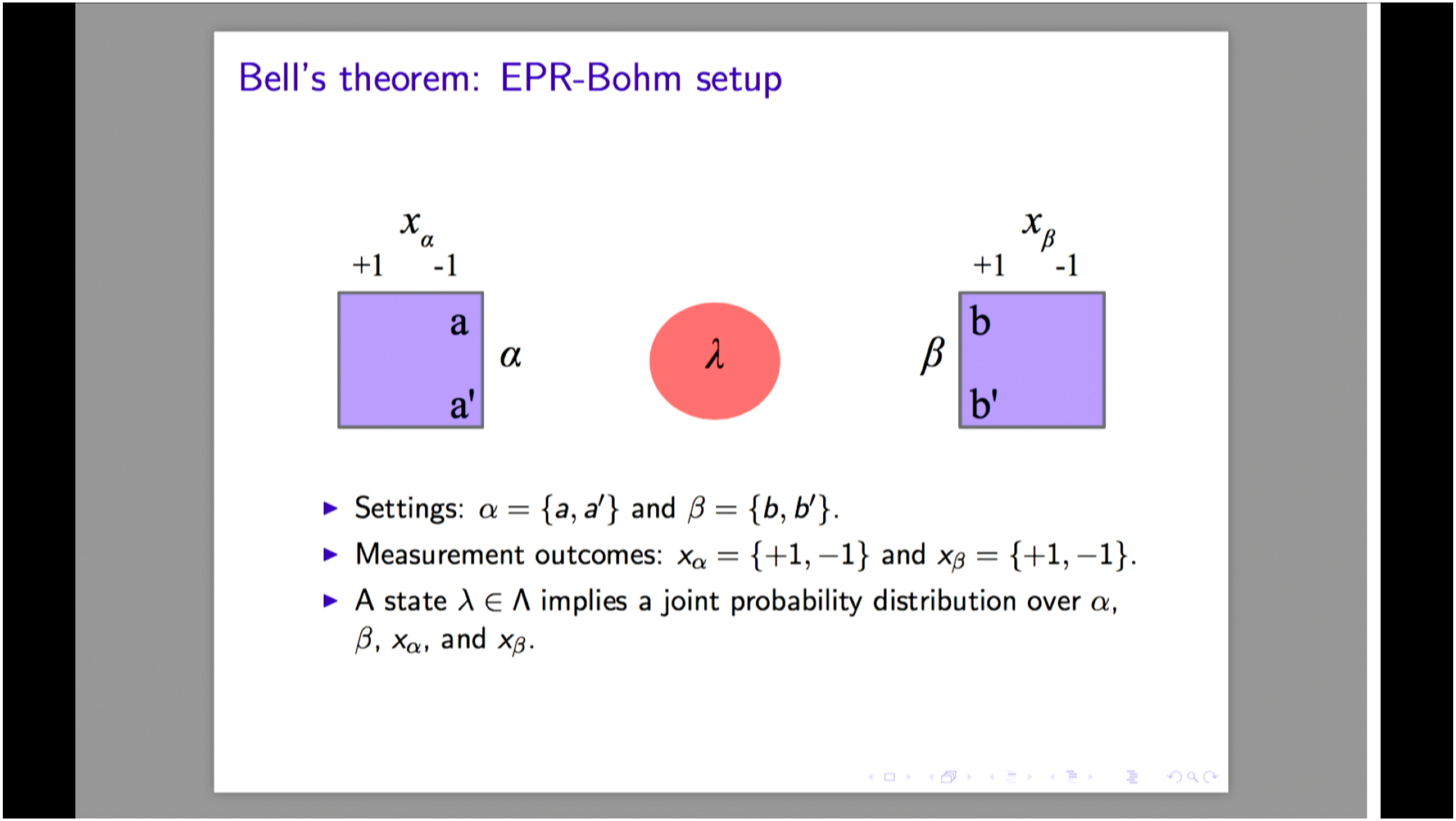


RBM: Conditional independence

The functional form of Eq. (5) implies that the units within each layer are conditionally independent. The conditional probabilities $P(\mathbf{v}|\mathbf{h})$ and $P(\mathbf{h}|\mathbf{v})$ take the simple product form

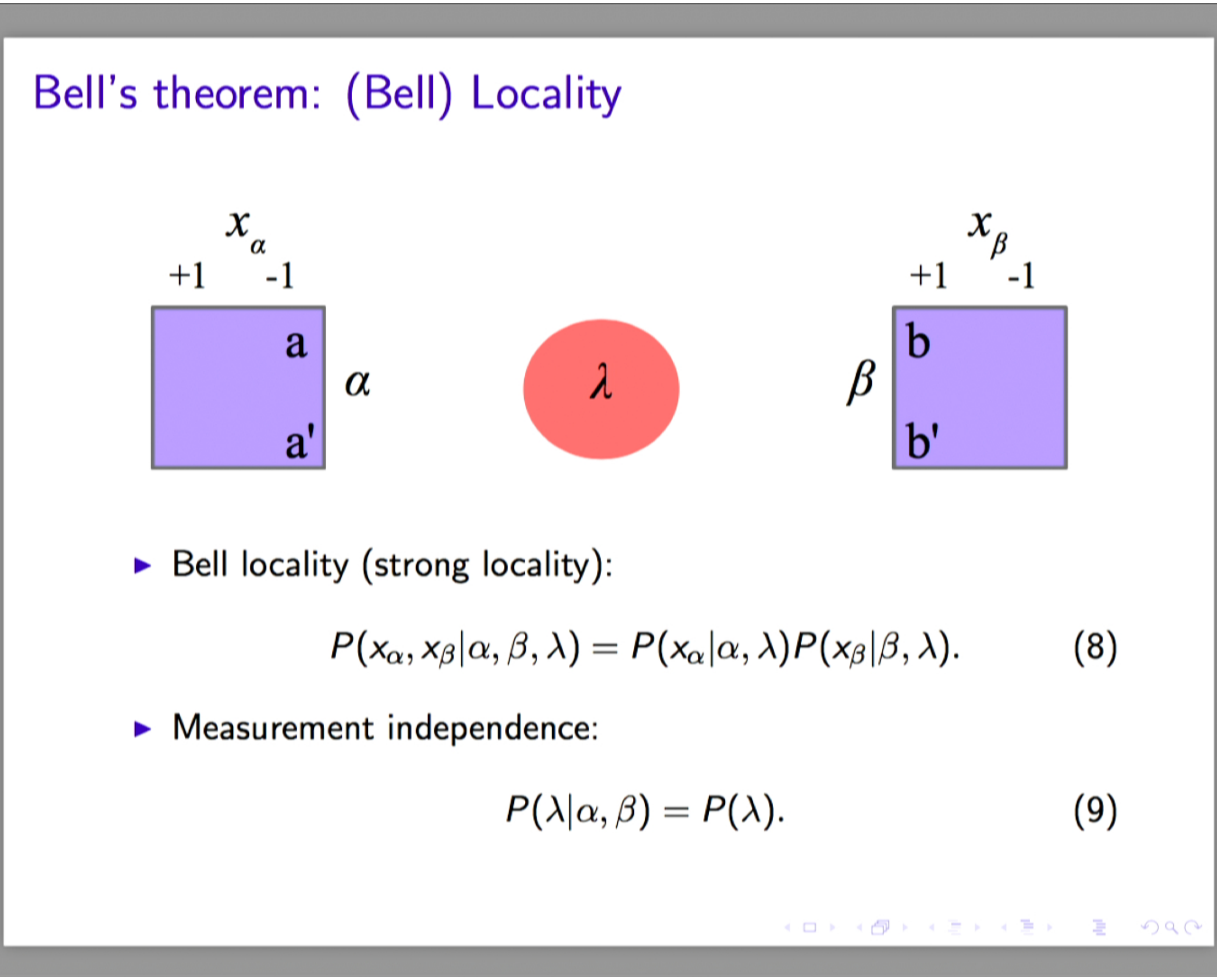
$$P(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{m} P(v_i|\mathbf{h}) \tag{6}$$

$$P(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{n} P(h_j|\mathbf{v}). \tag{7}$$



Bell's theorem: EPR-Bohm setup

- Settings: $\alpha = \{a, a'\}$ and $\beta = \{b, b'\}$.
- Measurement outcomes: $x_\alpha = \{+1, -1\}$ and $x_\beta = \{+1, -1\}$.
- A state $\lambda \in \Lambda$ implies a joint probability distribution over α, β, x_α, and x_β.



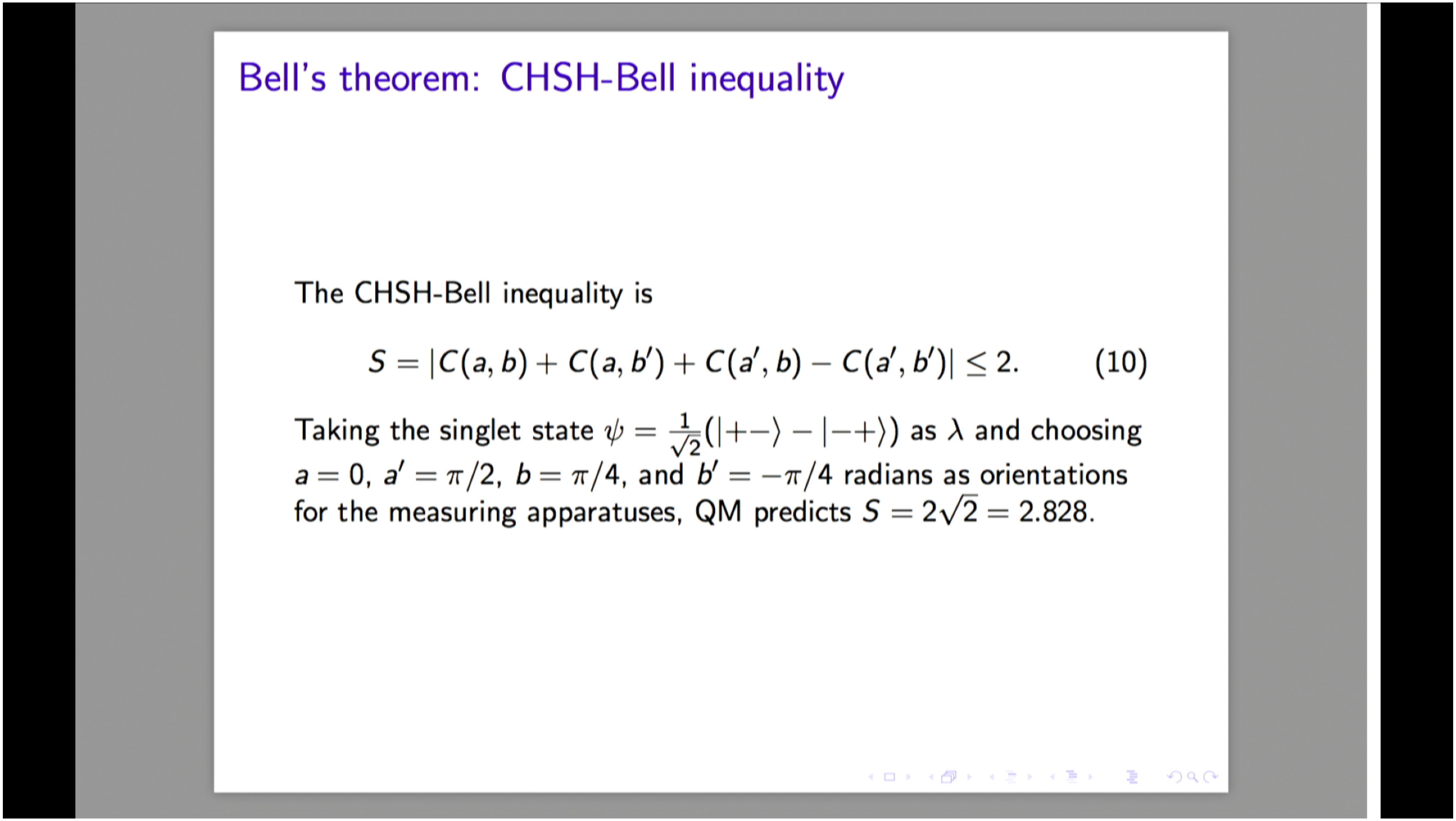
Bell's theorem: (Bell) Locality

- Bell locality (strong locality):

$$P(x_\alpha, x_\beta | \alpha, \beta, \lambda) = P(x_\alpha | \alpha, \lambda) P(x_\beta | \beta, \lambda). \tag{8}$$

- Measurement independence:

$$P(\lambda | \alpha, \beta) = P(\lambda). \tag{9}$$

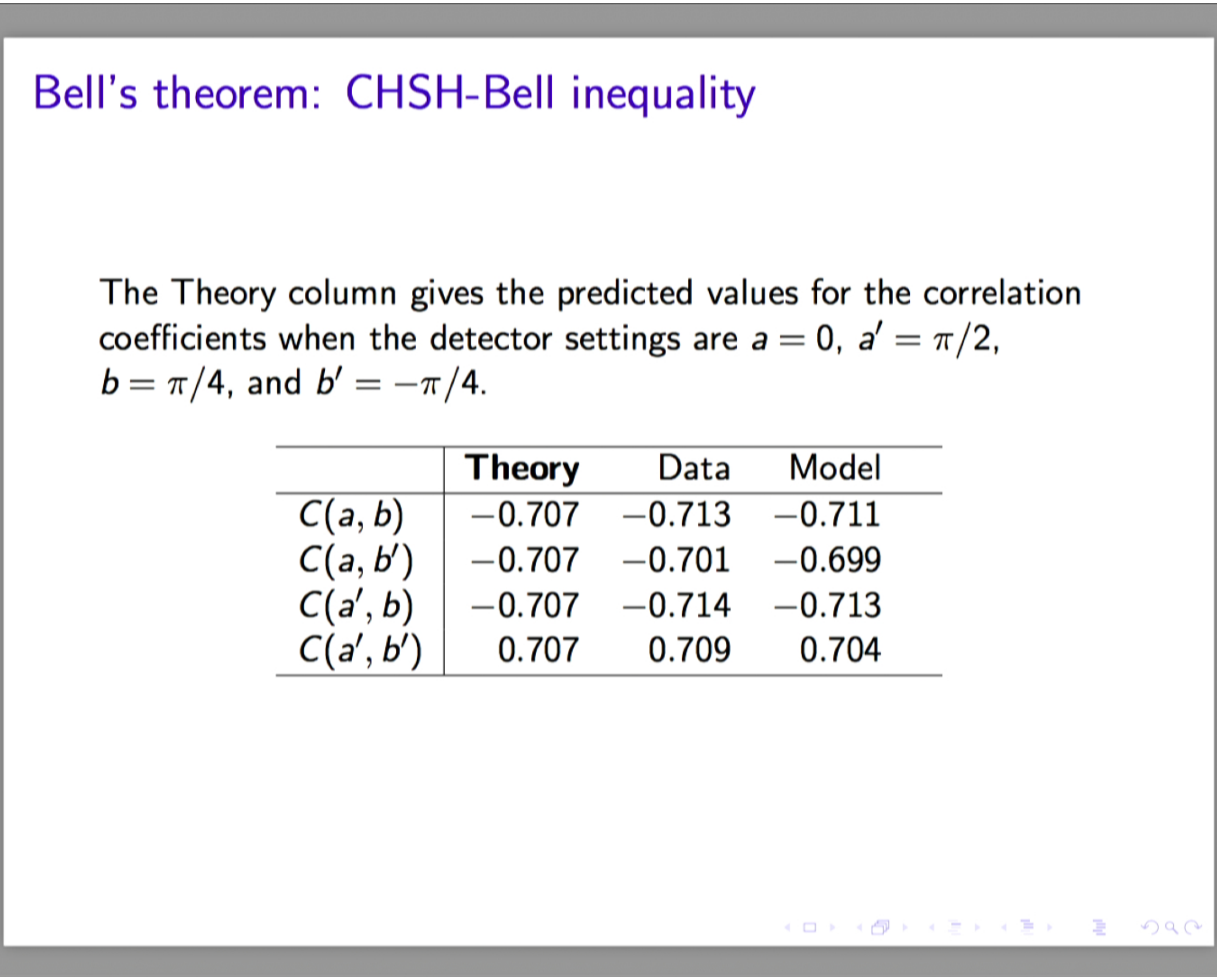


Bell's theorem: CHSH-Bell inequality

The CHSH-Bell inequality is

$$S = |C(a,b) + C(a,b') + C(a',b) - C(a',b')| \leq 2. \tag{10}$$

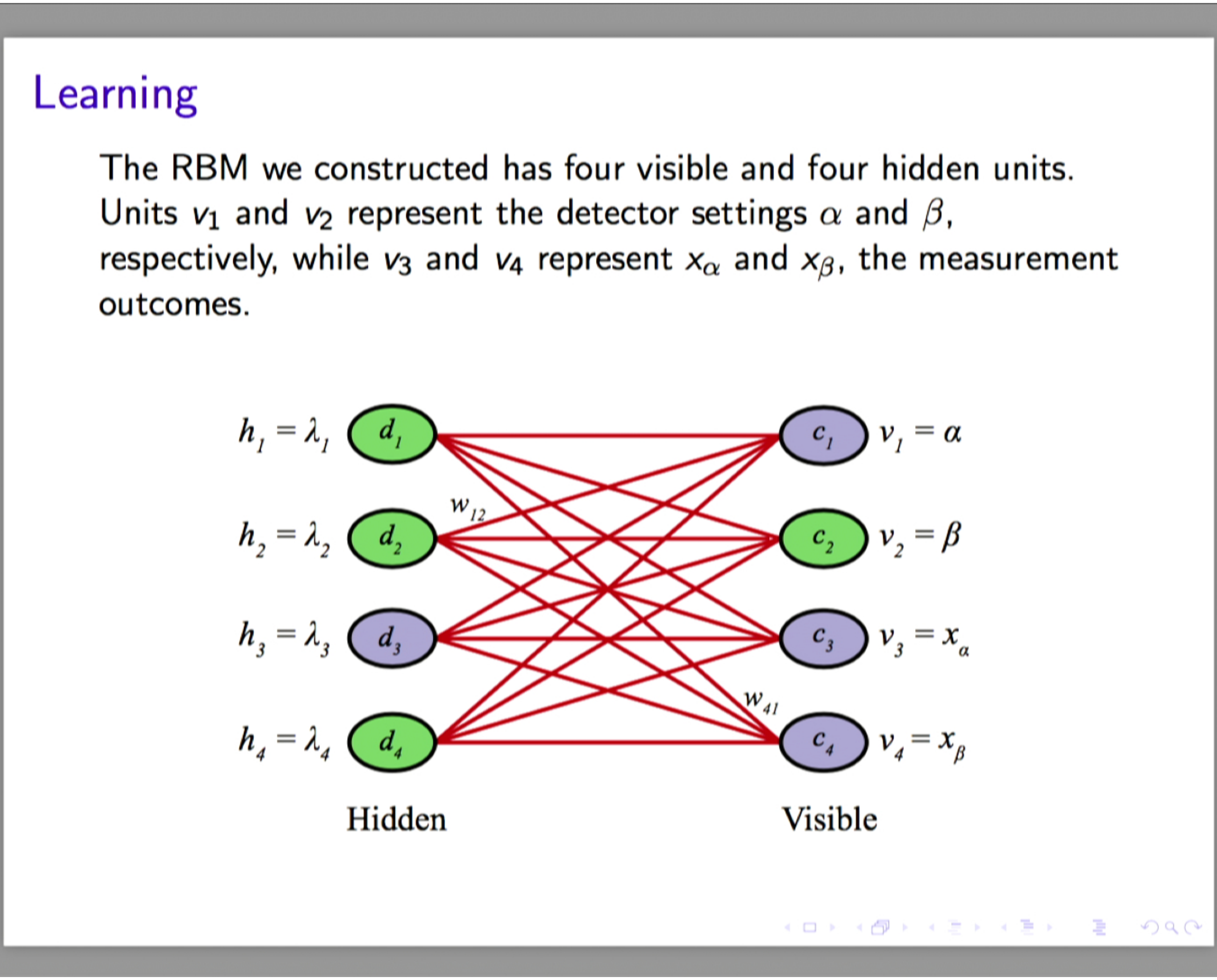
Taking the singlet state $\psi = \frac{1}{\sqrt{2}}(|+-\rangle - |-+\rangle)$ as λ and choosing $a = 0$, $a' = \pi/2$, $b = \pi/4$, and $b' = -\pi/4$ radians as orientations for the measuring apparatuses, QM predicts $S = 2\sqrt{2} = 2.828$.



Bell's theorem: CHSH-Bell inequality

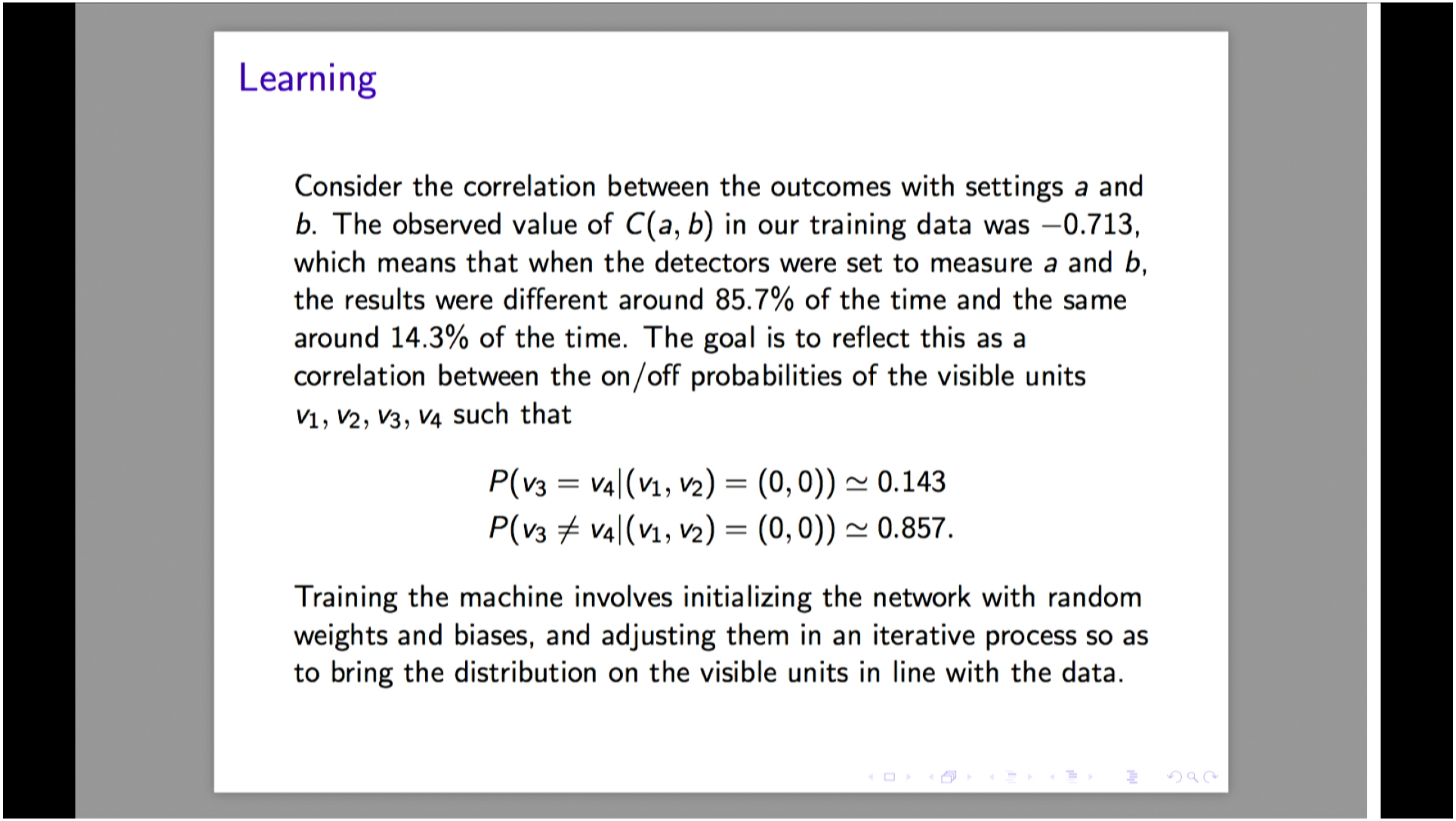
The Theory column gives the predicted values for the correlation coefficients when the detector settings are $a = 0$, $a' = \pi/2$, $b = \pi/4$, and $b' = -\pi/4$.

	Theory	Data	Model
$C(a, b)$	-0.707	-0.713	-0.711
$C(a, b')$	-0.707	-0.701	-0.699
$C(a', b)$	-0.707	-0.714	-0.713
$C(a', b')$	0.707	0.709	0.704



Learning

The RBM we constructed has four visible and four hidden units. Units v_1 and v_2 represent the detector settings α and β, respectively, while v_3 and v_4 represent x_α and x_β, the measurement outcomes.

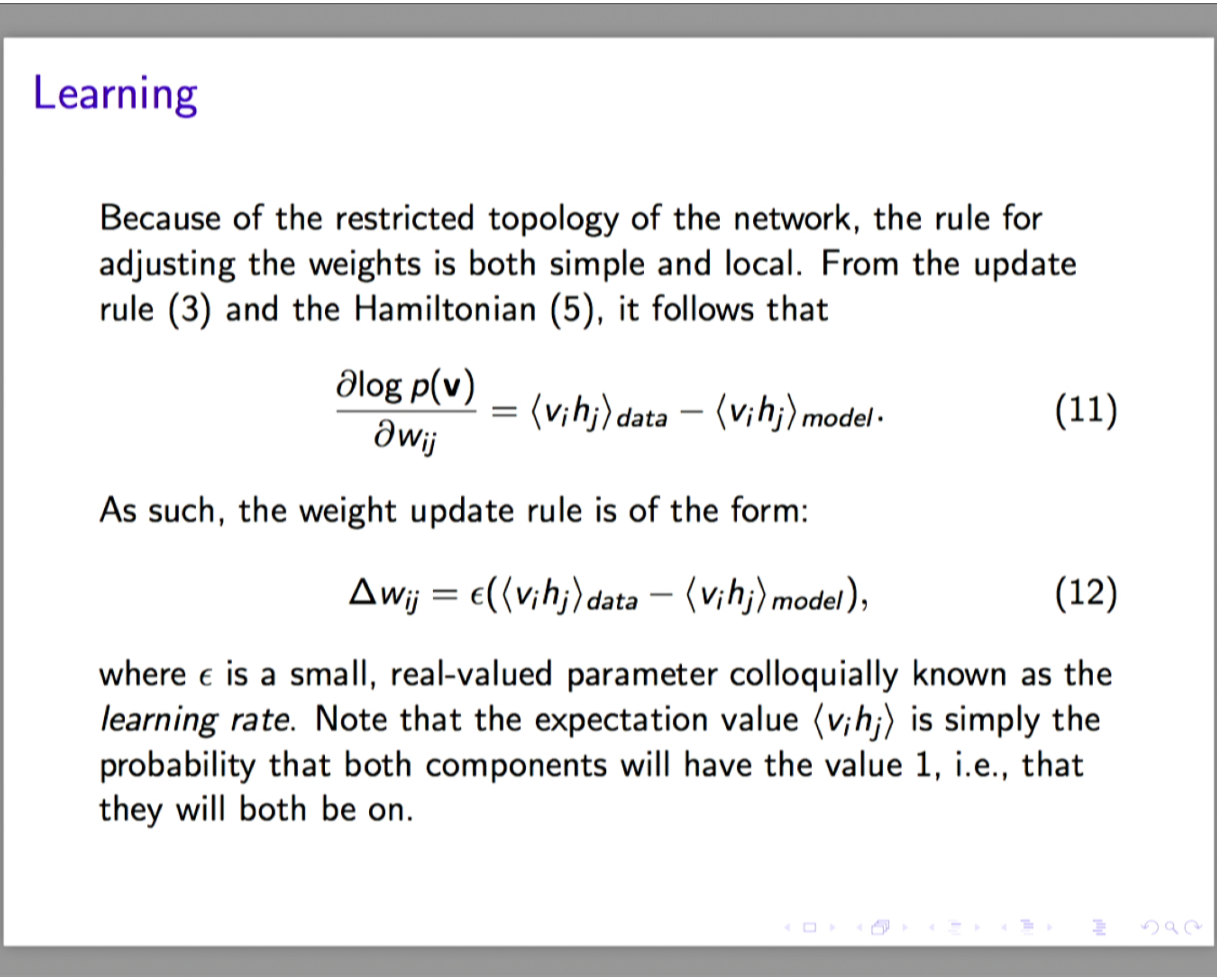


Learning

Consider the correlation between the outcomes with settings a and b. The observed value of $C(a, b)$ in our training data was -0.713, which means that when the detectors were set to measure a and b, the results were different around 85.7% of the time and the same around 14.3% of the time. The goal is to reflect this as a correlation between the on/off probabilities of the visible units v_1, v_2, v_3, v_4 such that

$$P(v_3 = v_4 | (v_1, v_2) = (0, 0)) \simeq 0.143$$
$$P(v_3 \neq v_4 | (v_1, v_2) = (0, 0)) \simeq 0.857.$$

Training the machine involves initializing the network with random weights and biases, and adjusting them in an iterative process so as to bring the distribution on the visible units in line with the data.



Learning

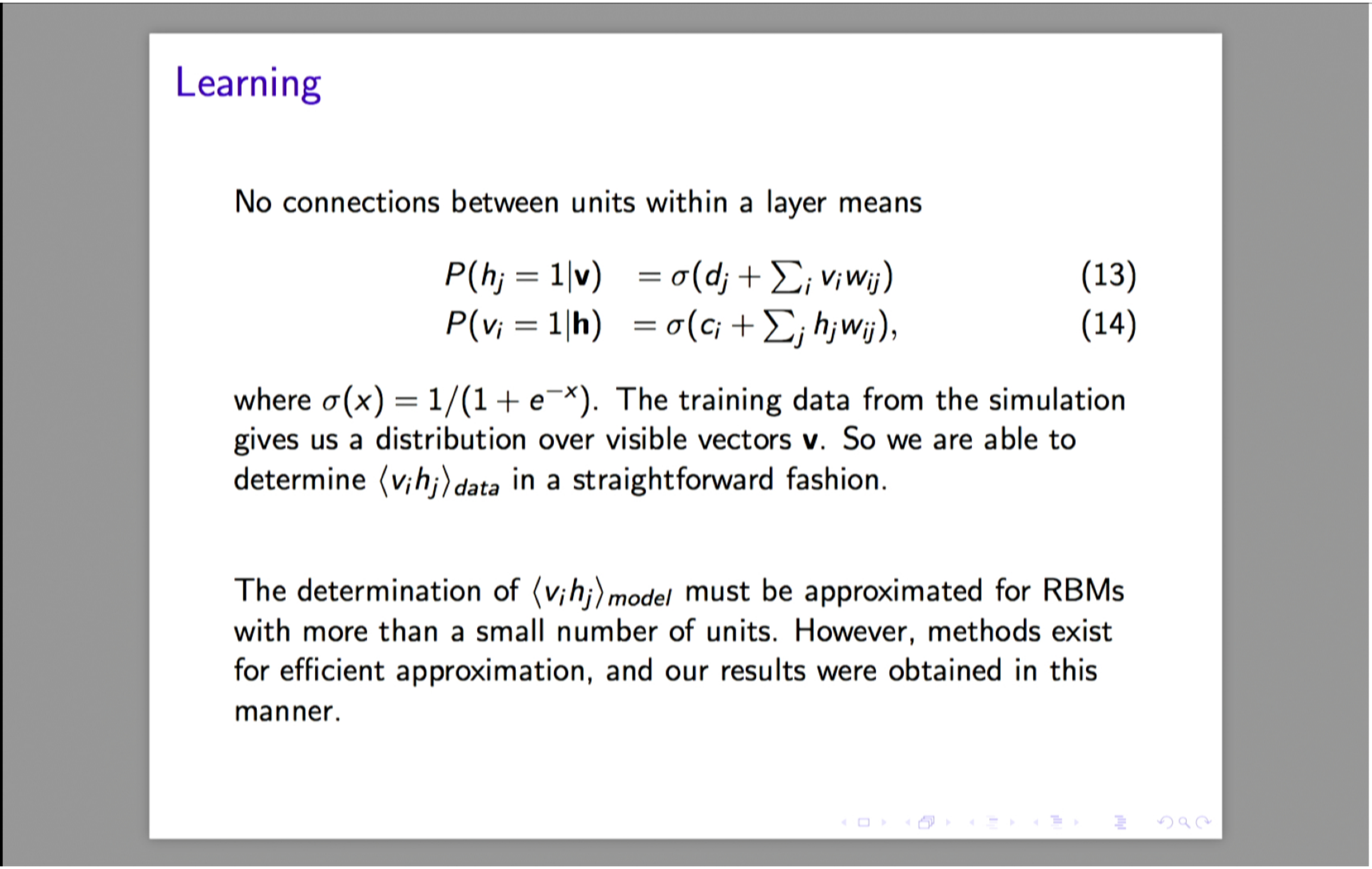
Because of the restricted topology of the network, the rule for adjusting the weights is both simple and local. From the update rule (3) and the Hamiltonian (5), it follows that

$$\frac{\partial \log p(\mathbf{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}. \qquad (11)$$

As such, the weight update rule is of the form:

$$\Delta w_{ij} = \epsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}), \qquad (12)$$

where ϵ is a small, real-valued parameter colloquially known as the *learning rate*. Note that the expectation value $\langle v_i h_j \rangle$ is simply the probability that both components will have the value 1, i.e., that they will both be on.



Learning

No connections between units within a layer means

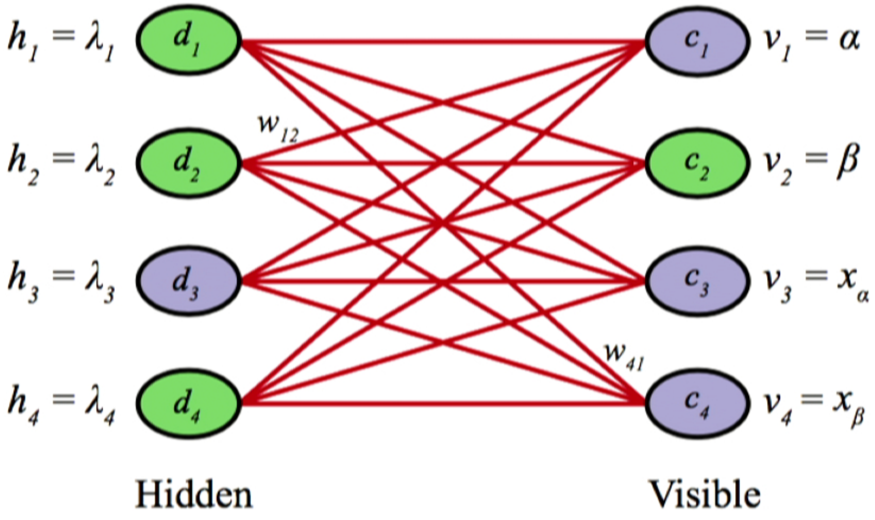
$$\begin{aligned} P(h_j = 1|\mathbf{v}) &= \sigma(d_j + \textstyle\sum_i v_i w_{ij}) && (13) \\ P(v_i = 1|\mathbf{h}) &= \sigma(c_i + \textstyle\sum_j h_j w_{ij}), && (14) \end{aligned}$$

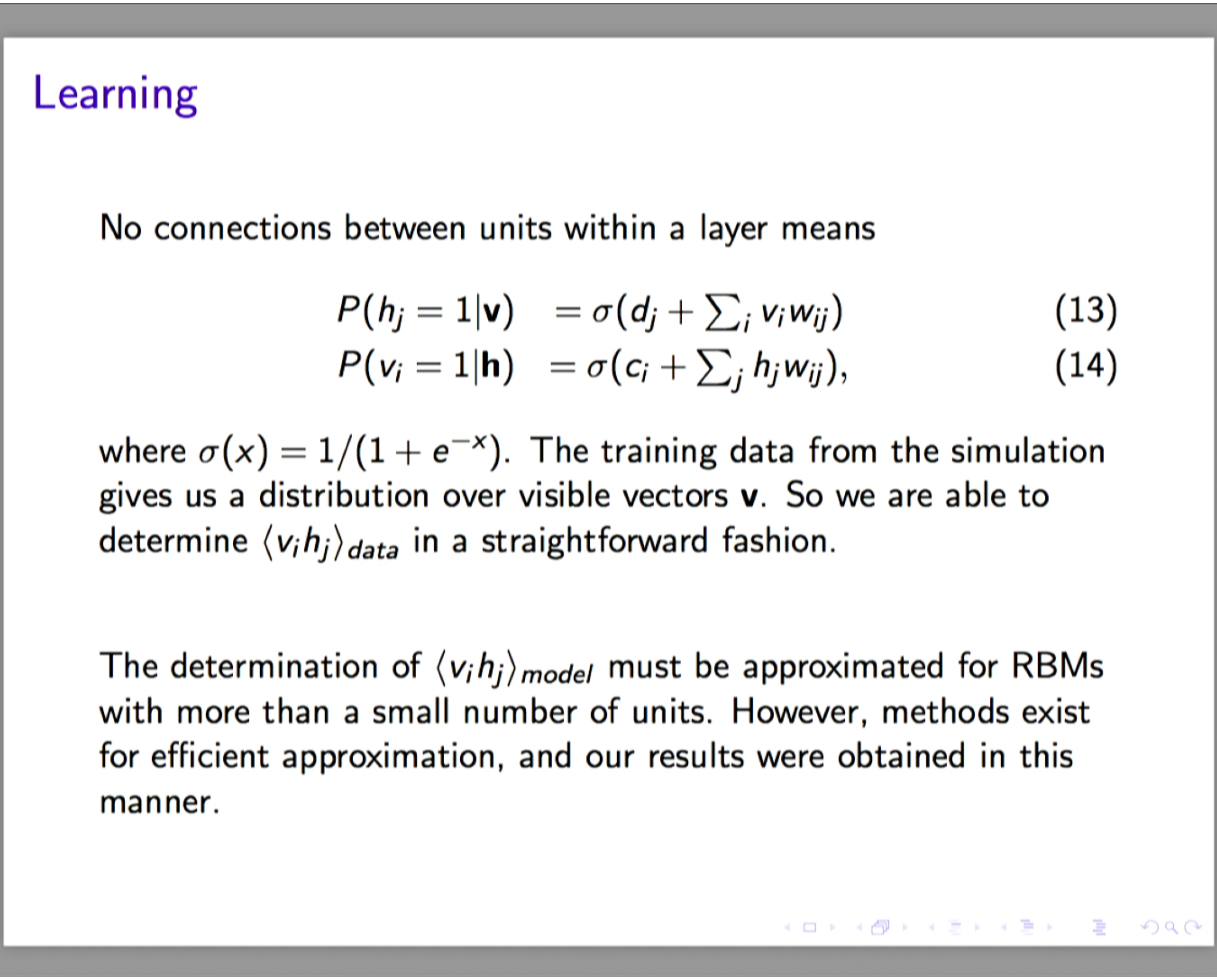
where $\sigma(x) = 1/(1 + e^{-x})$. The training data from the simulation gives us a distribution over visible vectors $\mathbf{v}$. So we are able to determine $\langle v_i h_j \rangle_{data}$ in a straightforward fashion.

The determination of $\langle v_i h_j \rangle_{model}$ must be approximated for RBMs with more than a small number of units. However, methods exist for efficient approximation, and our results were obtained in this manner.

Learning

The RBM we constructed has four visible and four hidden units. Units v_1 and v_2 represent the detector settings α and β, respectively, while v_3 and v_4 represent x_α and x_β, the measurement outcomes.





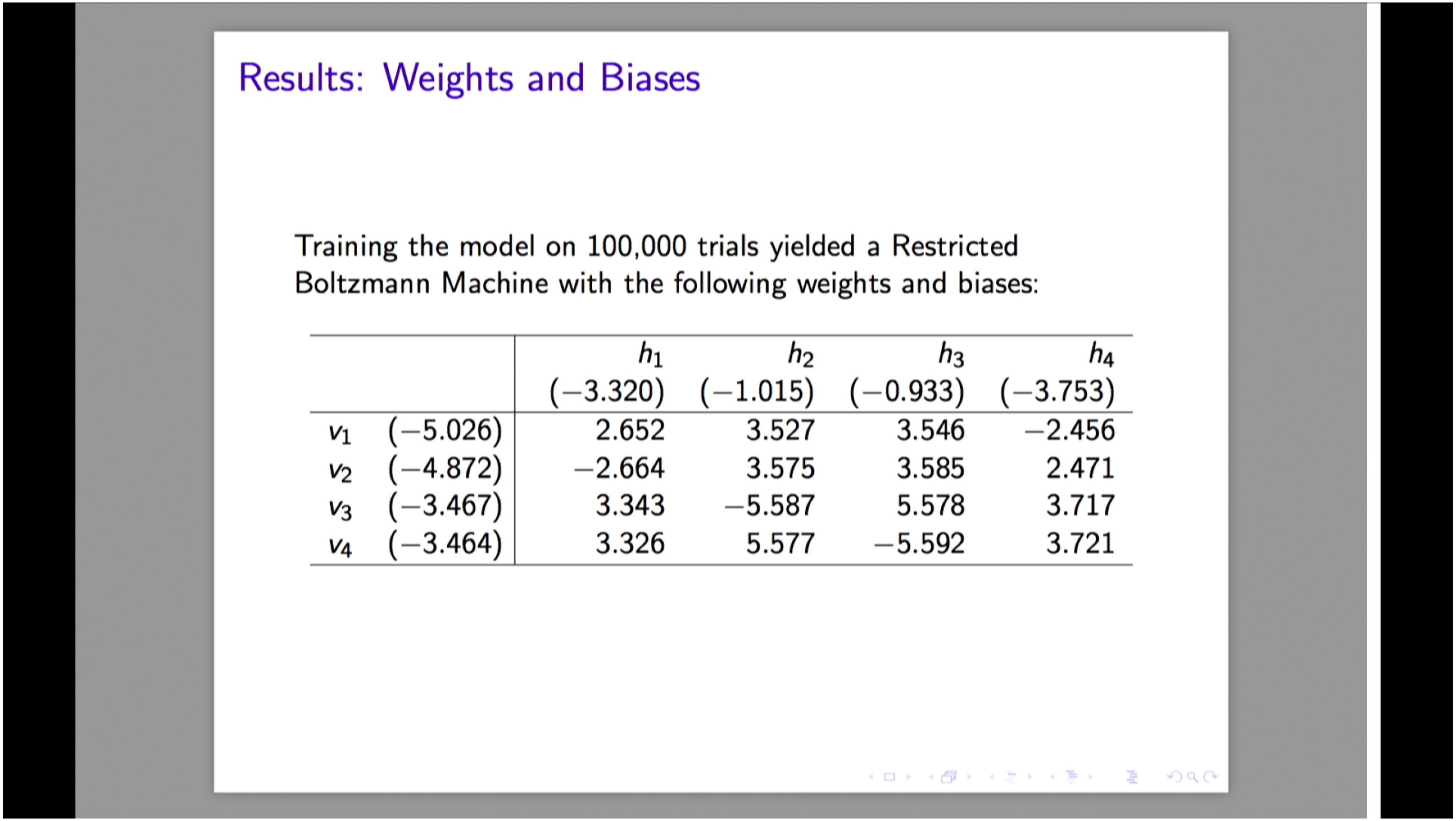
Learning

No connections between units within a layer means

$$\begin{aligned} P(h_j = 1|\mathbf{v}) &= \sigma(d_j + \textstyle\sum_i v_i w_{ij}) && (13) \\ P(v_i = 1|\mathbf{h}) &= \sigma(c_i + \textstyle\sum_j h_j w_{ij}), && (14) \end{aligned}$$

where $\sigma(x) = 1/(1 + e^{-x})$. The training data from the simulation gives us a distribution over visible vectors $\mathbf{v}$. So we are able to determine $\langle v_i h_j \rangle_{data}$ in a straightforward fashion.

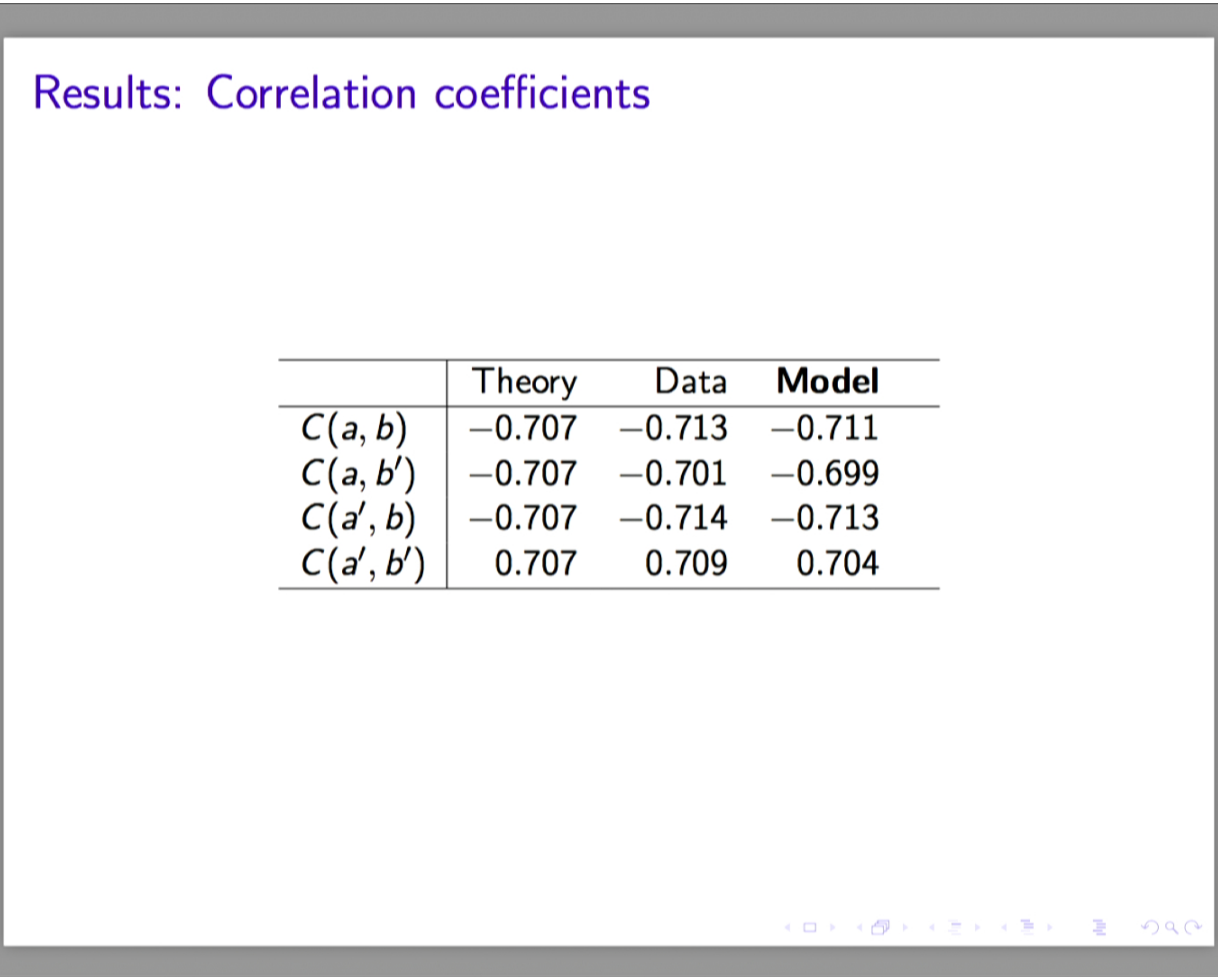
The determination of $\langle v_i h_j \rangle_{model}$ must be approximated for RBMs with more than a small number of units. However, methods exist for efficient approximation, and our results were obtained in this manner.



Results: Weights and Biases

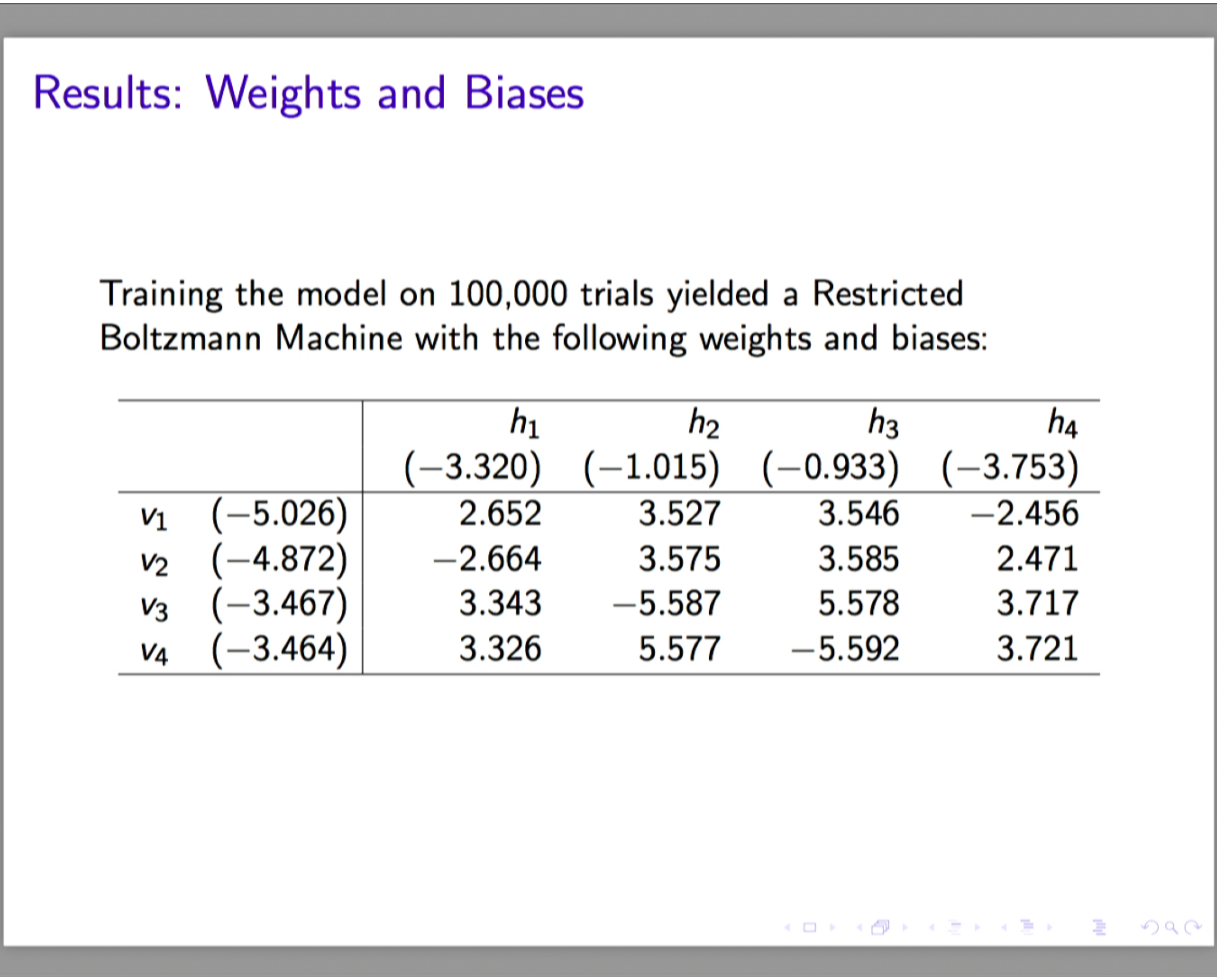
Training the model on 100,000 trials yielded a Restricted Boltzmann Machine with the following weights and biases:

		h_1	h_2	h_3	h_4
		(-3.320)	(-1.015)	(-0.933)	(-3.753)
v_1	(-5.026)	2.652	3.527	3.546	-2.456
v_2	(-4.872)	-2.664	3.575	3.585	2.471
v_3	(-3.467)	3.343	-5.587	5.578	3.717
v_4	(-3.464)	3.326	5.577	-5.592	3.721



Results: Correlation coefficients

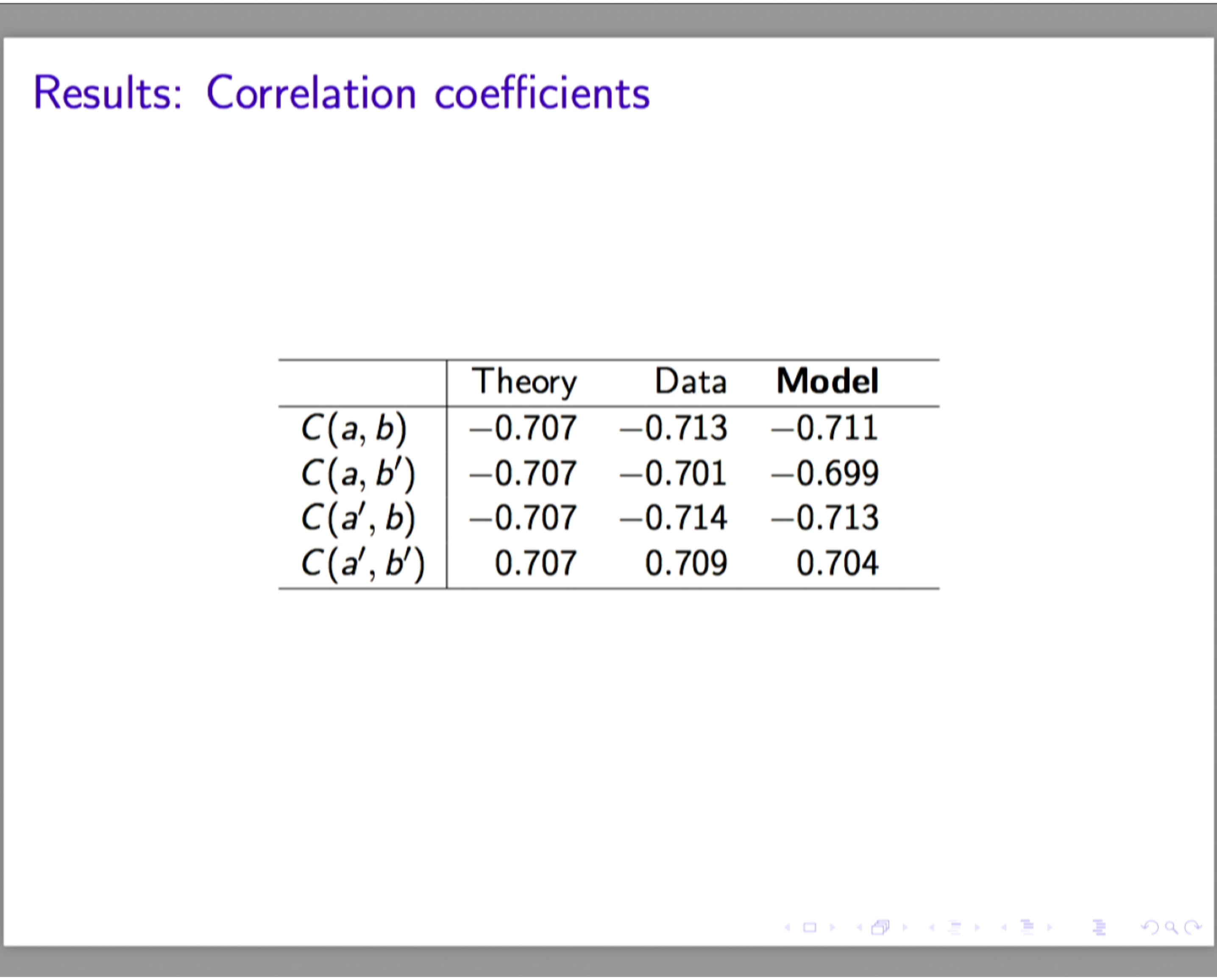
	Theory	Data	**Model**
$C(a,b)$	-0.707	-0.713	-0.711
$C(a,b')$	-0.707	-0.701	-0.699
$C(a',b)$	-0.707	-0.714	-0.713
$C(a',b')$	0.707	0.709	0.704



Results: Weights and Biases

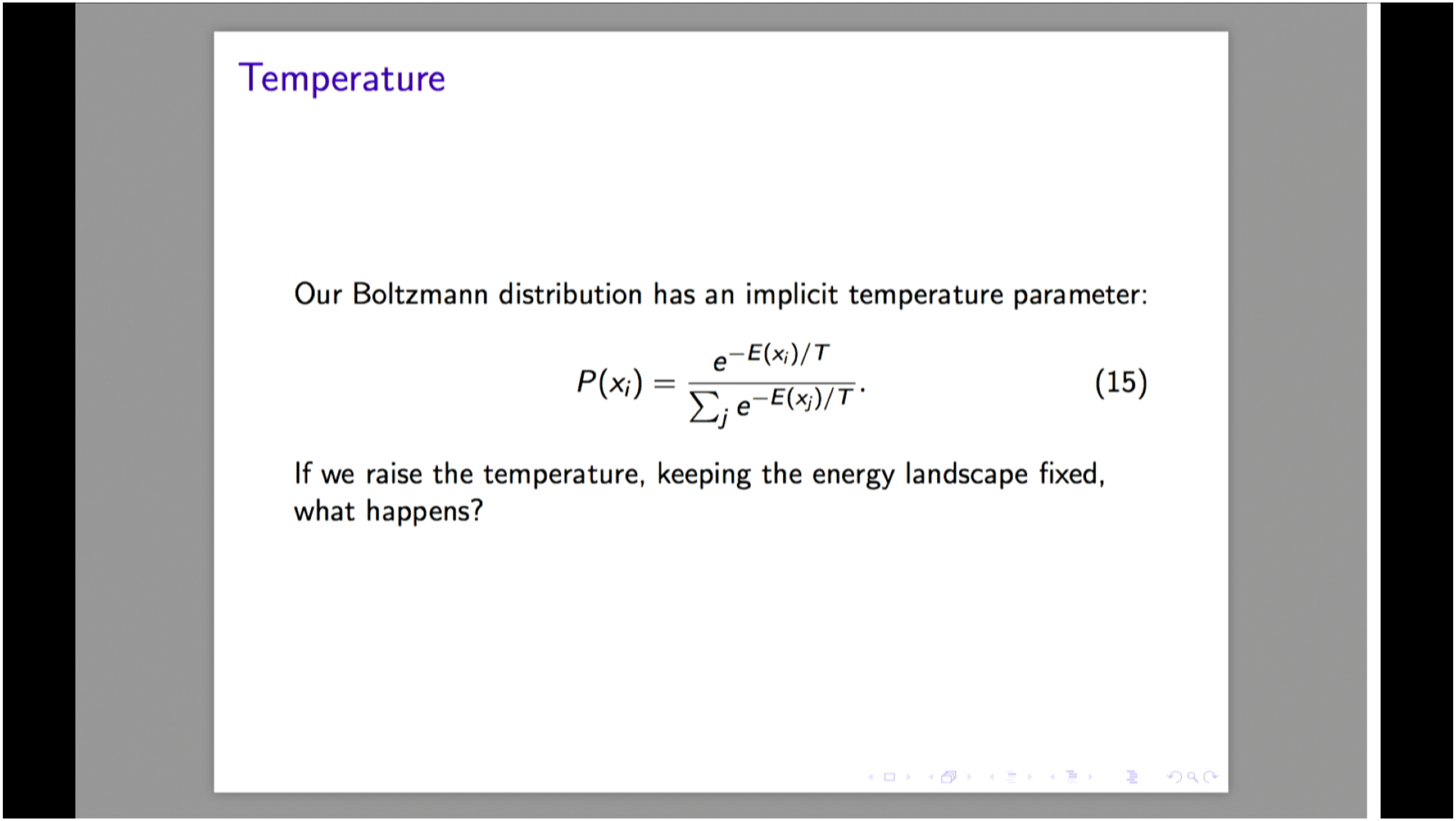
Training the model on 100,000 trials yielded a Restricted Boltzmann Machine with the following weights and biases:

		h_1	h_2	h_3	h_4
		(-3.320)	(-1.015)	(-0.933)	(-3.753)
v_1	(-5.026)	2.652	3.527	3.546	-2.456
v_2	(-4.872)	-2.664	3.575	3.585	2.471
v_3	(-3.467)	3.343	-5.587	5.578	3.717
v_4	(-3.464)	3.326	5.577	-5.592	3.721



Results: Correlation coefficients

	Theory	Data	**Model**
$C(a, b)$	-0.707	-0.713	-0.711
$C(a, b')$	-0.707	-0.701	-0.699
$C(a', b)$	-0.707	-0.714	-0.713
$C(a', b')$	0.707	0.709	0.704

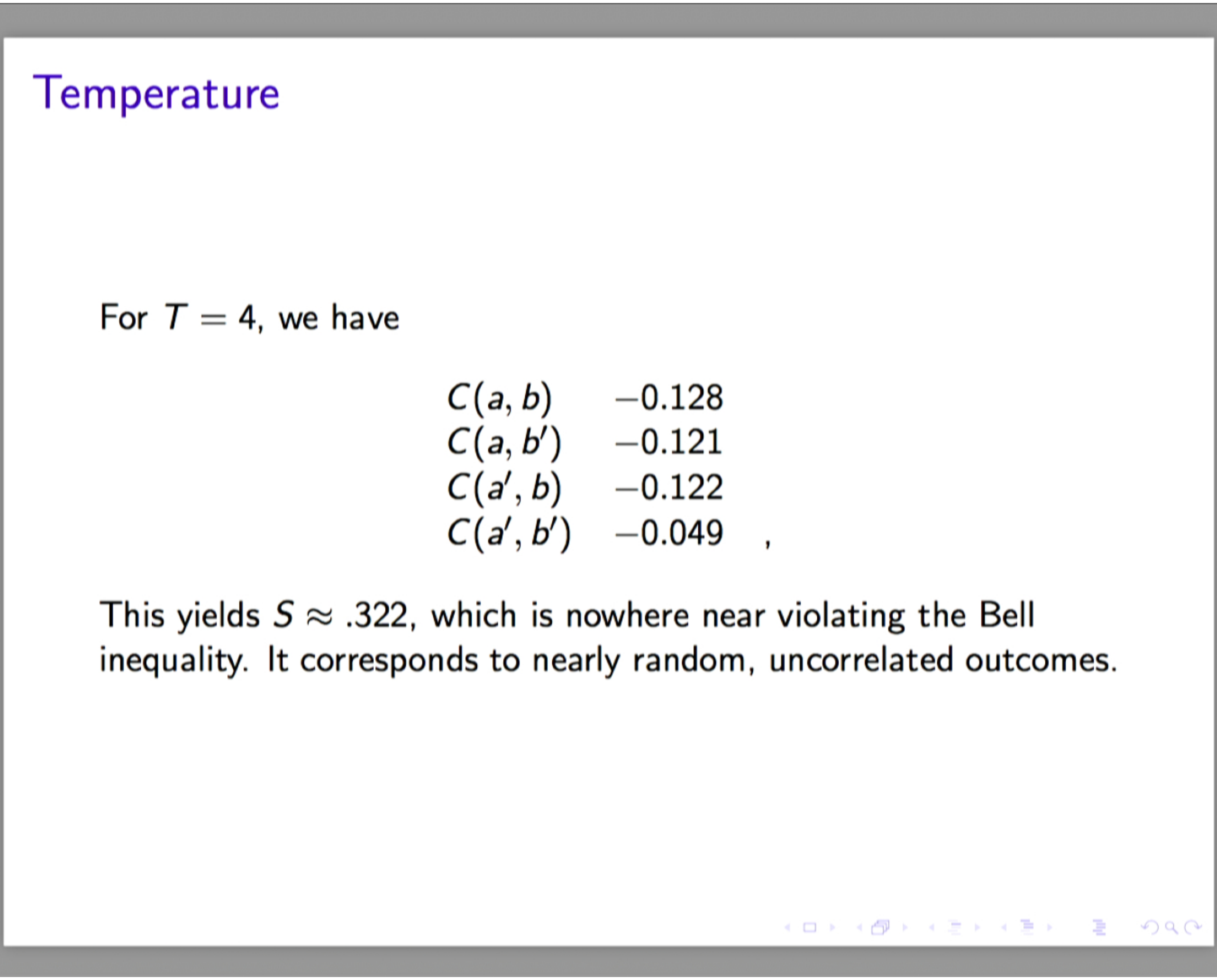


Temperature

Our Boltzmann distribution has an implicit temperature parameter:

$$P(x_i) = \frac{e^{-E(x_i)/T}}{\sum_j e^{-E(x_j)/T}}. \tag{15}$$

If we raise the temperature, keeping the energy landscape fixed, what happens?

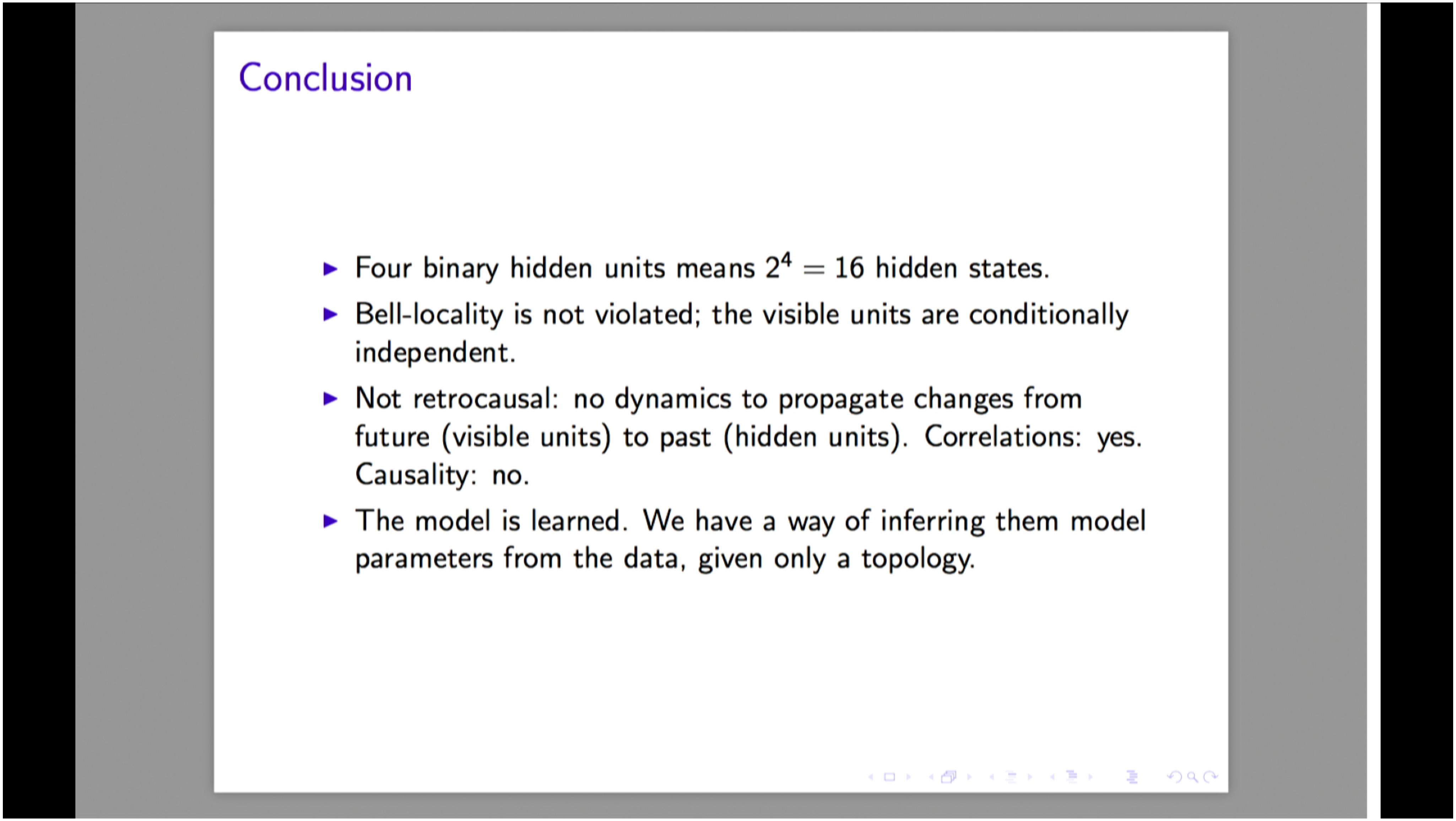


Temperature

For $T = 4$, we have

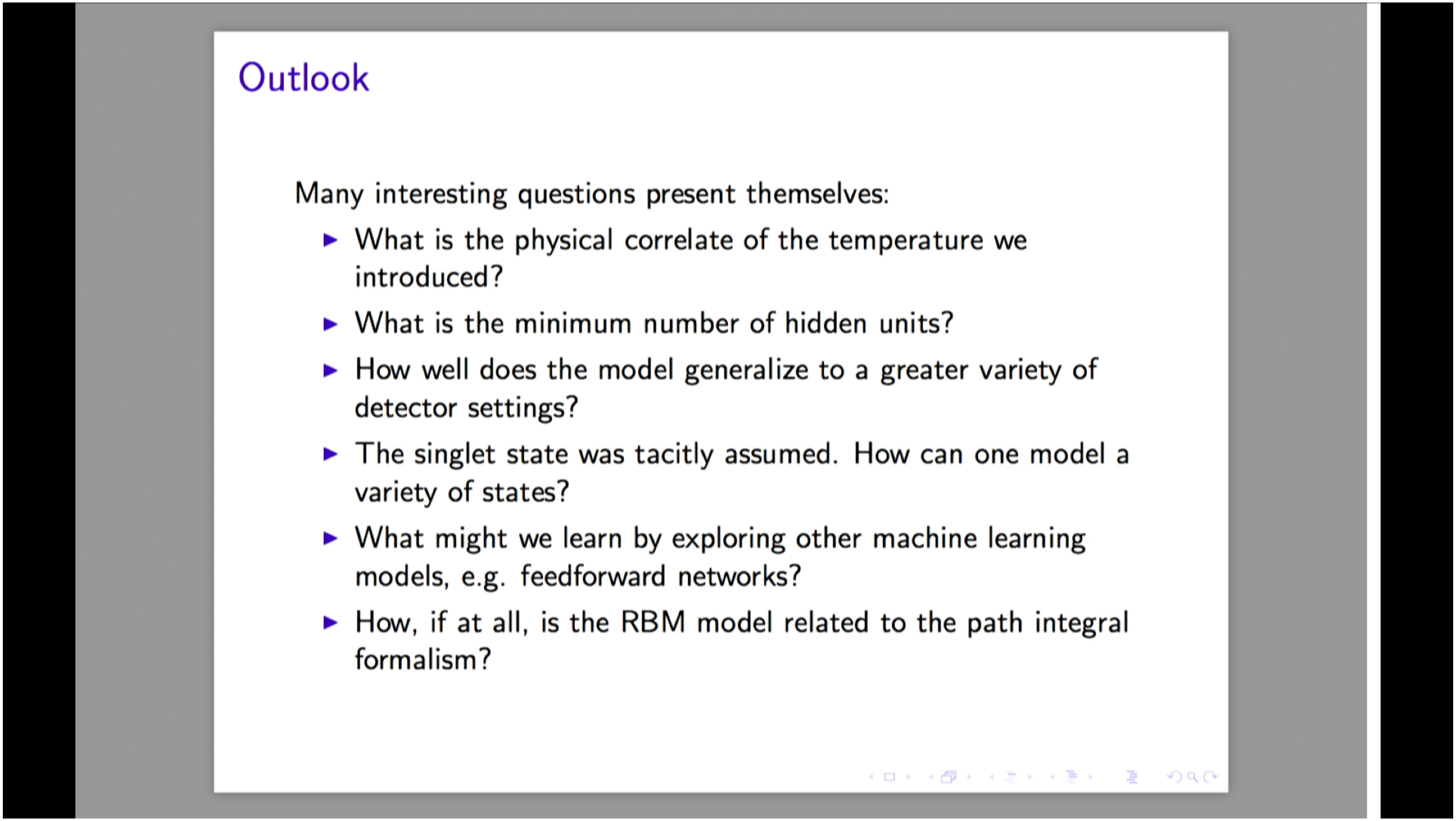
$C(a, b)$	-0.128	
$C(a, b')$	-0.121	
$C(a', b)$	-0.122	
$C(a', b')$	-0.049	,

This yields $S \approx .322$, which is nowhere near violating the Bell inequality. It corresponds to nearly random, uncorrelated outcomes.



Conclusion

- Four binary hidden units means $2^4 = 16$ hidden states.
- Bell-locality is not violated; the visible units are conditionally independent.
- Not retrocausal: no dynamics to propagate changes from future (visible units) to past (hidden units). Correlations: yes. Causality: no.
- The model is learned. We have a way of inferring them model parameters from the data, given only a topology.



Outlook

Many interesting questions present themselves:

- What is the physical correlate of the temperature we introduced?
- What is the minimum number of hidden units?
- How well does the model generalize to a greater variety of detector settings?
- The singlet state was tacitly assumed. How can one model a variety of states?
- What might we learn by exploring other machine learning models, e.g. feedforward networks?
- How, if at all, is the RBM model related to the path integral formalism?