Title: PSI 2017/2018 - Machine Learning for Many Body Physics - Lecture 12

Date: Apr 23, 2018 09:00 AM

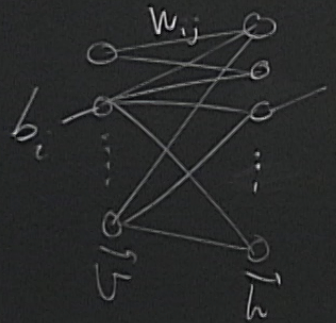
URL: http://pirsa.org/18040064

Abstract:

RBM:

$$E_\lambda = -b^T v - c^T h - v^T W h$$

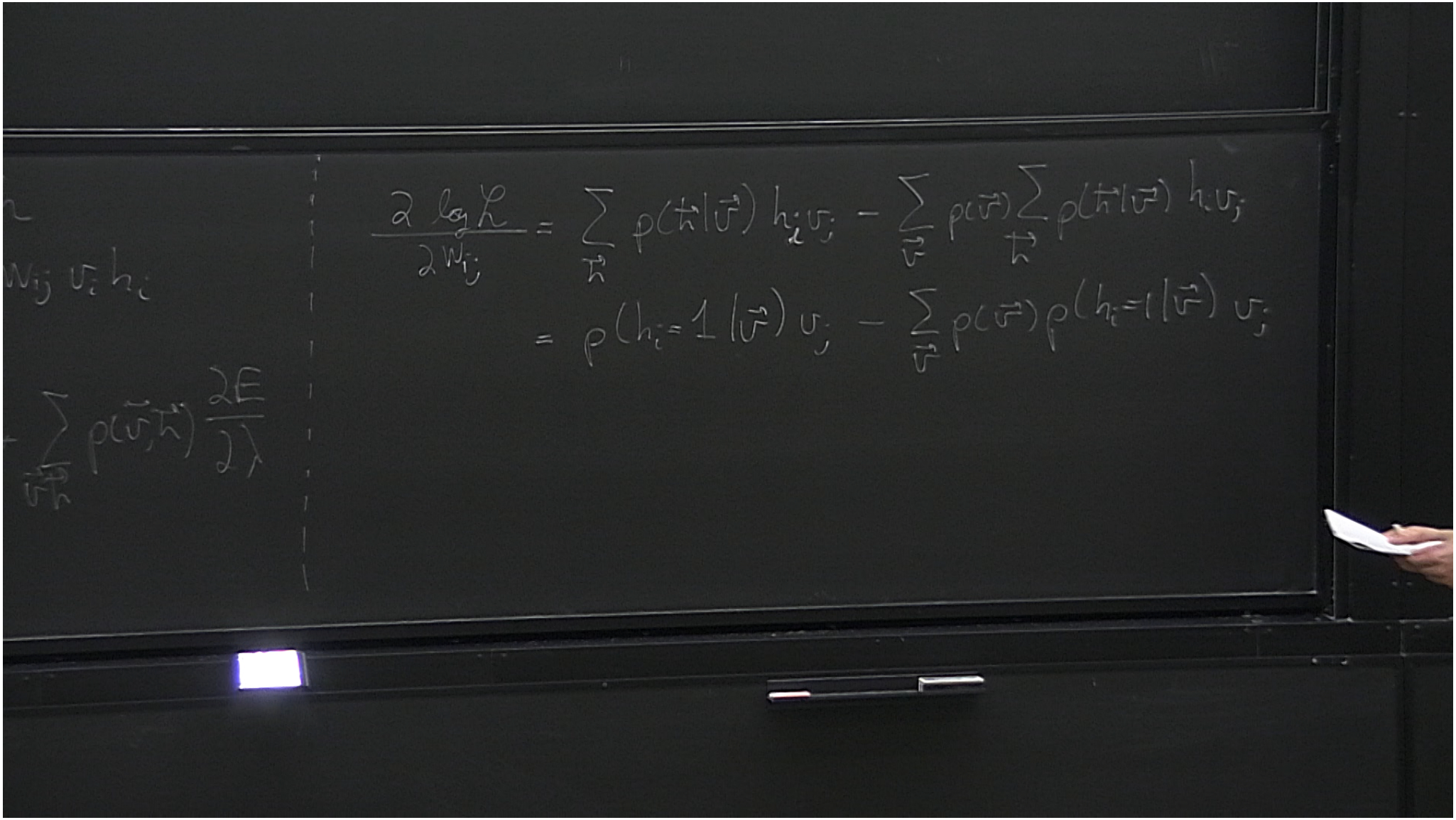
$$= -\sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j - \sum_{ij} w_{ij} v_i h_j$$



$p(\vec{v}, \vec{h})$

$$\frac{\partial \log p}{\partial \lambda} = -\sum_{\vec{h}} p(\vec{h}|\vec{v}) \frac{\partial E}{\partial \lambda} + \sum_{\vec{v},\vec{h}} p(\vec{v},\vec{h}) \frac{\partial E}{\partial \lambda}$$

$\lambda = (w, b, c)$


$\frac{2 \log \mathcal{L}}{2 W_{ij}} = \sum_{\vec{h}} p(\vec{h}|\vec{v})\, h_i v_j - \sum_{\vec{v}} p(\vec{v}) \sum_{\vec{h}} p(\vec{h}|\vec{v})\, h_i v_j$
$= p(h_i = 1|\vec{v})\, v_j - \sum_{\vec{v}} p(\vec{v})\, p(h_i = 1|\vec{v})\, v_j$

$\lambda = (W, b, c)$

In practice this is averaged over a mini-batch, size $b << l$

CD$_k$ can be used

$$\frac{1}{b} \sum_{\vec{v} \in \text{mini batch}} \frac{\partial p}{\partial w_{ij}} = \langle v_i h_j \rangle_{p(\vec{h}|\vec{v})} - \langle v_i h_j \rangle_{p(\vec{v},\vec{h})}$$

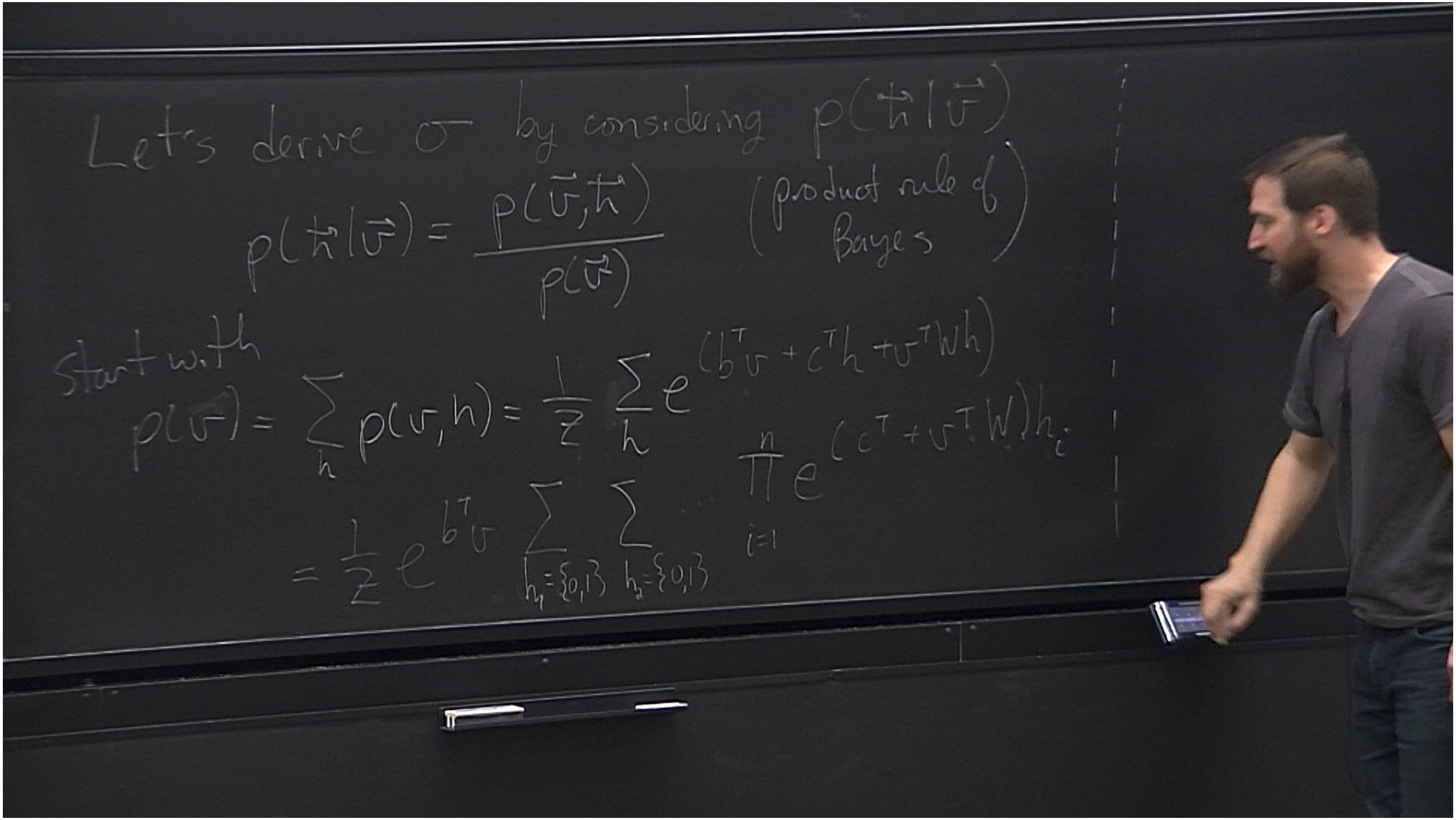
v_i is "clamped" to the data

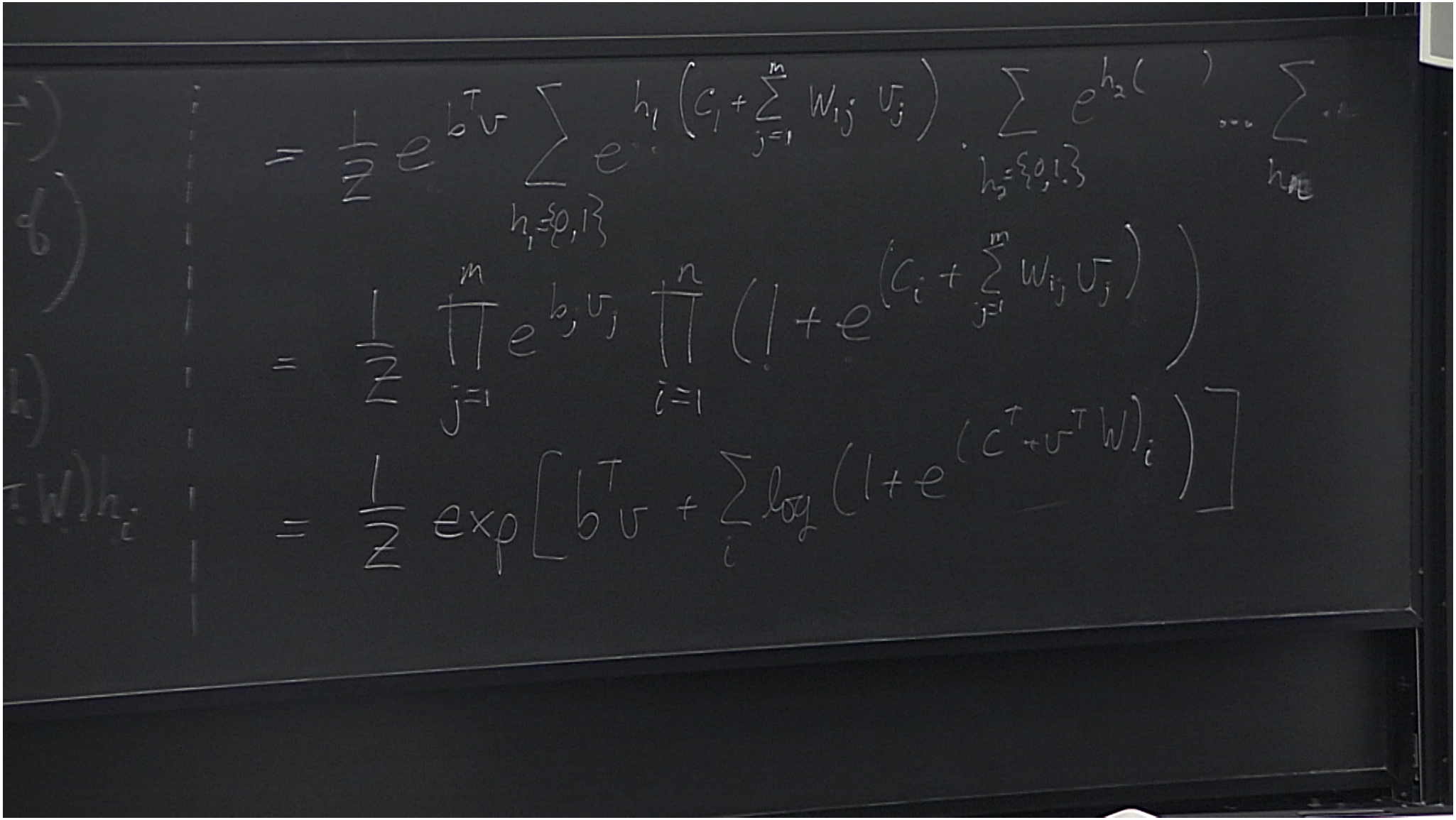
expectation value for the stationary distribution of the model

This calculation has to be completed for all weights & biases

$$\nabla_\lambda(\log_2 P_h) = \begin{cases} \langle v h^T \rangle_{p(h|v)} - \langle v h^T \rangle_{p(v,h)} & \lambda = W \\ \langle v \rangle_{p(h|v)} - \langle v \rangle_{p(v,h)} & \lambda = b \\ \langle h \rangle_{p(h|v)} - \langle h \rangle_{p(v,h)} & \lambda = c \end{cases}$$

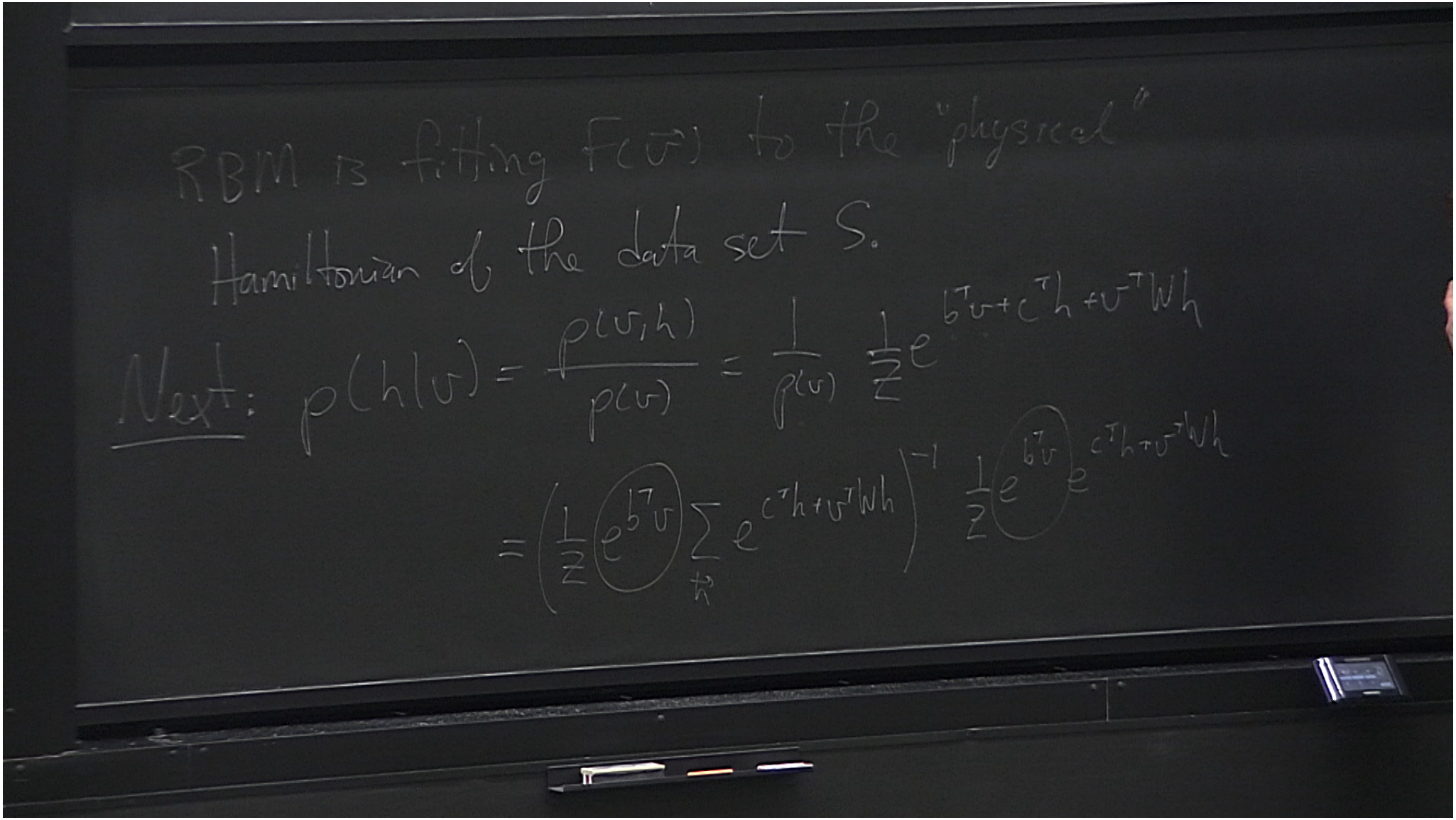
and $\lambda' = \lambda - \eta \nabla KL$


Let's derive σ by considering $p(\vec{h}|\vec{v})$
$p(\vec{h}|\vec{v}) = \frac{p(\vec{v},\vec{h})}{p(\vec{v})}$ (product rule of Bayes)
Start with
$p(v) = \sum_h p(v,h) = \frac{1}{Z}\sum_h e^{(b^T v + c^T h + v^T W h)}$
$= \frac{1}{Z} e^{b^T v} \sum_{h_1=\{0,1\}} \sum_{h_2=\{0,1\}} \prod_{i=1}^{n} e^{(c^T + v^T W)_i h_i}$


$= \frac{1}{Z} e^{b^T v} \sum_{h_1 \in \{0,1\}} e^{h_1 \left(c_1 + \sum_{j=1}^{m} W_{1j} v_j\right)} \sum_{h_2 \in \{0,1\}} e^{h_2 (\ )} \cdots \sum_{h_n}$
$= \frac{1}{Z} \prod_{j=1}^{m} e^{b_j v_j} \prod_{i=1}^{n} \left(1 + e^{\left(c_i + \sum_{j=1}^{m} W_{ij} v_j\right)}\right)$
$= \frac{1}{Z} \exp\left[b^T v + \sum_i \log\left(1 + e^{(c^T + v^T W)_i}\right)\right]$

RBM is fitting $F(\vec{v})$ to the "physical" Hamiltonian of the data set S.

Next: $p(h|v) = \frac{p(v,h)}{p(v)} = \frac{1}{p(v)} \frac{1}{Z} e^{b^T v + c^T h + v^T W h}$

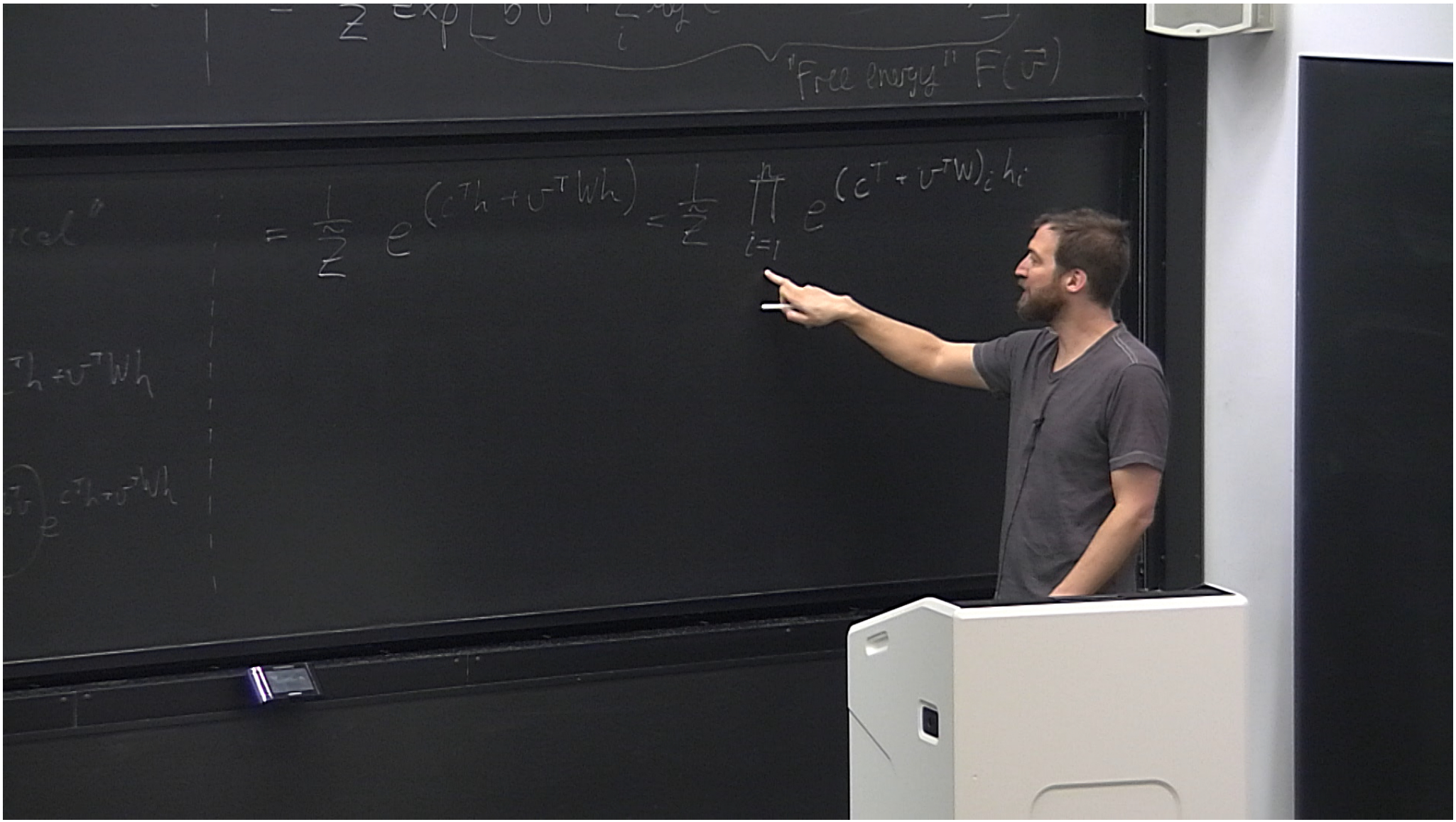

RBM is fitting F(v) to the "physical"
Hamiltonian of the data set S.
Next: $p(h|v) = \frac{p(v,h)}{p(v)} = \frac{1}{p(v)} \frac{1}{Z} e^{b^T v + c^T h + v^T W h}$
$= \left(\frac{1}{Z} e^{b^T v} \sum_{h'} e^{c^T h + v^T W h} \right)^{-1} \frac{1}{Z} e^{b^T v} e^{c^T h + v^T W h}$

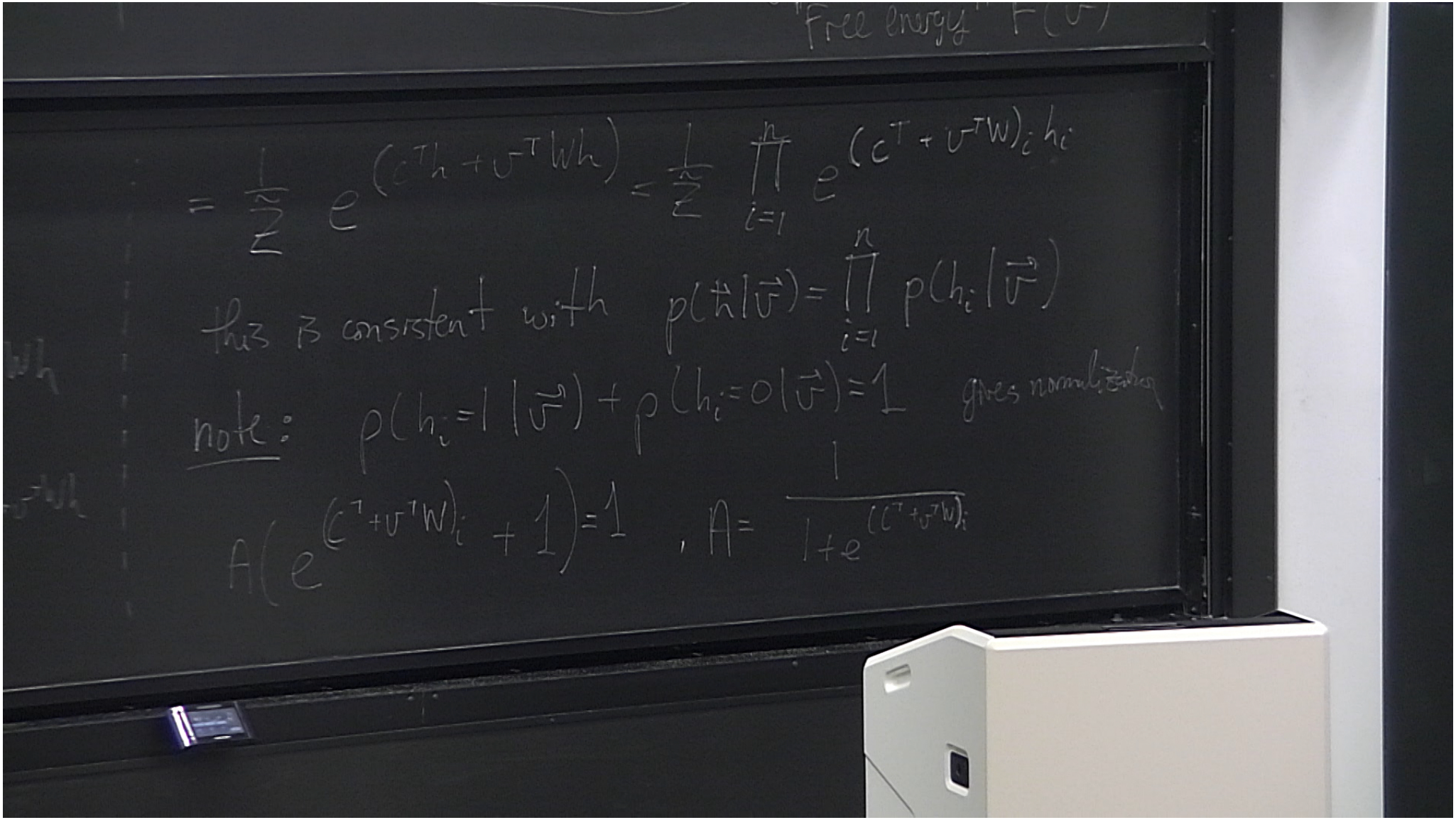
$h_1 = \{0,1\}$ $h_2 = \{0,1\}$

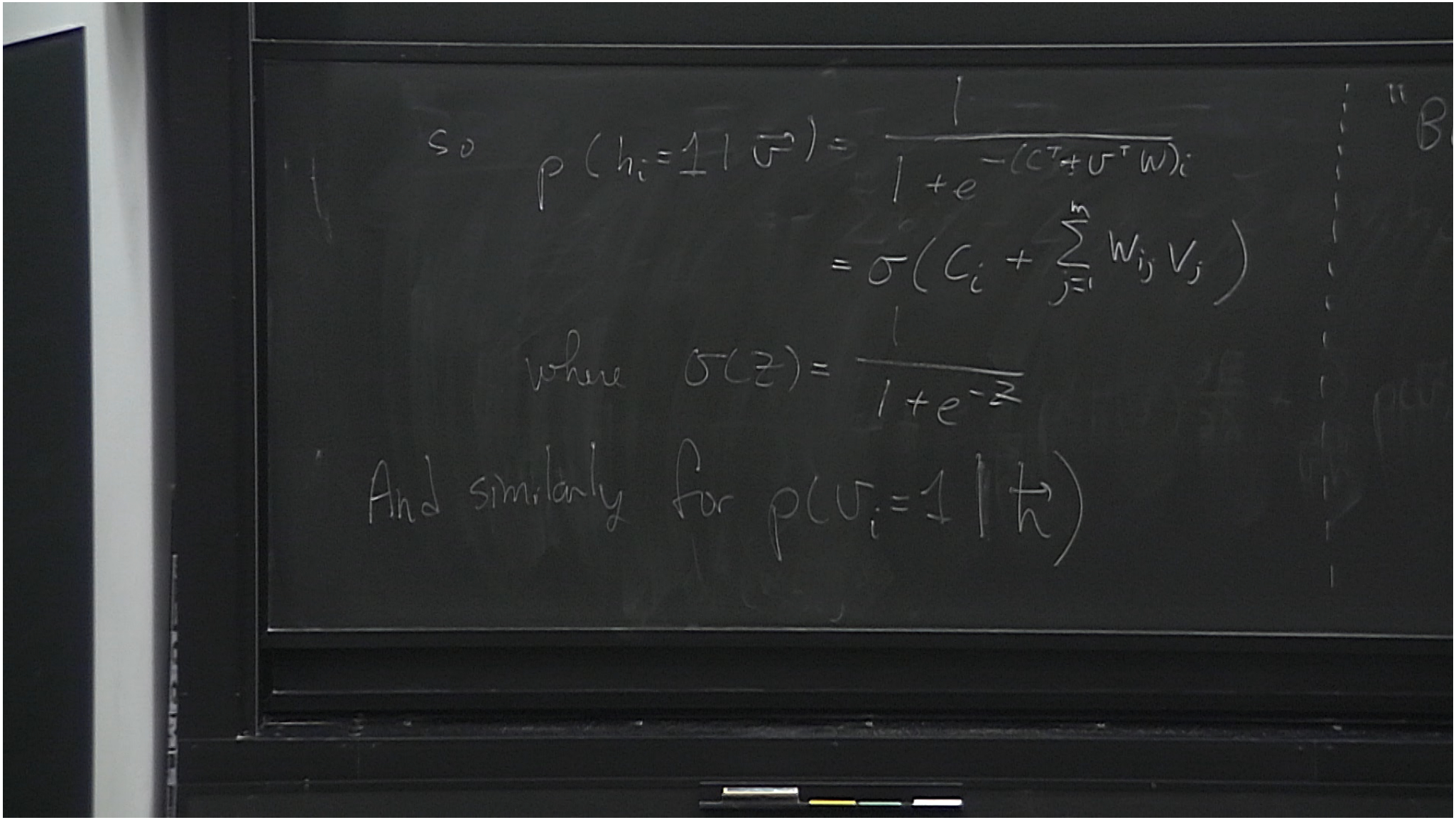
RBM is fitting $F(v)$ to the "physical" Hamiltonian of the data set S.

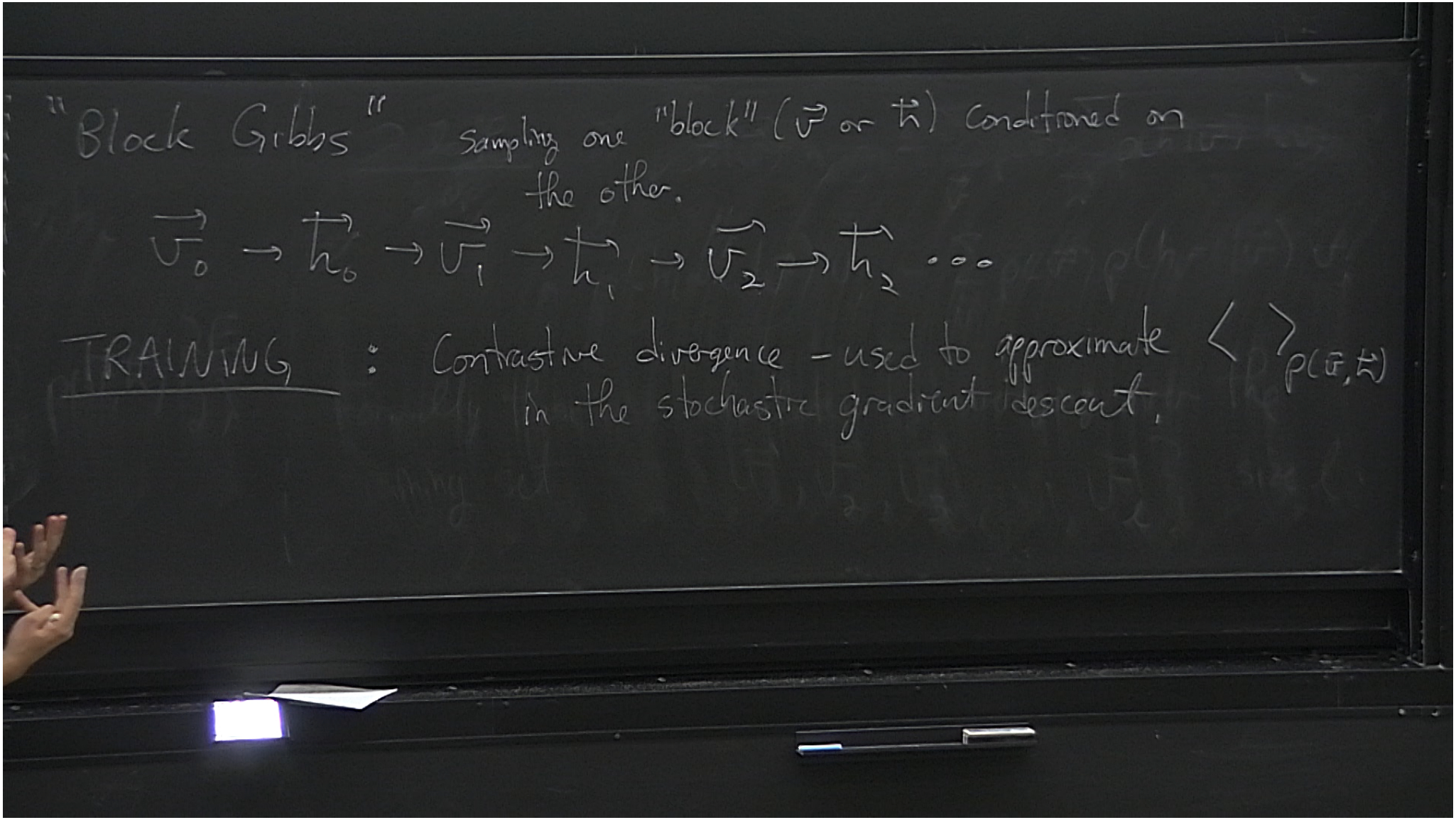
Next: $p(h|v) = \frac{p(v,h)}{p(v)} = \frac{1}{p(v)} \frac{1}{Z} e^{b^T v + c^T h + v^T W h}$

$= \left(\frac{1}{Z} e^{b^T v} \sum_{h'} e^{c^T h + v^T W h} \right)^{-1} \frac{1}{Z} e^{b^T v} e^{c^T h + v^T W h}$


"Free energy" F(v)


Free energy F(v)
$= \frac{1}{Z} e^{(c^T h + v^T W h)} = \frac{1}{Z} \prod_{i=1}^{n} e^{(c^T + v^T W)_i h_i}$
thus B consistent with $p(\vec{h}|\vec{v}) = \prod_{i=1}^{n} p(h_i|\vec{v})$
note: $p(h_i=1|\vec{v}) + p(h_i=0|\vec{v}) = 1$ gives normalization
$A(e^{(c^T+v^TW)_i} + 1) = 1$, $A = \frac{1}{1+e^{(c^T+v^TW)_i}}$


So $p(h_i = 1 \mid \vec{v}) = \frac{1}{1 + e^{-(C^T + v^T W)_i}}$
$= \sigma\left(C_i + \sum_{j=1}^{m} W_{ij} V_j\right)$
where $\sigma(z) = \frac{1}{1 + e^{-z}}$
And similarly for $p(v_i = 1 \mid \vec{h})$


"Block Gibbs" sampling one "block" ($\vec{v}$ or $\vec{h}$) conditioned on the other.
$\vec{v}_0 \to \vec{h}_0 \to \vec{v}_1 \to \vec{h}_1 \to \vec{v}_2 \to \vec{h}_2 \cdots$
TRAINING: Contrastive divergence - used to approximate $\langle \rangle_{p(\vec{v},\vec{h})}$ in the stochastic gradient descent.

"Block Gibbs" sampling one "block" ($\vec{v}$ or $\vec{h}$) conditioned on the other.

$$\vec{v}_0 \to \vec{h}_0 \to \vec{v}_1 \to \vec{h}_1 \to \vec{v}_2 \to \vec{h}_2 \cdots \vec{v}_k \to \vec{h}_k$$

TRAINING: Contrastive divergence - used to approximate $\langle \rangle_{p(\vec{v},\vec{h})}$ in the stochastic gradient descent.

Can calculate this by equilibrating the RBM and sampling $\vec{v}, \vec{h}$

CDk can be used: This calculation has to be completed for all weights & biases

$$\lambda = W \qquad \langle v h^T \rangle_{p(h|v)} - \langle v h^T \rangle_{p(v,h)}$$

Much simpler: run this Gibbs chain only for $k \sim O(1)$ steps ($k=1$ sometimes works)

- $\vec{v}_0$ initialized from a training example
- $\vec{h}_0$ obtained from $p(\vec{h} | \vec{v}_0)$
- repeat for k-steps

CD_k expression for ∇KL is

$$\frac{2\log p}{2\lambda} \simeq -\sum_{\vec{h}} p(\vec{h}|\vec{v}_0)\,\frac{2E_0}{2\lambda} + \sum_{\vec{h}} p(\vec{h}|\vec{v}_k)\,\frac{2E_k}{2\lambda}$$

giving for example

$$W' = W + \frac{\eta}{b} \sum_{\vec{v} \in \text{mini batch}} \left(\vec{v}_0 \vec{h}_0^T - \vec{v}_k \vec{h}_k^T\right)$$

The trained RBM & sampling:

Once trained we have a "model"

$P_\lambda(\vec{v}, \vec{h})$ such that

$$P_\lambda(\vec{v}) = \sum_{\vec{h}} P_\lambda(\vec{v}, \vec{h}) \simeq q(\vec{v})$$

↑ underlies data in S

We sample through Block Gibbs sampling $\vec{v}$ and $\vec{h}$

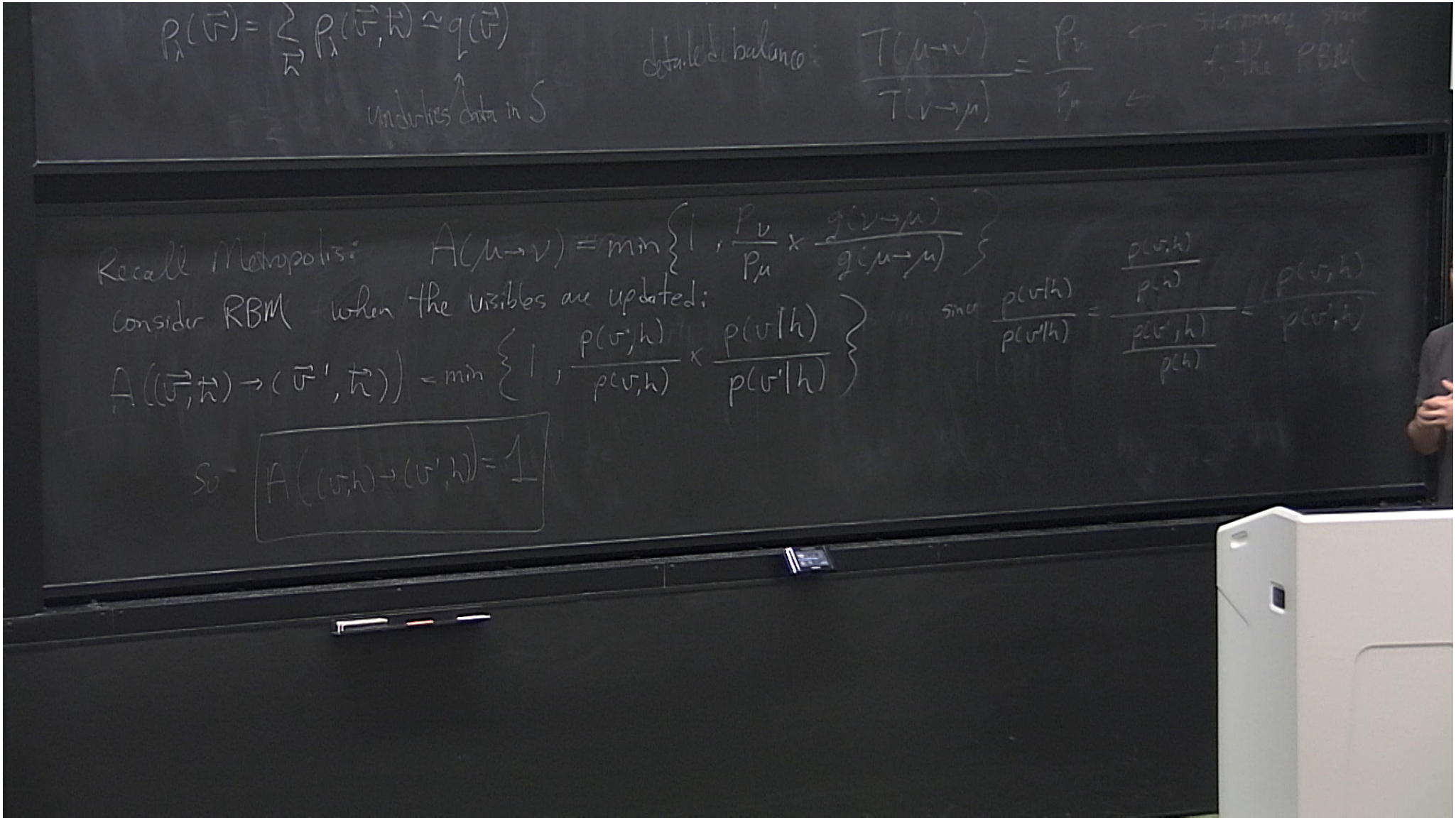
(we are interested in $\vec{v}$, $\langle O_v \rangle$)

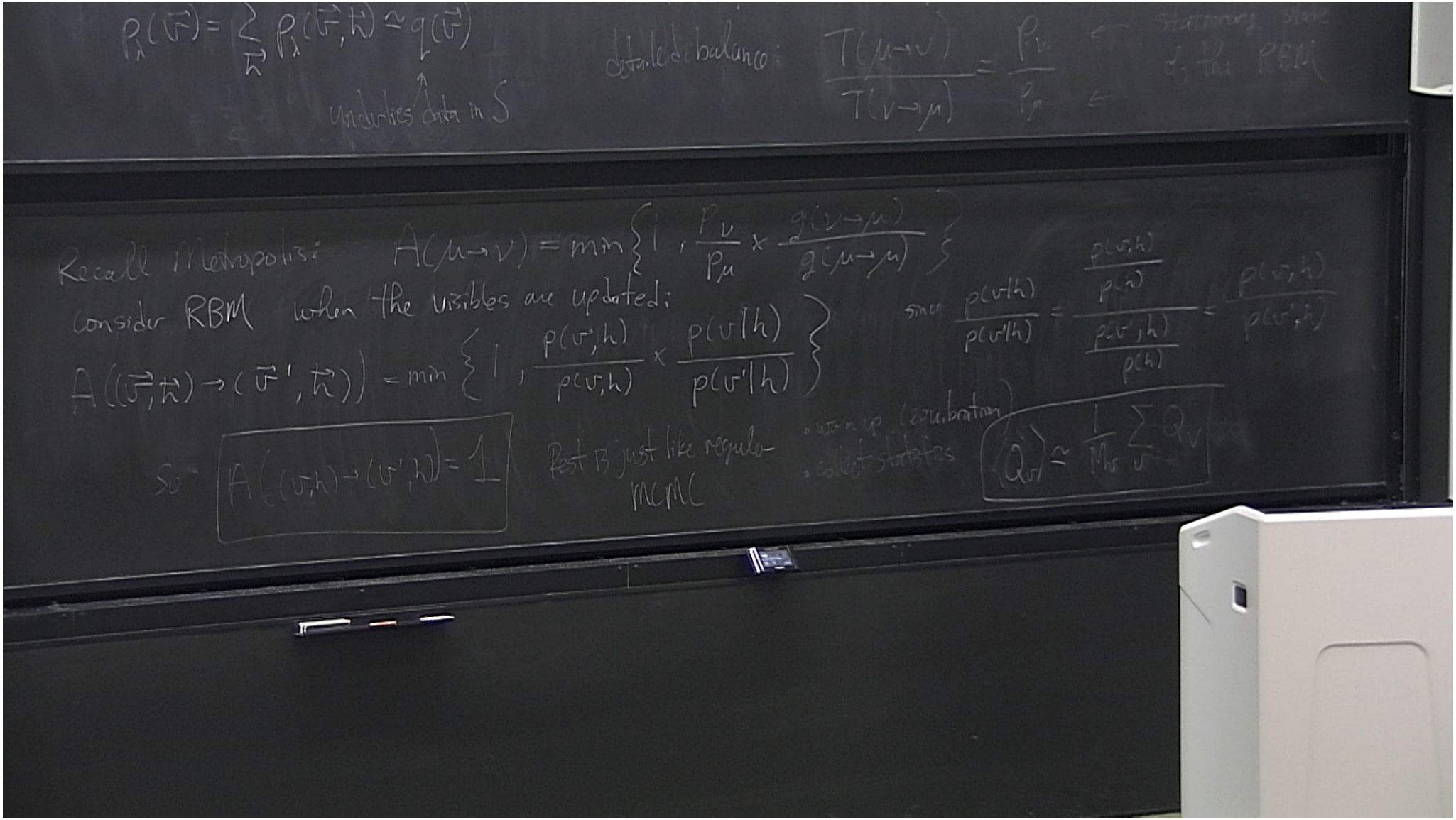
Recall: $T(\mu \to \nu) = g(\mu \to \nu) A(\mu \to \nu)$

"transition prob." "selection prob." "acceptance ratio"

detailed balance: $\frac{T(\mu \to \nu)}{T(\nu \to \mu)} = \frac{P_\nu}{P_\mu}$ ← stationary state of the RBM

$(c^T + v^T W)_i h_i$


$P_\lambda(\vec{v}) = \sum_{\vec{h}} P_\lambda(\vec{v},\vec{h}) \simeq q(\vec{v})$
underlies data in S
detailed balance: $\frac{T(\mu\to\nu)}{T(\nu\to\mu)} = \frac{P_\nu}{P_\mu}$
stationary state of the RBM
Recall Metropolis: $A(\mu\to\nu) = \min\left\{1, \frac{P_\nu}{P_\mu} \times \frac{q(\nu\to\mu)}{q(\mu\to\mu)}\right\}$
consider RBM when the visibles are updated:
$A((\vec{v},\vec{h})\to(\vec{v}',\vec{h})) = \min\left\{1, \frac{p(v',h)}{p(v,h)} \times \frac{p(v|h)}{p(v'|h)}\right\}$
since $\frac{p(v|h)}{p(v'|h)} = \frac{\frac{p(v,h)}{p(h)}}{\frac{p(v',h)}{p(h)}} = \frac{p(v,h)}{p(v',h)}$
So $A((v,h)\to(v',h)) = 1$


$P_\lambda(\vec{v}) = \sum_{\vec{h}} P_\lambda(\vec{v},\vec{h}) \simeq q(\vec{v})$
marginalizes
data in S
detailed balance: $\frac{T(\mu\to\nu)}{T(\nu\to\mu)} = \frac{P_\nu}{P_\mu}$
stationary state of the RBM
Recall Metropolis: $A(\mu\to\nu) = \min\left\{1, \frac{P_\nu}{P_\mu} \times \frac{q(\nu\to\mu)}{q(\mu\to\mu)}\right\}$
consider RBM when the visibles are updated:
$A((\vec{v},\vec{h})\to(\vec{v}',\vec{h})) = \min\left\{1, \frac{p(v',h)}{p(v,h)} \times \frac{p(v|h)}{p(v'|h)}\right\}$
since $\frac{p(v|h)}{p(v'|h)} = \frac{\frac{p(v,h)}{p(h)}}{\frac{p(v',h)}{p(h)}} = \frac{p(v,h)}{p(v',h)}$
So $A((v,h)\to(v',h)) = 1$
Rest B just like regular MCMC
• warm up (equilibration)
• collect statistics
$\langle Q_{ij}\rangle \simeq \frac{1}{M_{\text{MC}}} \sum_{\vec{v}} Q_{ij}$