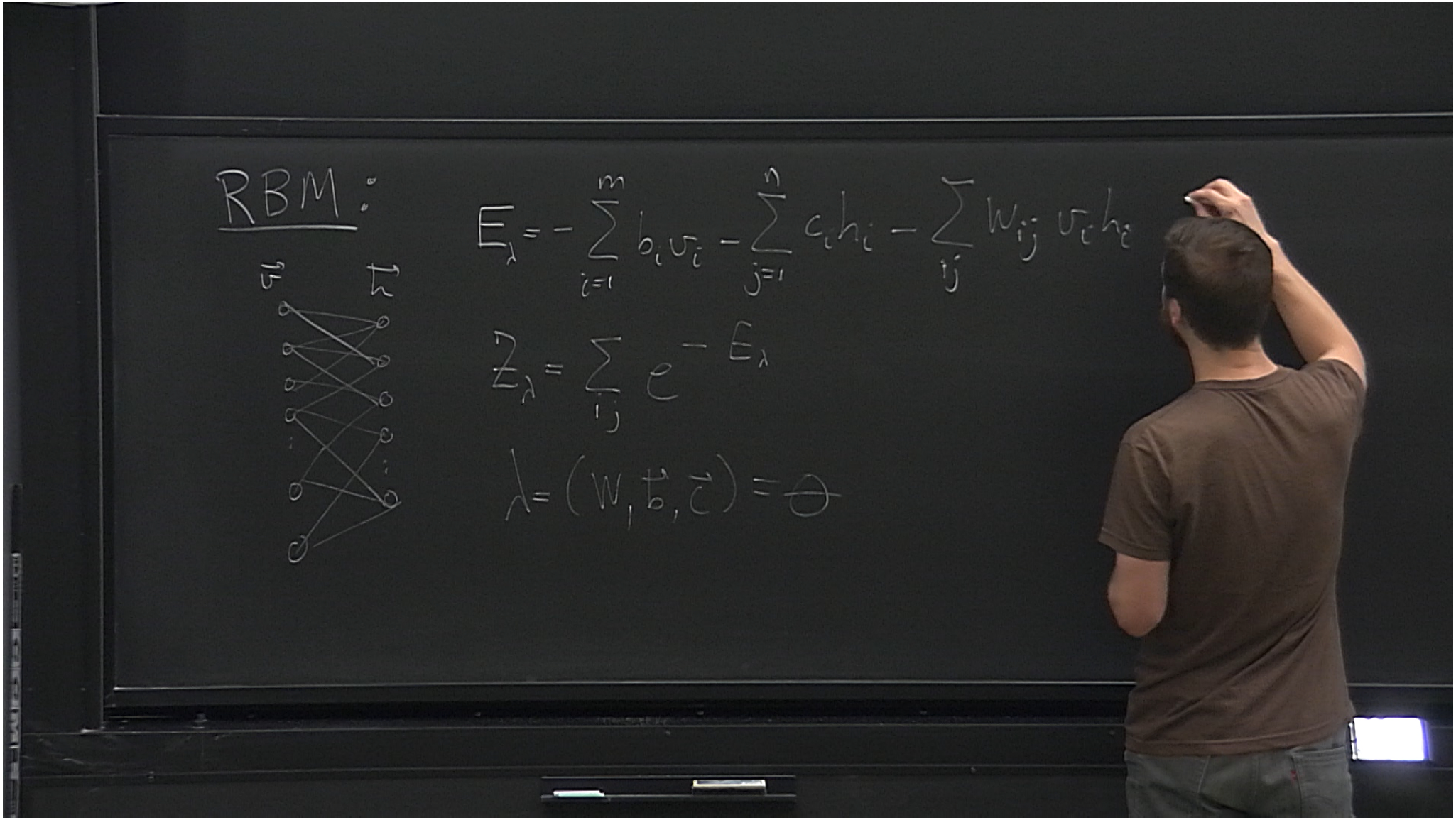
Title: PSI 2017/2018 - Machine Learning for Many Body Physics - Lecture 11

Date: Apr 20, 2018 09:00 AM

URL: http://pirsa.org/18040063

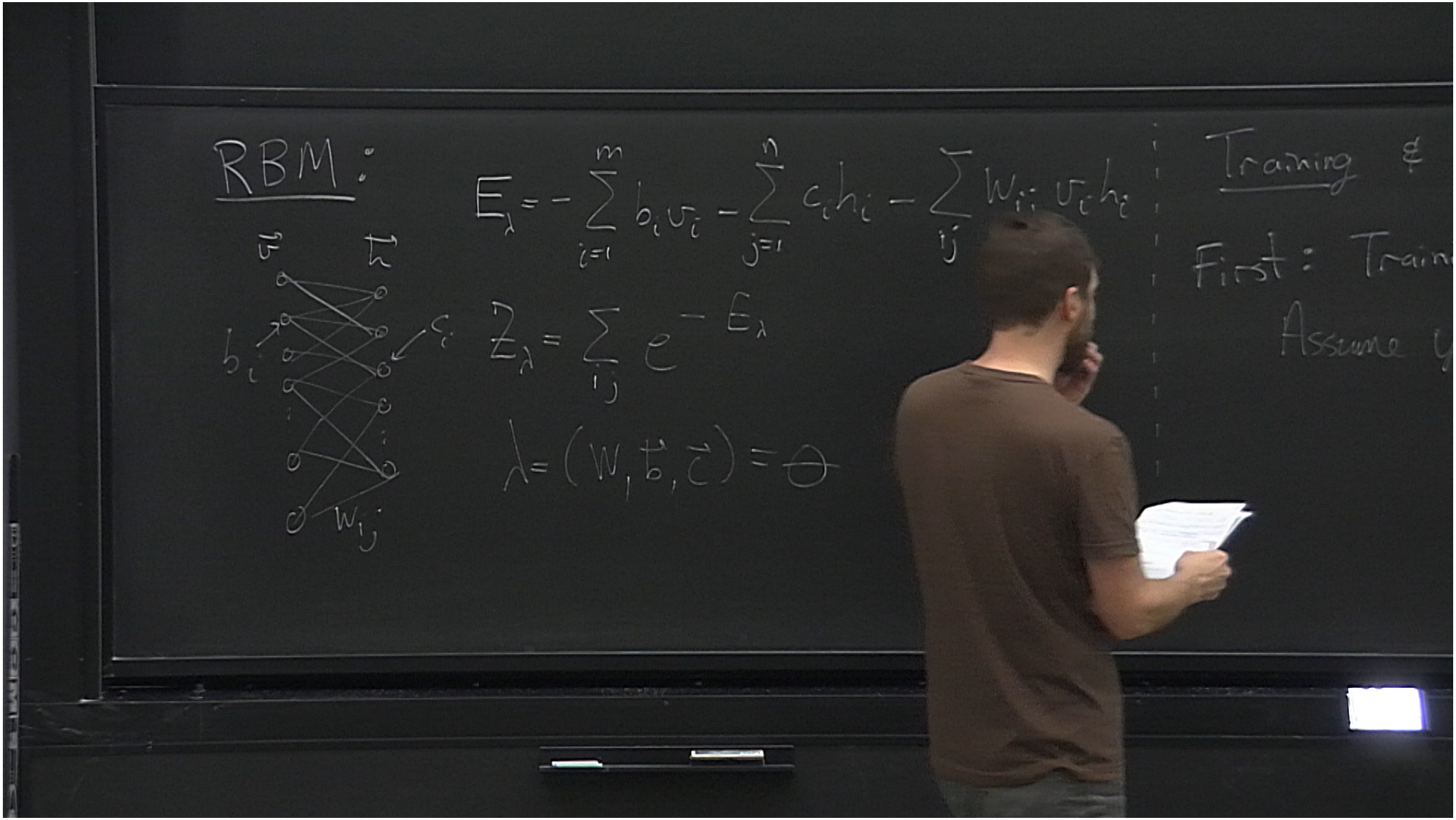
Abstract:

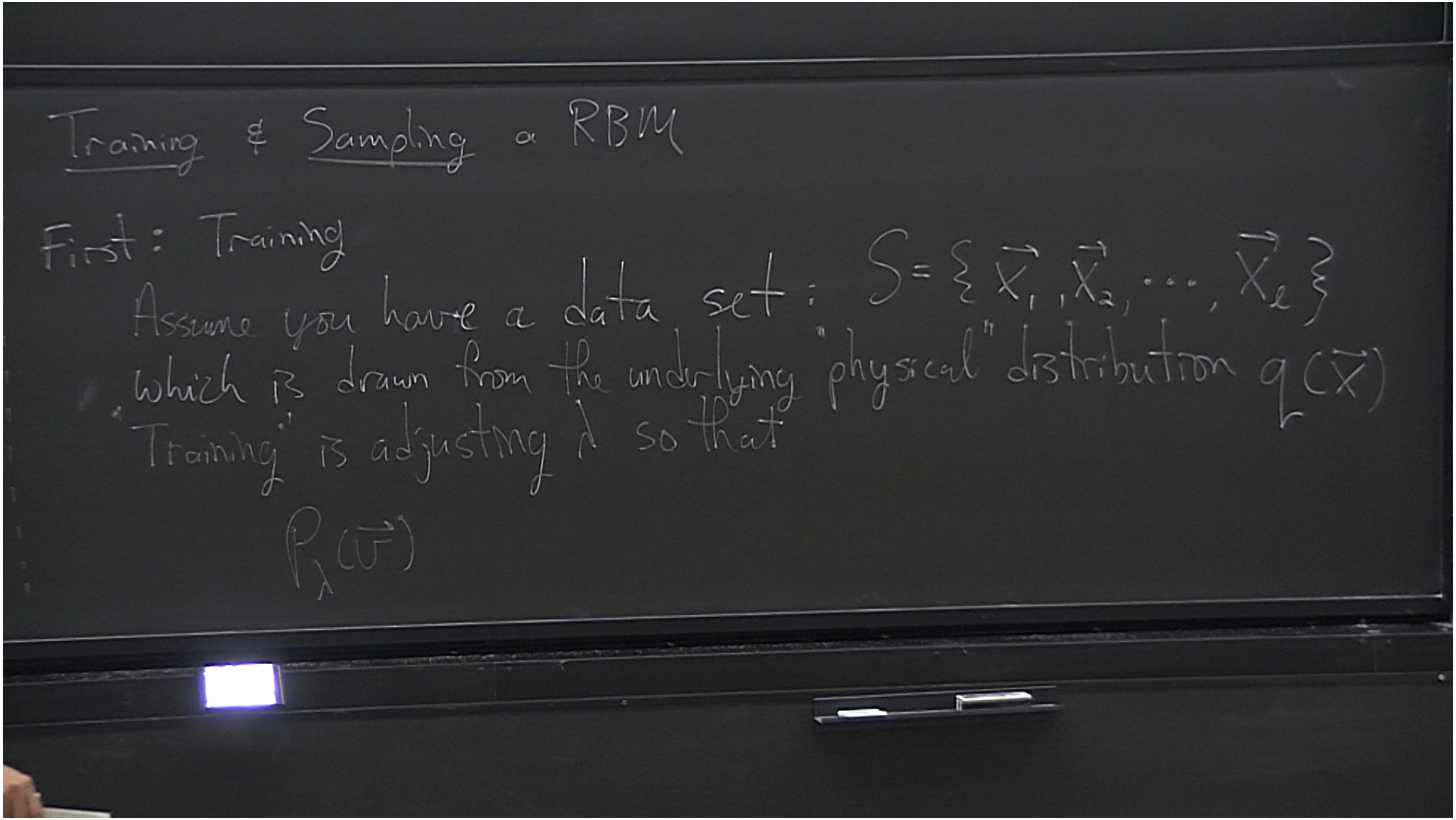

RBM:
$E_\lambda = -\sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j - \sum_{ij} w_{ij} v_i h_j$
$Z_\lambda = \sum_{ij} e^{-E_\lambda}$
$\lambda = (W, \vec{b}, \vec{c}) = \theta$

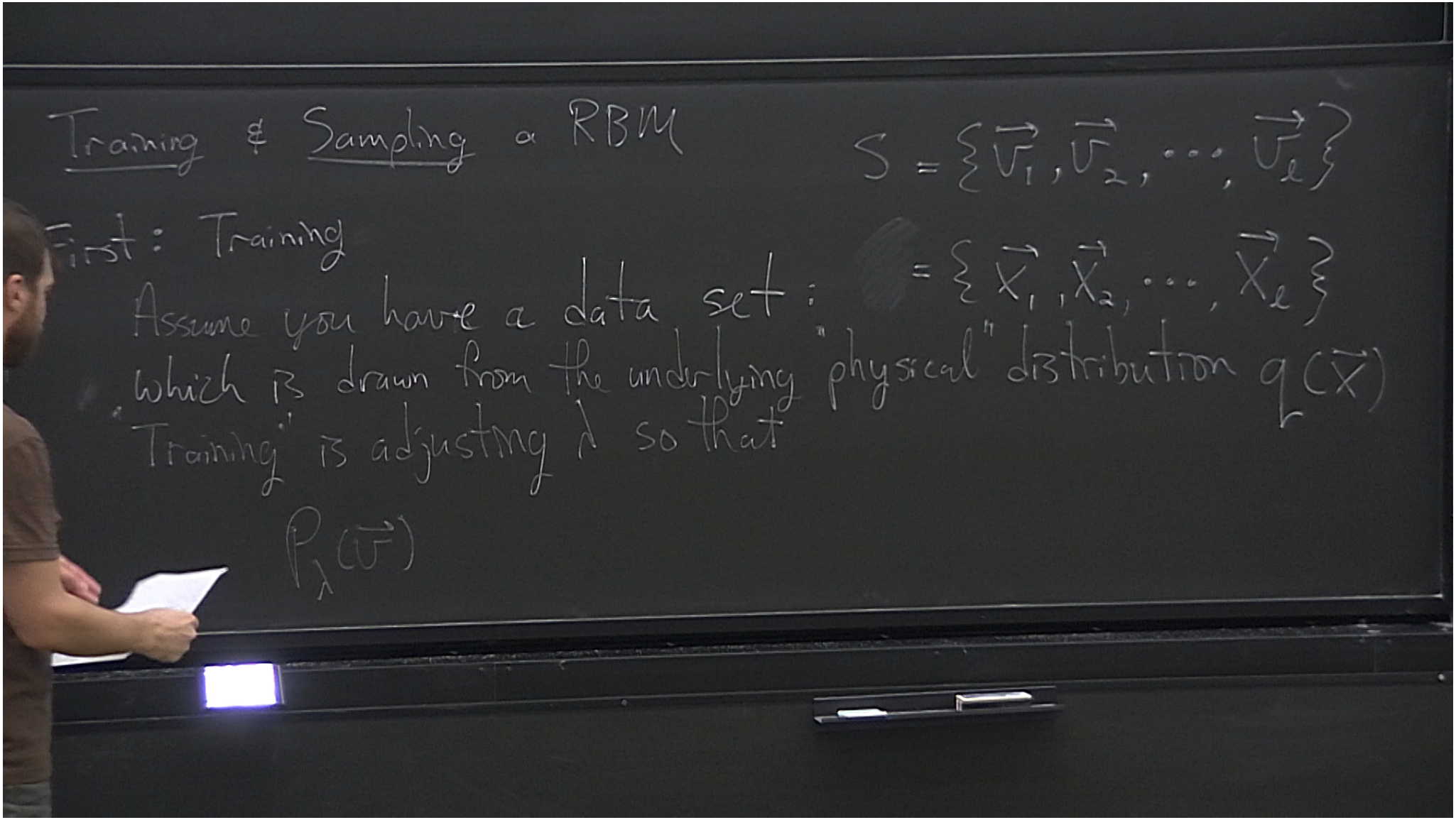
Training & Sampling a RBM

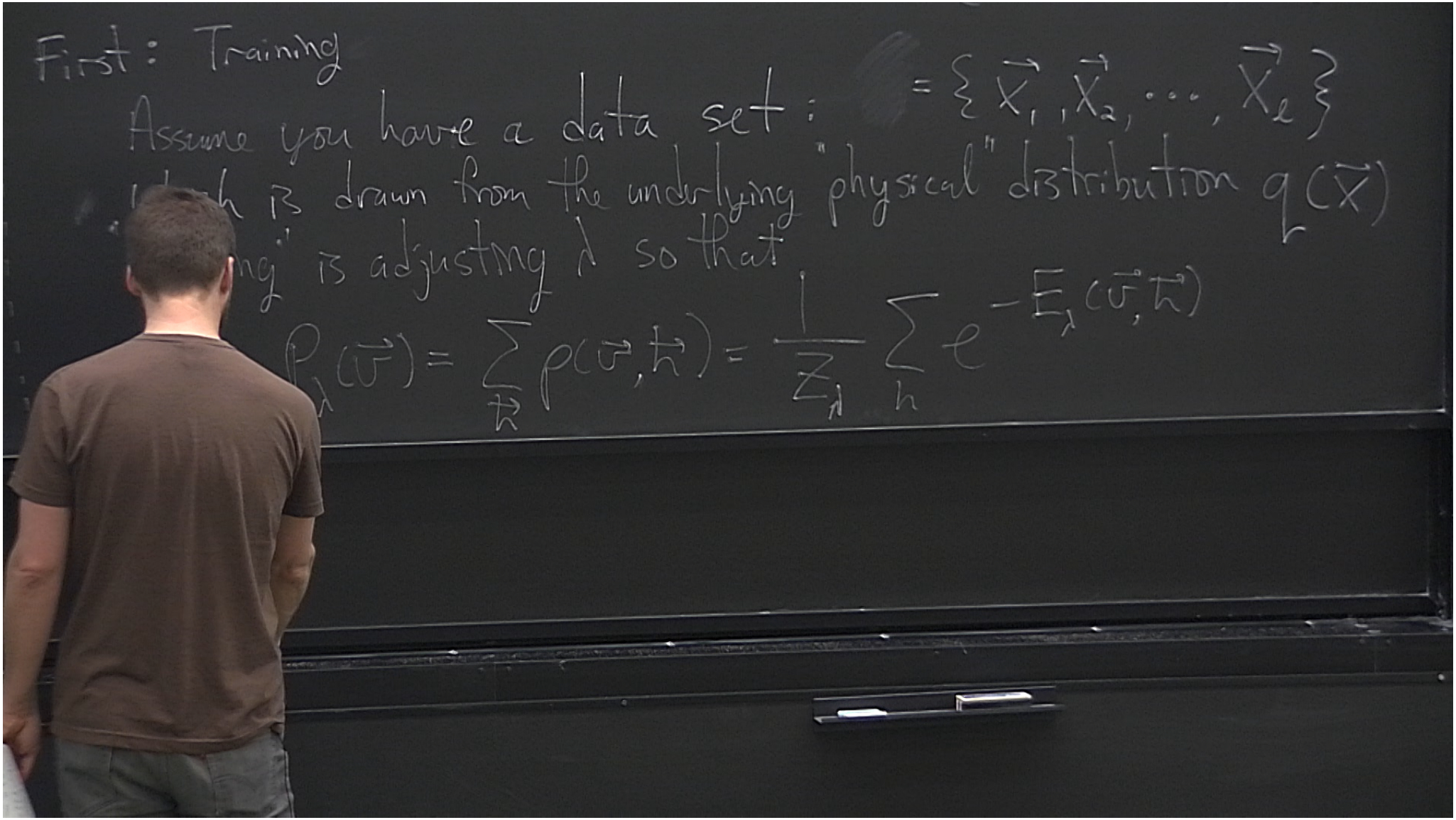
First: Training

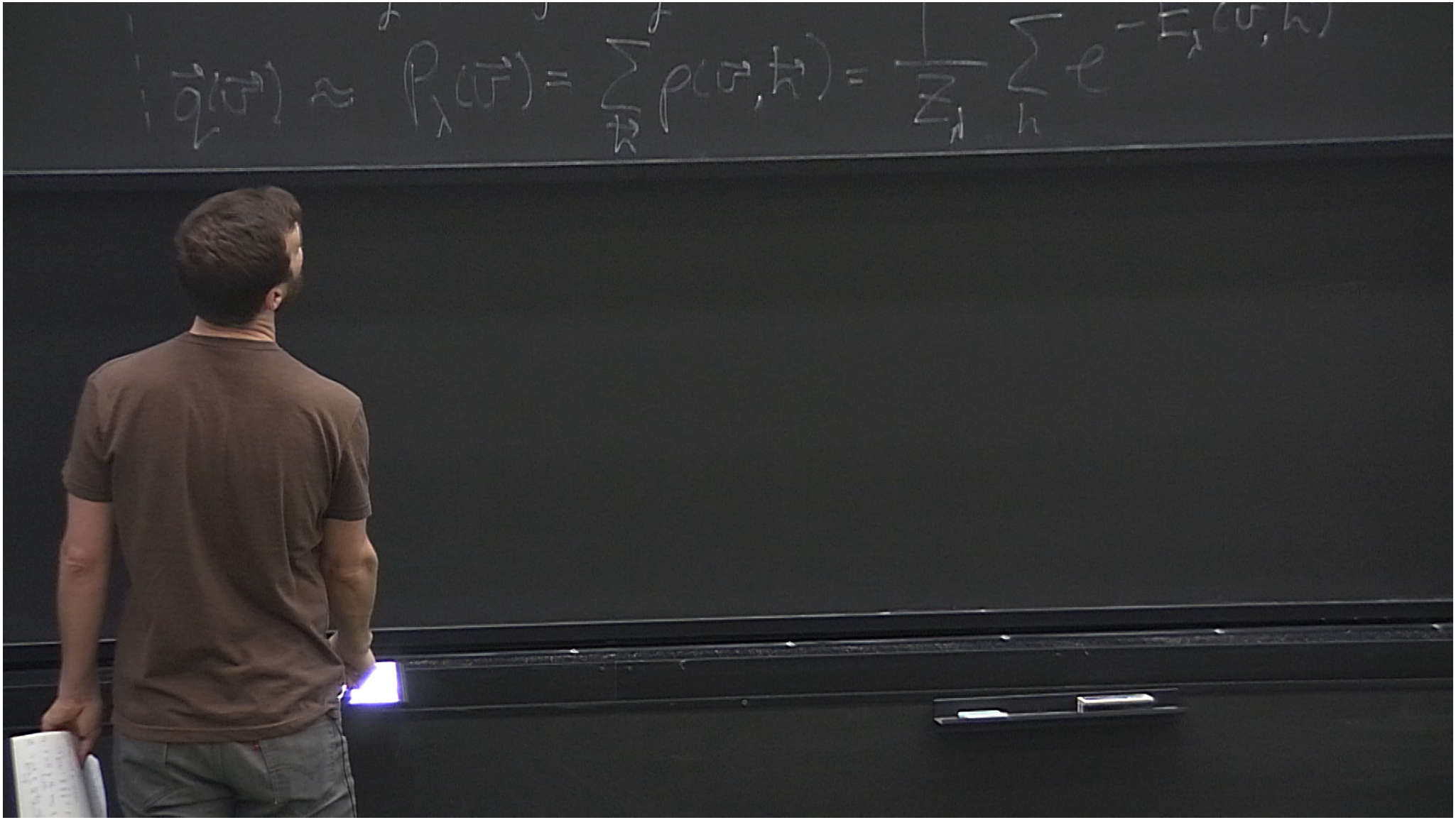
Assume you have a data set: $S = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_L\}$

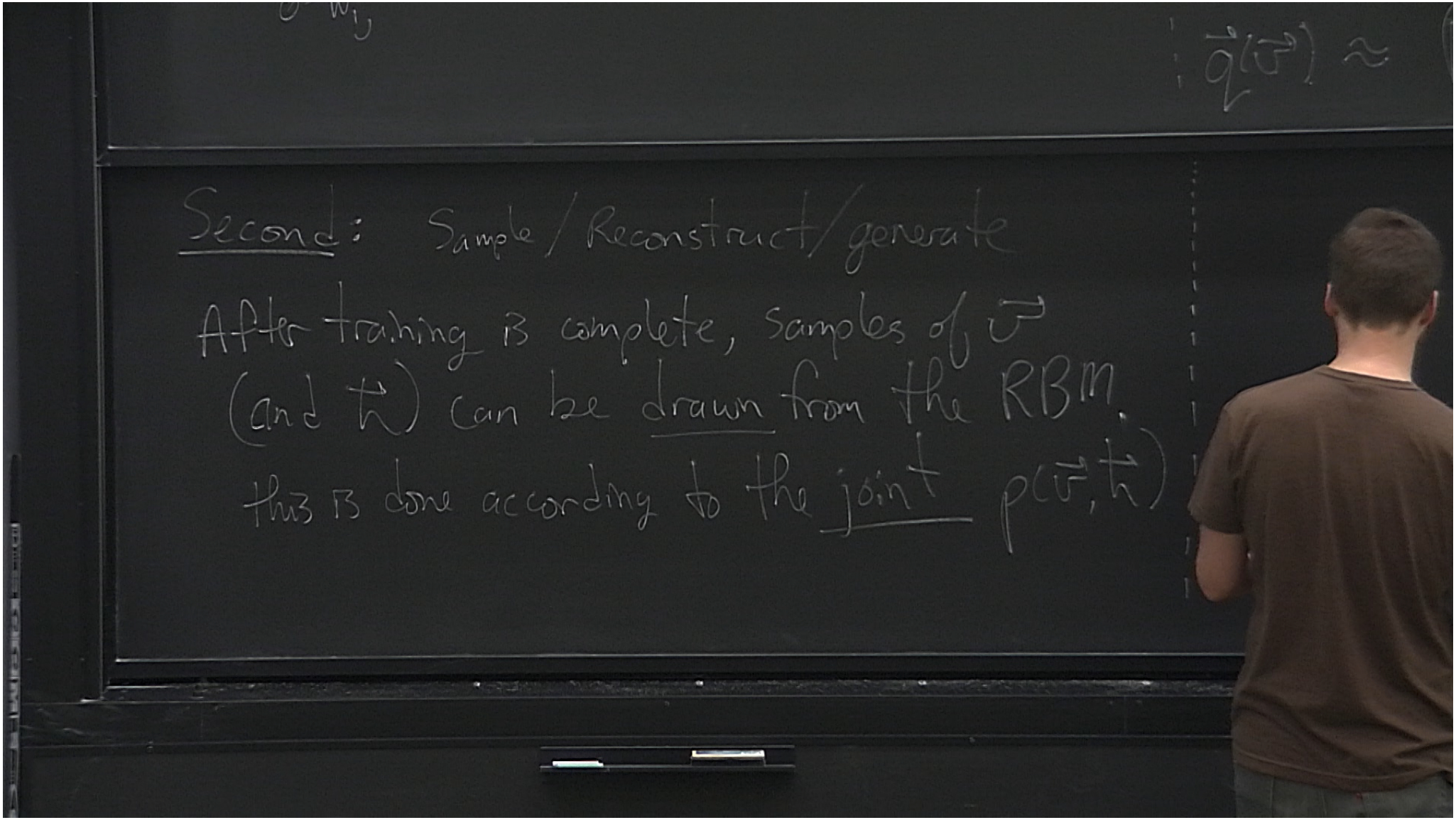

RBM:
$E_\lambda = -\sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j - \sum_{ij} W_{ij} v_i h_j$
$Z_\lambda = \sum_{ij} e^{-E_\lambda}$
$\lambda = (W, \vec{b}, \vec{c}) = \Theta$
b_i
c_j
W_{ij}
Training &
First: Train
Assume

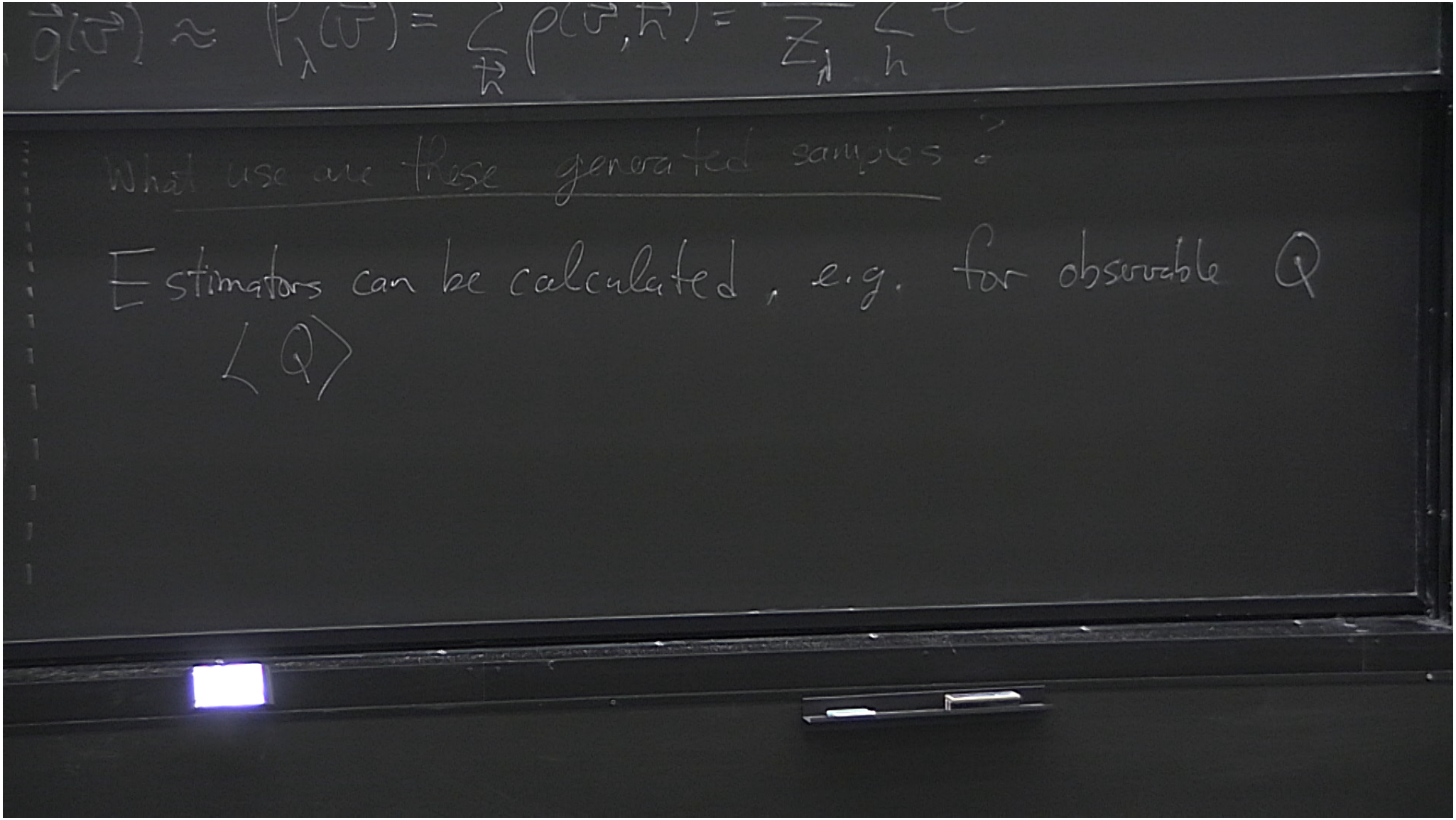

Training & Sampling a RBM
First: Training
Assume you have a data set: $S = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_L\}$
which is drawn from the underlying "physical" distribution $q(\vec{x})$
"Training" is adjusting λ so that
$P_\lambda(\vec{v})$


Training & Sampling a RBM
$S = \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_L\}$
First: Training
Assume you have a data set: $= \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_L\}$
which is drawn from the underlying "physical" distribution $q(\vec{x})$
Training is adjusting λ so that
$P_\lambda(\vec{v})$


First: Training
Assume you have a data set: $= \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_e\}$
which is drawn from the underlying "physical" distribution $q(\vec{x})$
ning" is adjusting λ so that
$P_\lambda(\vec{v}) = \sum_{\vec{h}} p(\vec{v}, \vec{h}) = \frac{1}{Z_\lambda} \sum_h e^{-E_\lambda(\vec{v}, \vec{h})}$


$\vec{q}(\vec{r}) \approx P_\lambda(\vec{r}) = \sum_h p(\vec{r}, h) = \frac{1}{Z_\lambda} \sum_h e^{-E_\lambda(\vec{r}, h)}$


$\vec{q}(\vec{v}) \approx$
Second: Sample / Reconstruct / generate
After training is complete, samples of $\vec{v}$
(and $\vec{h}$) can be drawn from the RBM.
this is done according to the joint $p(\vec{v}, \vec{h})$


What use are these generated samples?
Estimators can be calculated, e.g. for observable Q
⟨Q⟩

w_{ij}

$\vec{q}(\vec{v}) \approx$

Second: Sample / Reconstruct / generate

After training is complete, samples of $\vec{v}$ (and $\vec{h}$) can be drawn from the RBM.

This is done according to the joint $p(\vec{v}, \vec{h})$

What use

Estimato

$L($

$$q(\vec{v}) \approx p_\lambda(\vec{v}) = \sum_{\vec{h}} p(\vec{v},\vec{h}) = \frac{1}{Z_\lambda} \sum_{h}$$

What use are these generated samples?

Estimators can be calculated, e.g. for observable Q

$$\langle Q \rangle_{p_\lambda(\vec{v},\vec{h})} = \frac{1}{Z_\lambda} \sum_{\vec{v}} \sum_{\vec{h}} Q\, p_\lambda(\vec{v},\vec{h})$$

$\vec{q}(\vec{v}) \approx P_\lambda(\vec{v}) = \sum_{\vec{h}} p(\vec{v},\vec{h}) = \frac{1}{Z_\lambda} \sum_{h}$

What use are these generated samples?

Estimators can be calculated, e.g. for observable Q

$$\langle Q \rangle_{p_\lambda(\vec{v},\vec{h})} = \frac{1}{Z_\lambda} \sum_{\vec{v}} \sum_{\vec{h}} Q\, p_\lambda(\vec{v},\vec{h})$$

e.g. imagine defining a magnetization, etc ($v_i = 0,1$, $h_i = 0,1$)

$\vec{q}(\vec{v}) \approx P_\lambda(\vec{v}) = \sum_{\vec{h}} p(\vec{v},\vec{h}) = \frac{1}{Z_\lambda} \sum_h e^{-}$

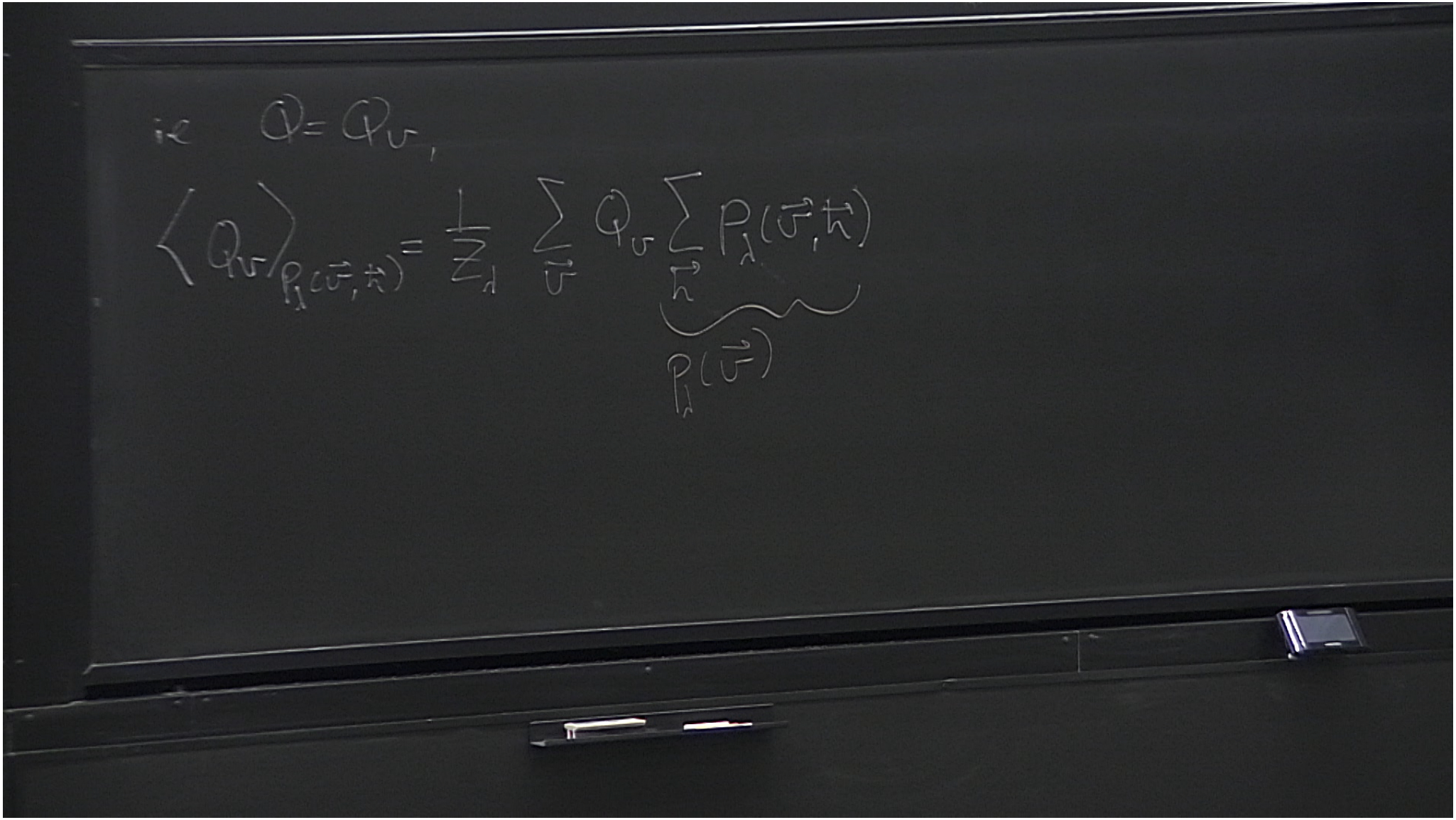
What use are these generated samples?

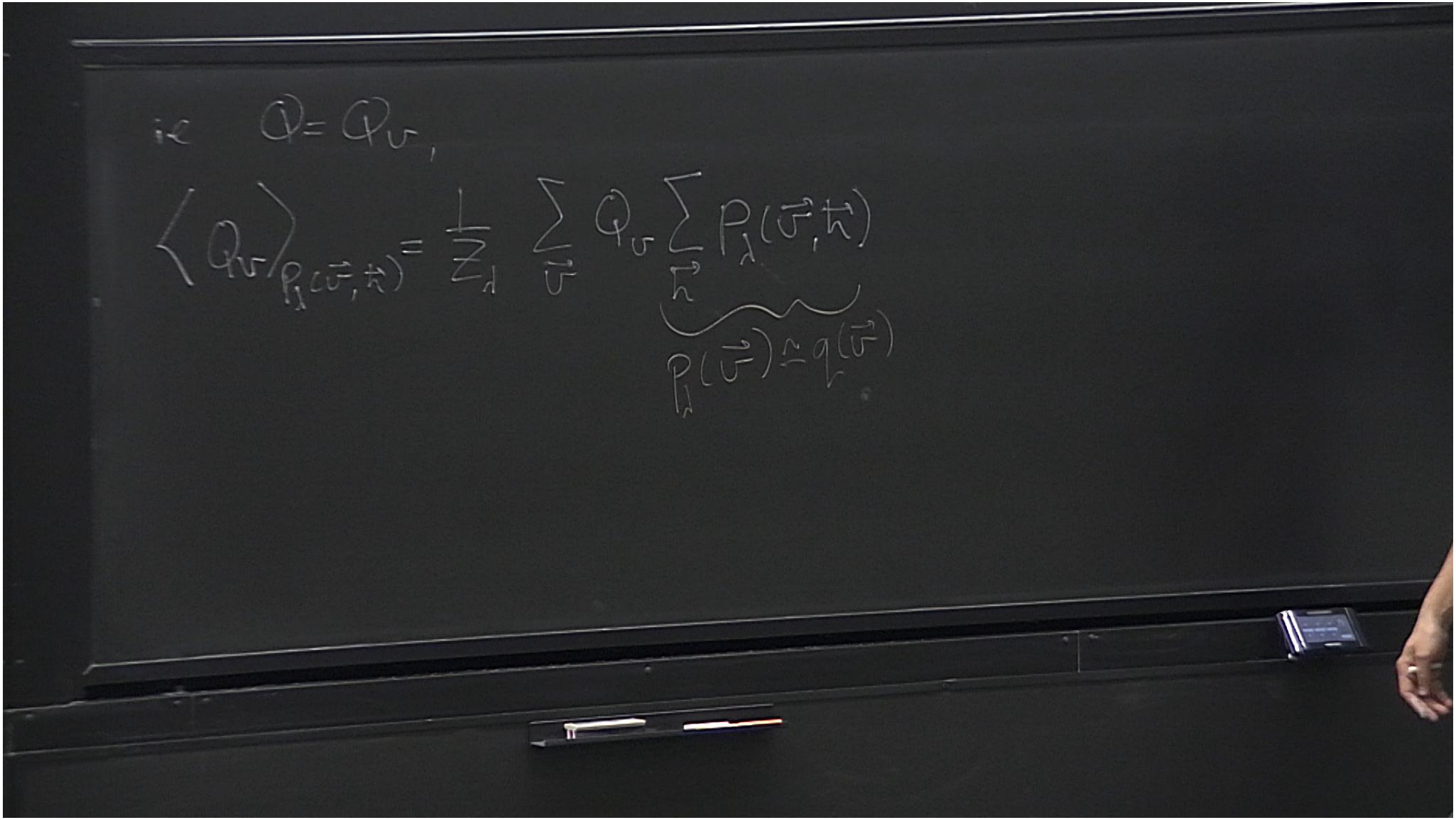
Estimators can be calculated, e.g. for observable Q

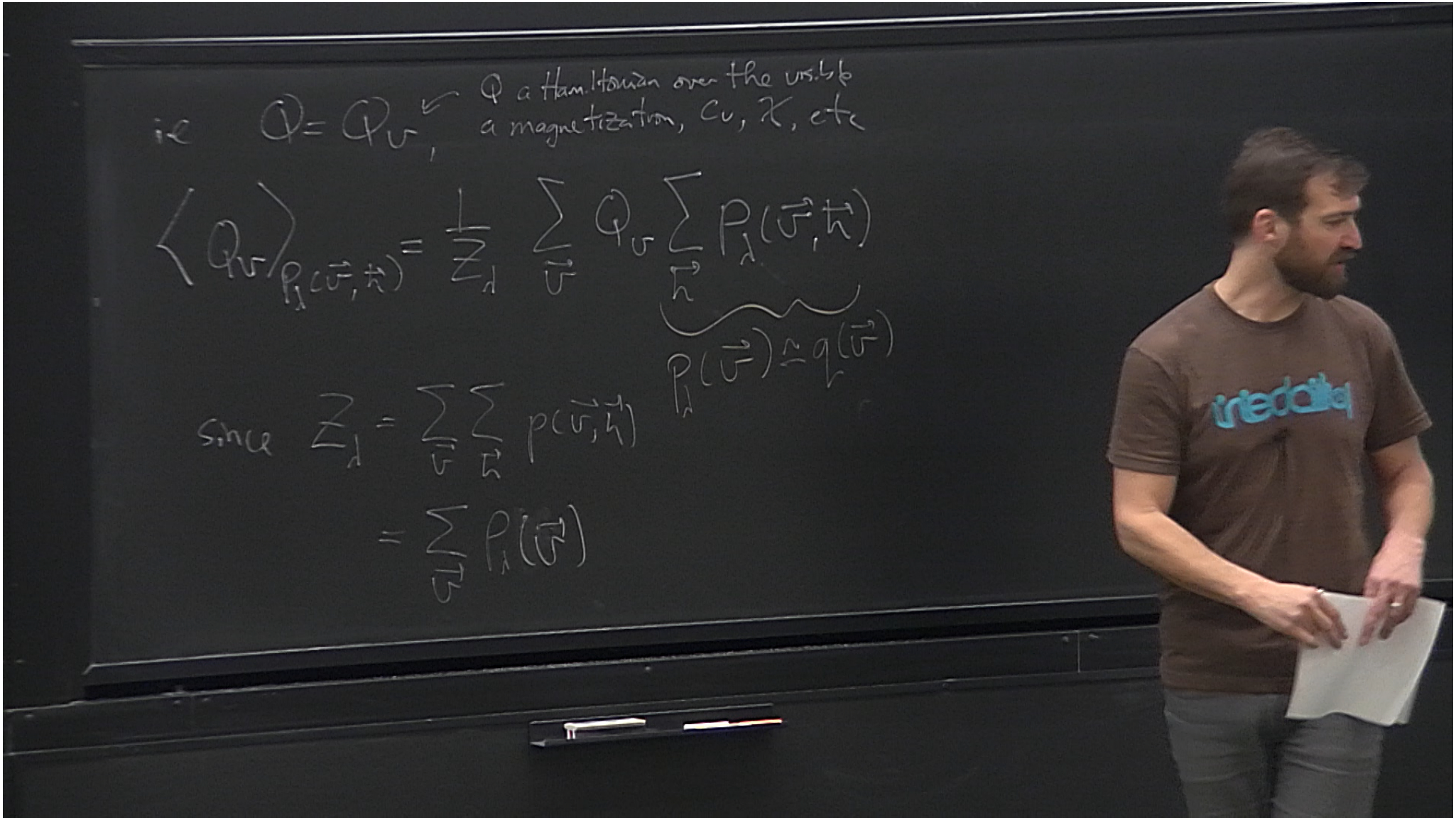
$$\langle Q \rangle_{p_\lambda(\vec{v},\vec{h})} = \frac{1}{Z_\lambda} \sum_{\vec{v}} \sum_{\vec{h}} Q\, p_\lambda(\vec{v},\vec{h})$$

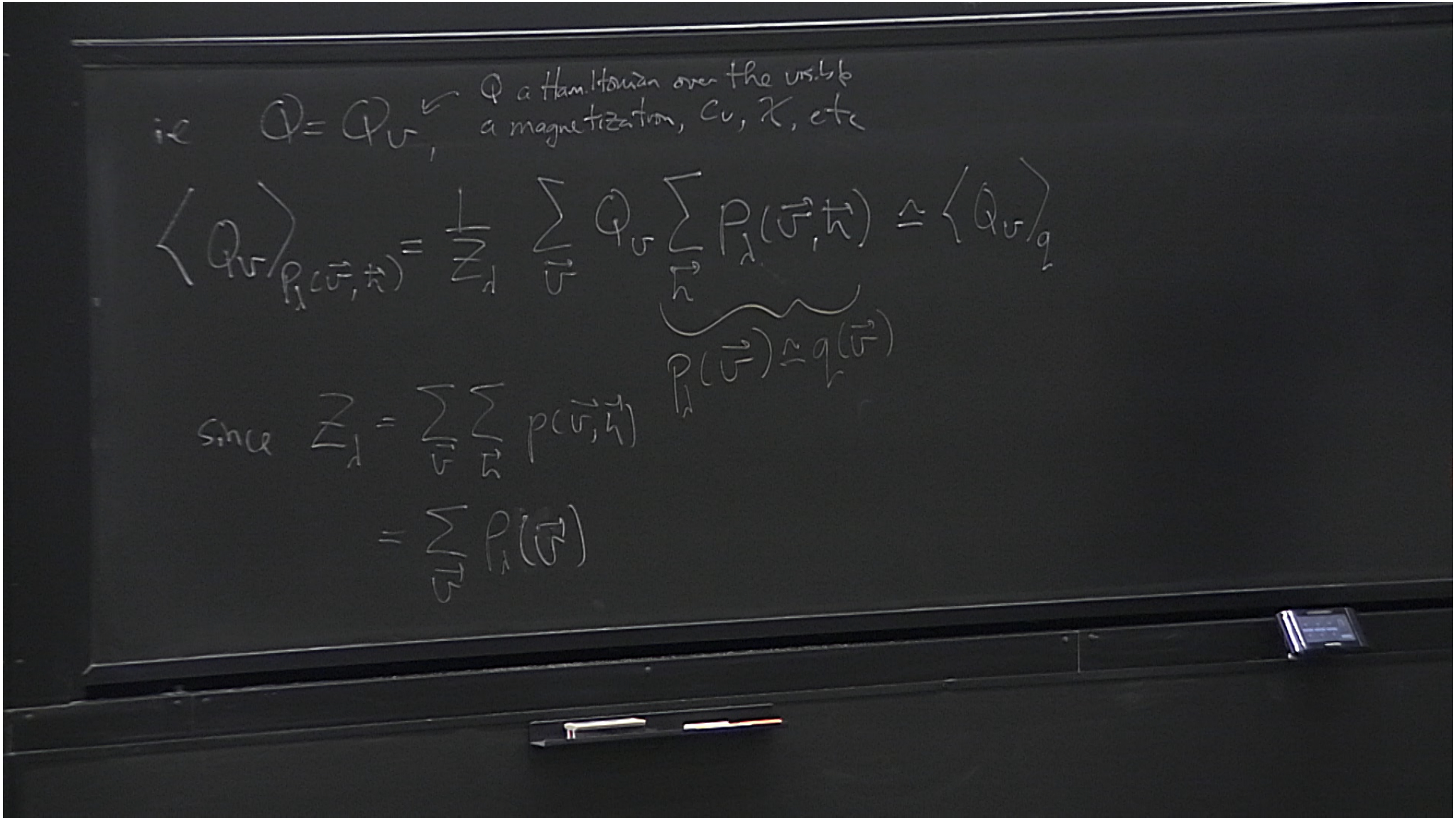
e.g. imagine defining a magnetization, etc ($v_i = 0,1$, $h_i = 0,1$)

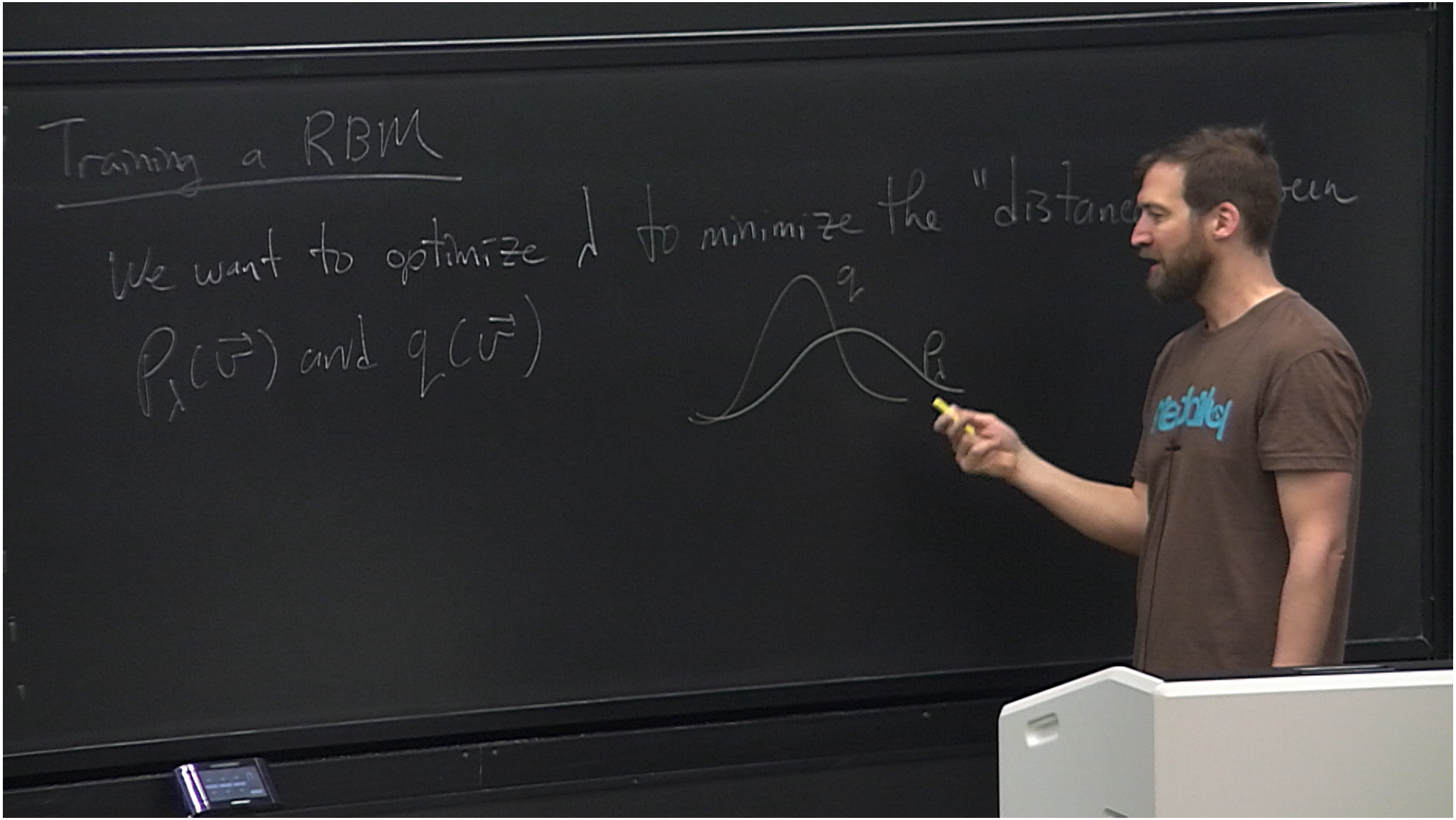
OR restrict the definition of Q to act on visible:


ie $Q = Q_\sigma$,
$\langle Q_\sigma \rangle_{P_\lambda(\vec{\sigma},\vec{h})} = \frac{1}{Z_\lambda} \sum_{\vec{\sigma}} Q_\sigma \underbrace{\sum_{\vec{h}} P_\lambda(\vec{\sigma},\vec{h})}_{P_\lambda(\vec{\sigma})}$


if $Q = Q_{\vec{v}}$,
$\langle Q_{\vec{v}} \rangle_{P_\lambda(\vec{v},\vec{h})} = \frac{1}{Z_\lambda} \sum_{\vec{v}} Q_{\vec{v}} \underbrace{\sum_{\vec{h}} P_\lambda(\vec{v},\vec{h})}_{P_\lambda(\vec{v}) \triangleq q(\vec{v})}$

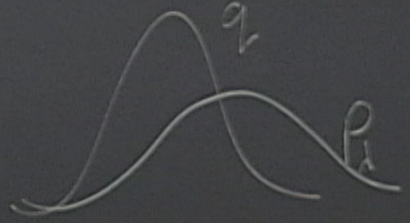

if $Q = Q_v$, ← Q a Hamiltonian over the visible
a magnetization, C_v, χ, etc
$\langle Q_v \rangle_{P_\lambda(\vec{v},\vec{h})} = \frac{1}{Z_\lambda} \sum_{\vec{v}} Q_v \underbrace{\sum_{\vec{h}} P_\lambda(\vec{v},\vec{h})}_{P_\lambda(\vec{v}) \triangleq q(\vec{v})}$
since $Z_\lambda = \sum_{\vec{v}} \sum_{\vec{h}} p(\vec{v},\vec{h})$
$= \sum_{\vec{v}} P_\lambda(\vec{v})$


if Q = Q_v ← Q a Hamiltonian over the visible a magnetization, C_v, χ, etc
⟨Q_v⟩_{P_λ(v,h)} = (1/Z_λ) Σ_v Q_v Σ_h P_λ(v,h) ≈ ⟨Q_v⟩_q
P_λ(v) ≈ q(v)
since Z_λ = Σ_v Σ_h p(v,h)
= Σ_v P_λ(v)


Training a RBM
We want to optimize λ to minimize the "distance
$P_\lambda(\vec{v})$ and $q(\vec{v})$
q
P_λ

Training a RBM

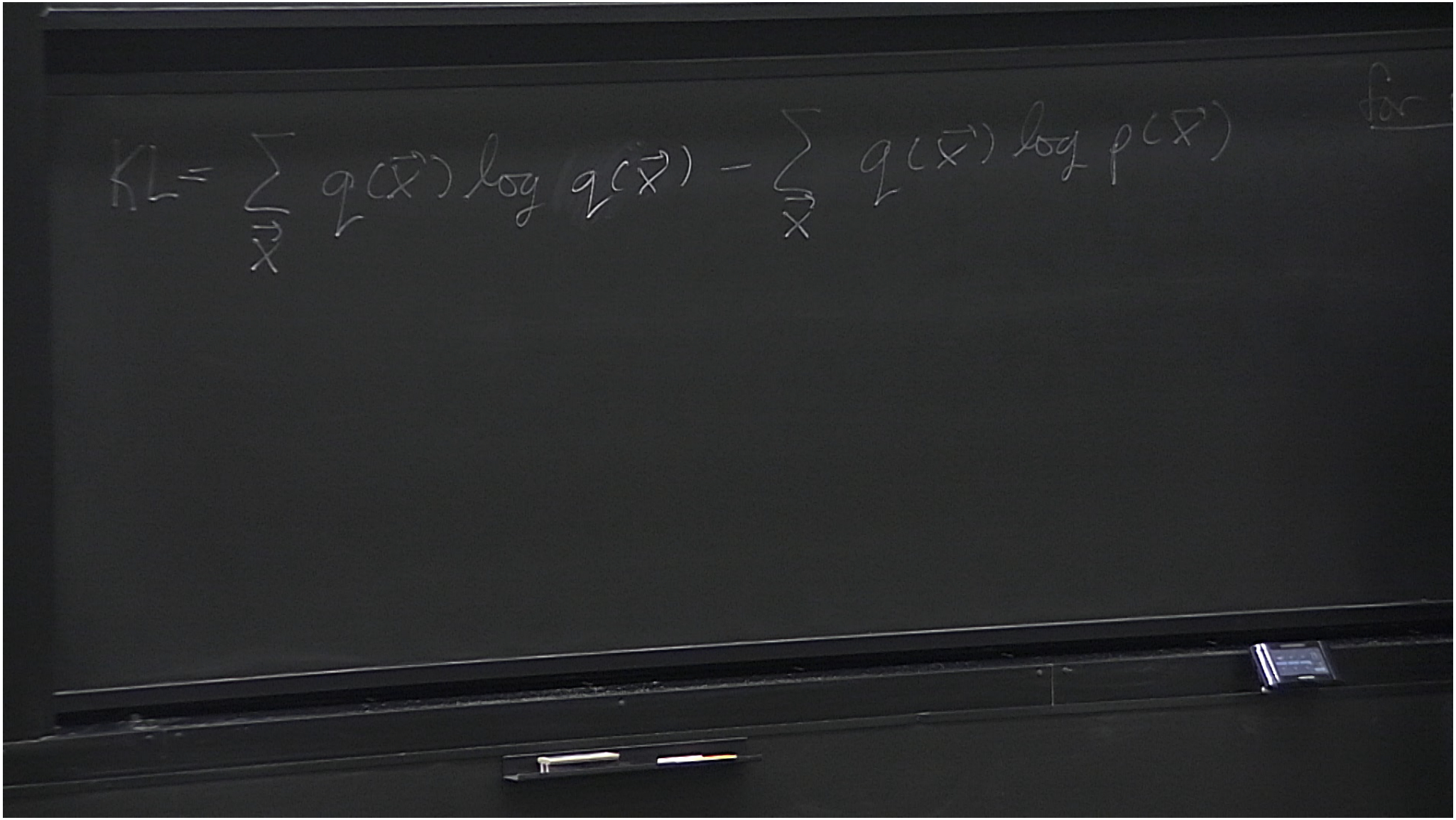
We want to optimize λ to minimize the "distance" between $P_\lambda(\vec{v})$ and $q(\vec{v})$

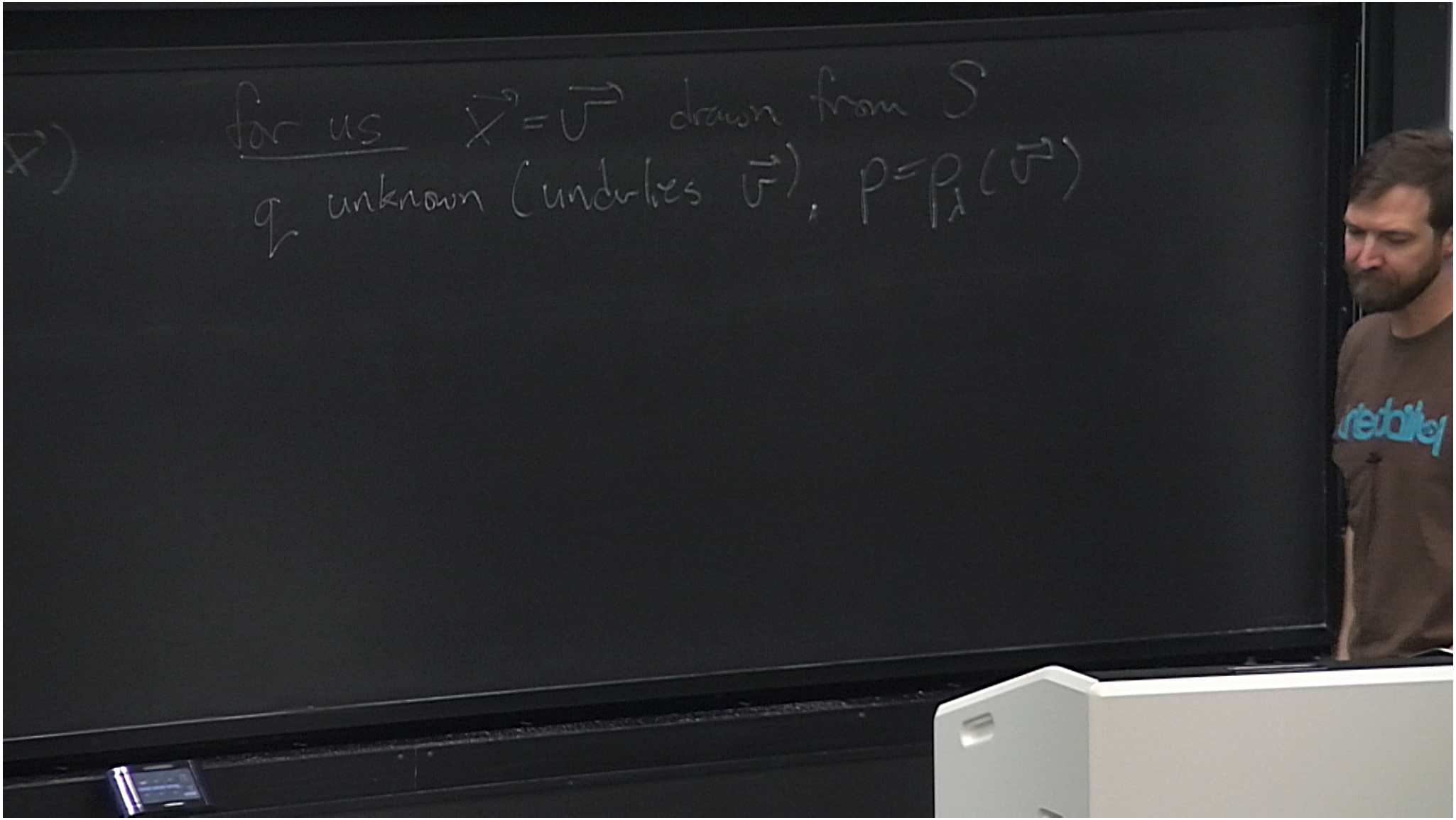


We do this by minimizing the Kullback-Leibler (KL) divergence:

$$KL = \sum_{\vec{x}} q(\vec{x}) \log \frac{q(\vec{x})}{P(\vec{x})}$$

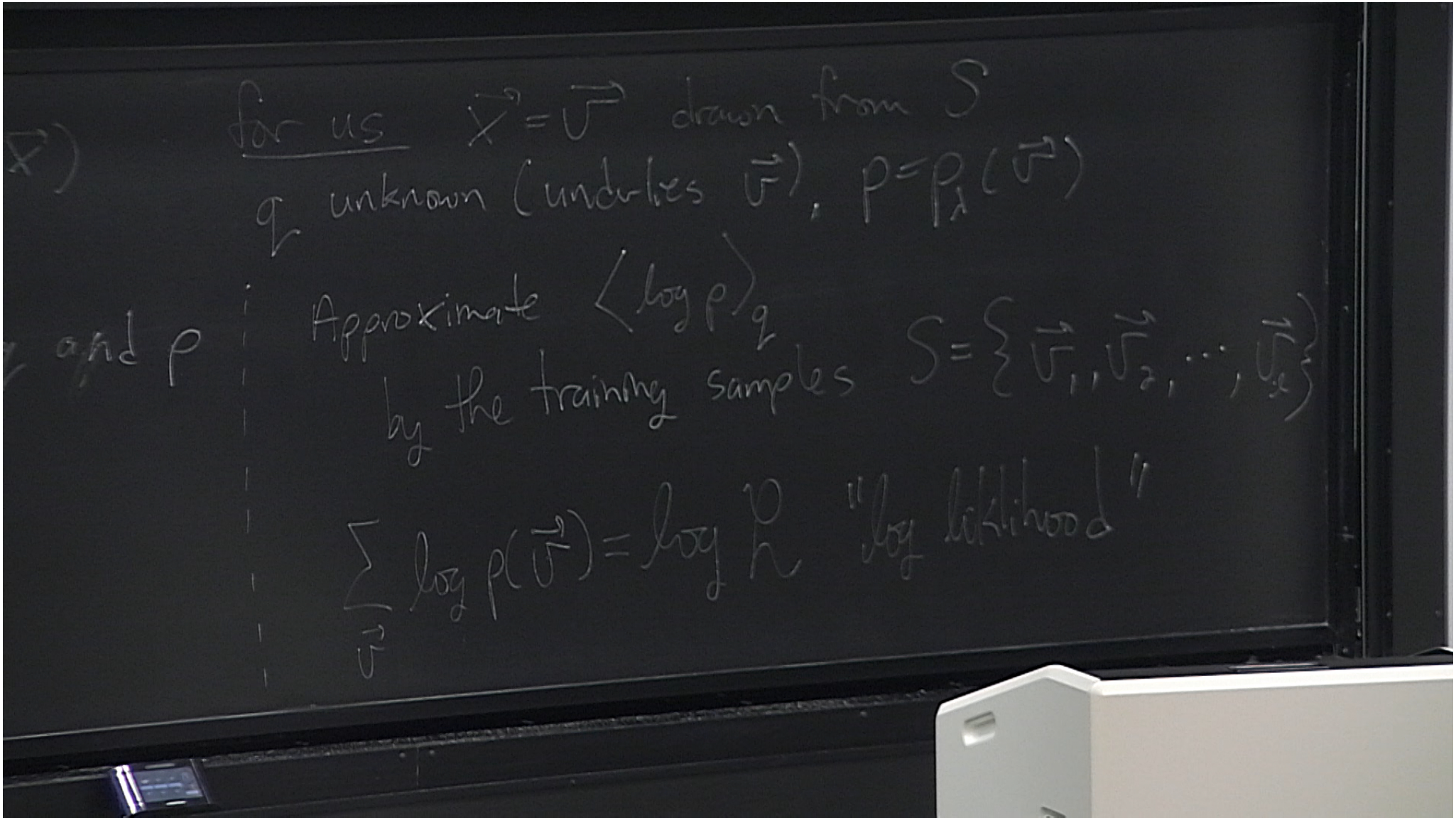
for any q, P


$KL = \sum_{\vec{x}} q(\vec{x}) \log q(\vec{x}) - \sum_{\vec{x}} q(\vec{x}) \log p(\vec{x})$
for


for us $\vec{x} = \vec{v}$ drawn from S
q unknown (underlies $\vec{v}$), $P = P_\lambda(\vec{v})$

$$KL = \sum_{\vec{x}} q(\vec{x}) \log q(\vec{x}) - \sum_{\vec{x}} q(\vec{x}) \log p(\vec{x})$$

- non-symmetric measure of the "distance" between q and p
- always positive; zero if $p = q$
- first term is the entropy of the data set
- second term contains parameters λ

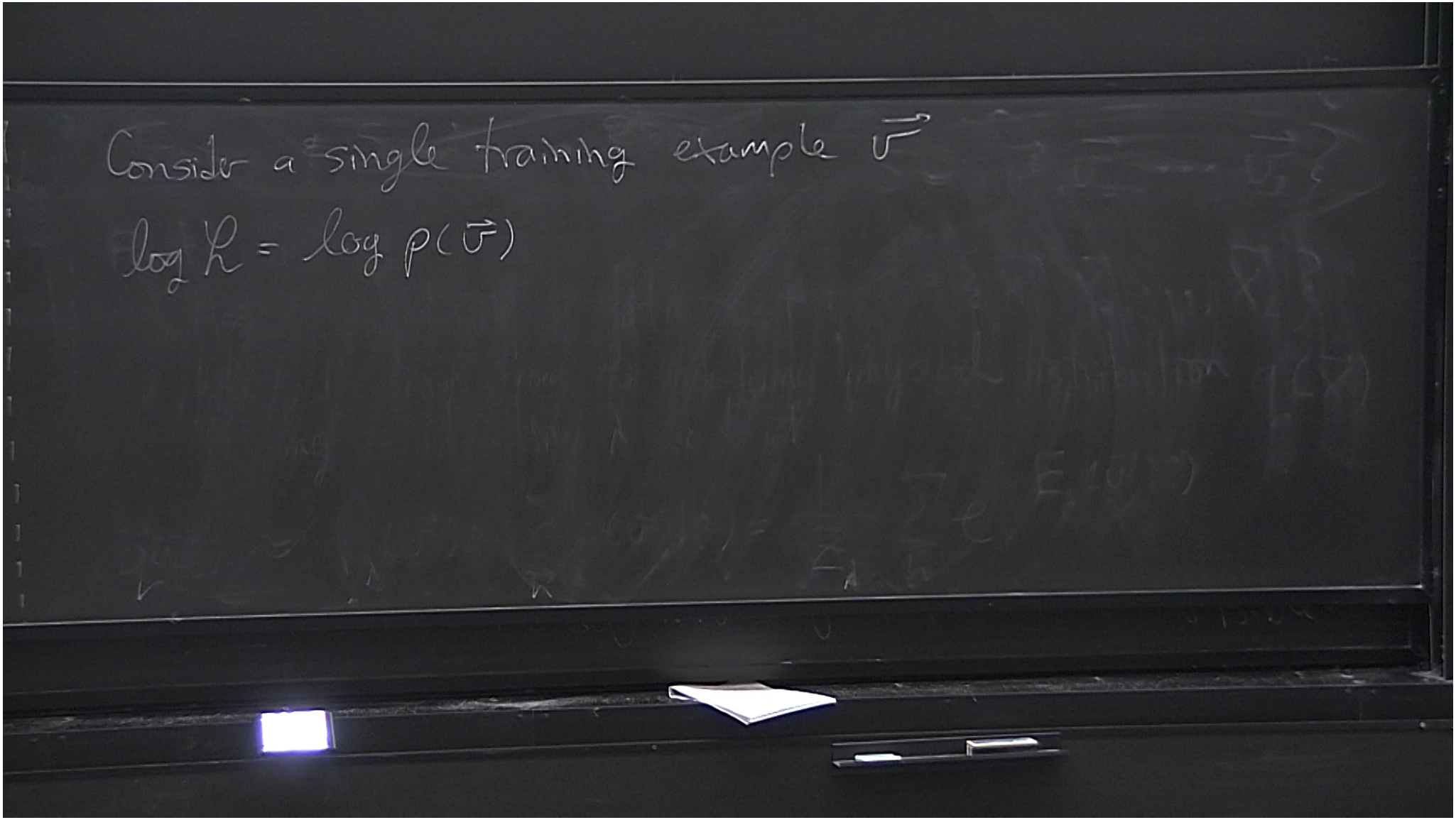

for us $\vec{x} = \vec{v}$ drawn from S
q unknown (underlies $\vec{v}$), $p = p_\lambda(\vec{v})$
Approximate $\langle \log p \rangle_q$
by the training samples $S = \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_N\}$
$\sum_{\vec{v}} \log p(\vec{v}) = \log \mathcal{L}$ "log likelihood"

Training is done by stochastic gradient descent
(pick a mini-batch out of S)

The gradient of the (negative) log liklihood
is used:

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

η learning rate


Consider a single training example $\vec{v}$
$\log \mathcal{L} = \log p(\vec{v})$

Consider a single training example $\vec{v}$

$$\log \mathcal{L} = \log p_\lambda(\vec{v}) = \log p(\vec{v})$$

$$= \log \frac{1}{Z} \sum_{\vec{h}} e^{-E(\vec{v},\vec{h})}$$

$$= \log \sum_{\vec{h}} e^{-E(\vec{v},\vec{h})} - \log \sum_{\vec{v}\vec{h}} e^{-E(\vec{v},\vec{h})}$$

What use are these generated samples?

Estimators can be calculated, e.g. for observable Q

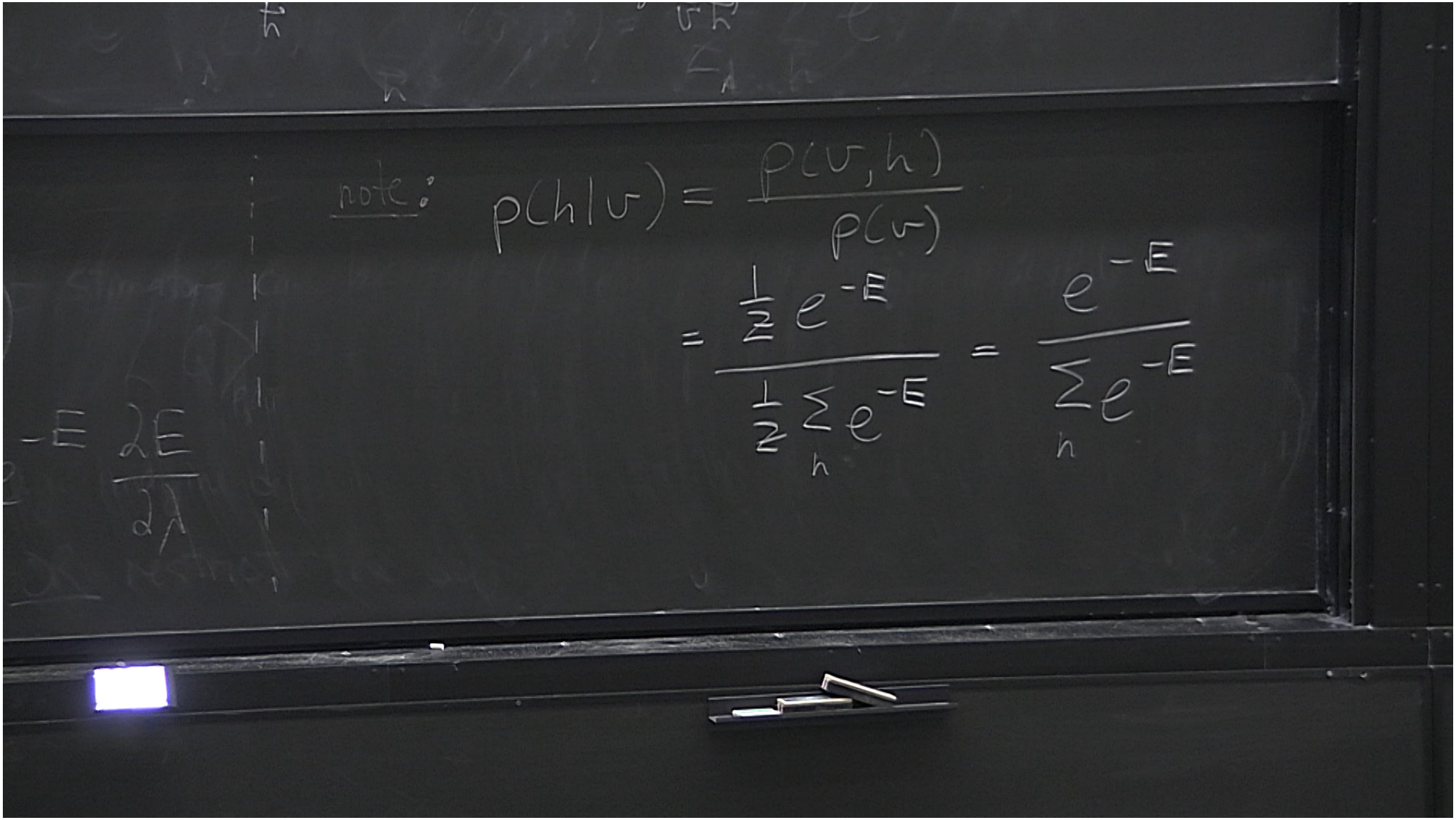
$\lambda \leftarrow \lambda - \eta \nabla_\lambda \mathcal{L}$

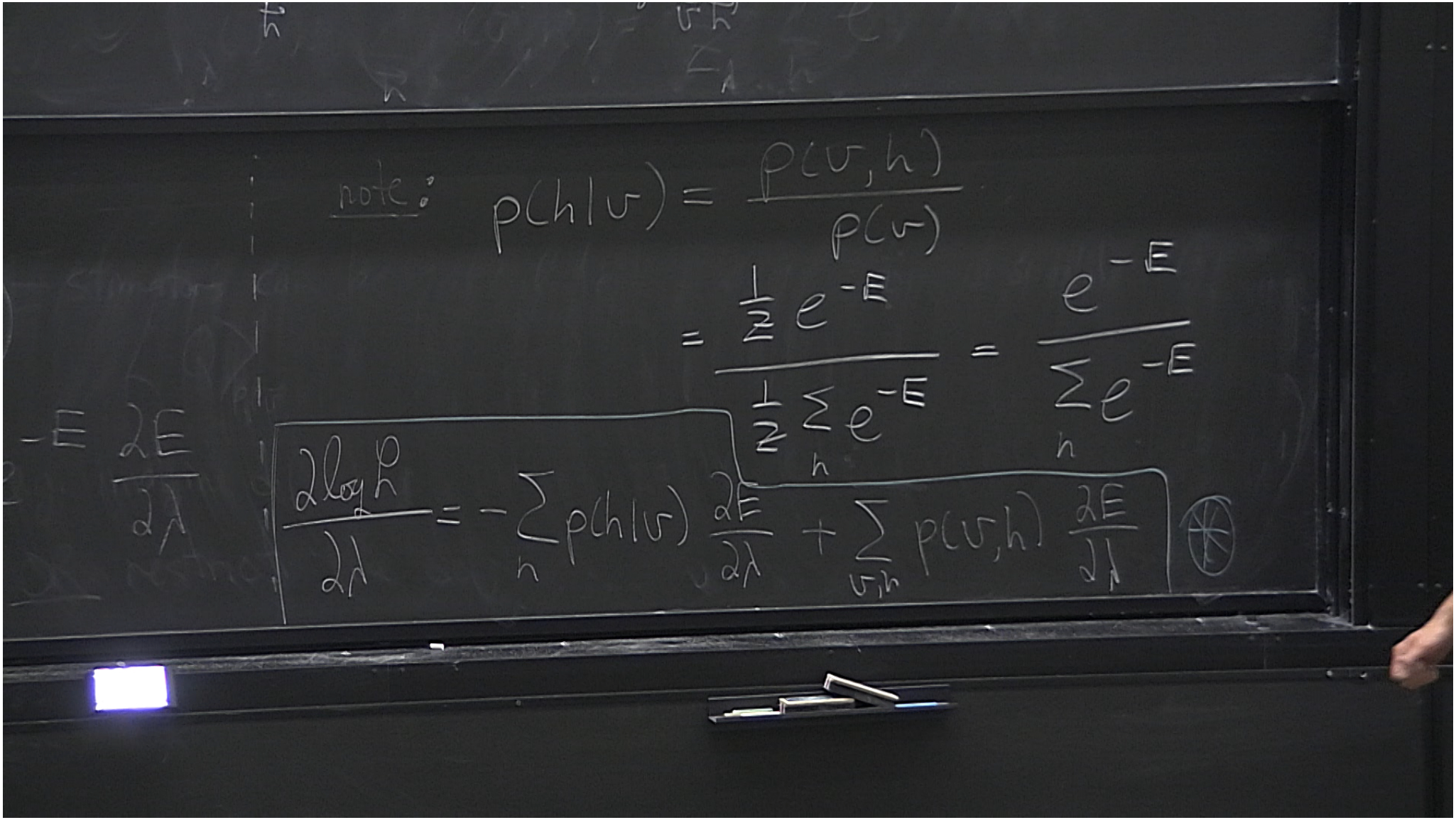
η learning rate

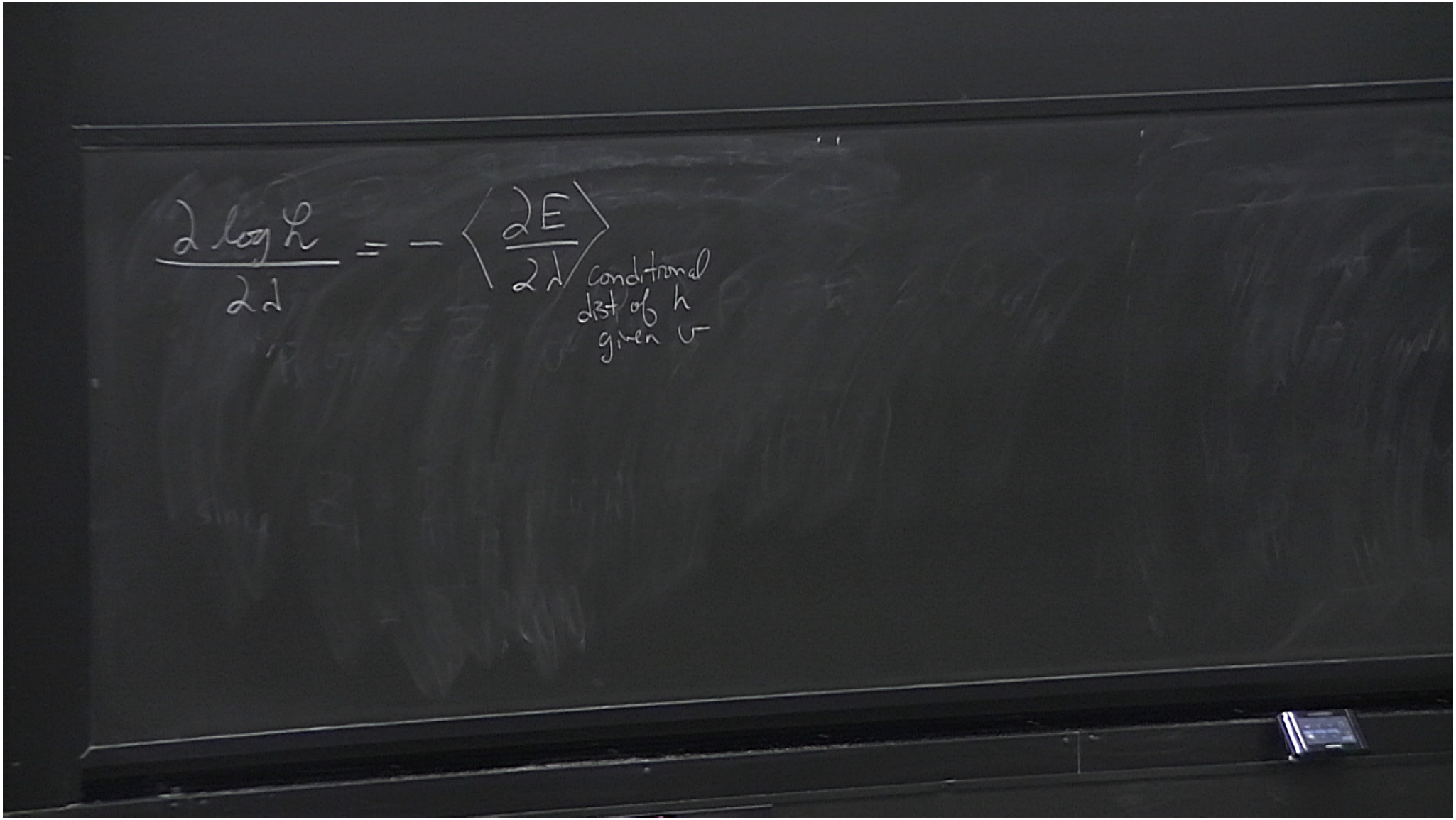
the gradient

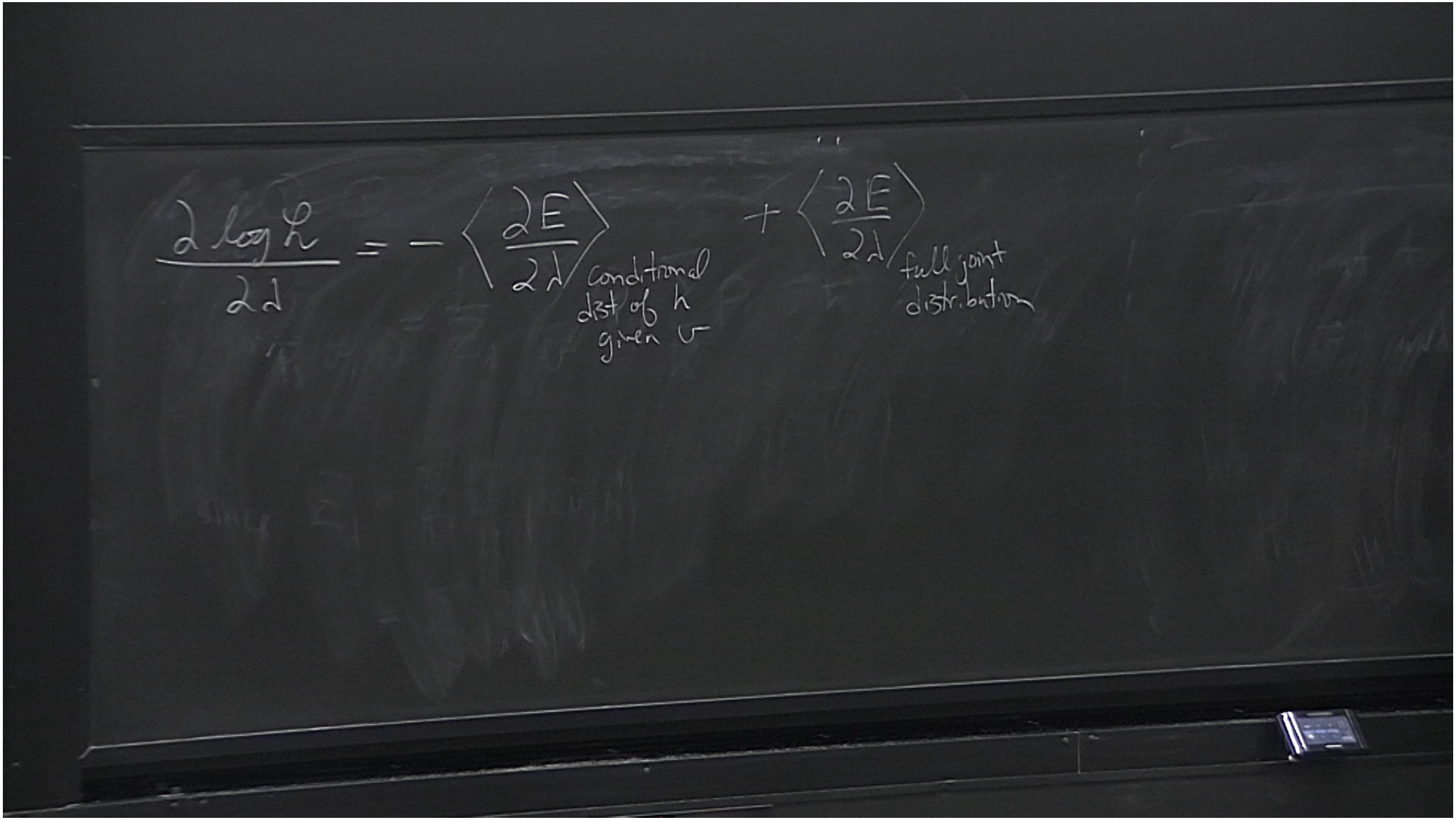
$$\frac{\partial \log \mathcal{L}}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(\log \sum_h e^{-E}\right) - \frac{\partial}{\partial \lambda}\left(\log \sum_{v,h} e^{-E}\right)$$

$$= \frac{-1}{\sum e^{-E}} \sum_h e^{-E} \frac{\partial E}{\partial \lambda} + \frac{1}{\sum_{v,h} e^{-E}} \sum_{v,h} e^{-E} \frac{\partial E}{\partial \lambda}$$


note: $p(h|v) = \frac{p(v,h)}{p(v)}$
$= \frac{\frac{1}{Z} e^{-E}}{\frac{1}{Z} \sum_h e^{-E}} = \frac{e^{-E}}{\sum_h e^{-E}}$
$-E \quad \frac{\partial E}{\partial \lambda}$


note: $p(h|v) = \frac{p(v,h)}{p(v)}$
$= \frac{\frac{1}{Z} e^{-E}}{\frac{1}{Z}\sum_h e^{-E}} = \frac{e^{-E}}{\sum_h e^{-E}}$
$\frac{\partial \log p}{\partial \lambda} = -\sum_h p(h|v) \frac{\partial E}{\partial \lambda} + \sum_{v,h} p(v,h) \frac{\partial E}{\partial \lambda}$ ⊛
$-E \frac{\partial E}{\partial \lambda}$


$\frac{\partial \log L}{\partial \lambda} = -\left\langle \frac{\partial E}{\partial \lambda} \right\rangle$
conditional dist of h given v

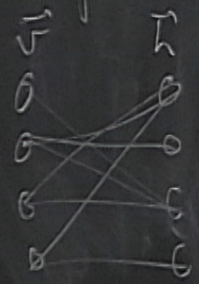

$\frac{\partial \log \mathcal{L}}{\partial \lambda} = -\left\langle \frac{\partial E}{\partial \lambda} \right\rangle + \left\langle \frac{\partial E}{\partial \lambda} \right\rangle$
conditional dist of h given v
full joint distribution

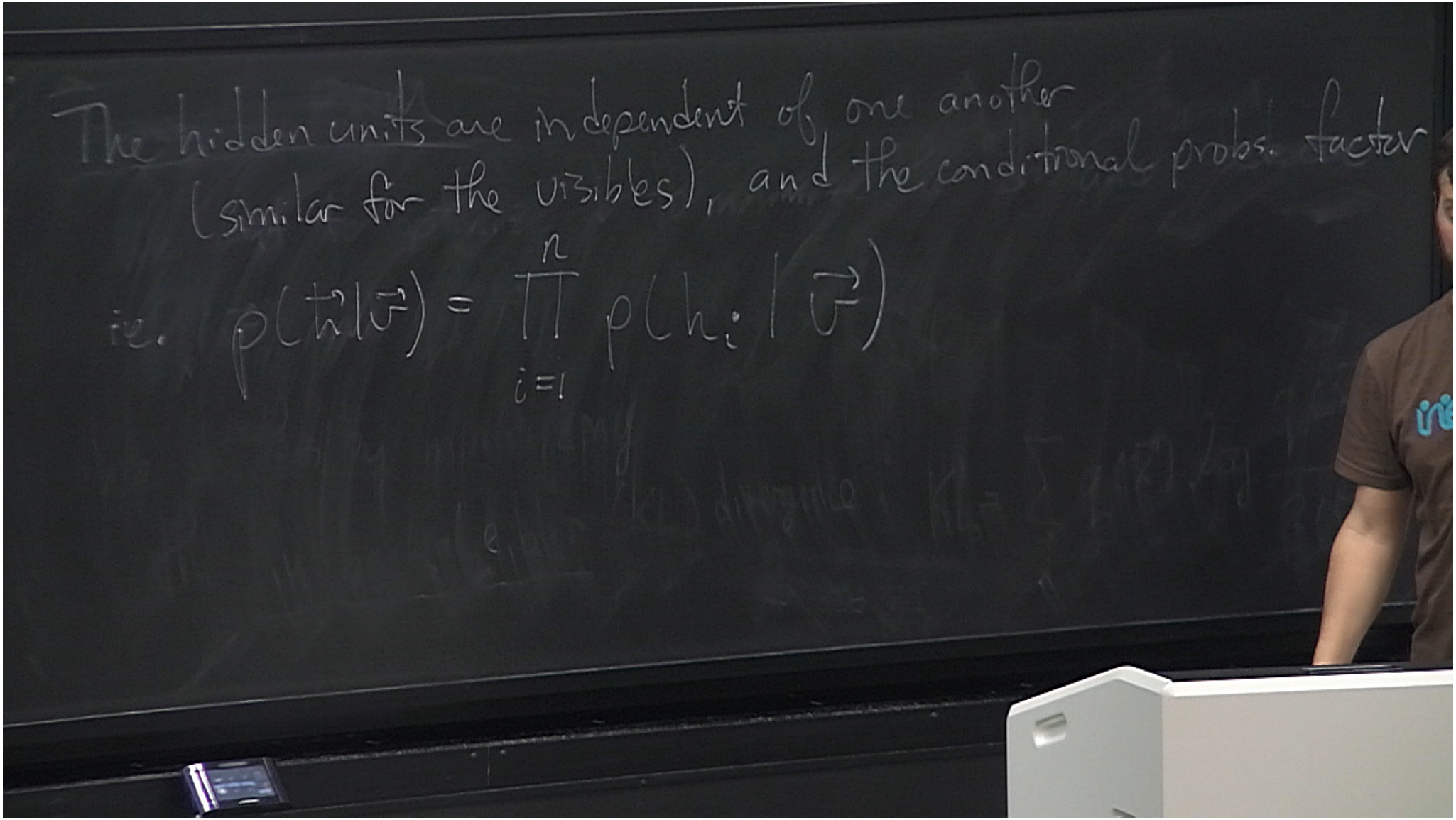
$$\frac{\partial \log \mathcal{L}}{\partial \lambda} = -\left\langle \frac{\partial E}{\partial \lambda} \right\rangle + \left\langle \frac{\partial E}{\partial \lambda} \right\rangle$$

conditional dist of $\vec{h}$ given $\vec{v}$

full joint distribution

To calculate $p(\vec{h}|\vec{v})$ recall



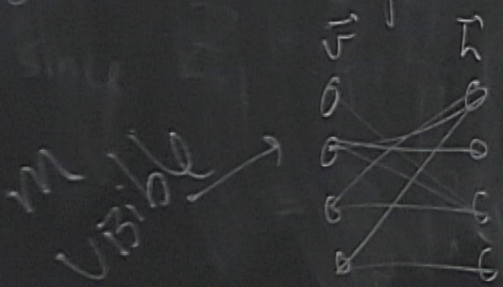

The hidden units are independent of one another
(similar for the visibles), and the conditional probs. factor
ie. $p(\vec{h}|\vec{v}) = \prod_{i=1}^{n} p(h_i|\vec{v})$

$$\frac{\partial \log \mathcal{L}}{\partial \theta} = -\left\langle \frac{\partial E}{\partial \theta} \right\rangle + \left\langle \frac{\partial E}{\partial \theta} \right\rangle$$

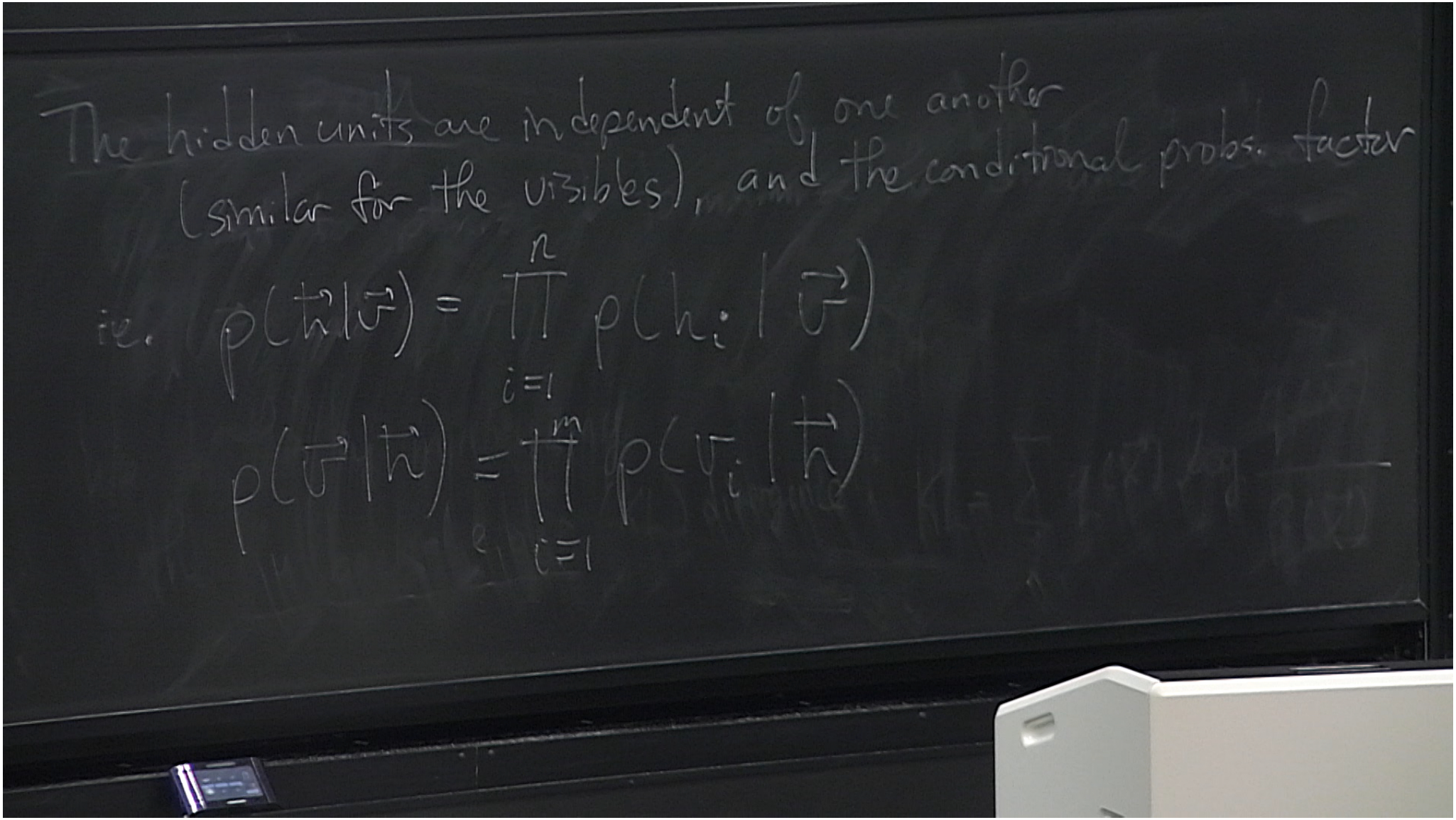
$\left\langle \frac{\partial E}{\partial \theta} \right\rangle$: conditional dist of h given v

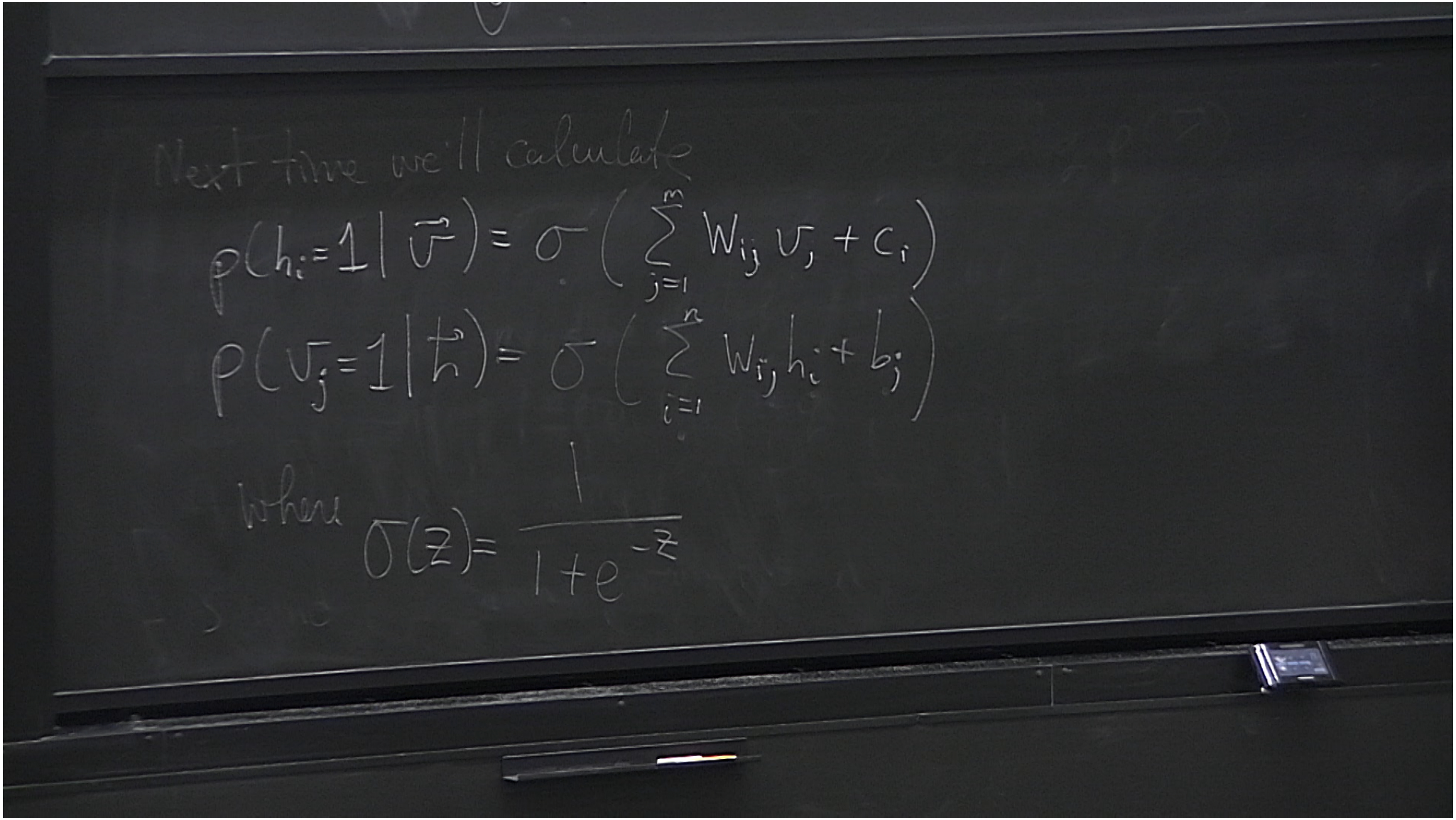
$\left\langle \frac{\partial E}{\partial \theta} \right\rangle$: full joint distribution

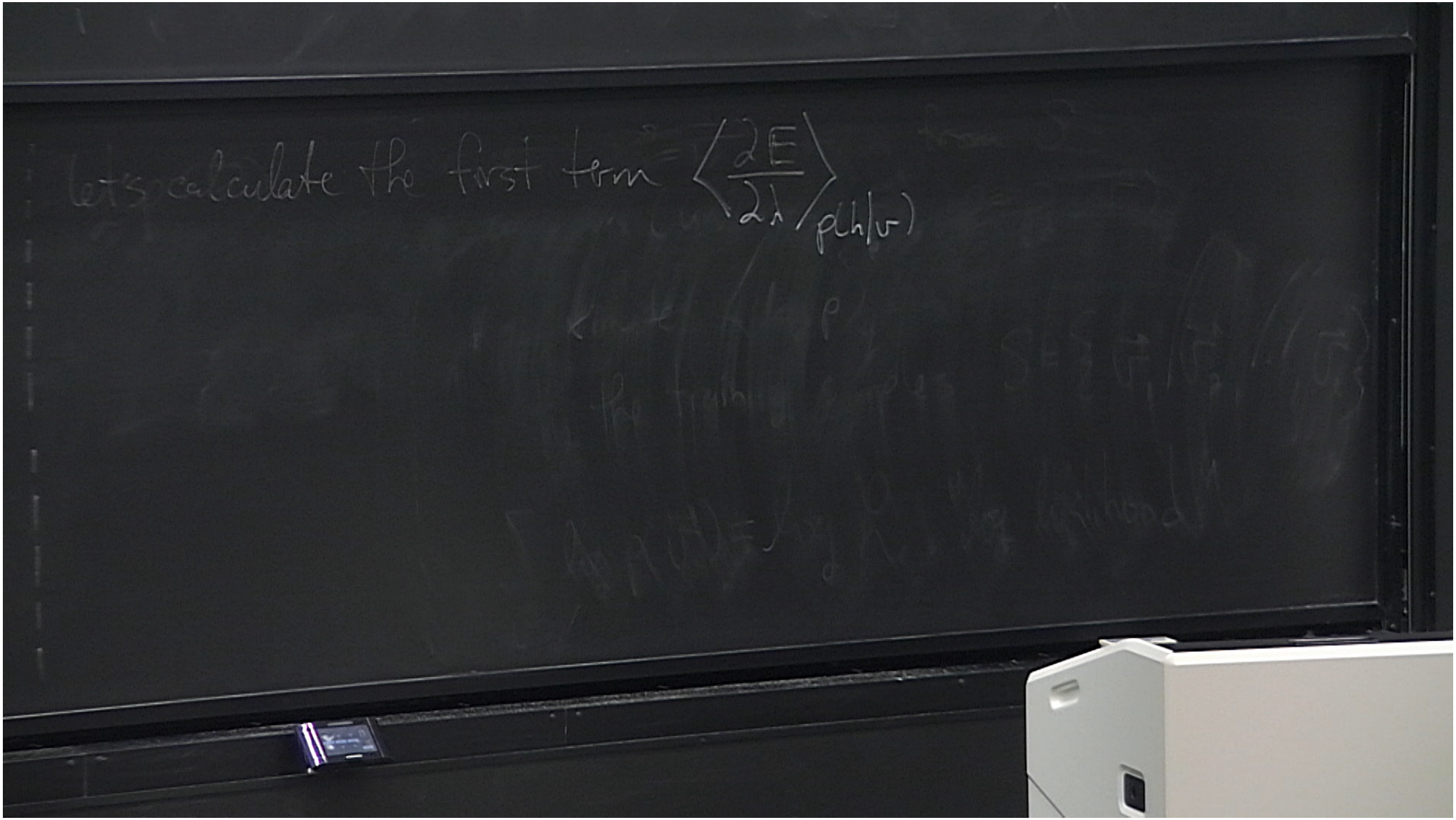
To calculate $p(\vec{h}|\vec{v})$ recall



The hidden uni
(simi lar for
ie. $p(\vec{h}|\vec{v}$
$p(\vec{v}|h$


The hidden units are independent of one another
(similar for the visibles), and the conditional probs. factor
ie. $p(\vec{h}|\vec{v}) = \prod_{i=1}^{n} p(h_i|\vec{v})$
$p(\vec{v}|\vec{h}) = \prod_{i=1}^{m} p(v_i|\vec{h})$

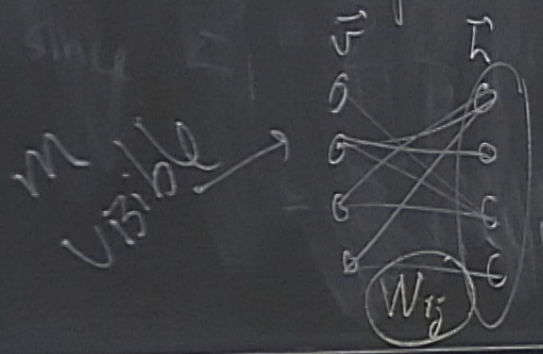

Next time we'll calculate
$p(h_i = 1 | \vec{v}) = \sigma\left(\sum_{j=1}^{m} W_{ij} v_j + c_i\right)$
$p(v_j = 1 | \vec{h}) = \sigma\left(\sum_{i=1}^{n} W_{ij} h_i + b_j\right)$
where $\sigma(z) = \frac{1}{1+e^{-z}}$


let's calculate the first term $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{p(h|v)}$

$\frac{\partial}{\partial d}$ = $\frac{\partial}{\partial d}$ conditional dist of h given v / $\frac{\partial}{\partial d}$ full joint distribution

(similar for

To calculate $p(\vec{h}|\vec{v})$ recall



ie. $p(\vec{h}|\vec{v}$

$p(\vec{v}|\vec{h}$

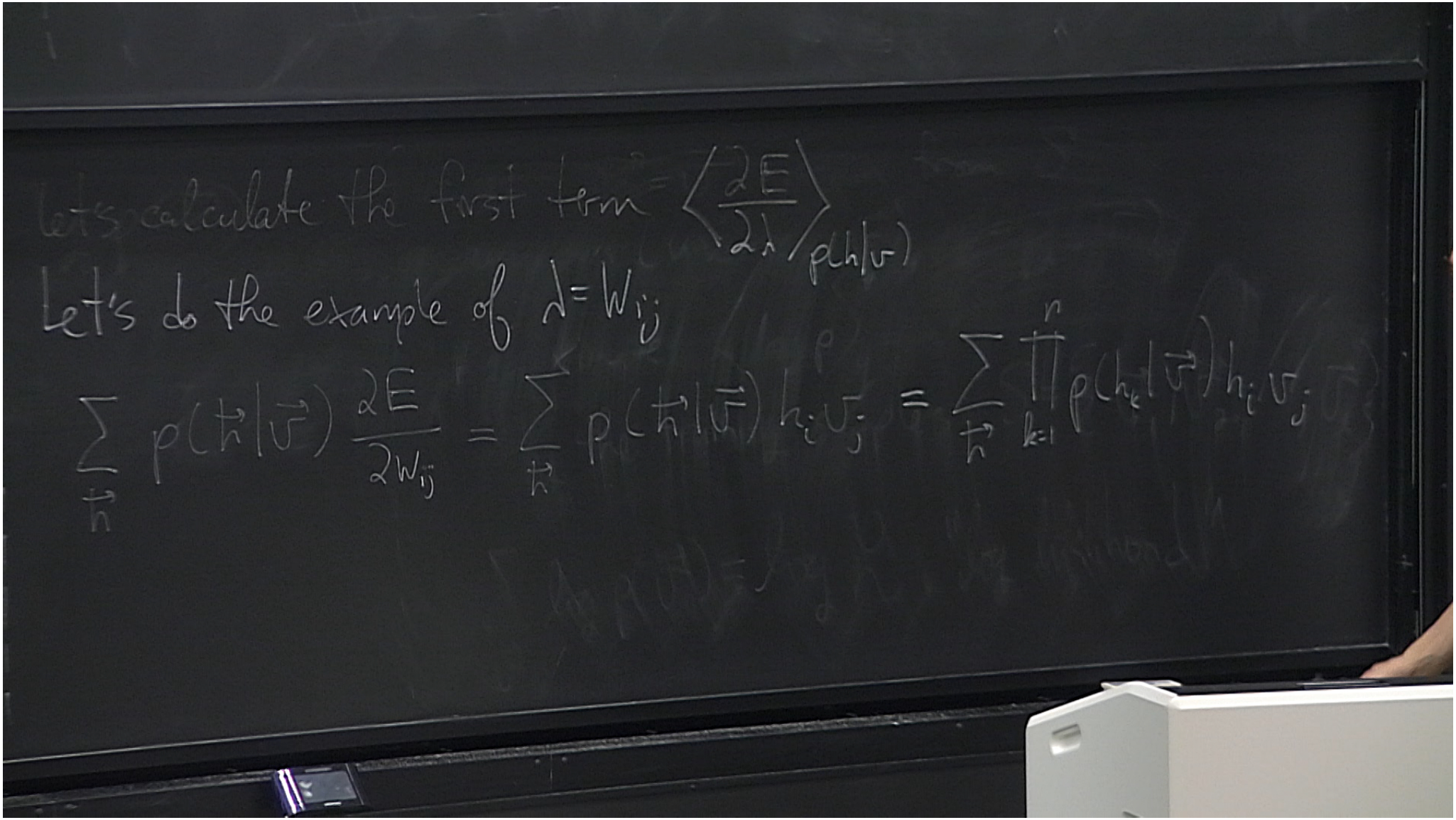
Next time we'll calculate

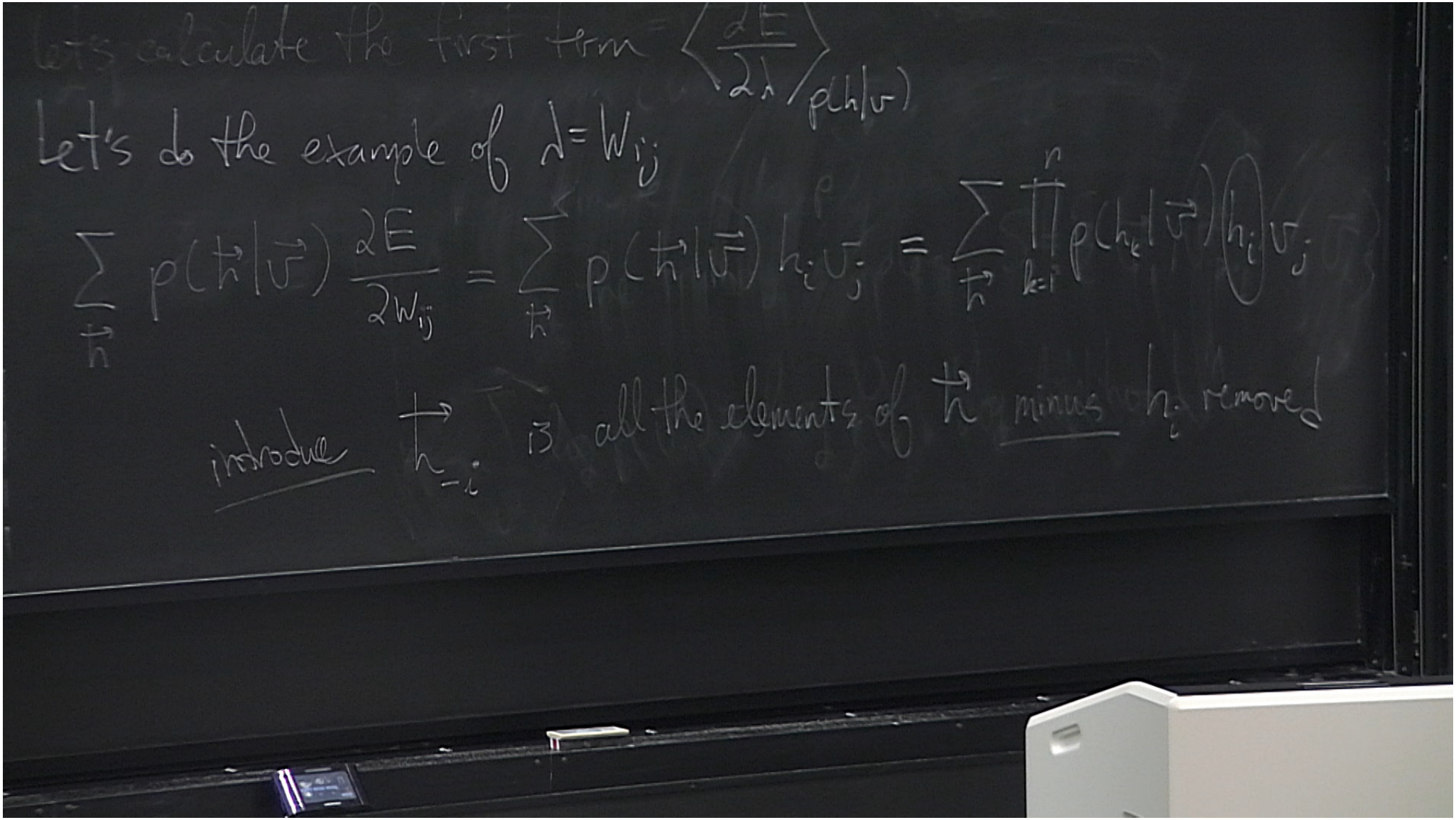
$$p(h_i=1|\vec{v}) = \sigma\left(\sum_{j=1}^{m} W_{ij} v_j + c_i\right)$$

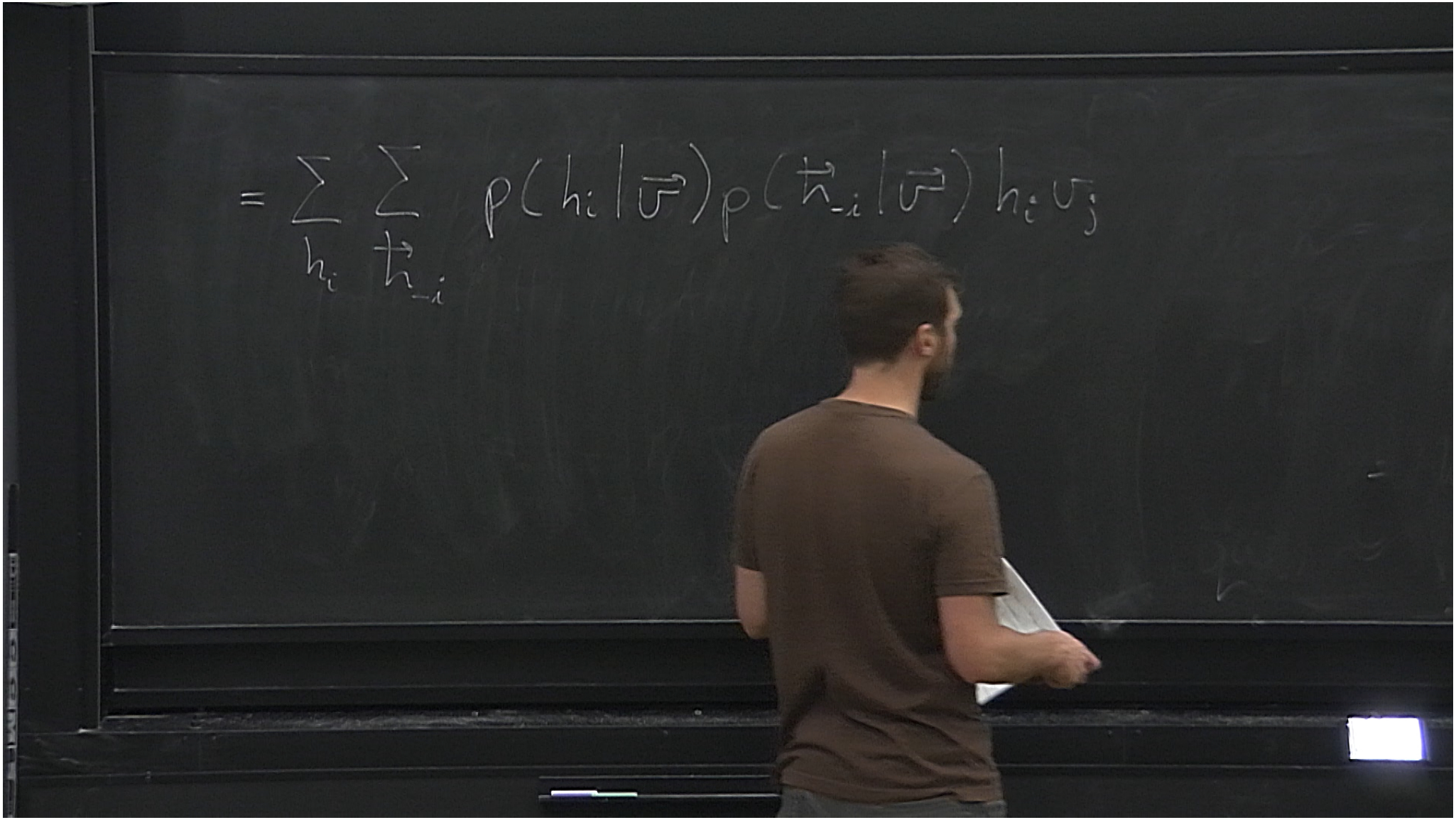
$$p(v_i=1|\vec{h}) = \sigma\left(\sum_{i=1}^{n} W_{ij} h_i + b_j\right)$$

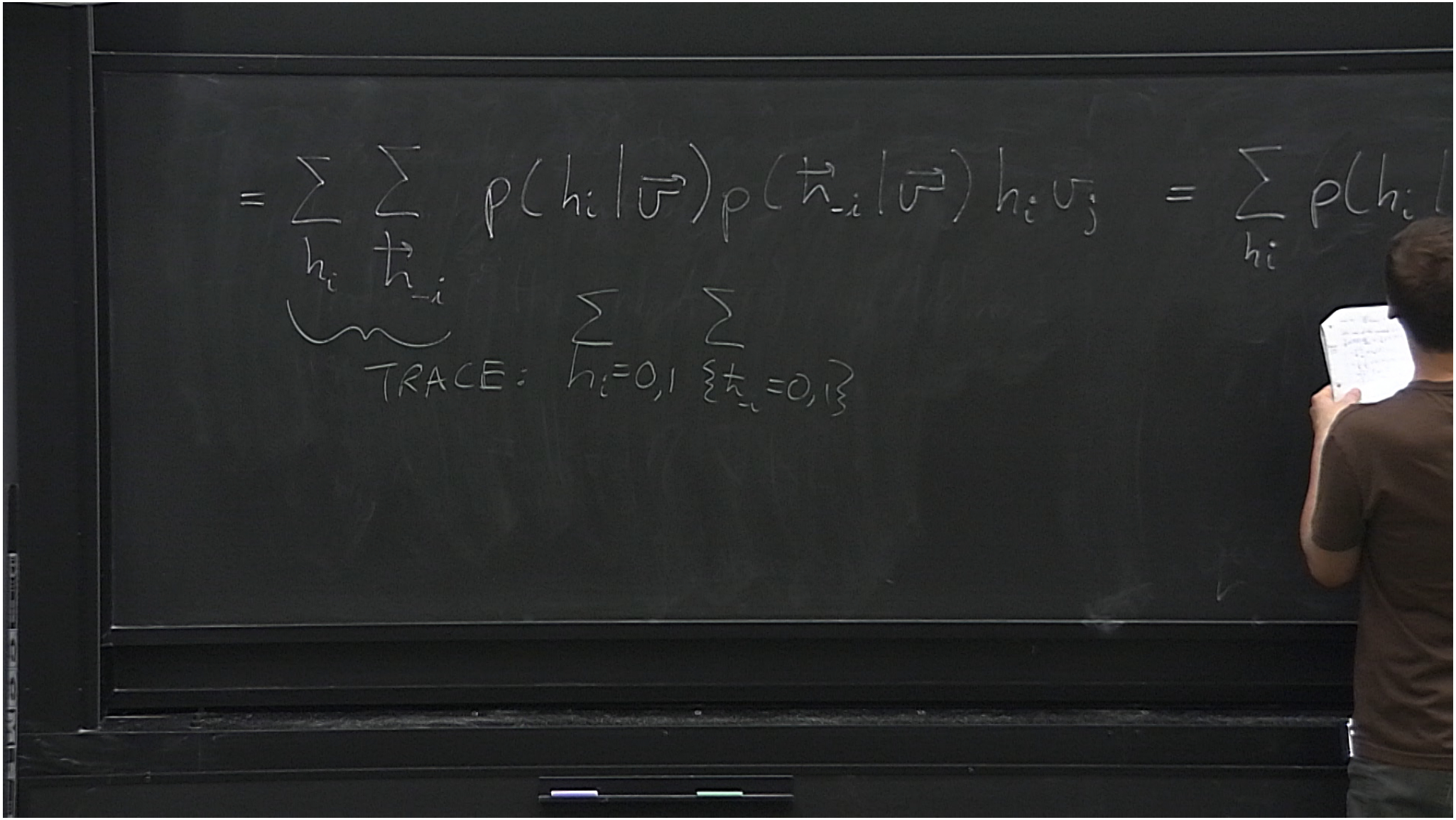
Let's calculate the fir

Let's do the example


Let's calculate the first term $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{p(h|v)}$
Let's do the example of $\lambda = W_{ij}$
$\sum_{\vec{h}} p(\vec{h}|\vec{v}) \frac{\partial E}{\partial W_{ij}} = \sum_{\vec{h}} p(\vec{h}|\vec{v}) h_i v_j = \sum_{\vec{h}} \prod_{k=1}^{r} p(h_k|\vec{v}) h_i v_j$


Let's calculate the first term $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{p(h|v)}$
Let's do the example of $\lambda = W_{ij}$
$\sum_{\vec{h}} p(\vec{h}|\vec{v}) \frac{\partial E}{\partial W_{ij}} = \sum_{\vec{h}} p(\vec{h}|\vec{v}) h_i v_j = \sum_{\vec{h}} \prod_{k=1}^{n} p(h_k|\vec{v}) h_i v_j$
introduce $\vec{h}_{-i}$: is all the elements of $\vec{h}$ minus h_i removed


$= \sum_{h_i} \sum_{\vec{h}_{-i}} p(h_i|\vec{v}) p(\vec{h}_{-i}|\vec{v}) h_i v_j$

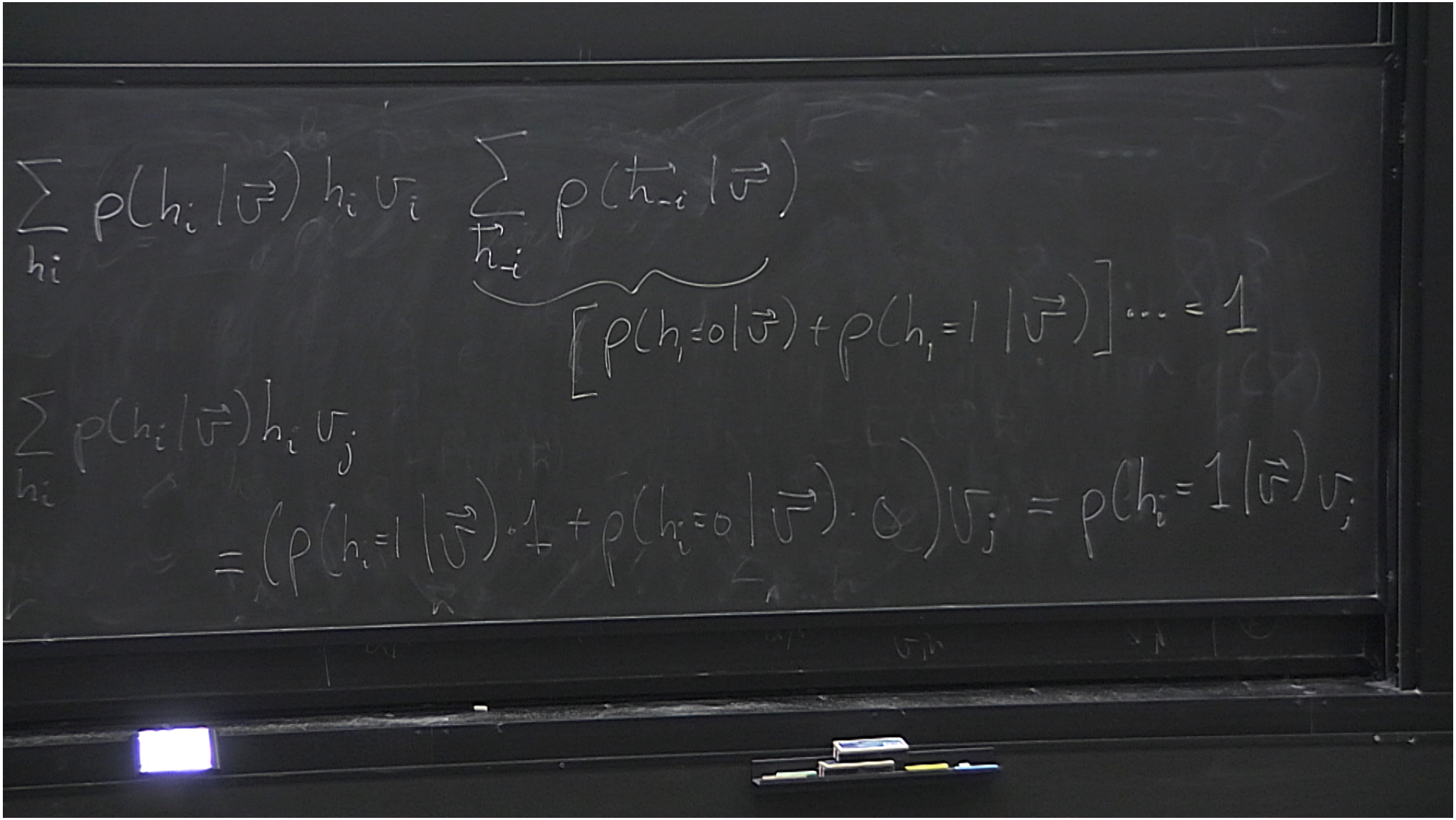

$= \sum_{h_i} \sum_{\vec{h}_{-i}} p(h_i|\vec{v}) p(\vec{h}_{-i}|\vec{v}) h_i v_j = \sum_{h_i} p(h_i$
$h_i \; \vec{h}_{-i}$
TRACE: $\sum_{h_i=0,1} \sum_{\{\vec{h}_{-i}=0,1\}}$

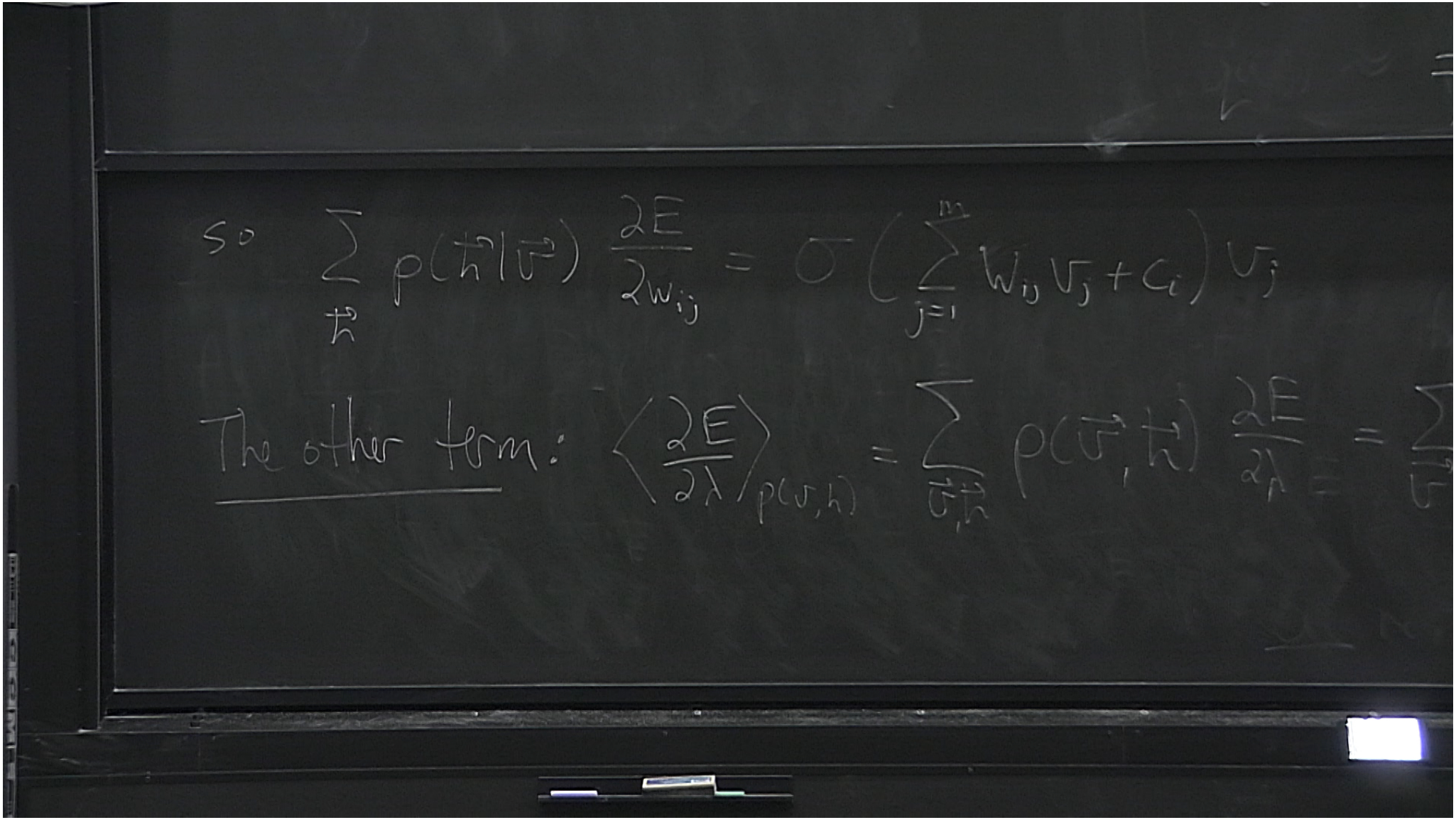
$$= \sum_{h_i} p(h_i | \vec{v}) \, h_i \, v_i \sum_{\vec{h}_{-i}} p(\vec{h}_{-i} | \vec{v})$$

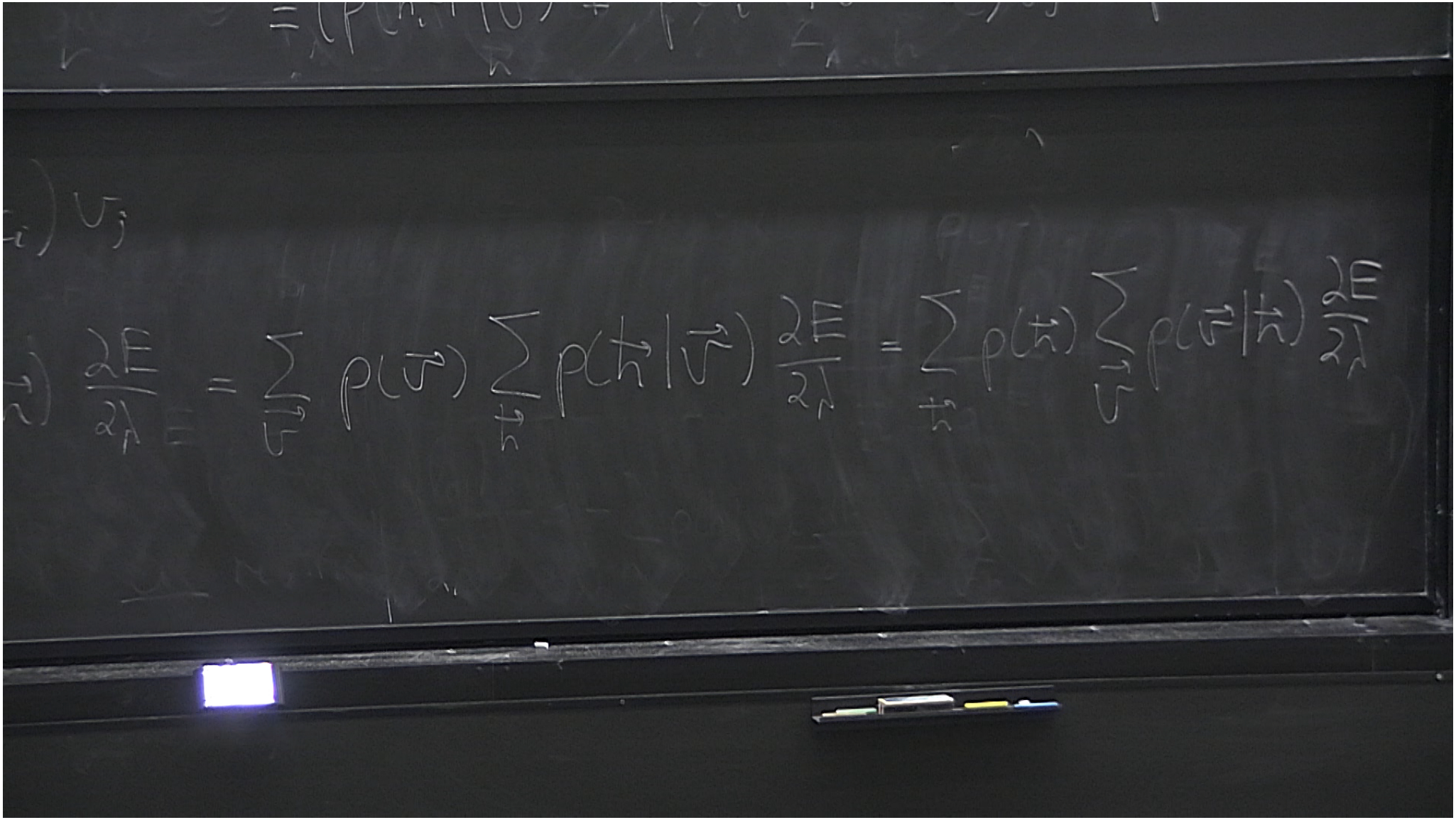
$$\left[p(h_1 = 0 | \vec{v}) + p(h_1 = 1 | \vec{v}) \right] \cdots$$

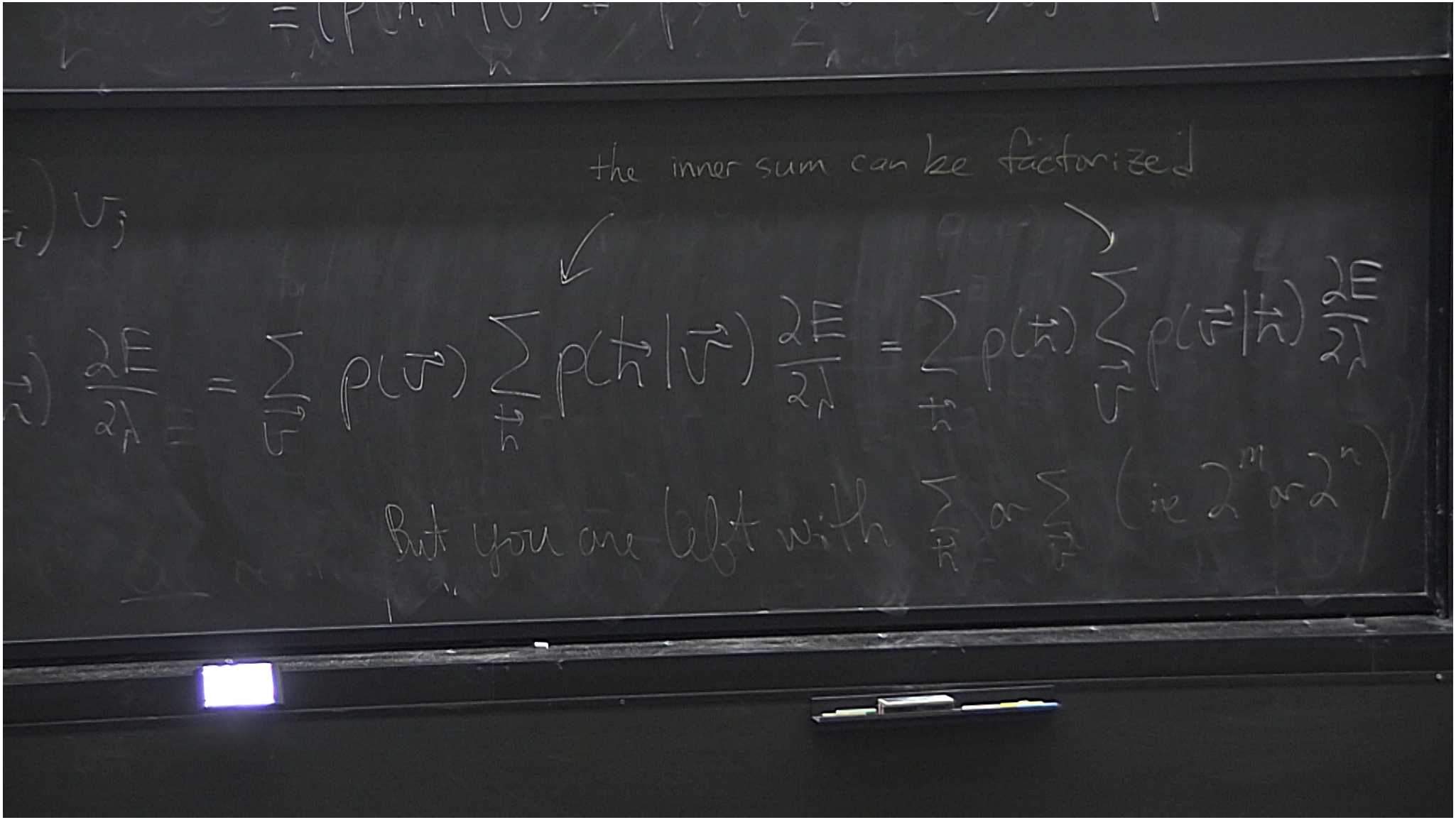
$$= \sum_{h_i} p(h_i | \vec{v}) \, h_i \, v_i \sum_{\vec{h}_{-i}} p(\vec{h}_{-i} | \vec{v})$$

$$\underbrace{\left[p(h_1 = 0 | \vec{v}) + p(h_1 = 1 | \vec{v}) \right] \cdots}_{= 1}$$


$\sum_{h_i} p(h_i|\vec{v}) h_i v_i \sum_{\vec{h}_{-i}} p(\vec{h}_{-i}|\vec{v})$
$[p(h_i=0|\vec{v}) + p(h_i=1|\vec{v})] \cdots = 1$
$\sum_{h_i} p(h_i|\vec{v}) h_i v_j$
$= (p(h_i=1|\vec{v}) \cdot 1 + p(h_i=0|\vec{v}) \cdot 0) v_j = p(h_i=1|\vec{v}) v_j$


So $\sum_{\vec{h}} p(\vec{h}|\vec{v}) \frac{\partial E}{\partial w_{ij}} = \sigma\left(\sum_{j=1}^{m} w_{ij} v_j + c_i\right) v_j$
The other term: $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_{p(v,h)} = \sum_{\vec{v},\vec{h}} p(\vec{v},\vec{h}) \frac{\partial E}{\partial \lambda} =$


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


the inner sum can be factorized
$\frac{\partial E}{\partial \lambda} = \sum_{\vec{v}} p(\vec{v}) \sum_{\vec{h}} p(\vec{h}|\vec{v}) \frac{\partial E}{\partial \lambda} = \sum_{\vec{h}} p(\vec{h}) \sum_{\vec{v}} p(\vec{v}|\vec{h}) \frac{\partial E}{\partial \lambda}$
But you are left with $\sum_{\vec{h}}$ or $\sum_{\vec{v}}$ (ie 2^m or 2^n)