Title: PSI 2017/2018 - Machine Learning for Many Body Physics - Lecture 8

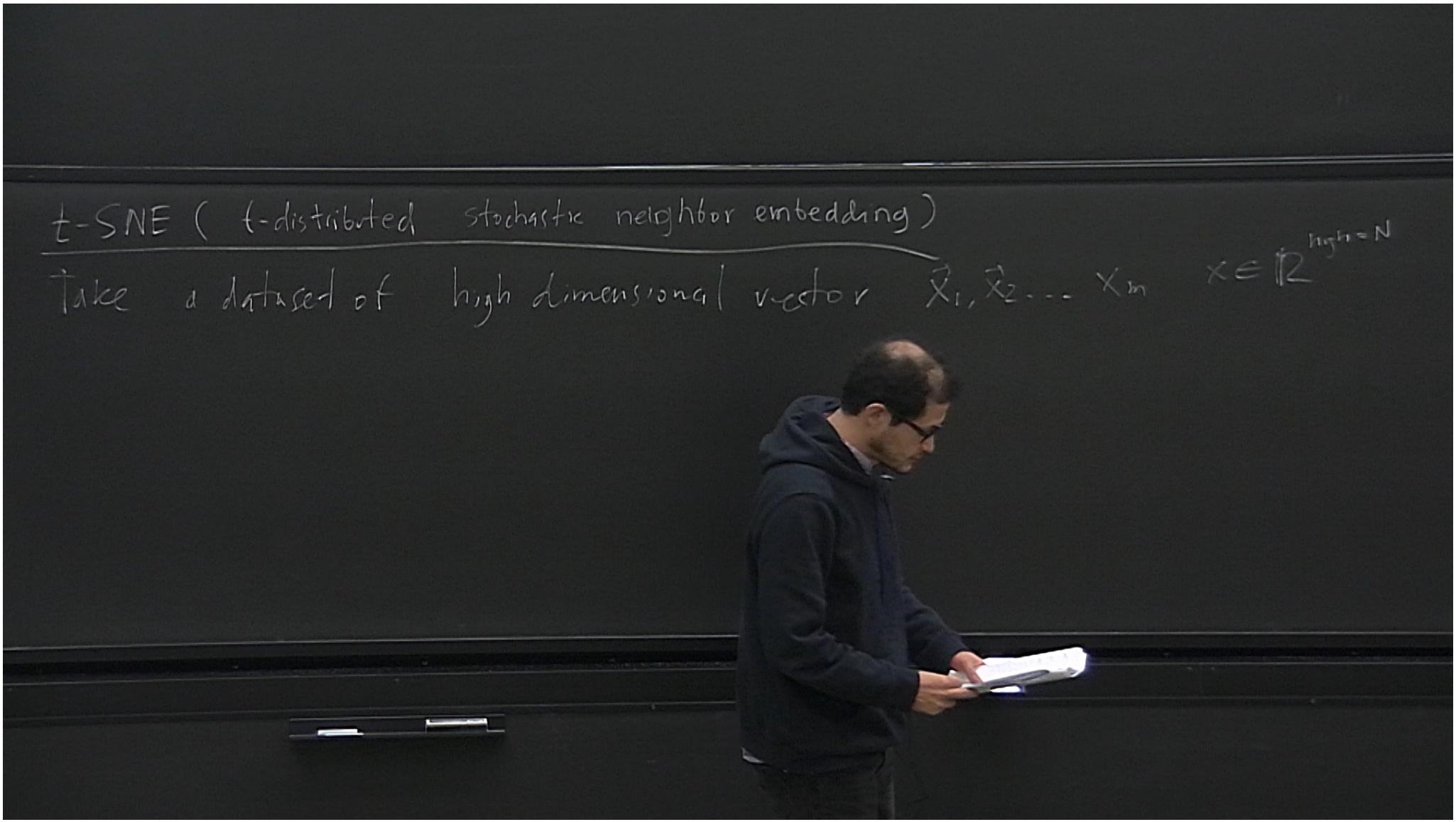
Date: Apr 18, 2018 09:00 AM

URL: http://pirsa.org/18040061

Abstract:

t-SNE (t-distributed stochastic neighbor embedding)

Take a dataset of high dimensional vector $\vec{x}_1, \vec{x}_2 \ldots x_m$ $x \in \mathbb{R}^{\text{high-N}}$



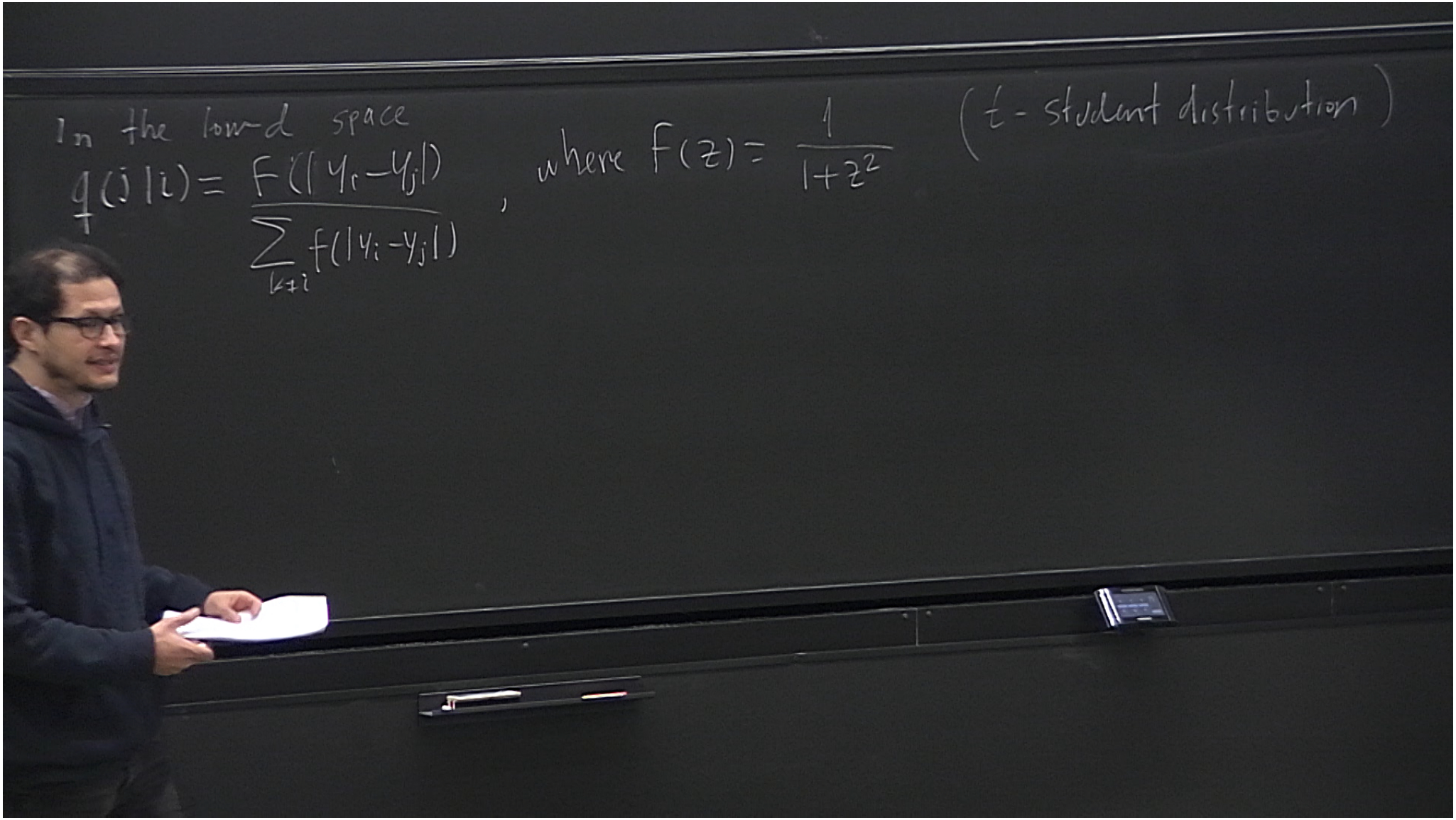
t-SNE (t-distributed stochastic neighbor embedding)

Take a dataset of high dimensional vector $\vec{X}_1, \vec{X}_2 \dots X_m$ $X \in \mathbb{R}^{\text{high} = N}$

- Create a 2-d representation of the dataset $\vec{Y}_1\ \vec{Y}_2 \dots \vec{Y}_m$ $Y \in \mathbb{R}^2$

How? Convert euclidean distance into conditional probabilities that represent similarity

Similarity $\rightarrow$ $P(j|i) = e^{\frac{-\|X_i - X_j\|^2}{2\sigma^2}} \Big/ \sum_{k \neq i} e^{-\|X_i - X_k\|^2/2\sigma^2}$


In the low-d space
$q(j|i) = \frac{f(|y_i - y_j|)}{\sum_{k \neq i} f(|y_i - y_j|)}$, where $f(z) = \frac{1}{1+z^2}$ (t-student distribution)

In the low-d space

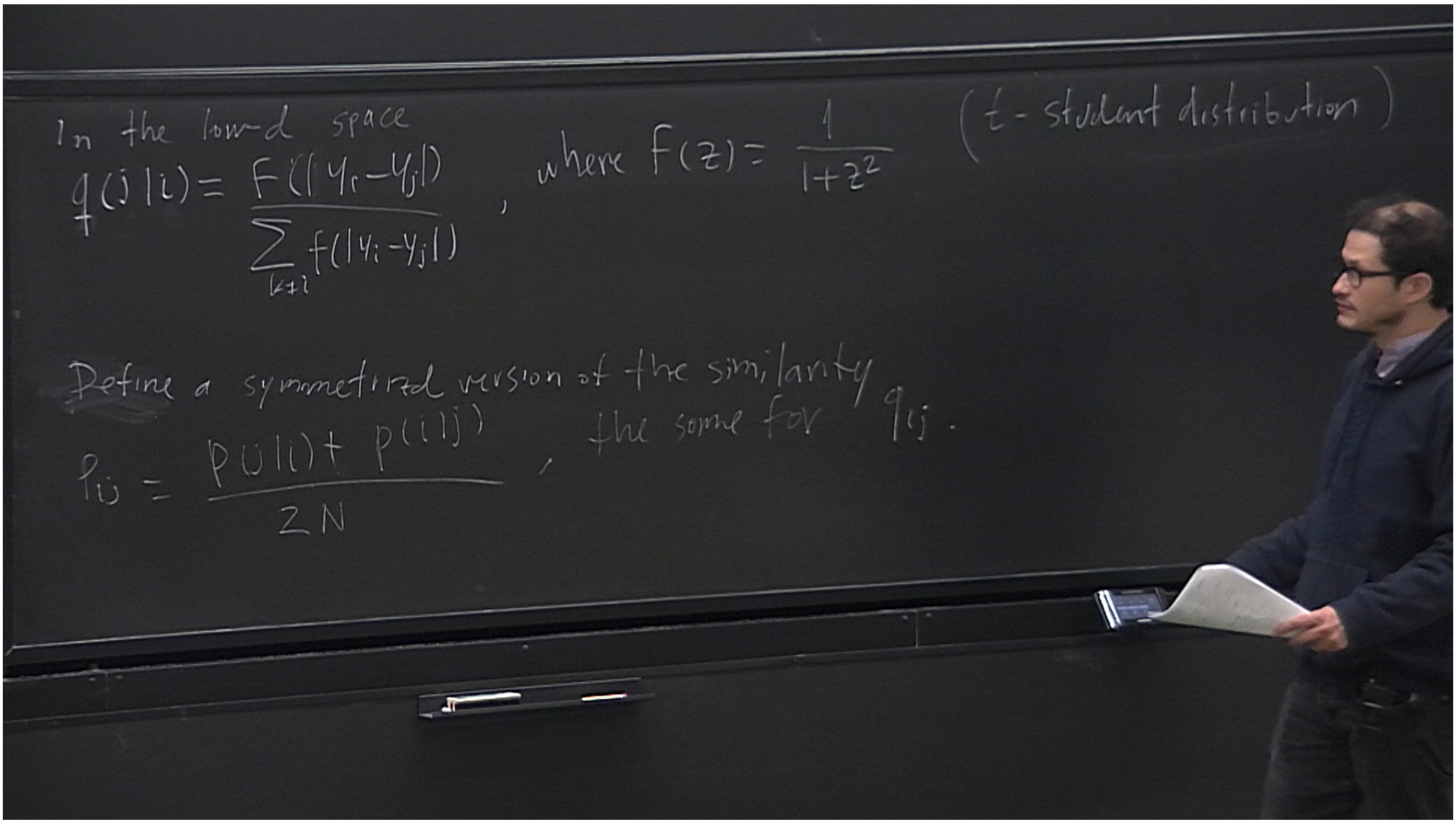
$$q(j|i) = \frac{f(|y_i - y_j|)}{\sum_{k \neq i} f(|y_i - y_j|)}, \quad \text{where } f(z) = \frac{1}{1+z^2}$$

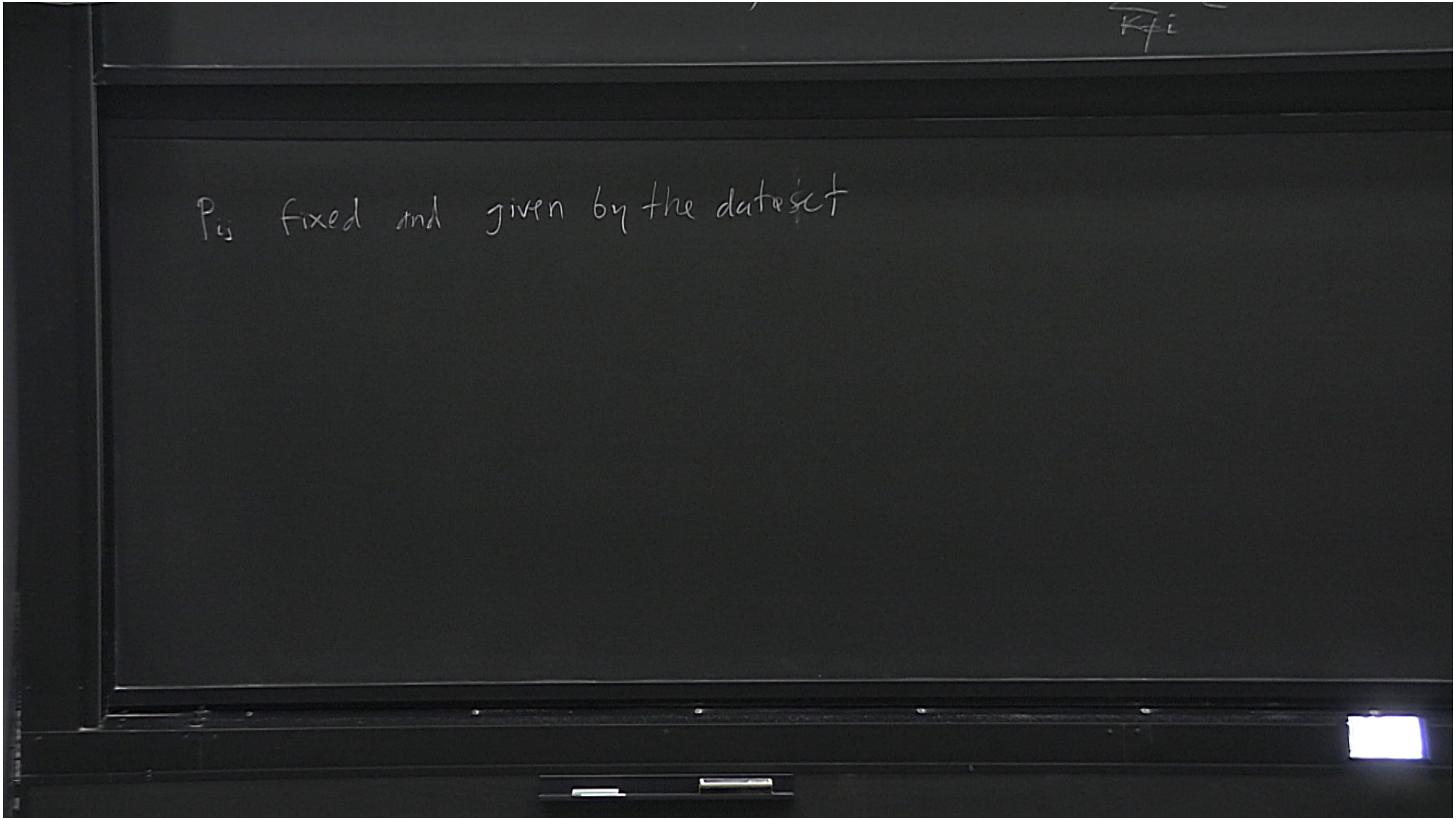
(t-student distribution)

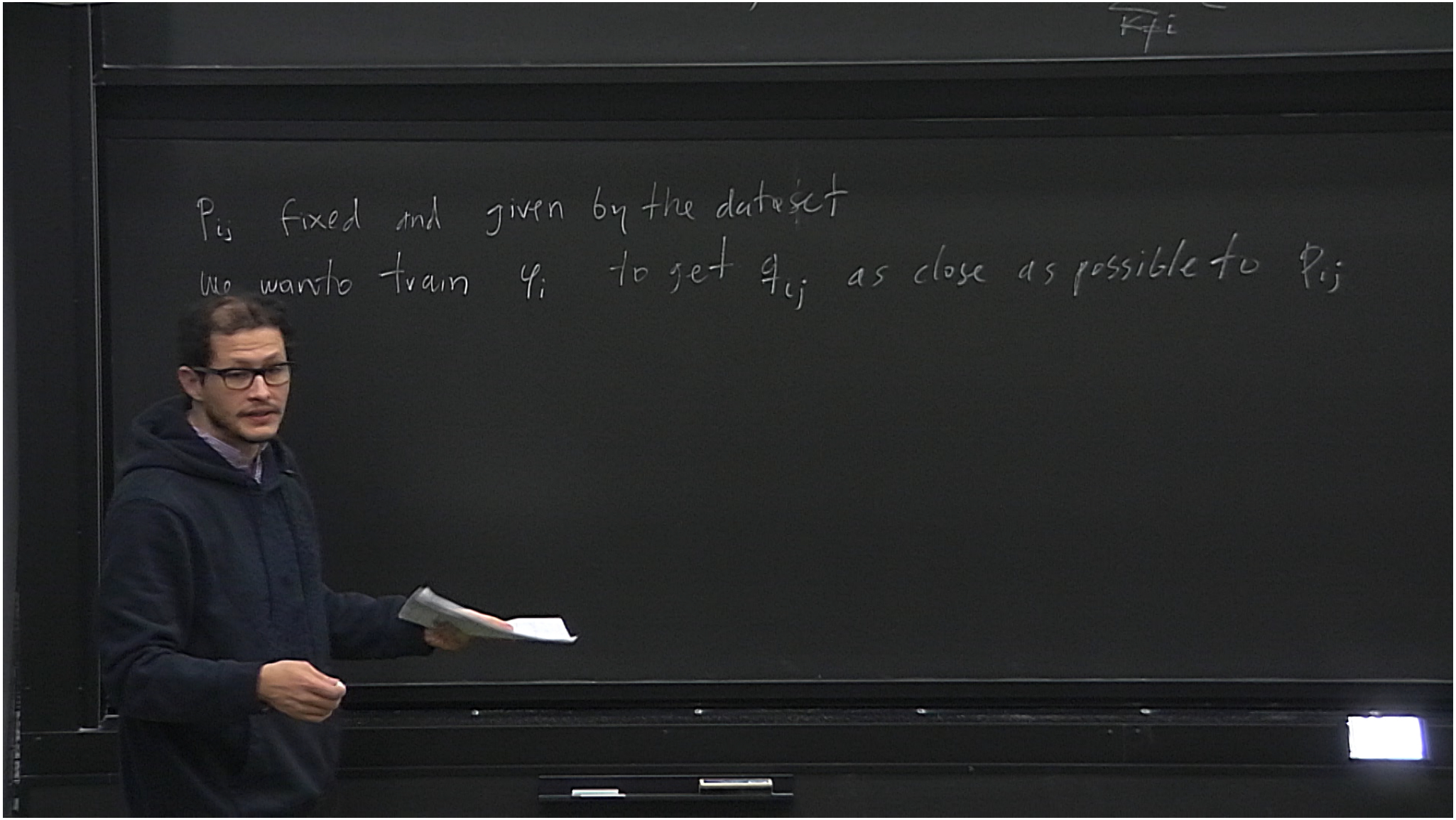
Define a symmetrized version of the similarity

$$p_{ij} = \frac{p(j|i) + p(i|j)}{2N},$$

the same for q_{ij}.




P_{ij} fixed and given by the dataset


$K+i$
P_{ij} fixed and given by the dataset
We want to train q_i to get q_{ij} as close as possible to P_{ij}

P_{ij} fixed and given by the dataset

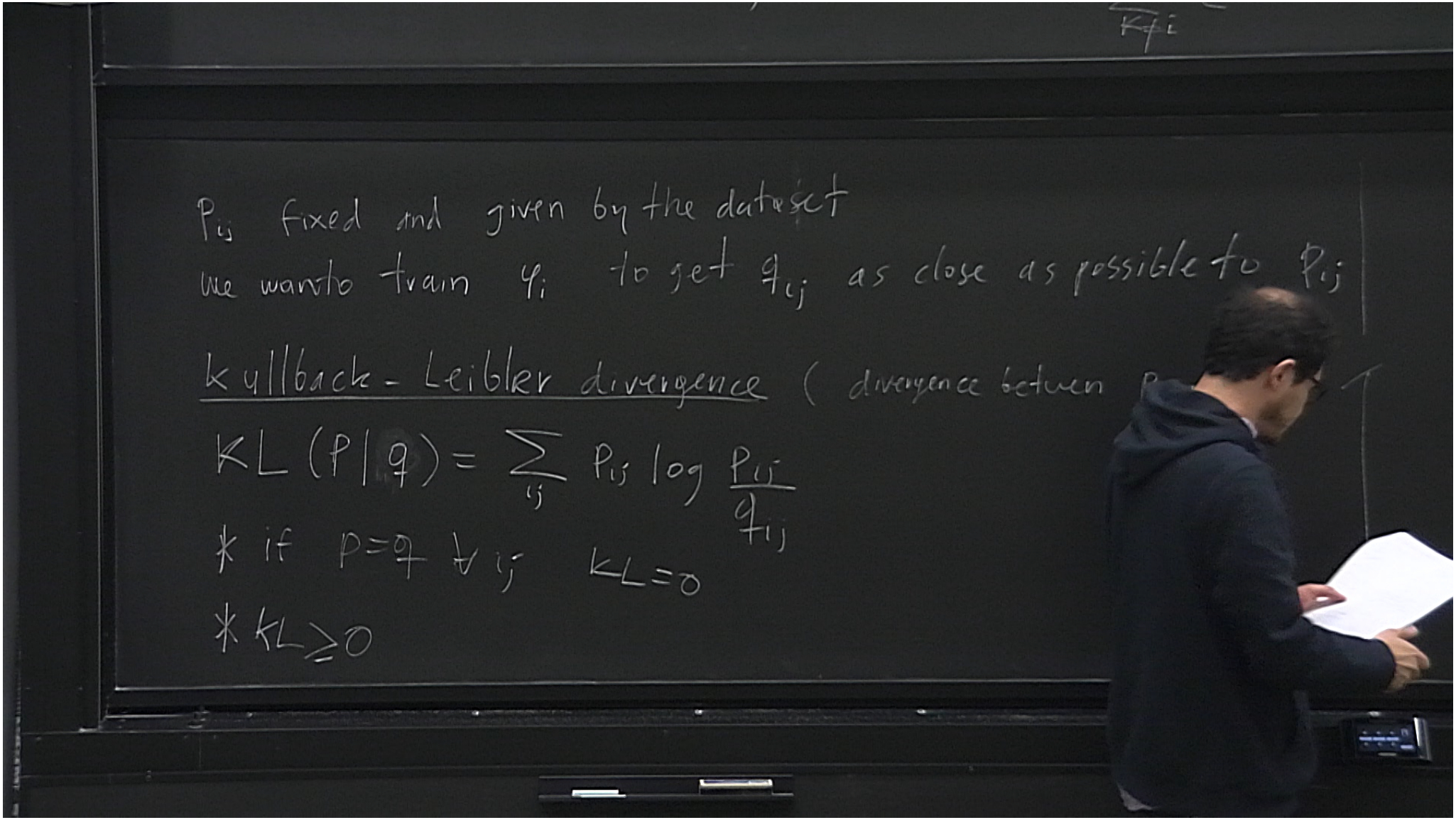
We want to train q_i to get q_{ij} as close as possible to P_{ij}

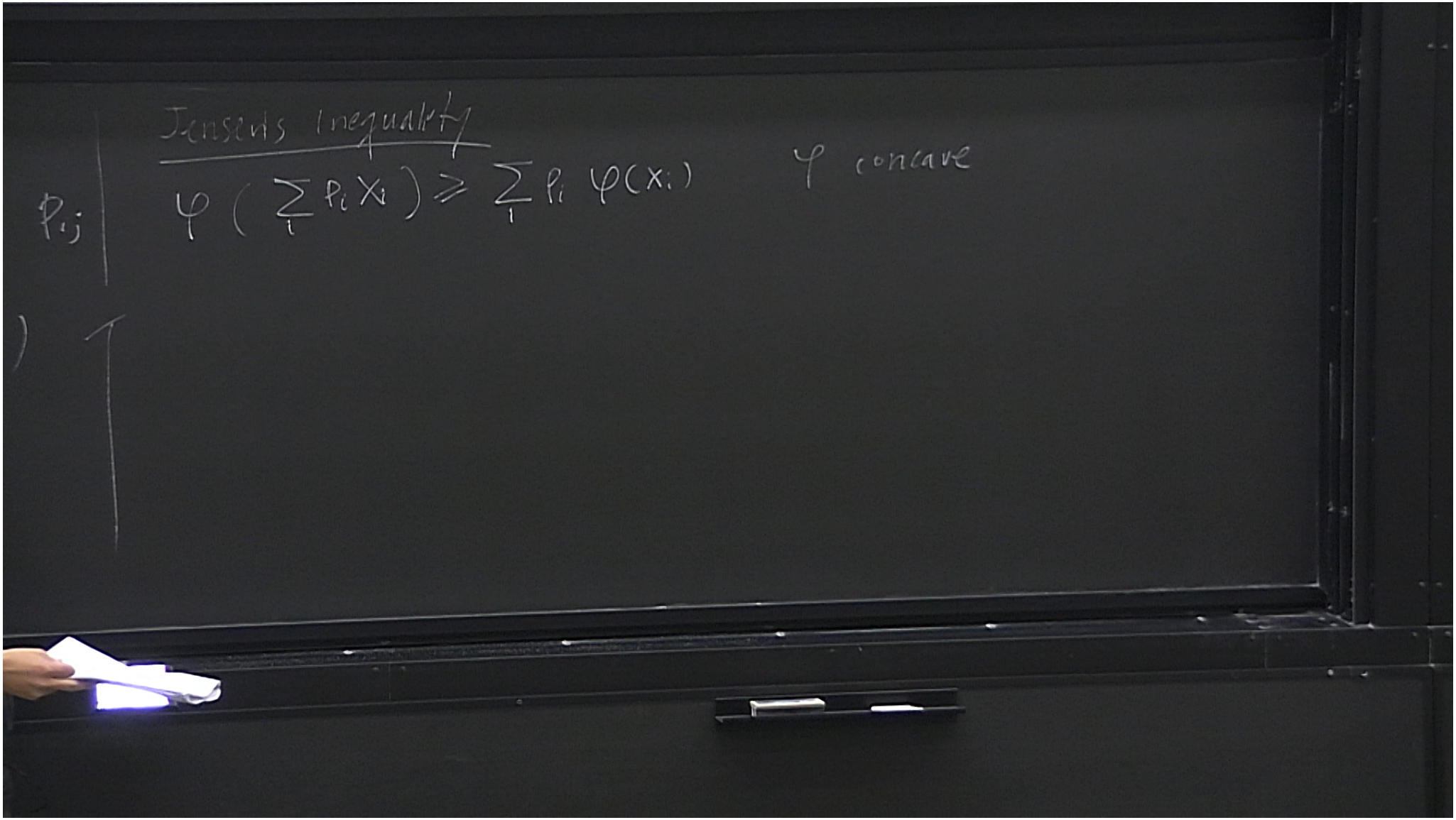
Kullback-Leibler divergence (divergence between P and q)

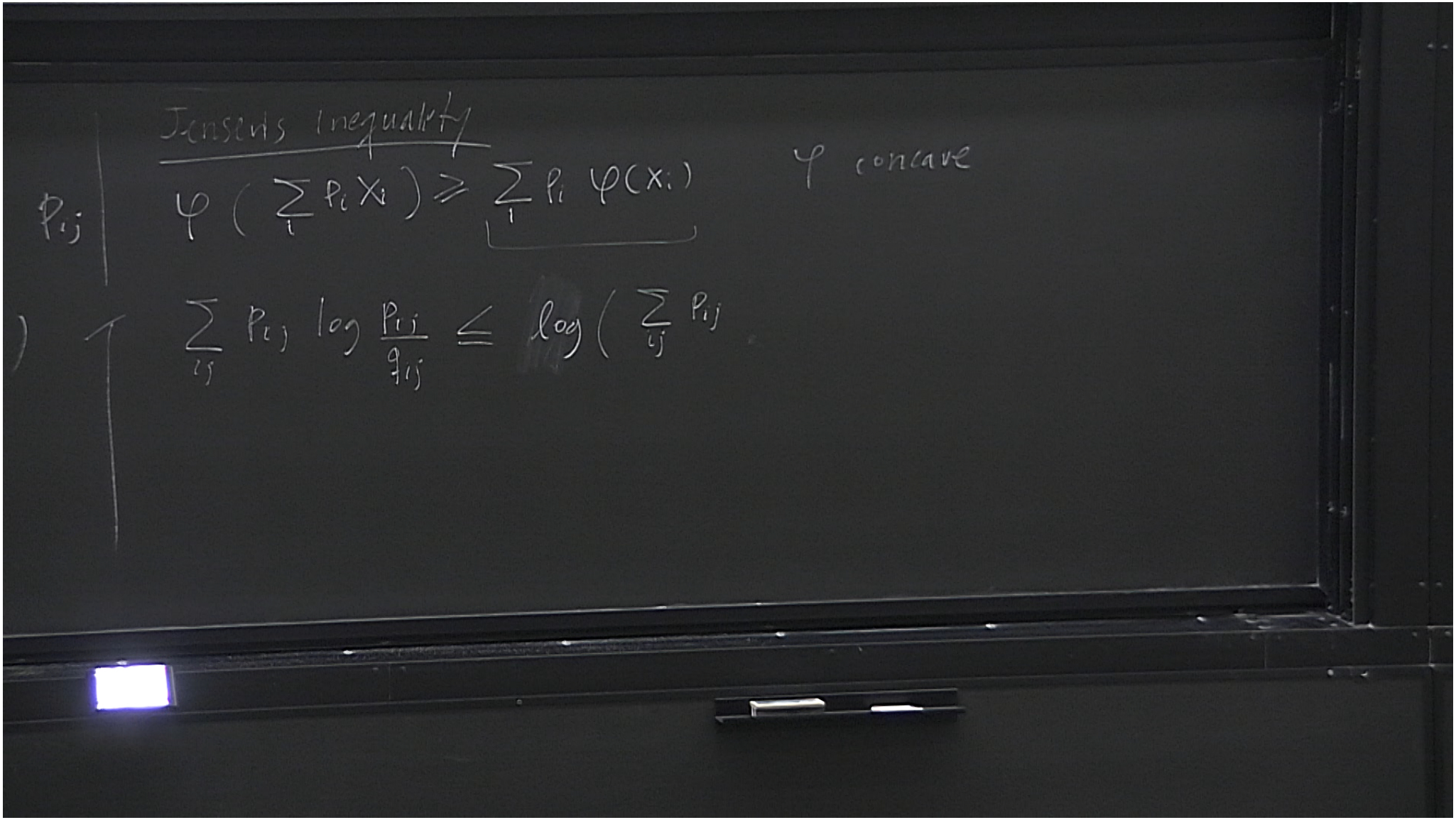
$$KL(P|q) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{q_{ij}}$$

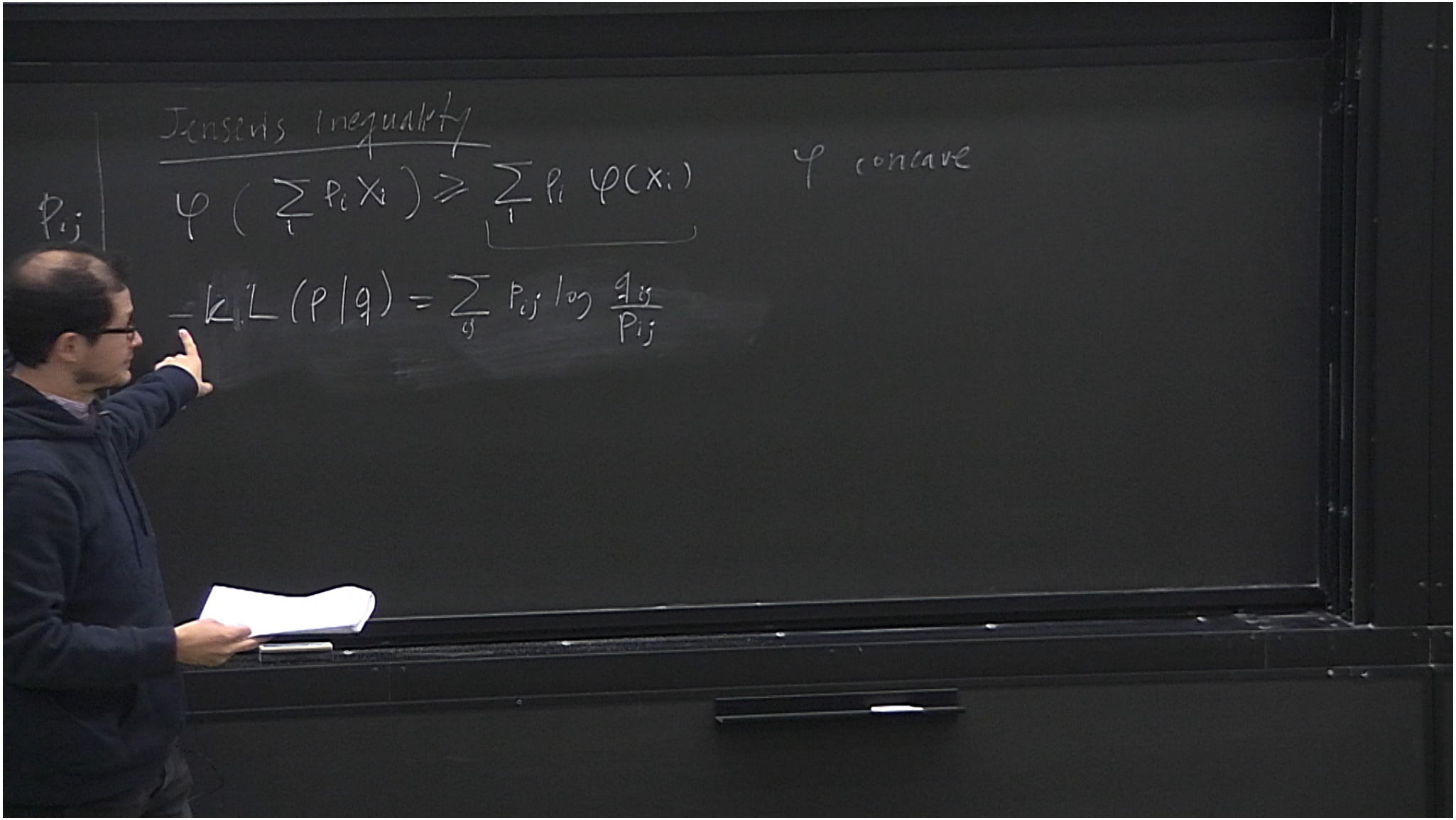
if $P = q$ $\forall\, ij$ $KL = 0$

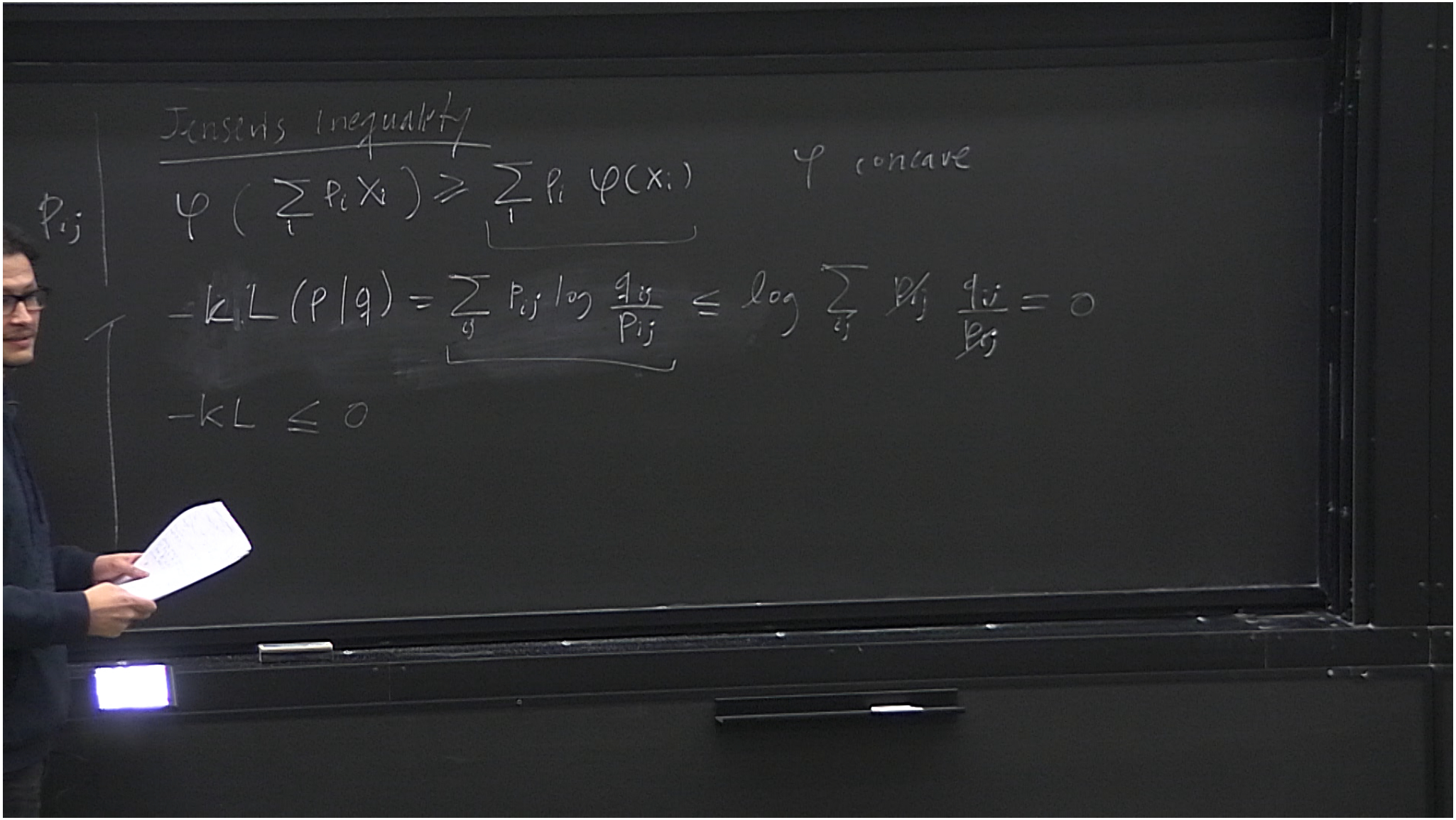
$KL \geq 0$

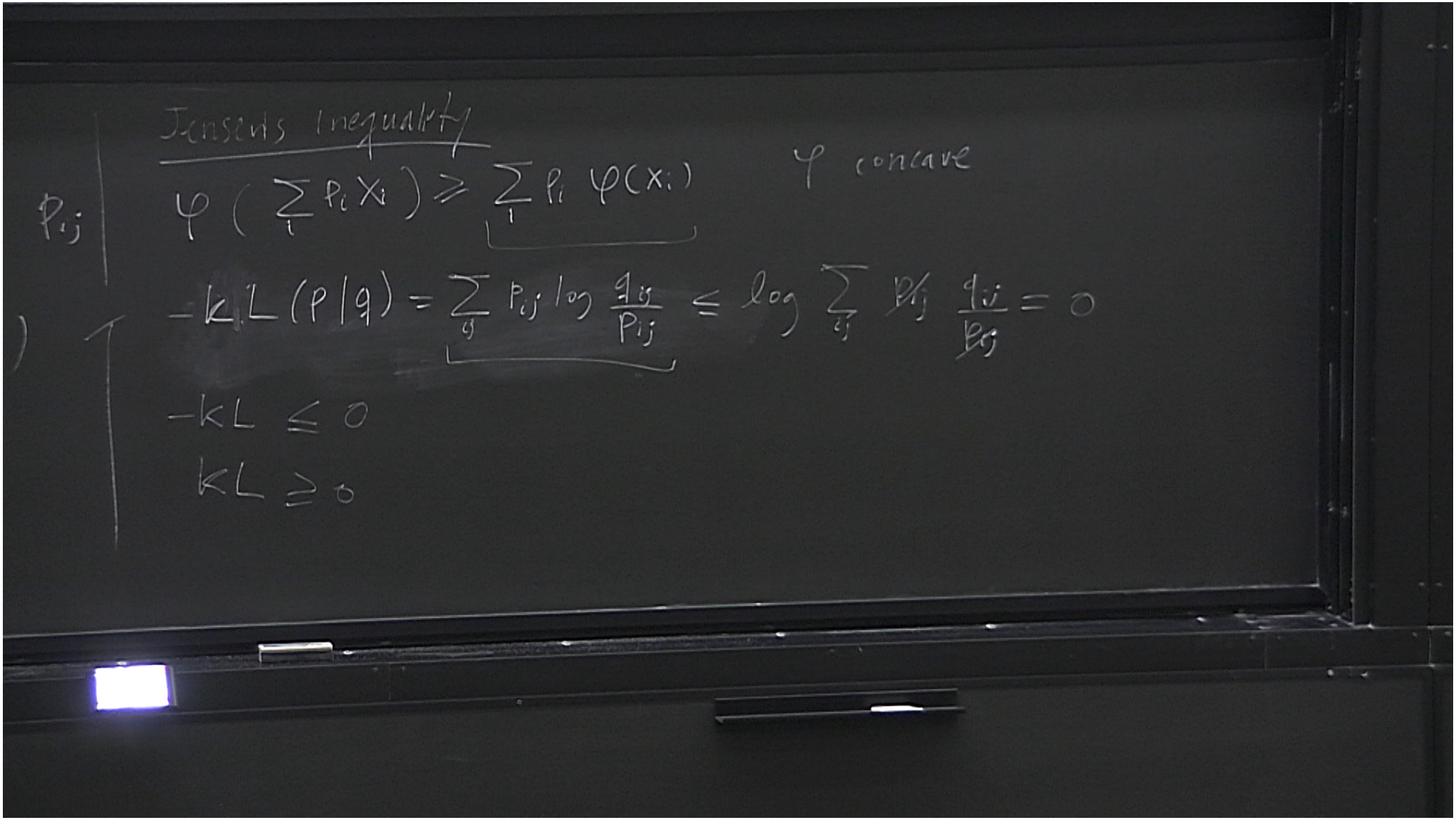

P_{ij} fixed and given by the dataset
We want to train q_i to get q_{ij} as close as possible to P_{ij}
Kullback - Leibler divergence (divergence between
$KL(P|q) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{q_{ij}}$
* if $P=q$ $\forall ij$ $KL=0$
* $KL \geq 0$

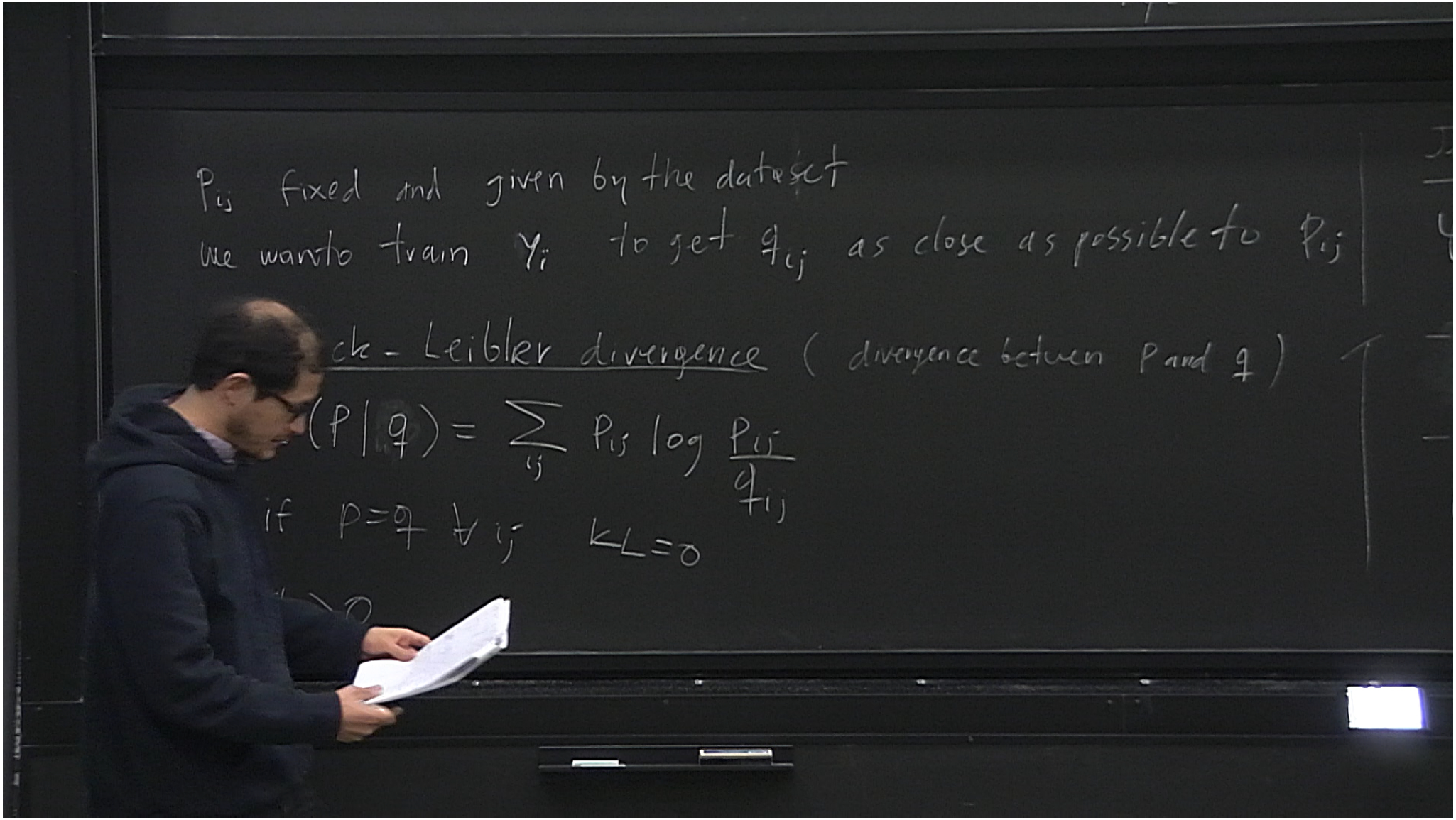

Jensens Inequality
$\varphi\left(\sum_i p_i x_i\right) \geq \sum_i p_i \varphi(x_i)$
φ concave
p_i

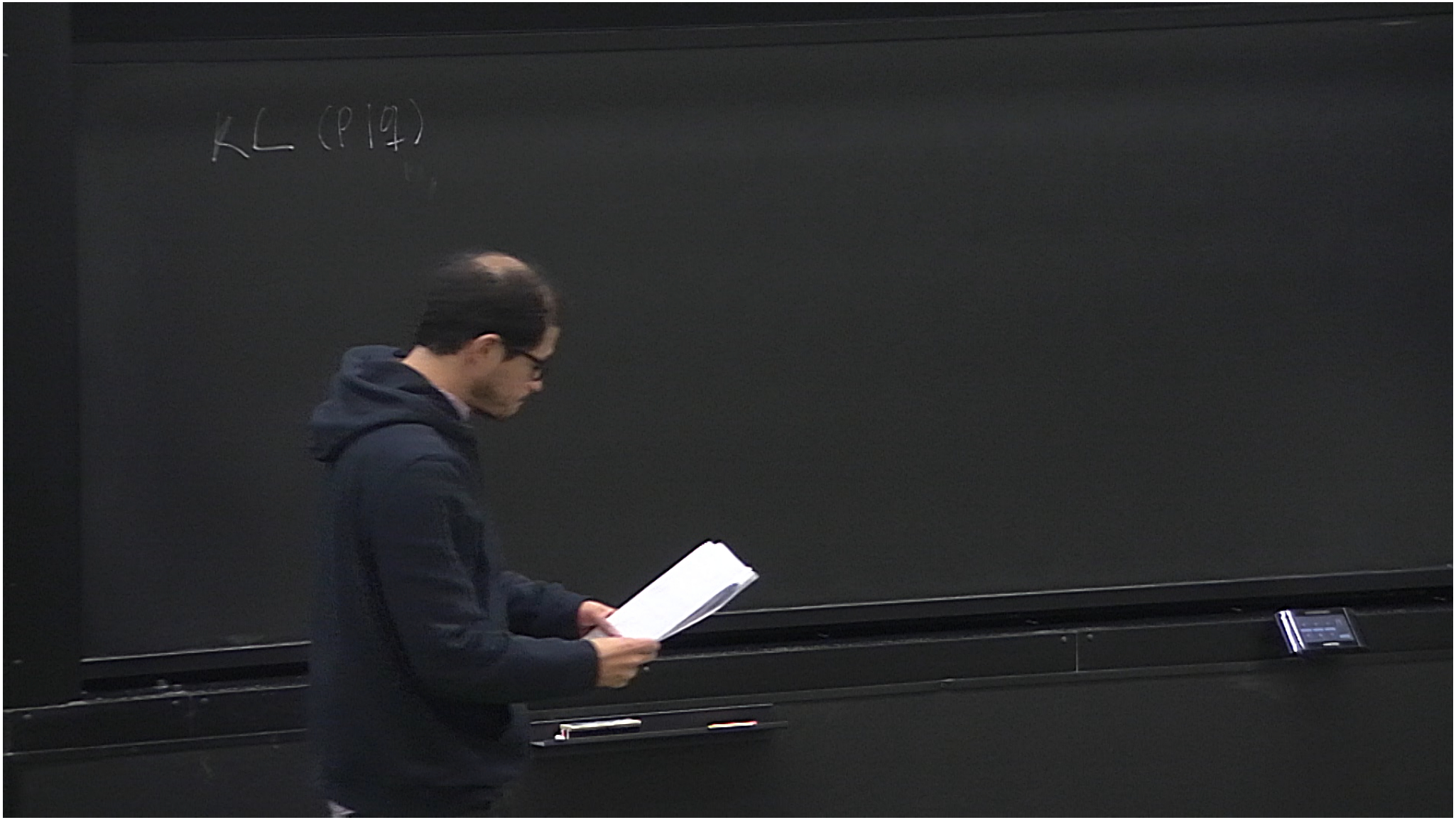

Jensens inequality
$\varphi\left(\sum_i p_i x_i\right) \geq \sum_i p_i \varphi(x_i)$
φ concave
p_{ij}
$\sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} \leq \log\left(\sum_{ij} p_{ij}\right.$

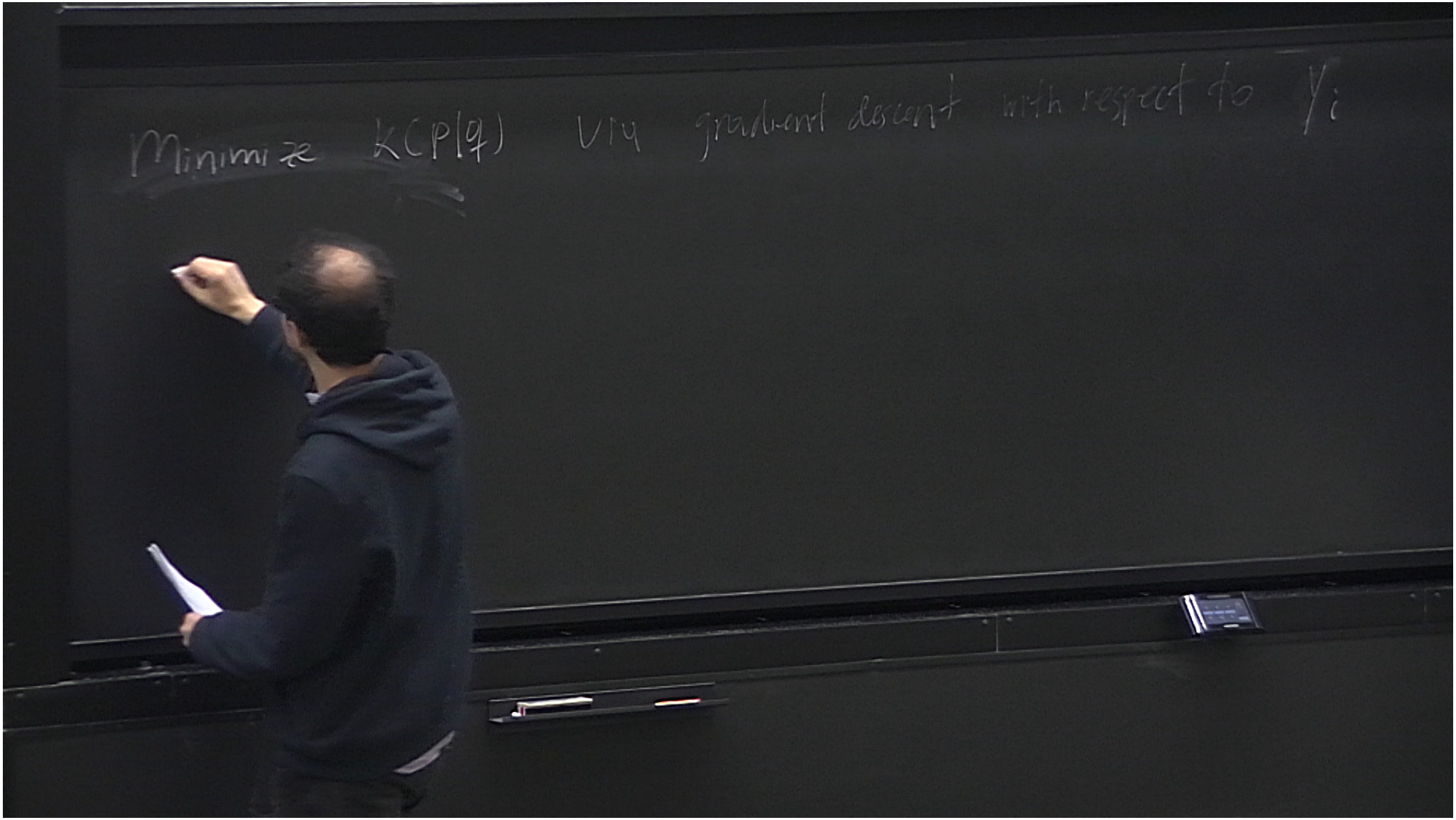

Jensens inequality
$\varphi\left(\sum_i p_i x_i\right) \geq \sum_i p_i \varphi(x_i)$
φ concave
p_{ij}
$-KL(P|Q) = \sum_{ij} p_{ij} \log \frac{q_{ij}}{p_{ij}}$

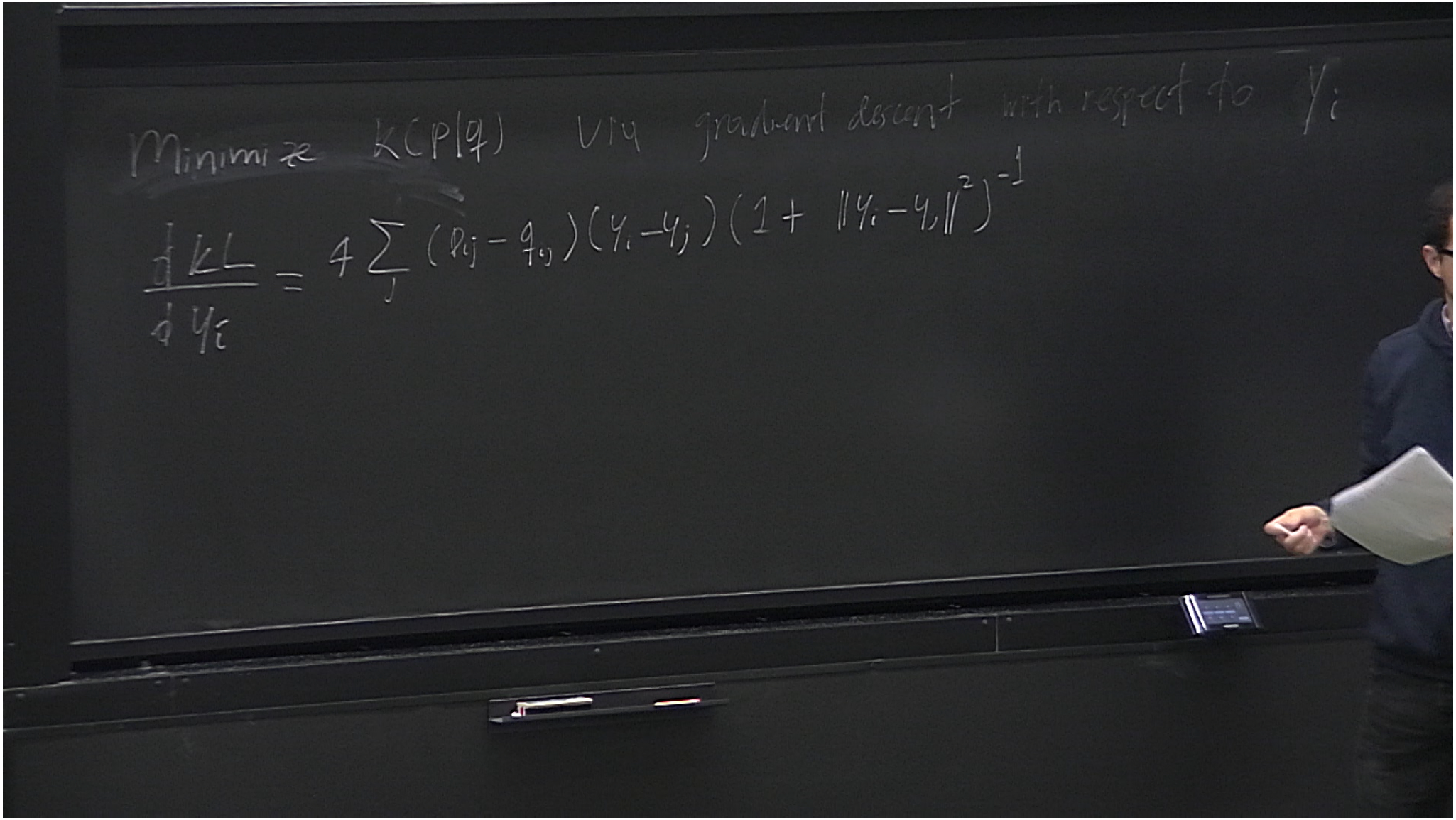

Jensens inequality
$\varphi\left(\sum_i p_i x_i\right) \geq \sum_i p_i \varphi(x_i)$
φ concave
p_{ij}
$-kL(P|q) = \sum_{ij} p_{ij} \log \frac{q_{ij}}{p_{ij}} \leq \log \sum_{ij} p_{ij} \frac{q_{ij}}{p_{ij}} = 0$
$-kL \leq 0$

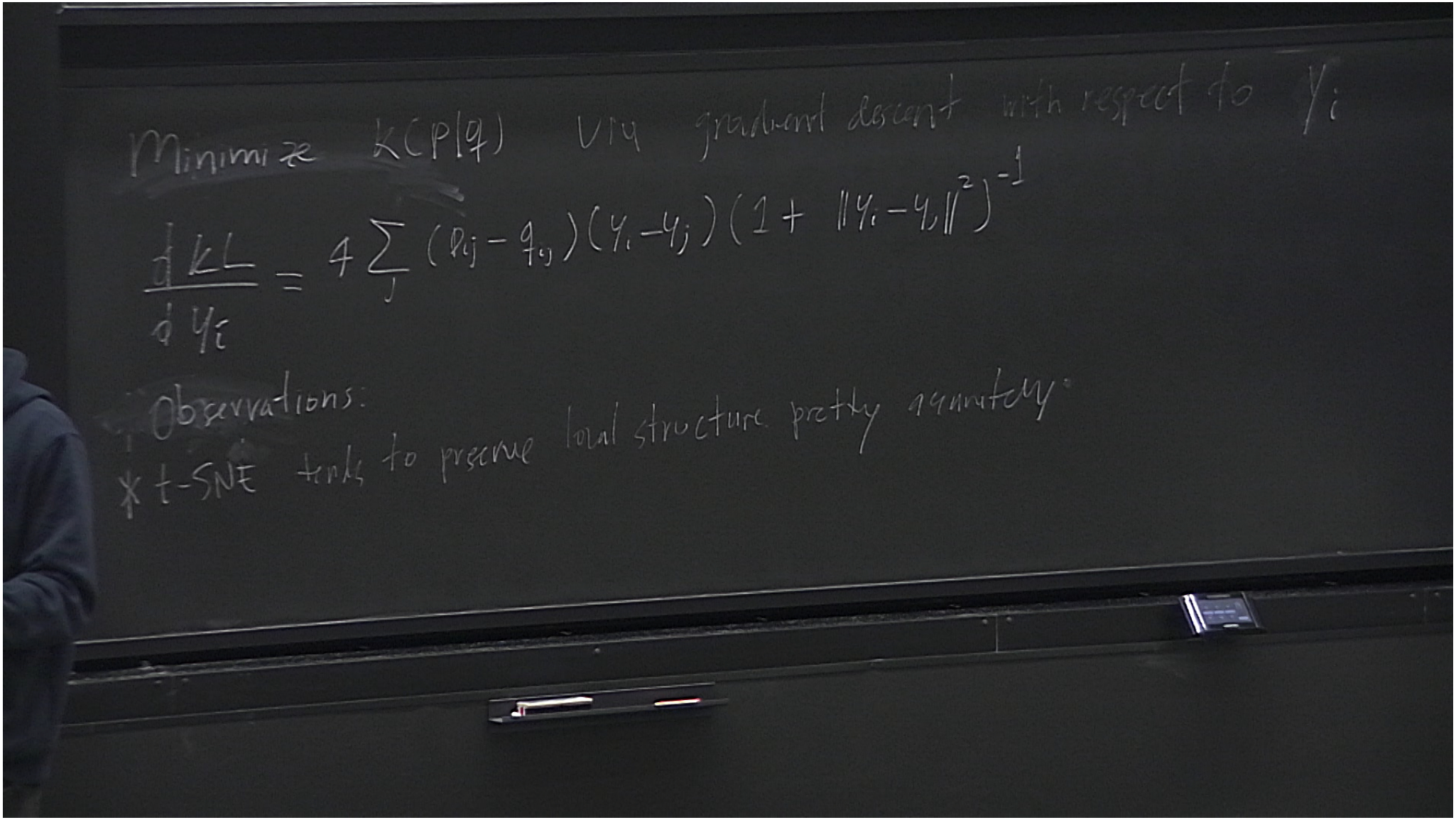

Jensens inequality
$\varphi\left(\sum_i p_i x_i\right) \geq \sum_i p_i \varphi(x_i)$
φ concave
p_{ij}
$-KL(p|q) = \sum_{ij} p_{ij} \log \frac{q_{ij}}{p_{ij}} \leq \log \sum_{ij} p_{ij} \frac{q_{ij}}{p_{ij}} = 0$
$-KL \leq 0$
$KL \geq 0$

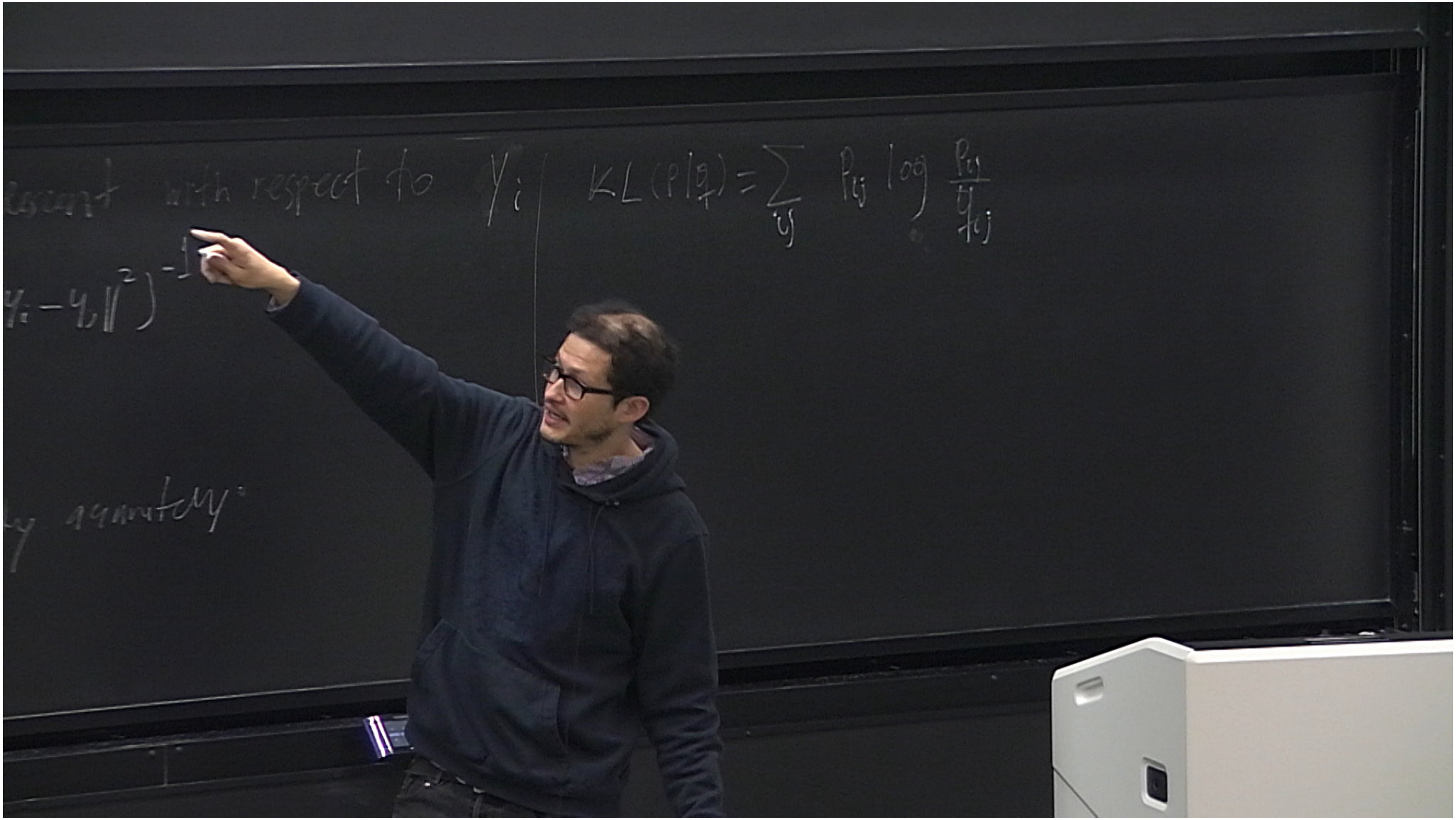

P_{ij} fixed and given by the dataset
We want to train Y_i to get q_{ij} as close as possible to P_{ij}
ck-Leibler divergence (divergence between P and q)
$(P|q) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{q_{ij}}$
if $P=q$ $\forall$ ij $KL=0$

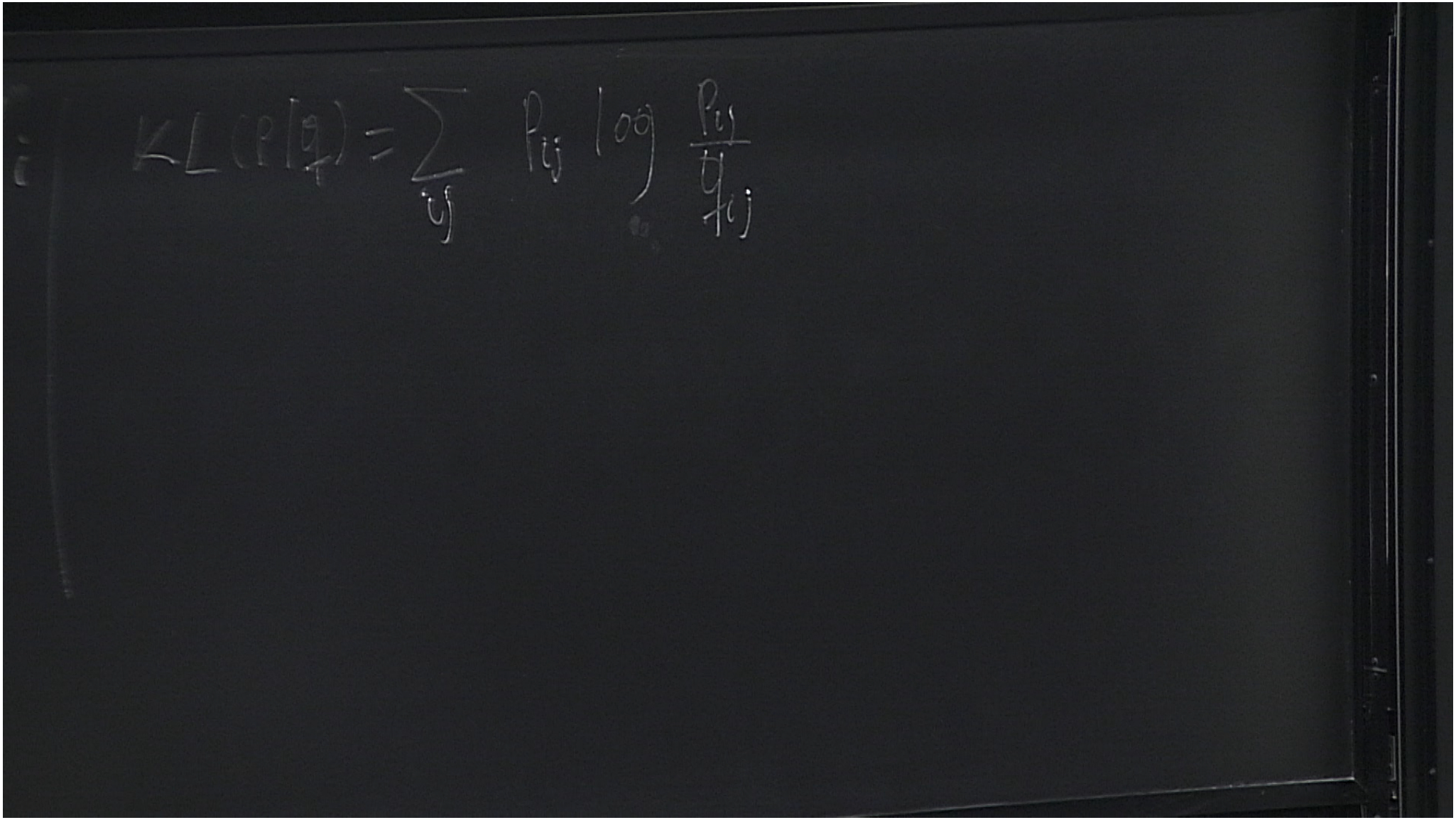

KL (P17)

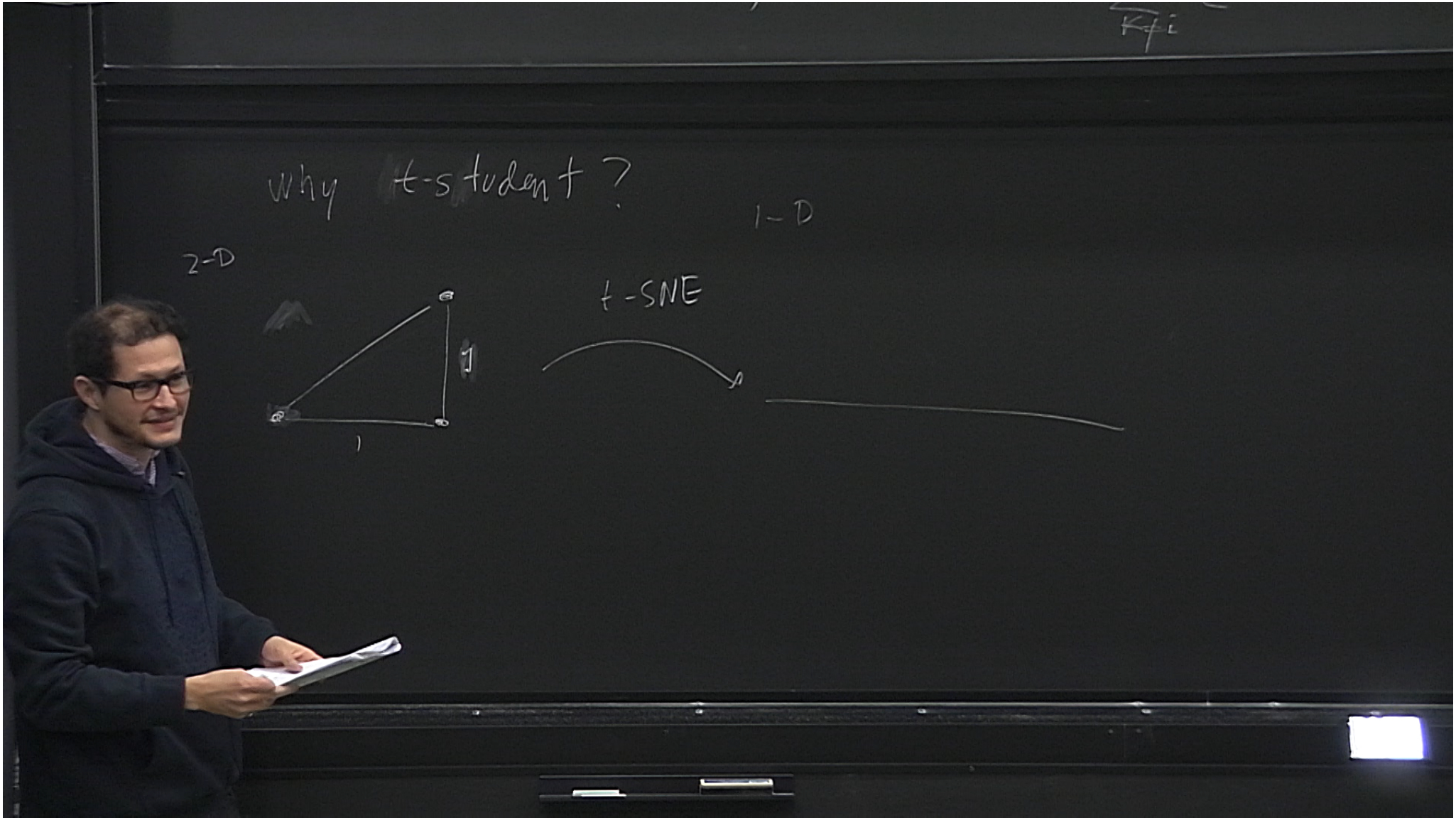

Minimize K(P|q) via gradient descent with respect to Yi

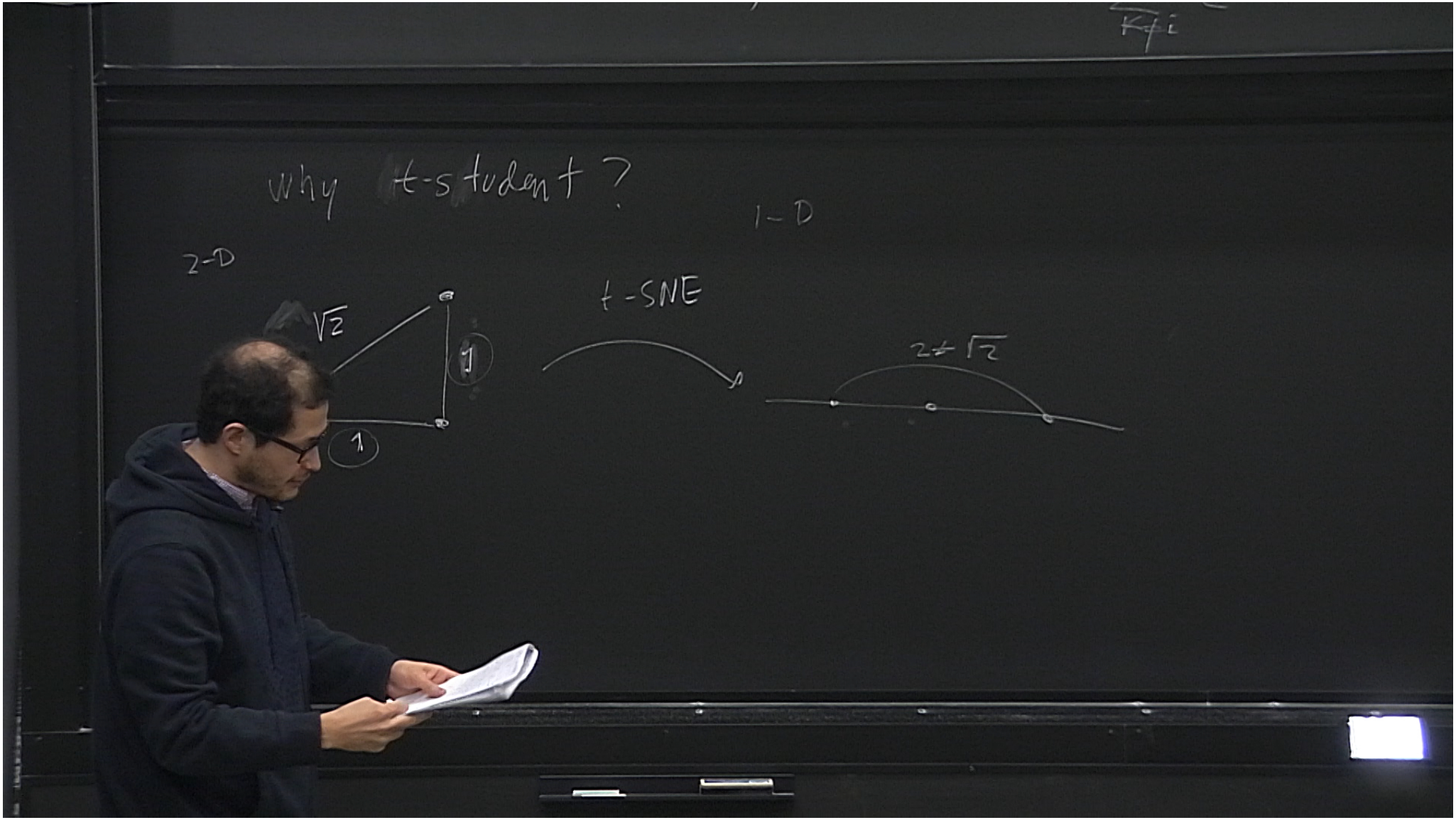

Minimize KC(P|Q) via gradient descent with respect to y_i
$\frac{\partial KL}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$

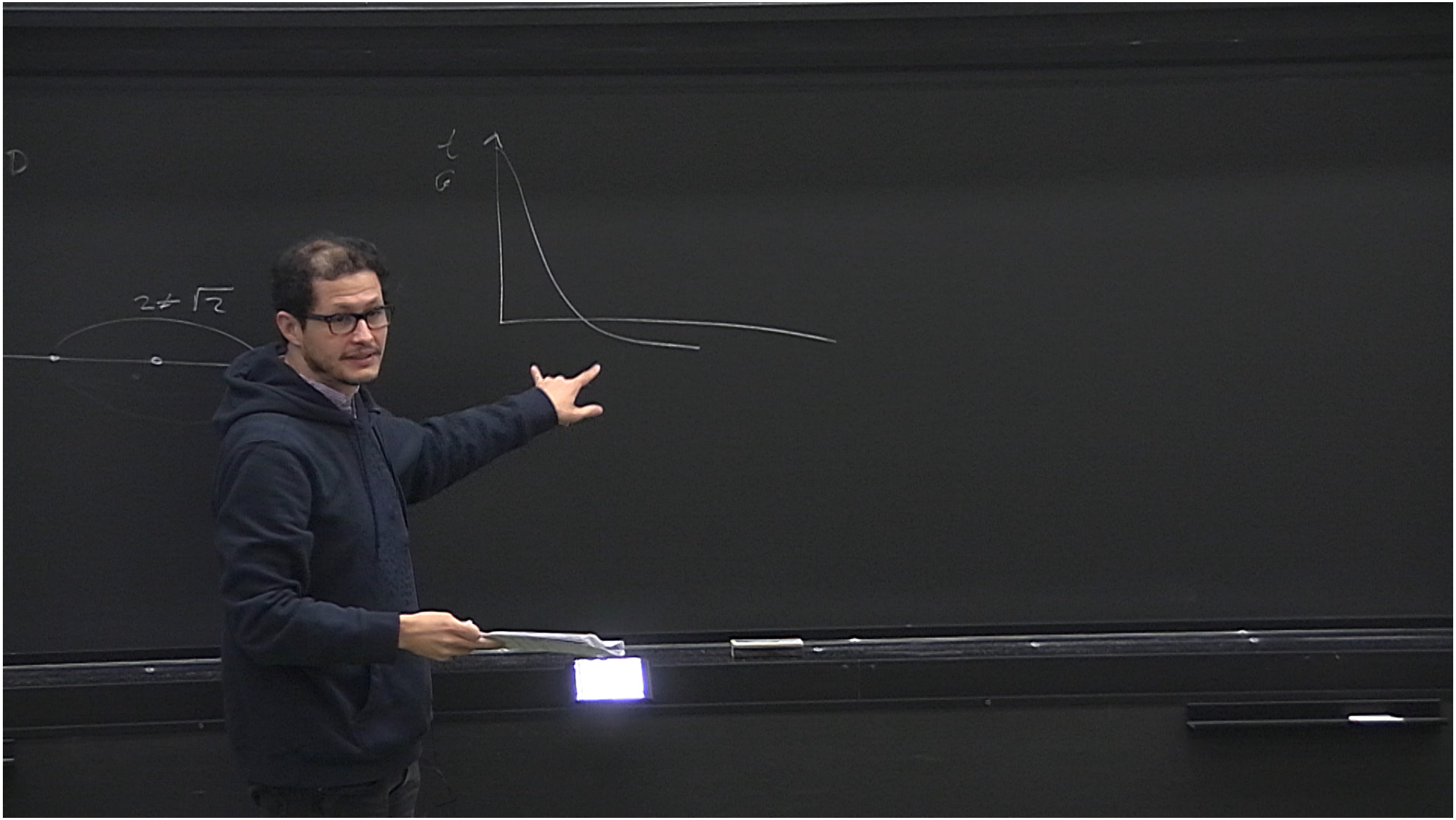

Minimize $KL(P|Q)$ via gradient descent with respect to y_i
$\frac{\partial KL}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$
Observations:
* t-SNE tends to preserve local structure pretty nicely.

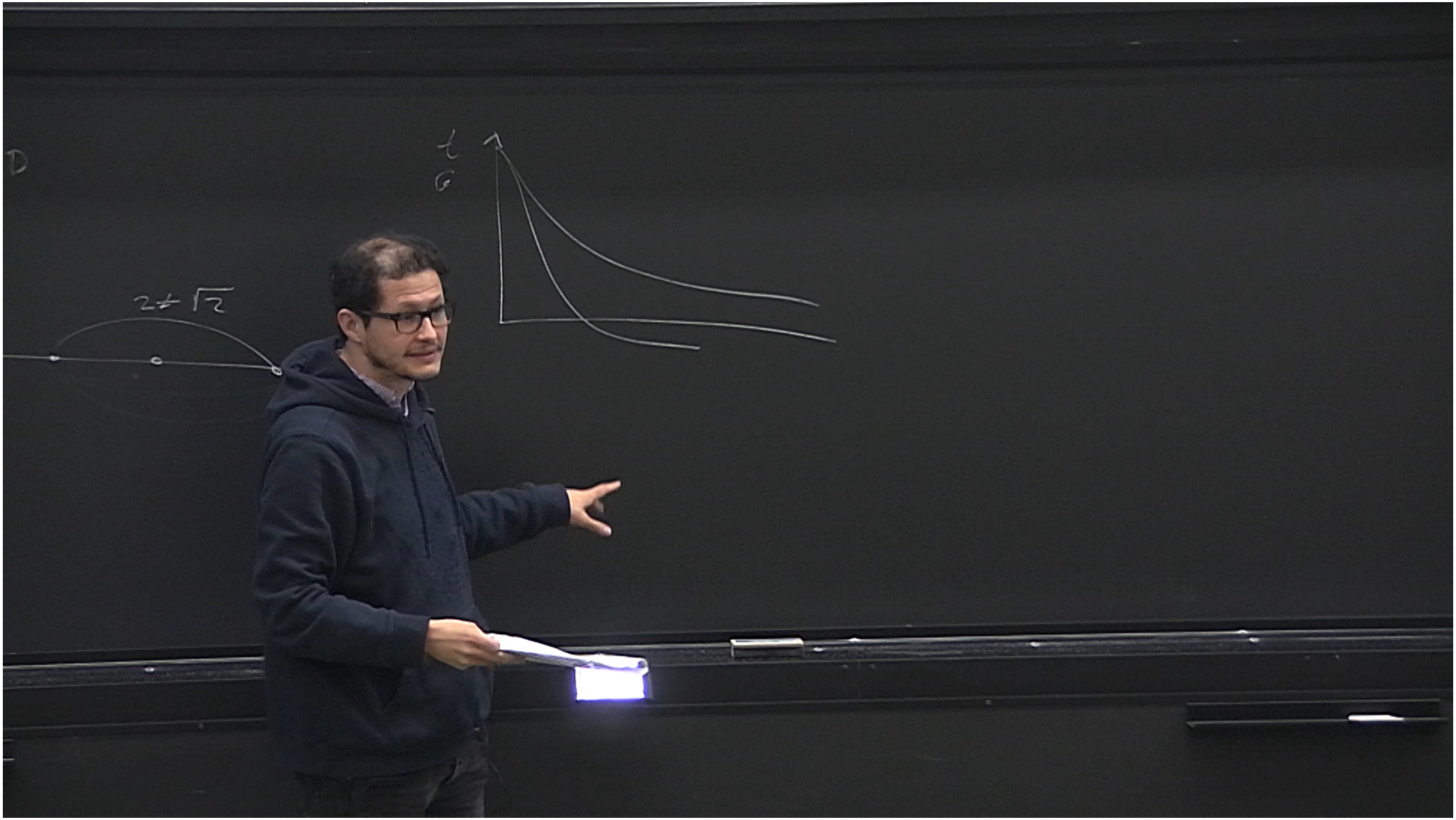

with respect to Y_i
$KL(P\|Q) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{Q_{ij}}$
$\|y_i - y_j\|^2)^{-1}$
by symmetry

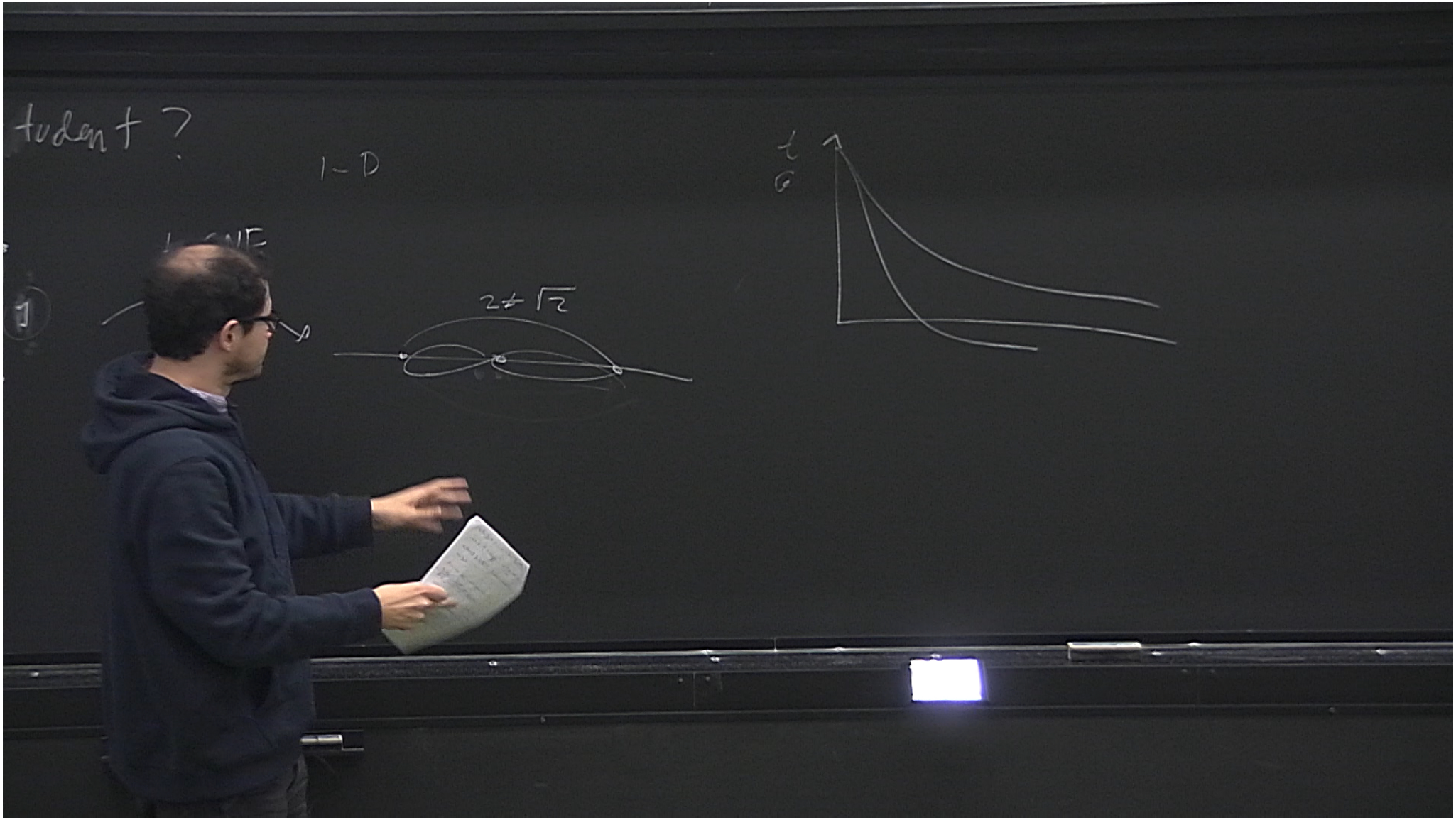

$KL(P\|Q) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{Q_{ij}}$

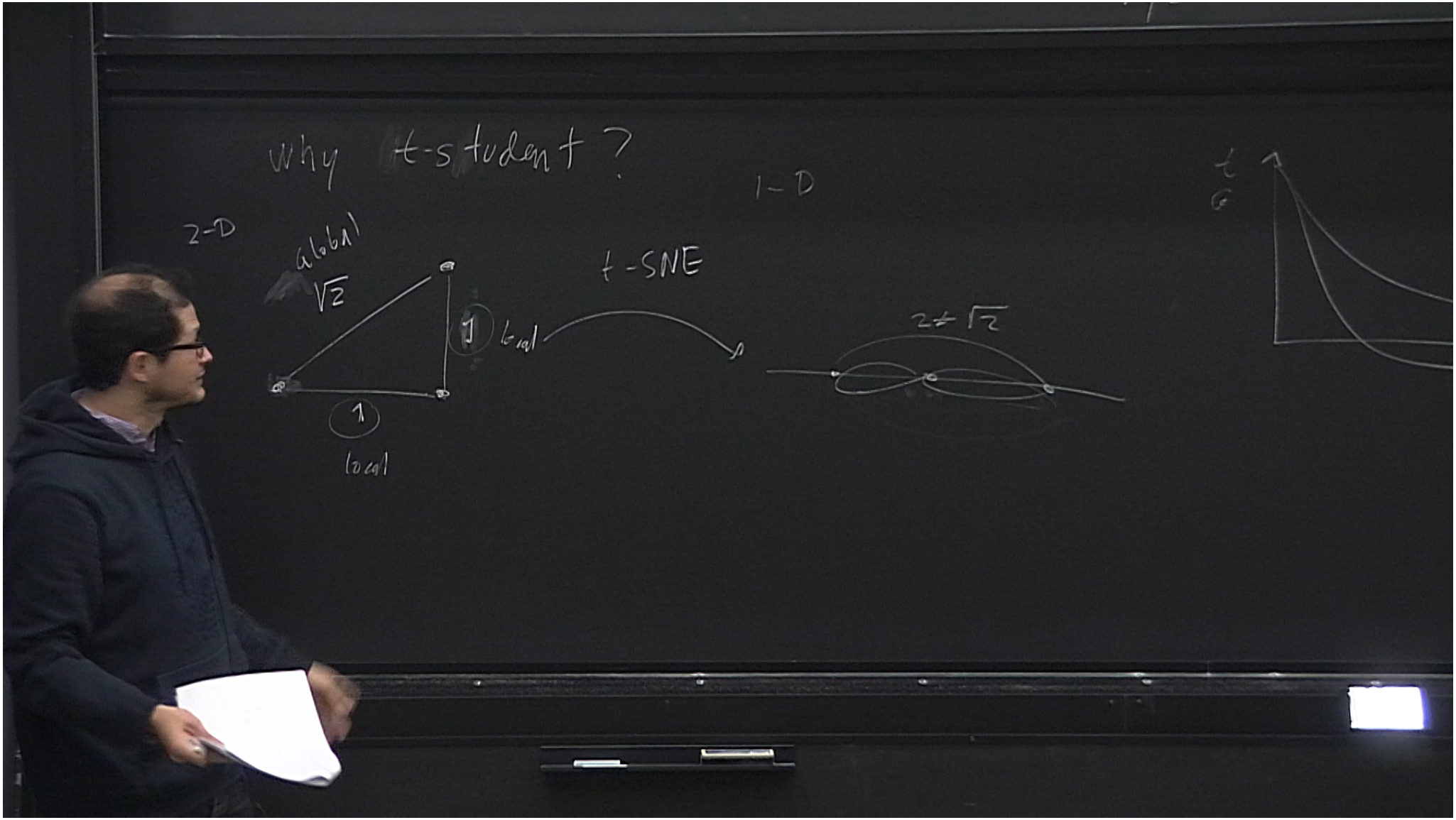

Why t-student?
2-D
1-D
t-SNE
1
1

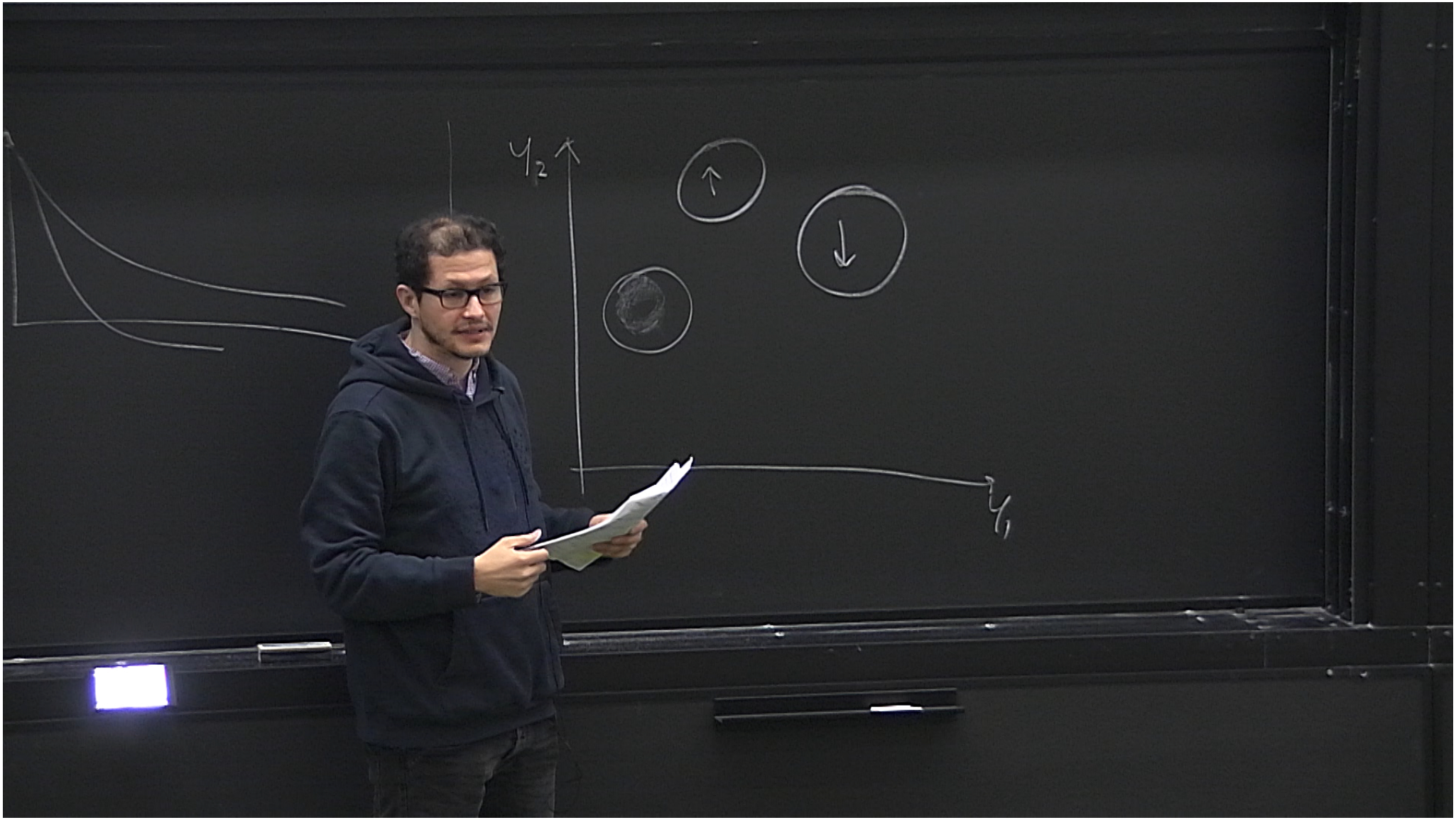

Why t-student ?
2-D
√2
1
1
t-SNE
1-D
2≠√2

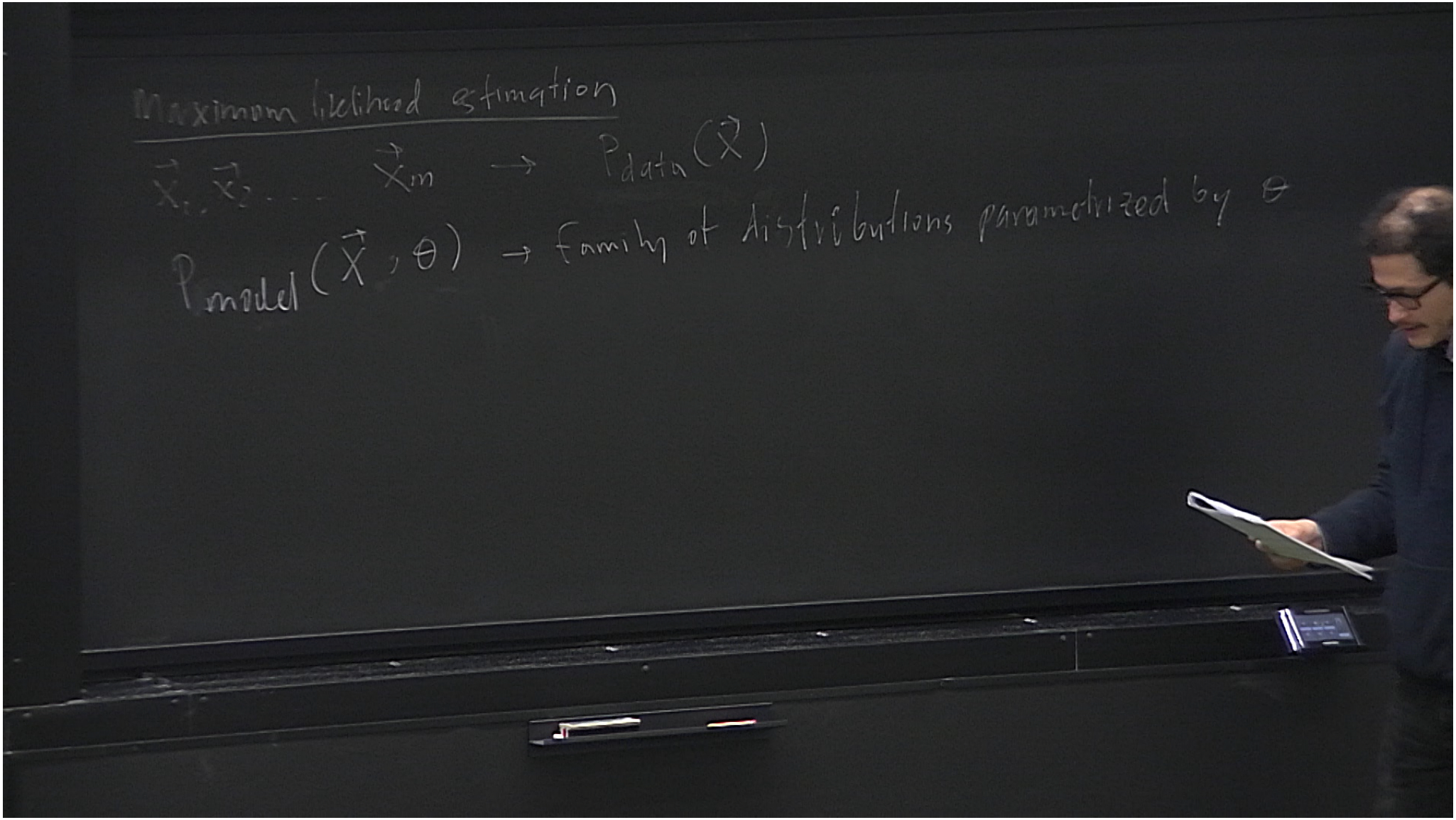

2+√2

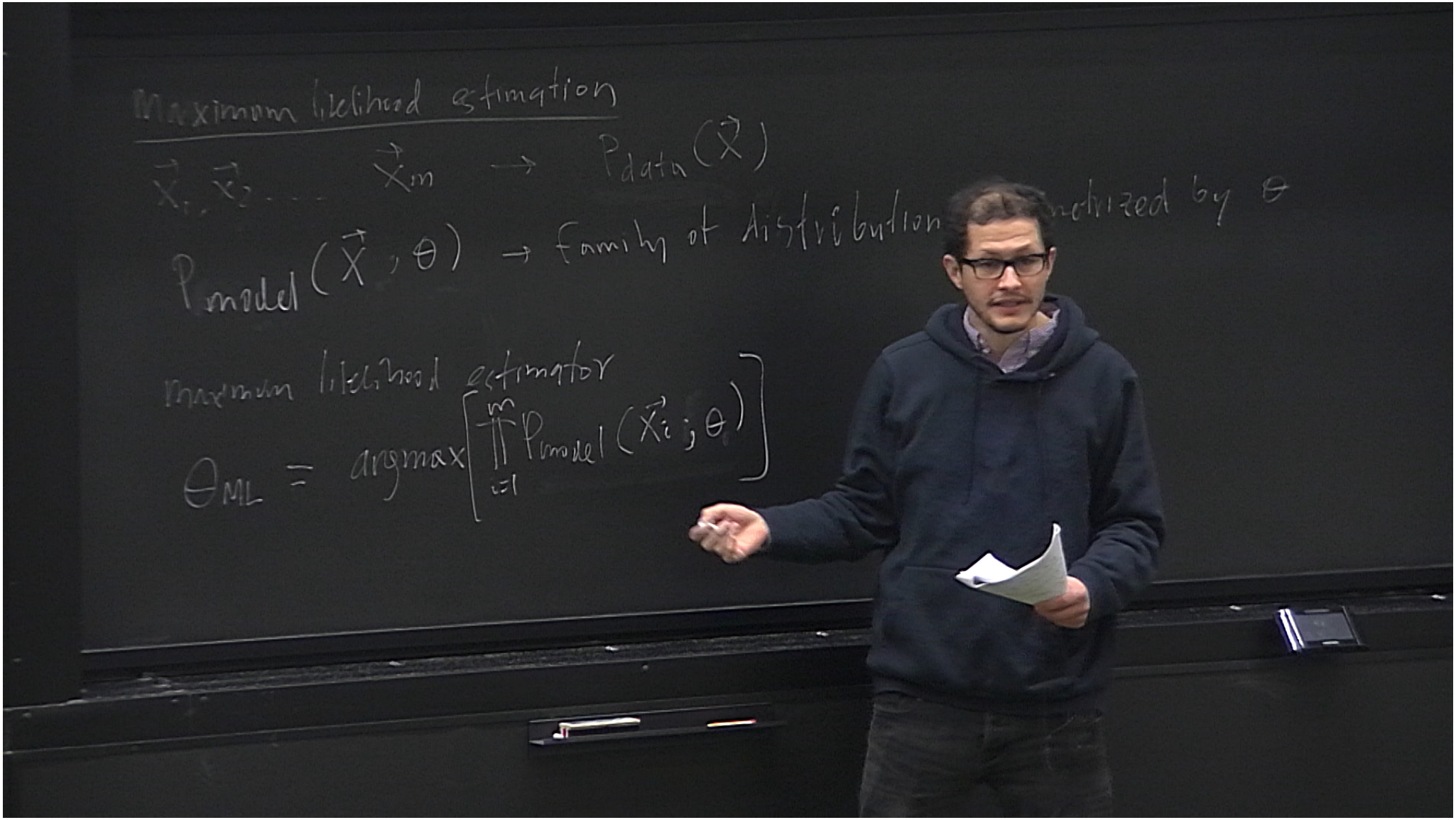

t
σ
2+√2

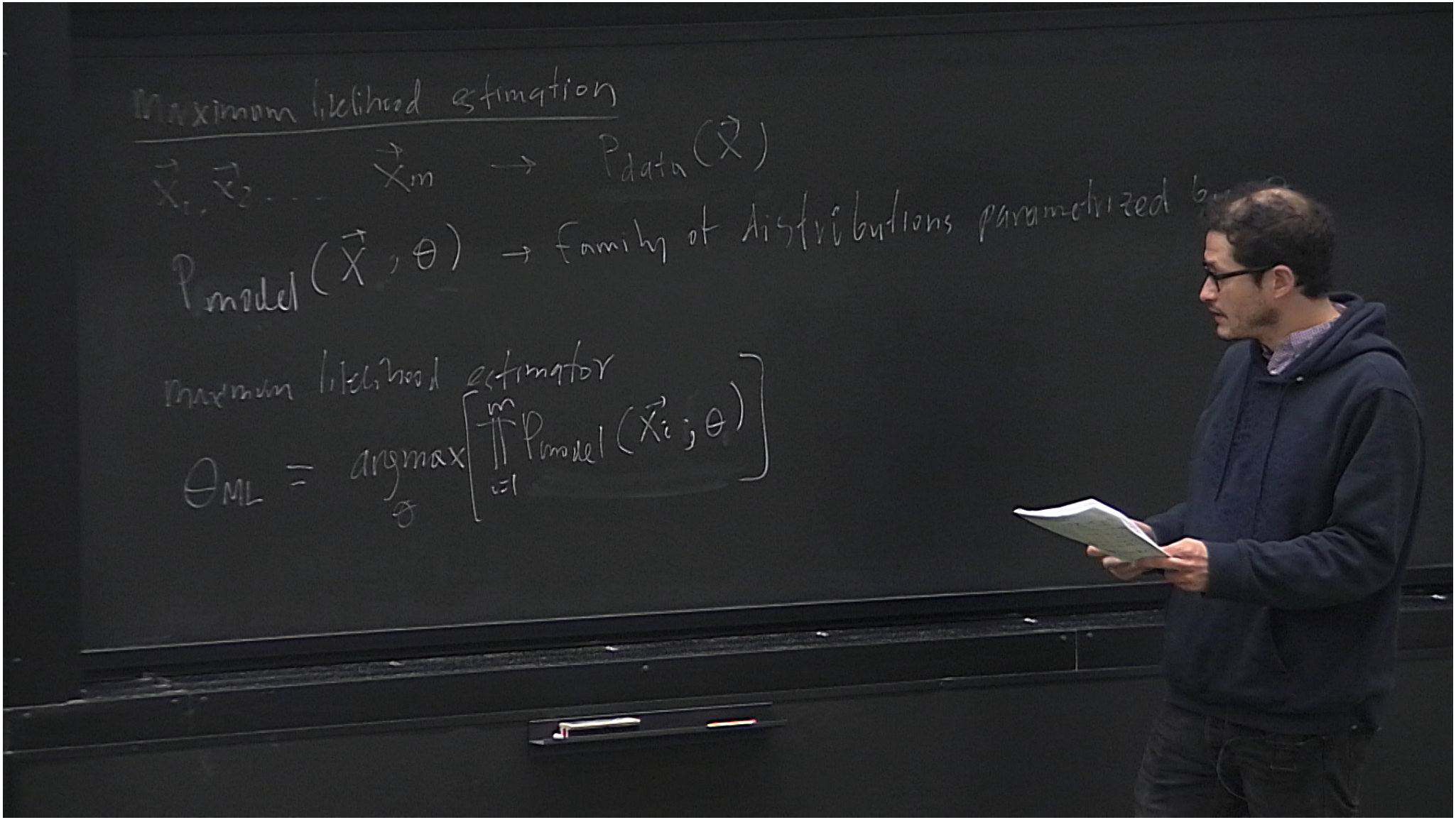

tudent?
1-D
2+√2
t
σ

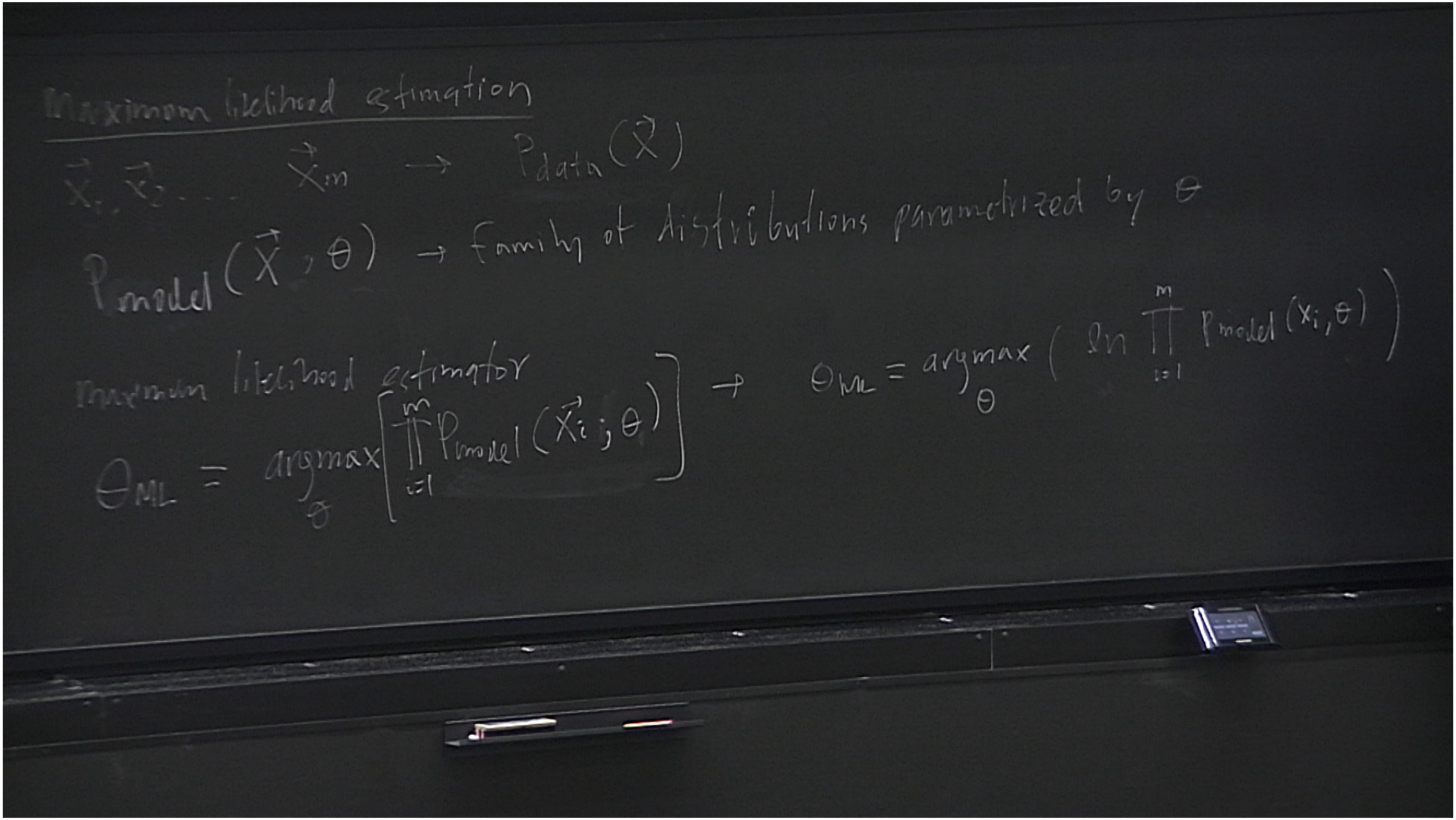

Why t-student ?
2-D
global
√2
1
local
1
local
t-SNE
1-D
2≠√2
t
G
1

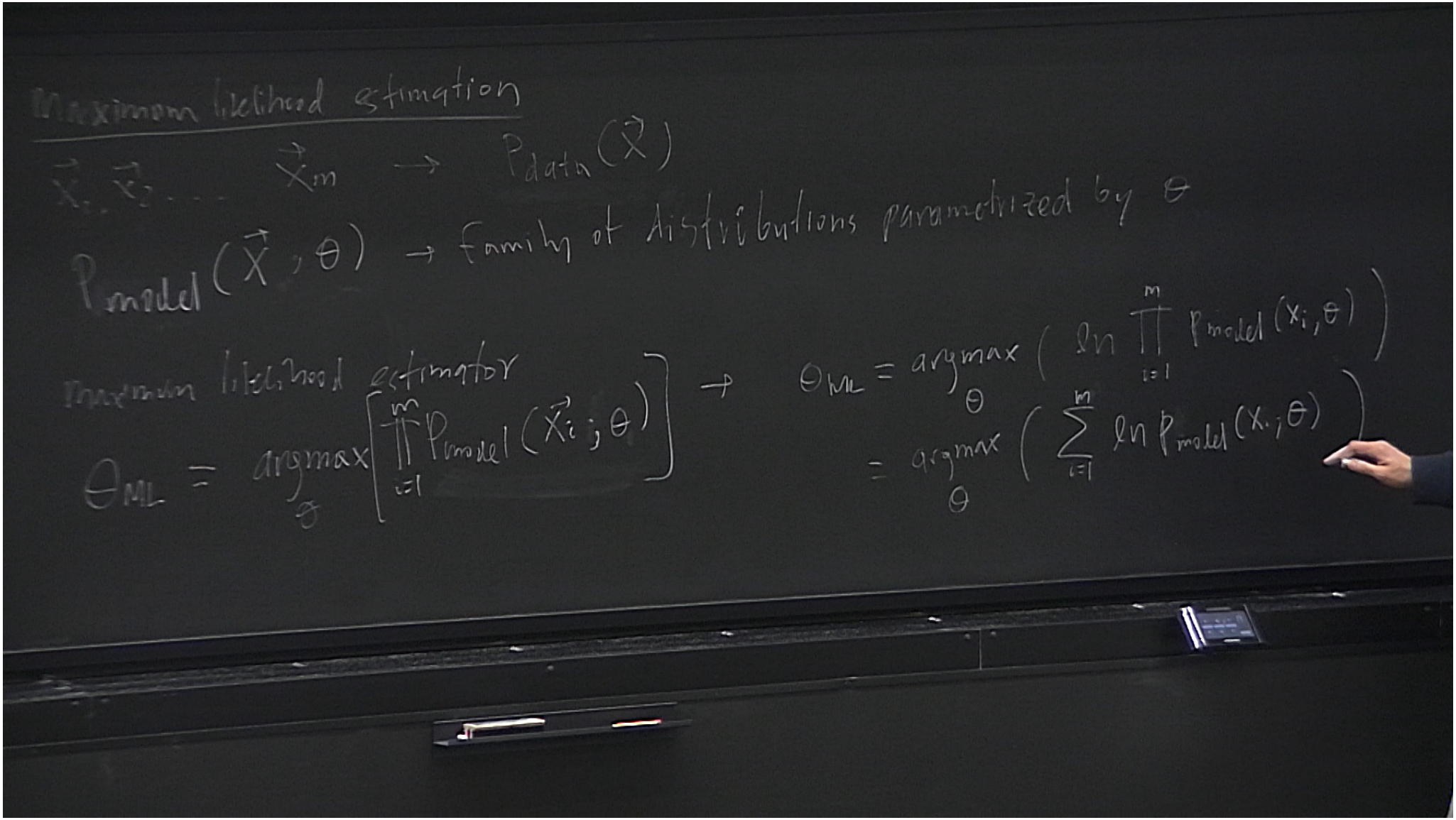

y_2
y_1

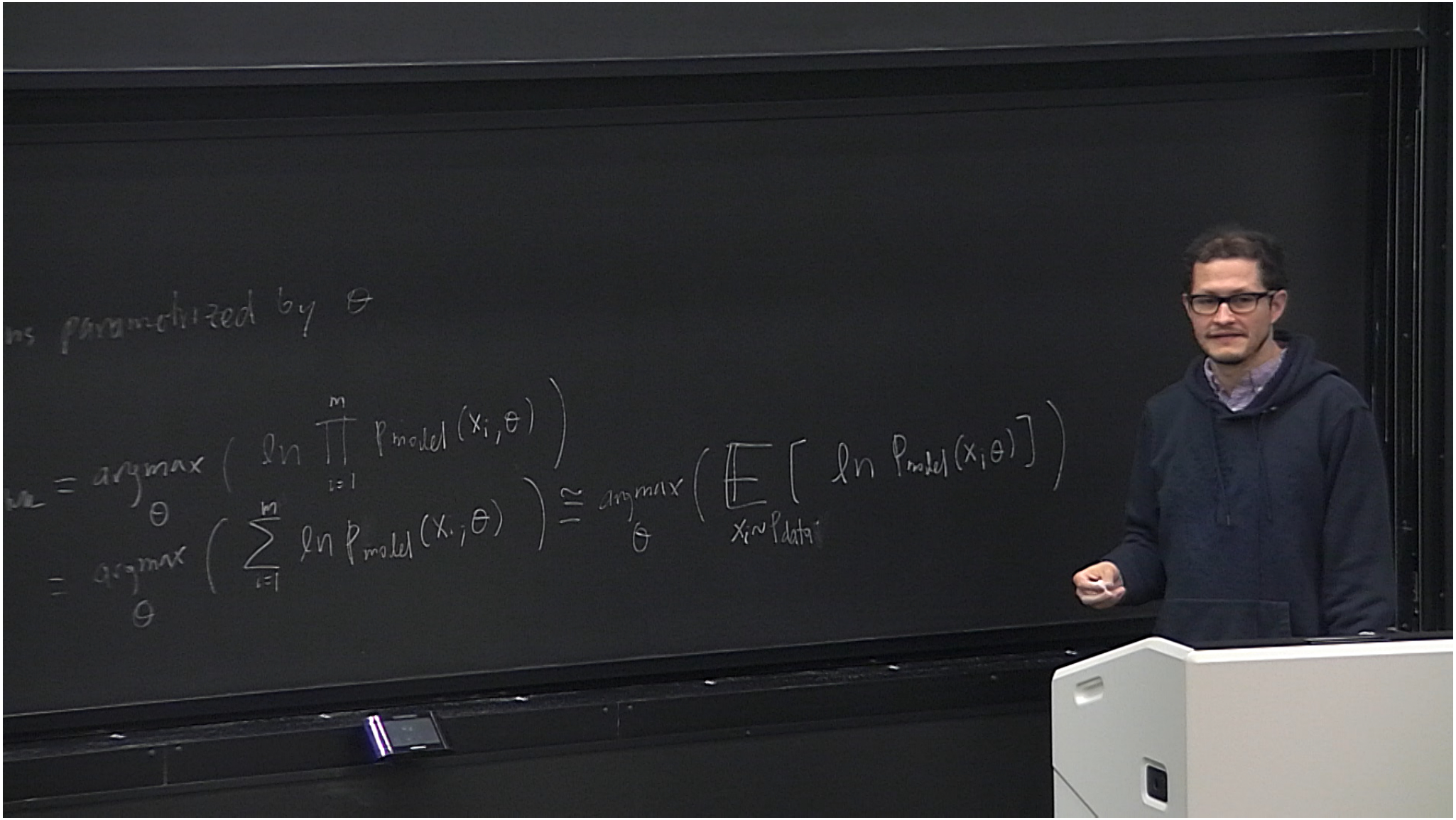

Maximum likelihood estimation
$\vec{x}_1, \vec{x}_2, \ldots \vec{x}_m \rightarrow P_{data}(\vec{x})$
$P_{model}(\vec{x}; \theta) \rightarrow$ family of distributions parametrized by θ

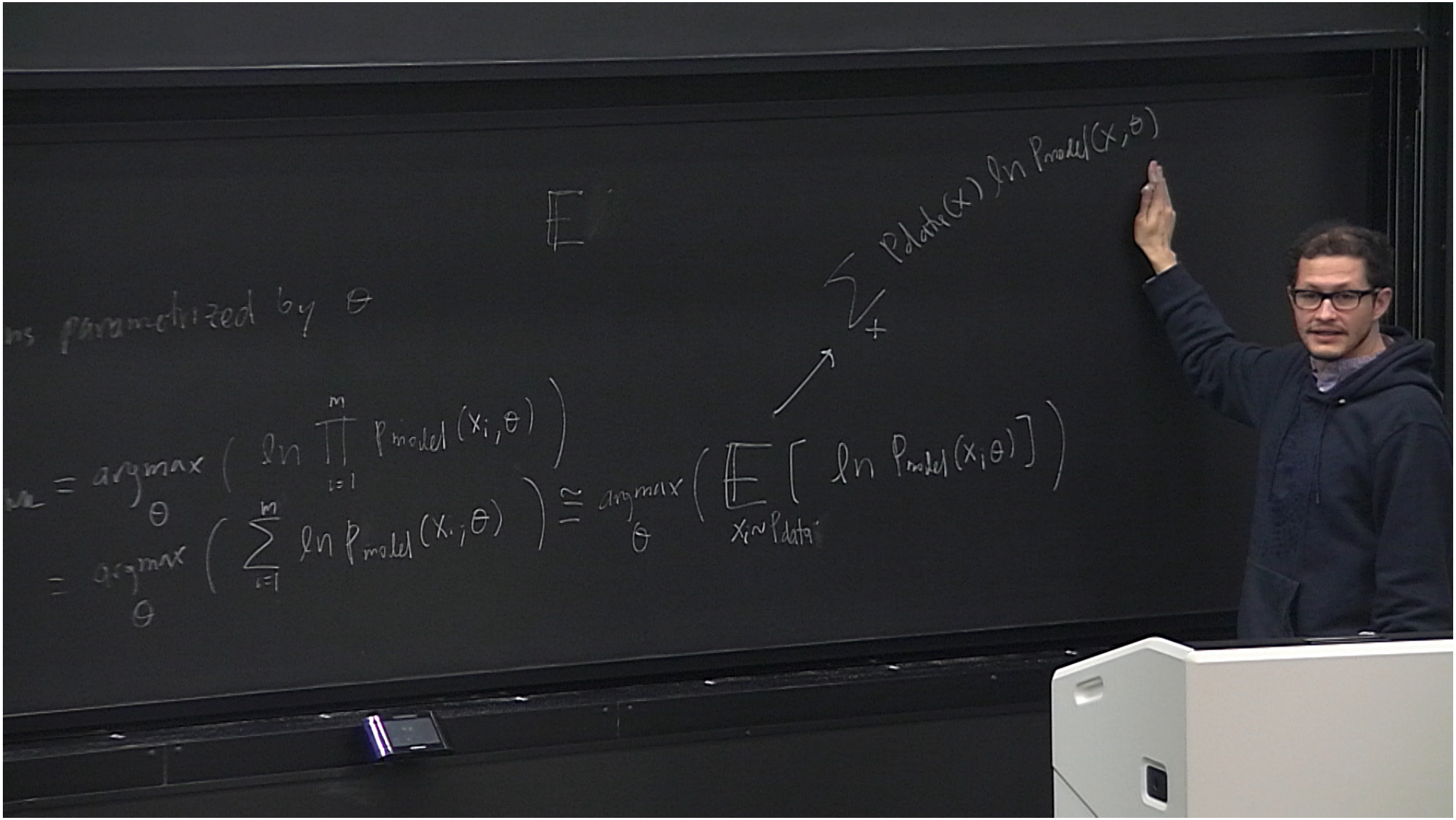

Maximum likelihood estimation
$\vec{x}_1, \vec{x}_2, \ldots \vec{x}_m \rightarrow p_{data}(\vec{x})$
$p_{model}(\vec{x}; \theta) \rightarrow$ family of distribution
parametrized by θ
maximum likelihood estimator
$\theta_{ML} = \operatorname{argmax}\left[\prod_{i=1}^{m} p_{model}(\vec{x}_i; \theta)\right]$

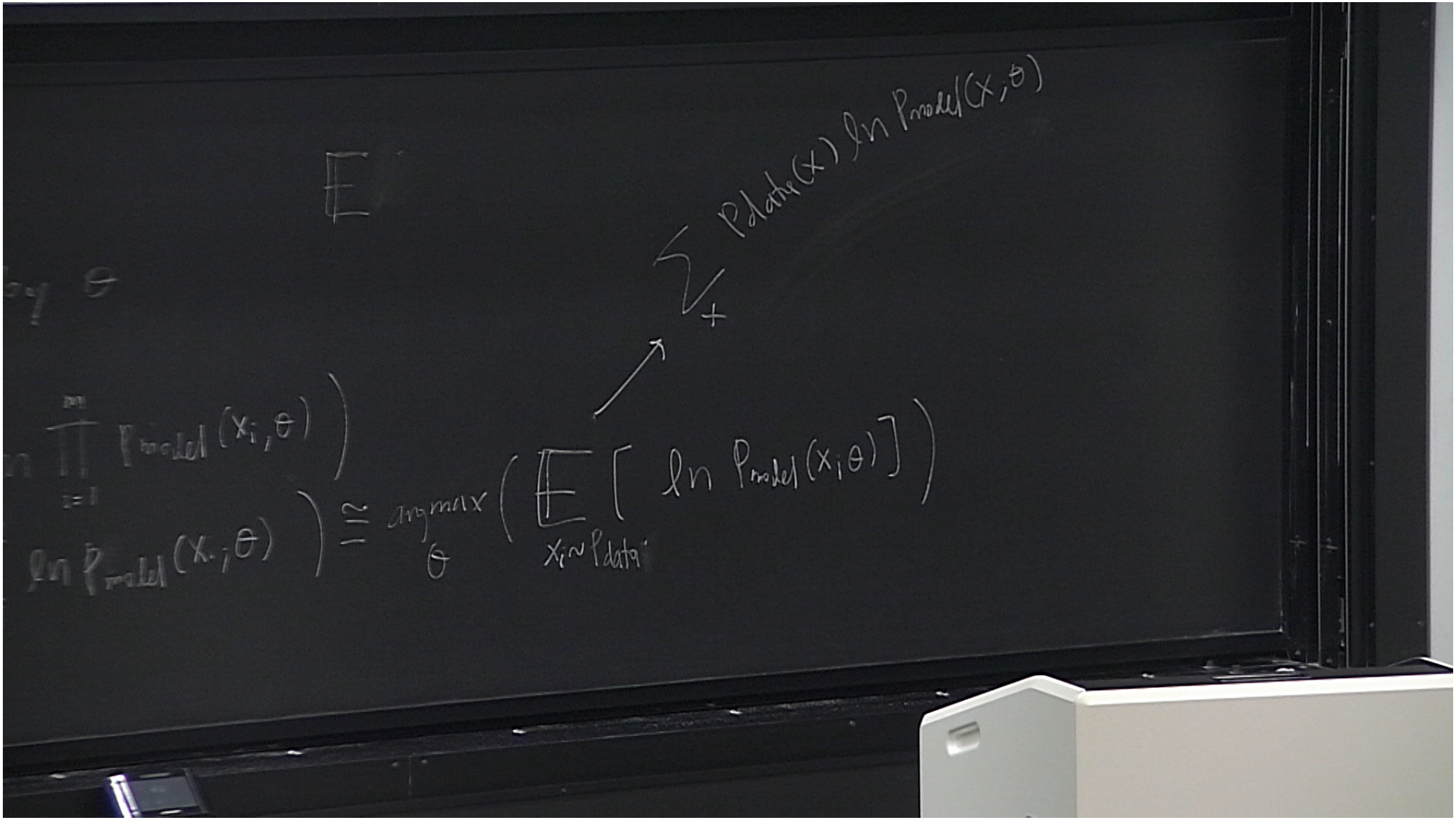

Maximum likelihood estimation
$\vec{x}_1, \vec{x}_2, \ldots \vec{x}_m \rightarrow p_{data}(\vec{x})$
$p_{model}(\vec{x}; \theta) \rightarrow$ family of distributions parametrized by
maximum likelihood estimator
$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{m} p_{model}(\vec{x}_i; \theta) \right]$

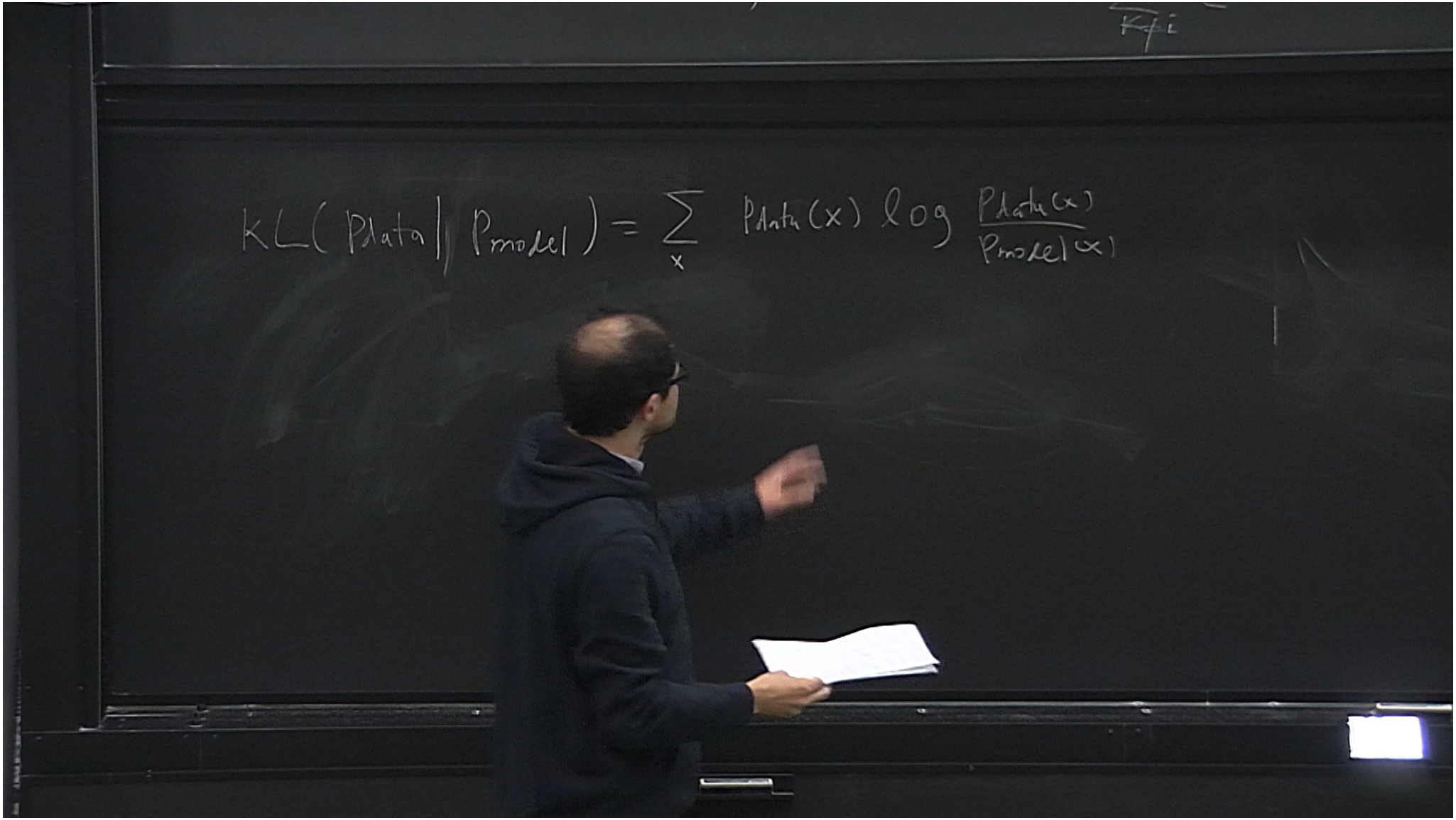

Maximum likelihood estimation
$\vec{x}_1, \vec{x}_2, \ldots \vec{x}_m \rightarrow p_{data}(\vec{x})$
$p_{model}(\vec{x}; \theta)$ → family of distributions parametrized by θ
Maximum likelihood estimator
$\theta_{ML} = \arg\max_{\theta} \left[\prod_{i=1}^{m} p_{model}(\vec{x}_i; \theta) \right]$ → $\theta_{ML} = \arg\max_{\theta} \left(\ln \prod_{i=1}^{m} p_{model}(x_i, \theta) \right)$

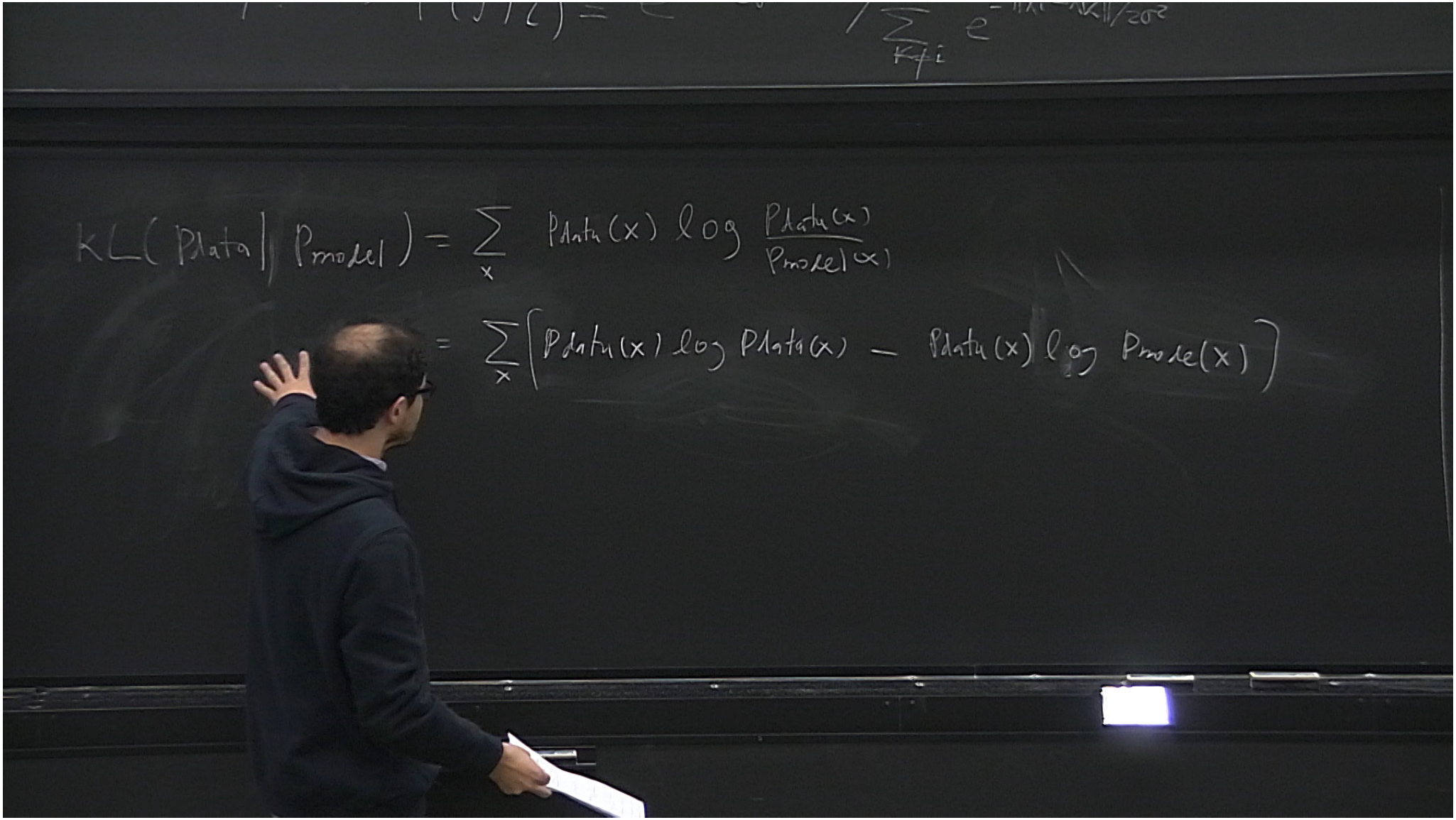

Maximum likelihood estimation
$\vec{X}_1, \vec{X}_2, \ldots \vec{X}_m \rightarrow P_{data}(\vec{X})$
$P_{model}(\vec{X}; \theta) \rightarrow$ family of distributions parametrized by θ
Maximum likelihood estimator
$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{m} P_{model}(\vec{X}_i; \theta) \right] \rightarrow \theta_{ML} = \underset{\theta}{\operatorname{argmax}} \left(\ln \prod_{i=1}^{m} P_{model}(X_i; \theta) \right)$
$= \underset{\theta}{\operatorname{argmax}} \left(\sum_{i=1}^{m} \ln P_{model}(X_i; \theta) \right)$

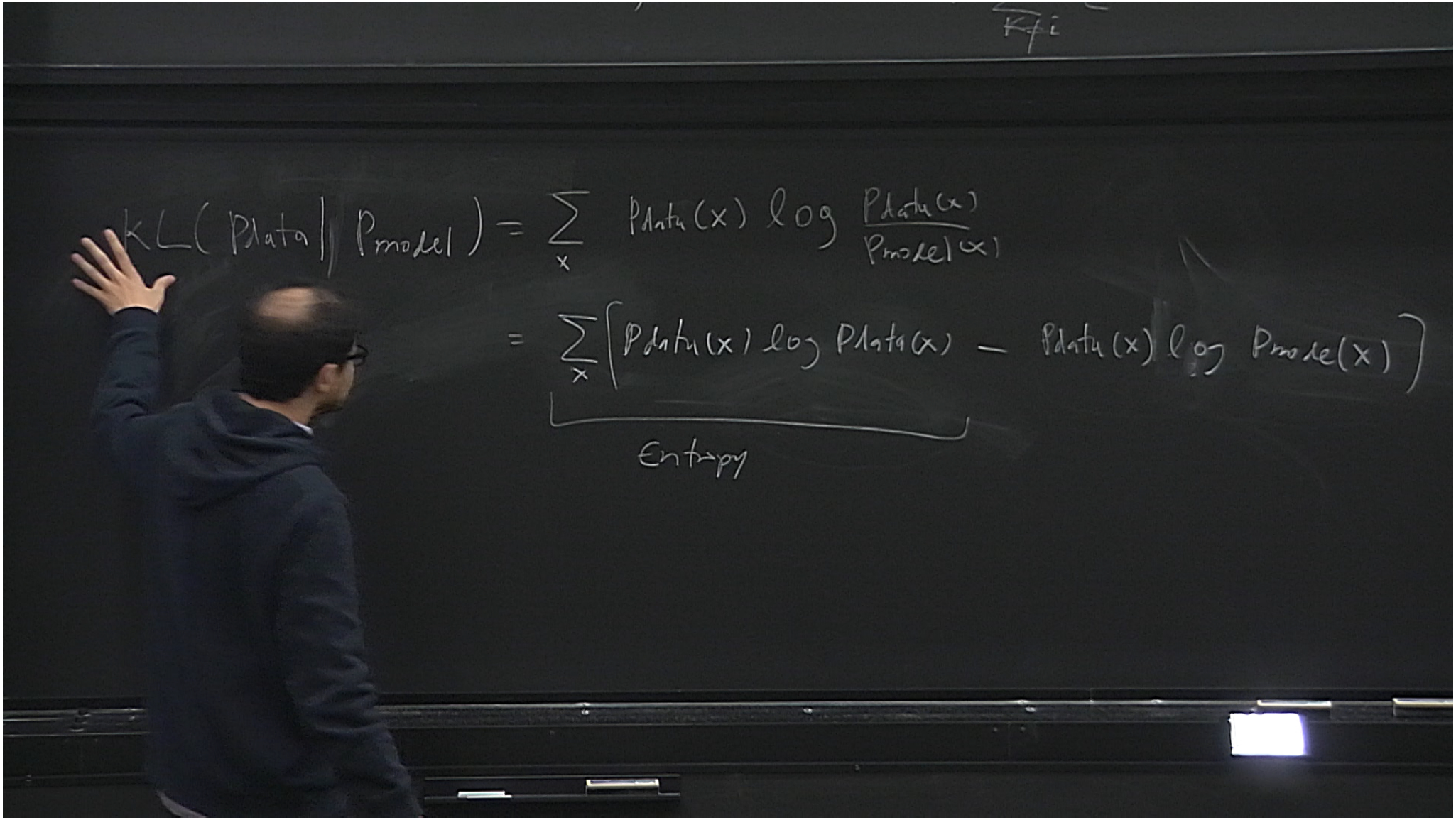

ns parametrized by θ
= argmax_θ (ln ∏_{i=1}^{m} P_model(x_i; θ))
= argmax_θ (Σ_{i=1}^{m} ln P_model(x_i; θ)) ≈ argmax_θ (E_{x_i ~ P_data} [ln P_model(x_i; θ)])

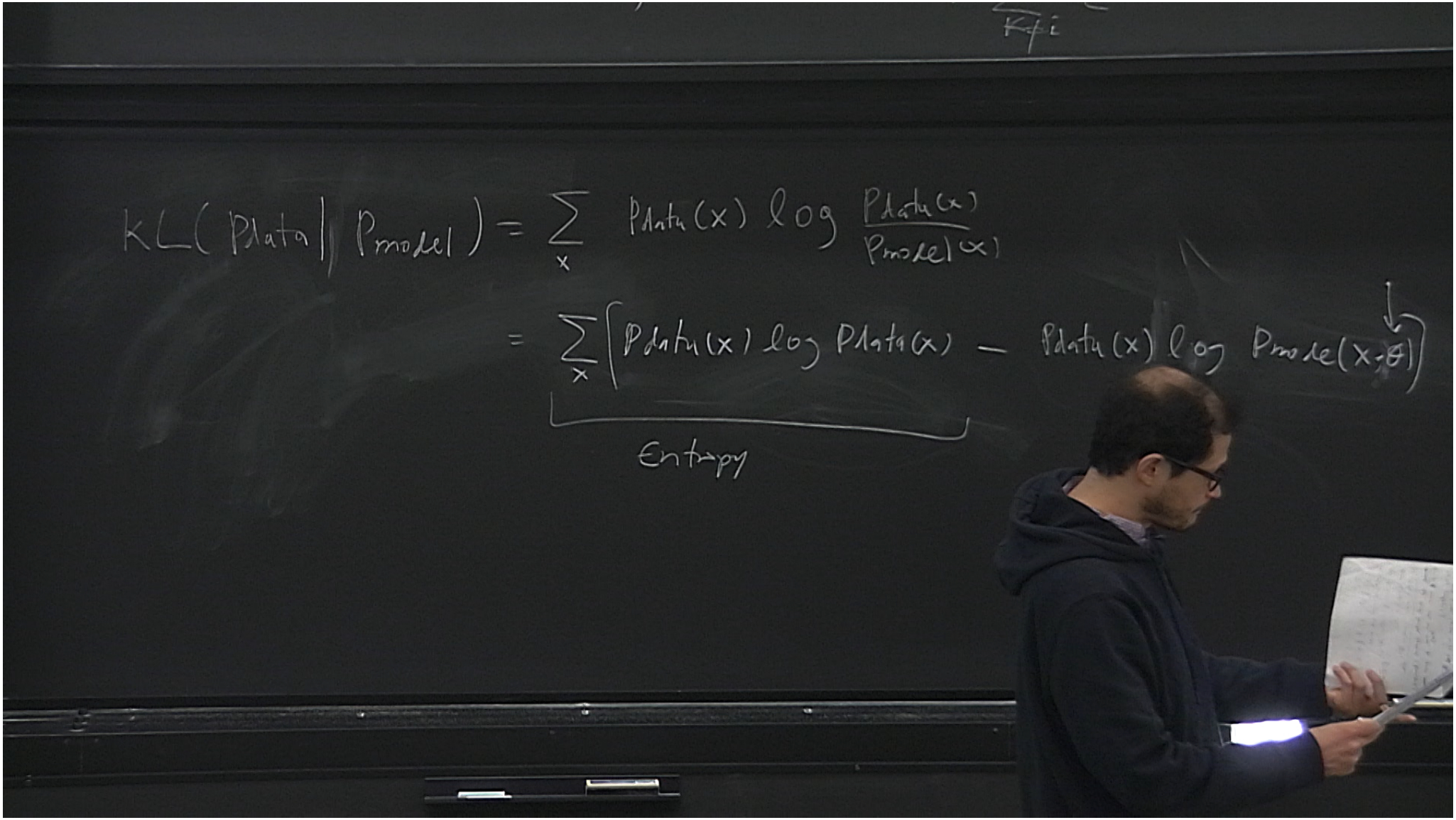

ns parametrized by θ
= argmax_θ (ln ∏_{i=1}^{m} P_model(x_i, θ))
= argmax_θ (Σ_{i=1}^{m} ln P_model(x_i, θ)) ≈ argmax_θ (E_{x_i ~ P_data} [ln P_model(x_i, θ)])
Σ_x P_data(x) ln P_model(x, θ)

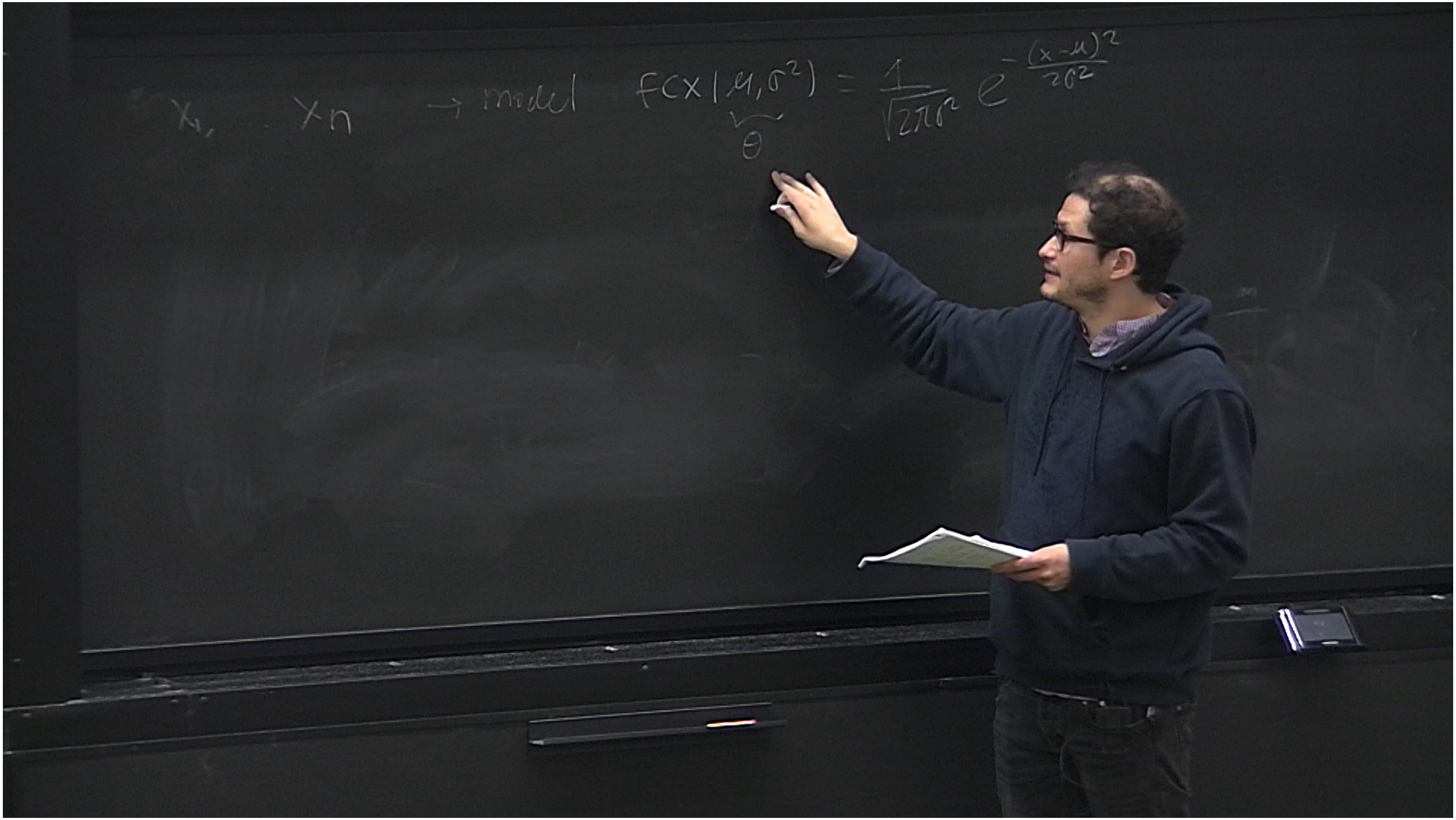

E
$\sum_x P_{data}(x) \ln P_{model}(x;\theta)$
$\prod_{i=1}^{m} P_{model}(x_i;\theta)$
$\ln P_{model}(x_i;\theta)$
$\cong \arg\max_\theta \left(\mathbb{E}_{x_i \sim P_{data}} [\ln P_{model}(x_i;\theta)] \right)$

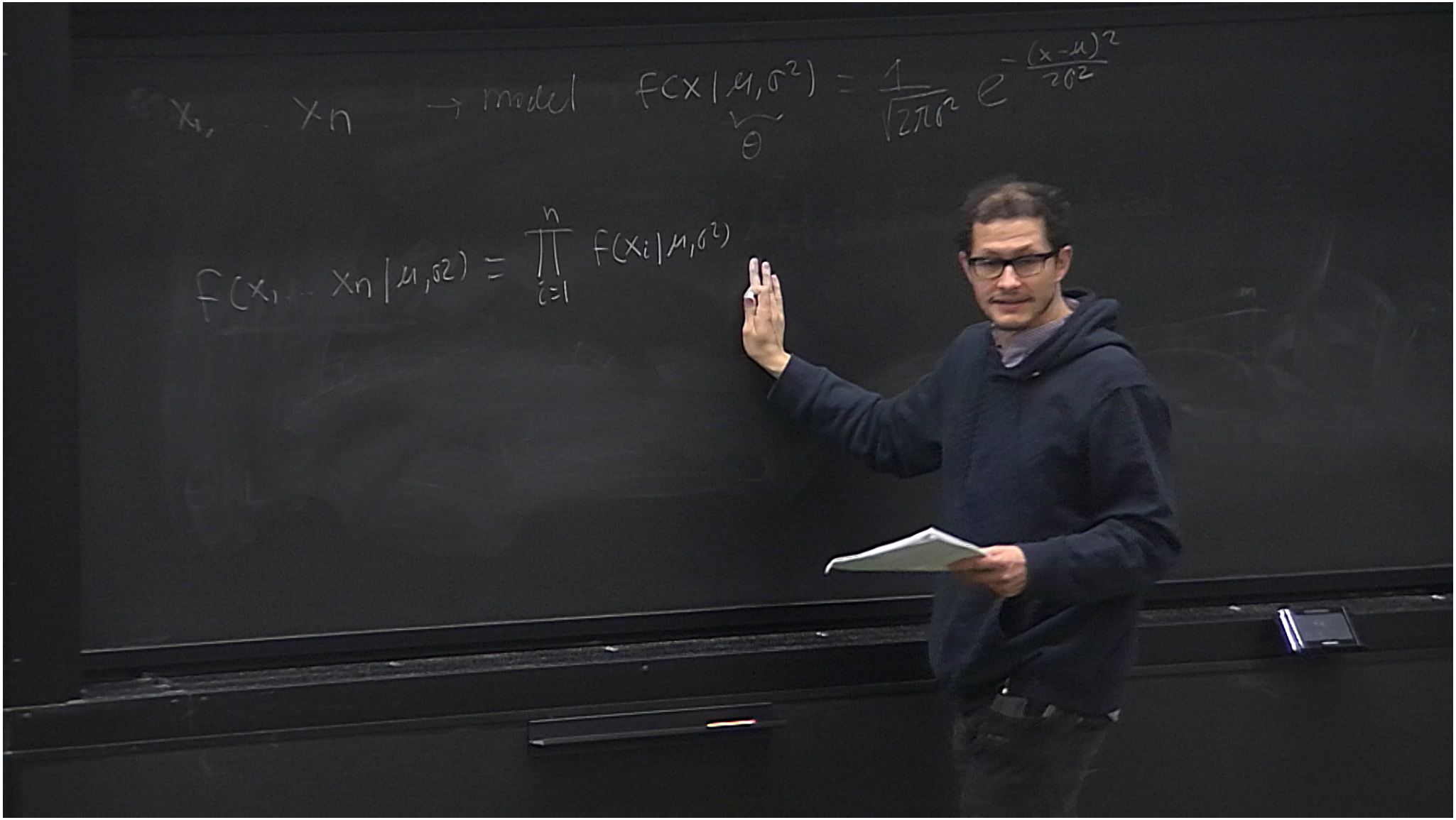

$KL(P_{data} \| P_{model}) = \sum_x P_{data}(x) \log \frac{P_{data}(x)}{P_{model}(x)}$

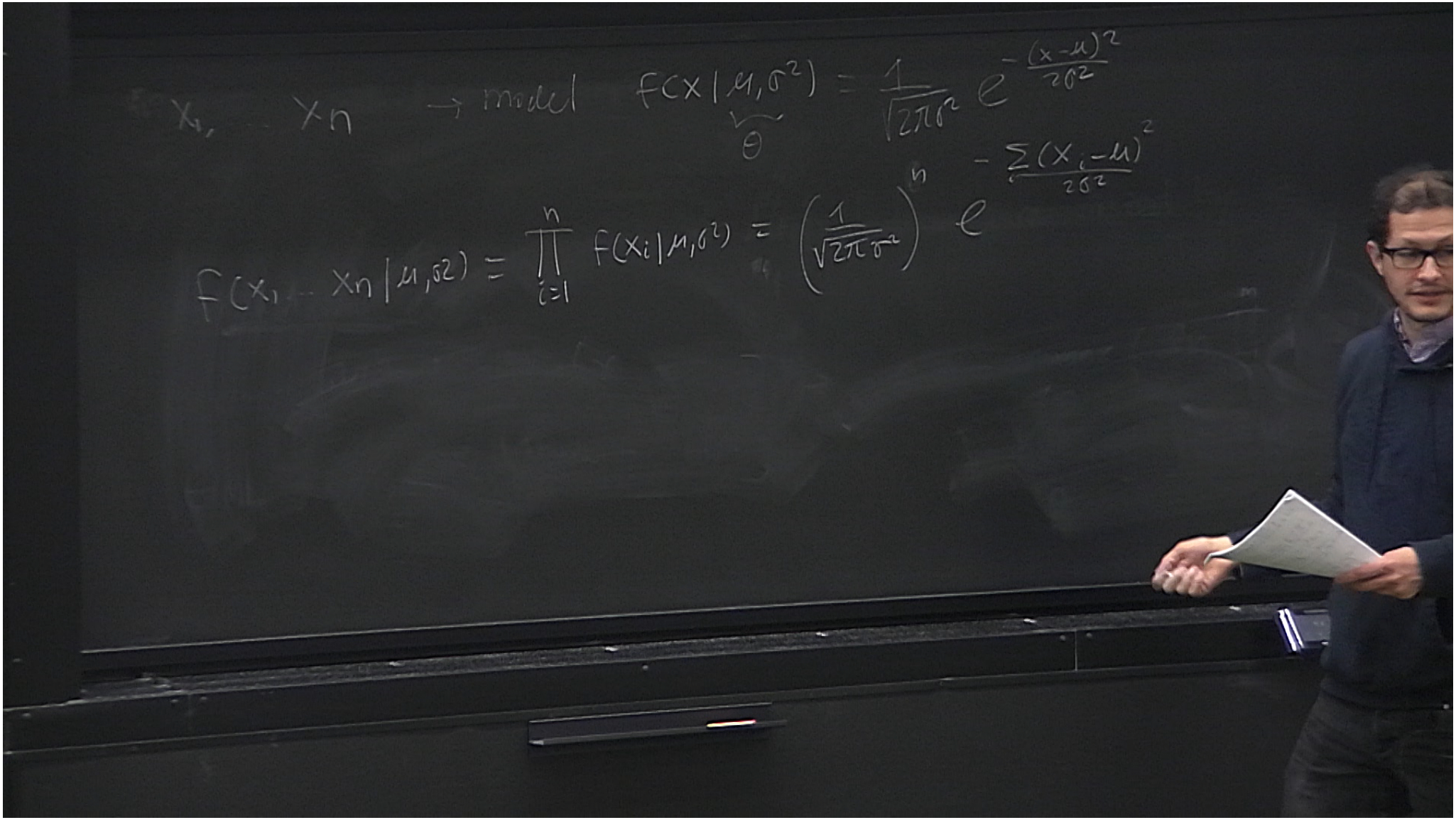

$KL(P_{data} \| P_{model}) = \sum_x P_{data}(x) \log \frac{P_{data}(x)}{P_{model}(x)}$
$= \sum_x \left(P_{data}(x) \log P_{data}(x) - P_{data}(x) \log P_{model}(x) \right)$

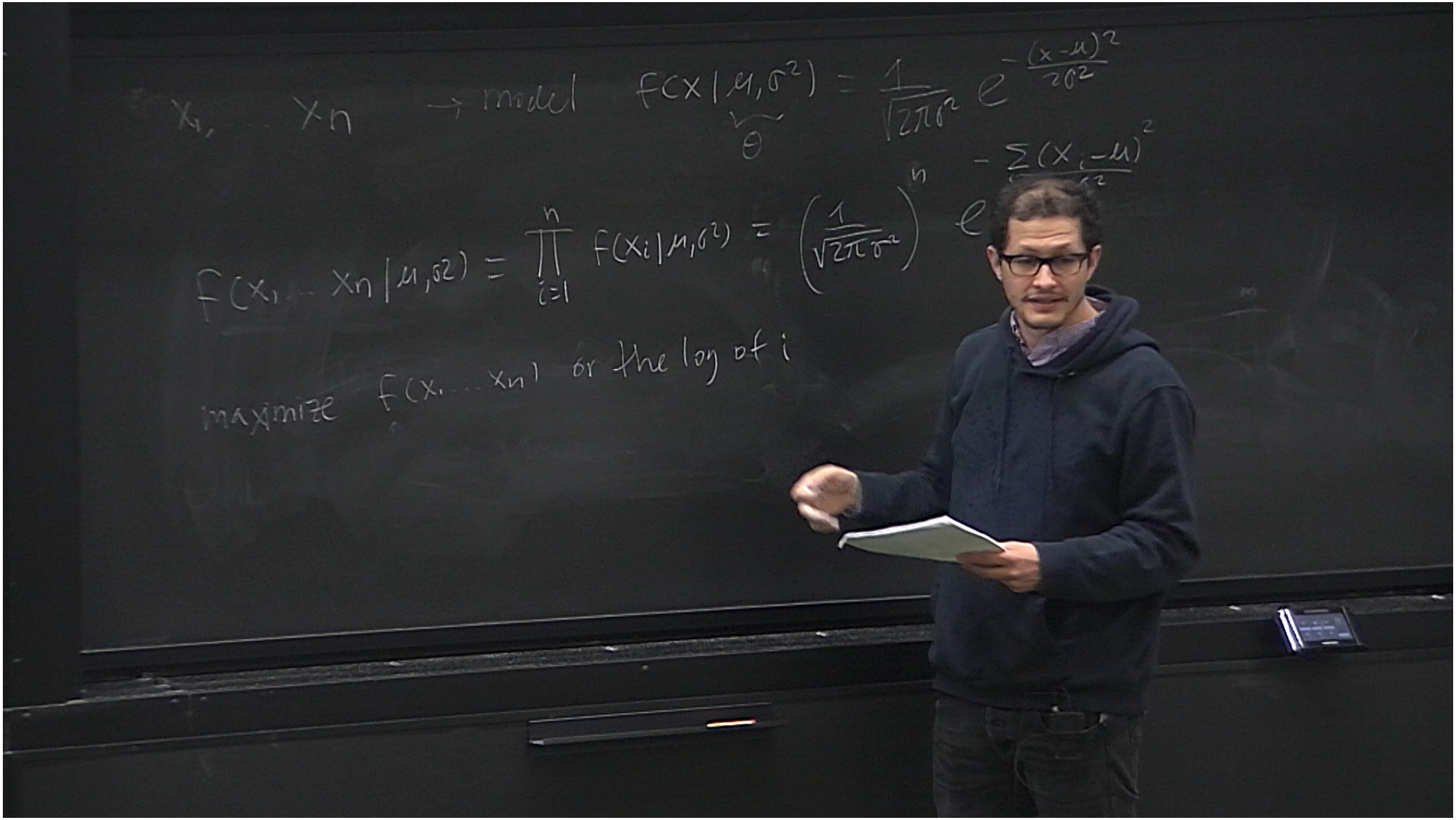

$KL(P_{data} \| P_{model}) = \sum_x P_{data}(x) \log \frac{P_{data}(x)}{P_{model}(x)}$
$= \sum_x \left[P_{data}(x) \log P_{data}(x) - P_{data}(x) \log P_{model}(x) \right]$
Entropy

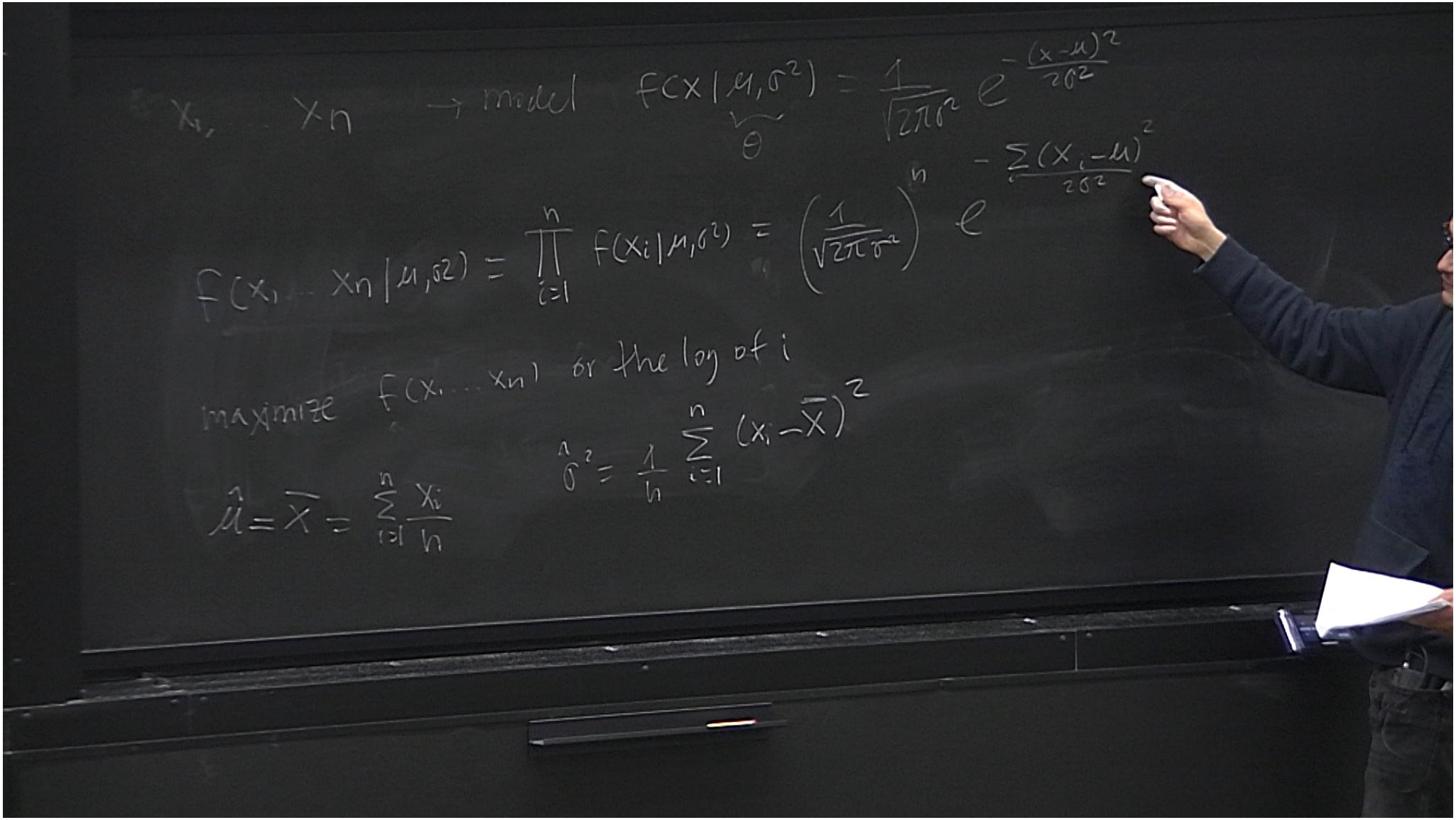

$KL(P_{data} \| P_{model}) = \sum_x P_{data}(x) \log \frac{P_{data}(x)}{P_{model}(x)}$
$= \sum_x \left[P_{data}(x) \log P_{data}(x) - P_{data}(x) \log P_{model}(x, \theta) \right]$
Entropy

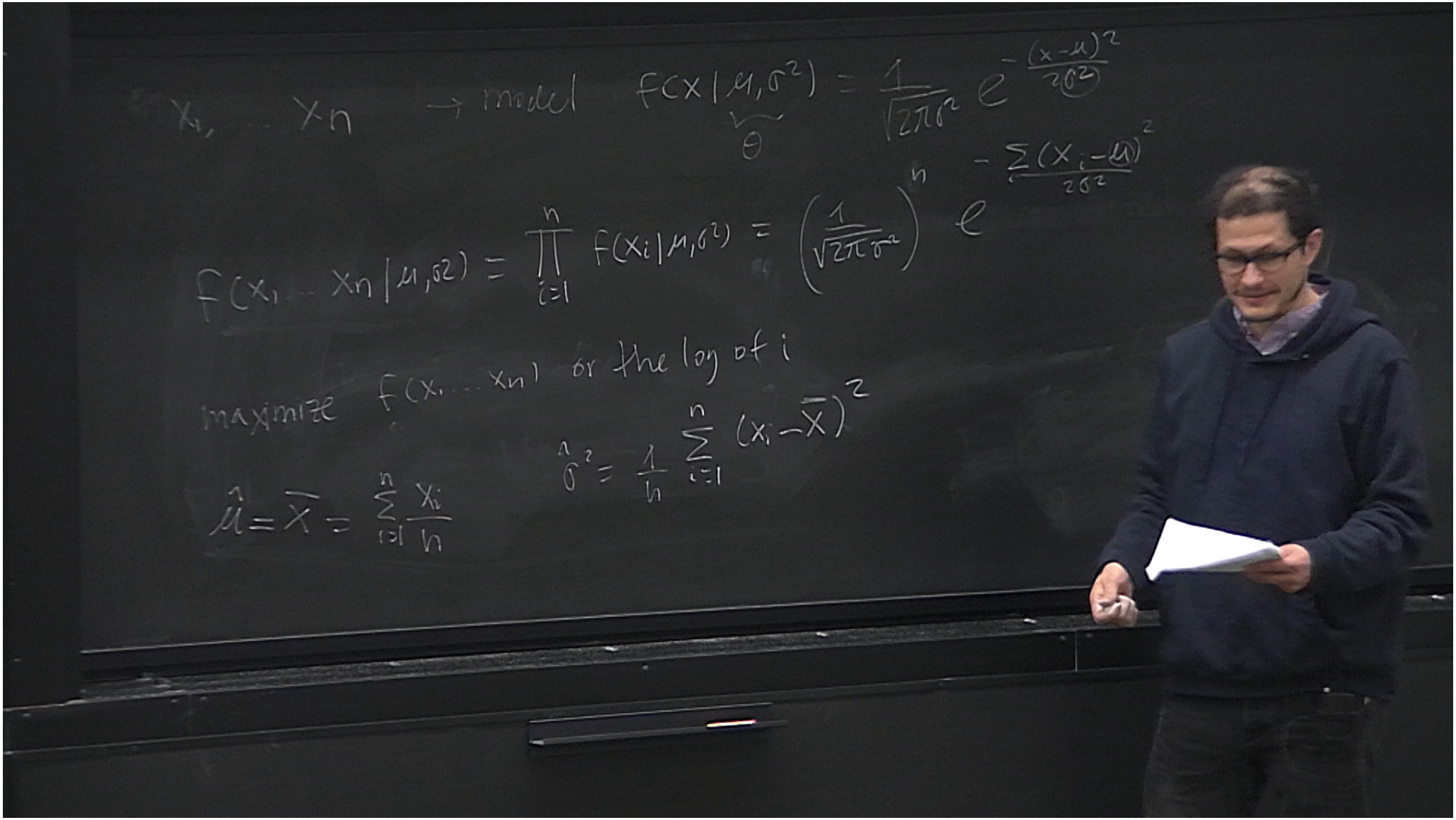

$X_1, \ldots X_n \rightarrow$ model $f(x|\underbrace{\mu,\sigma^2}_{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

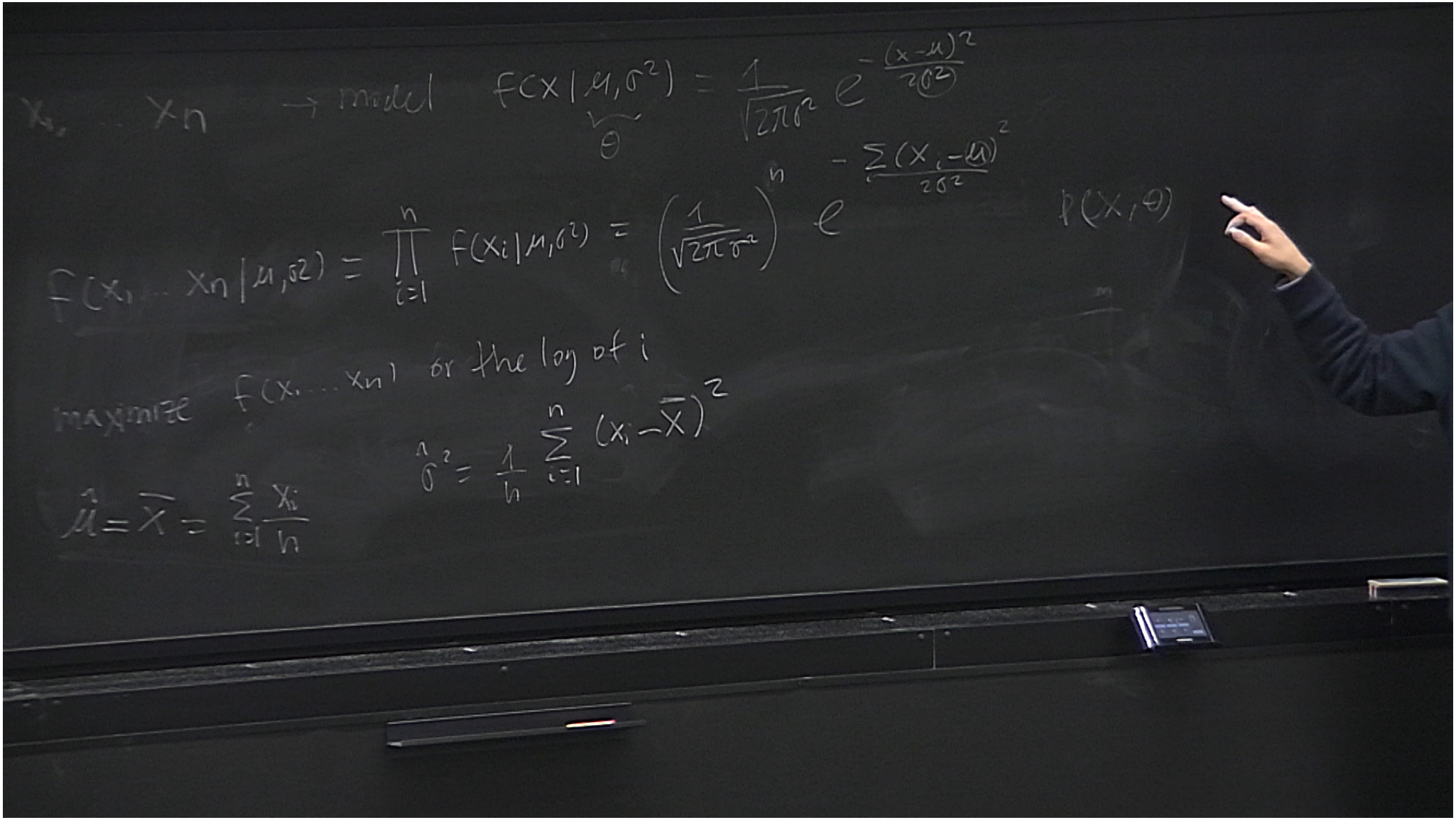

$X_1 \ldots X_n \rightarrow$ model $f(x|\underbrace{\mu,\sigma^2}_{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$f(x_1 \ldots x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2)$


$X_1 \dots X_n \to$ model $f(x|\underbrace{\mu,\sigma^2}_{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$f(x_1 \dots x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$


$X_1 \dots X_n \to$ model $f(x|\underbrace{\mu,\sigma^2}_{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$f(x_1 \dots x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$
maximize $f(x_1 \dots x_n)$ or the log of i


$X_1 \dots X_n \rightarrow$ model $f(x|\underbrace{\mu,\sigma^2}_{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$f(x_1 \dots x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$
maximize $f(x_1 \dots x_n)$ or the log of i
$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$
$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$


$X_1 \dots X_n \rightarrow$ model $f(x|\underbrace{\mu,\sigma^2}_{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$f(x_1 \dots x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$
maximize $f(x_1 \dots x_n)$ or the log of i
$\hat{\mu} = \bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$
$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2$


$X_1 \dots X_n \rightarrow$ model $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
θ
$f(x_1 \dots x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$
$P(X,\theta)$
maximize $f(x_1 \dots x_n)$ or the log of i
$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$
$\hat{\mu} = \bar{x} = \sum_{i=1}^{n}\frac{x_i}{n}$