Hidden values
Expectation Maximization
Mixture Models

- Marked A4: now
- A5: Wednesday
- PP: Monday
- A6: out tonight, due next Wed.
- MT: 2 weeks from today
  (study guide coming soon)

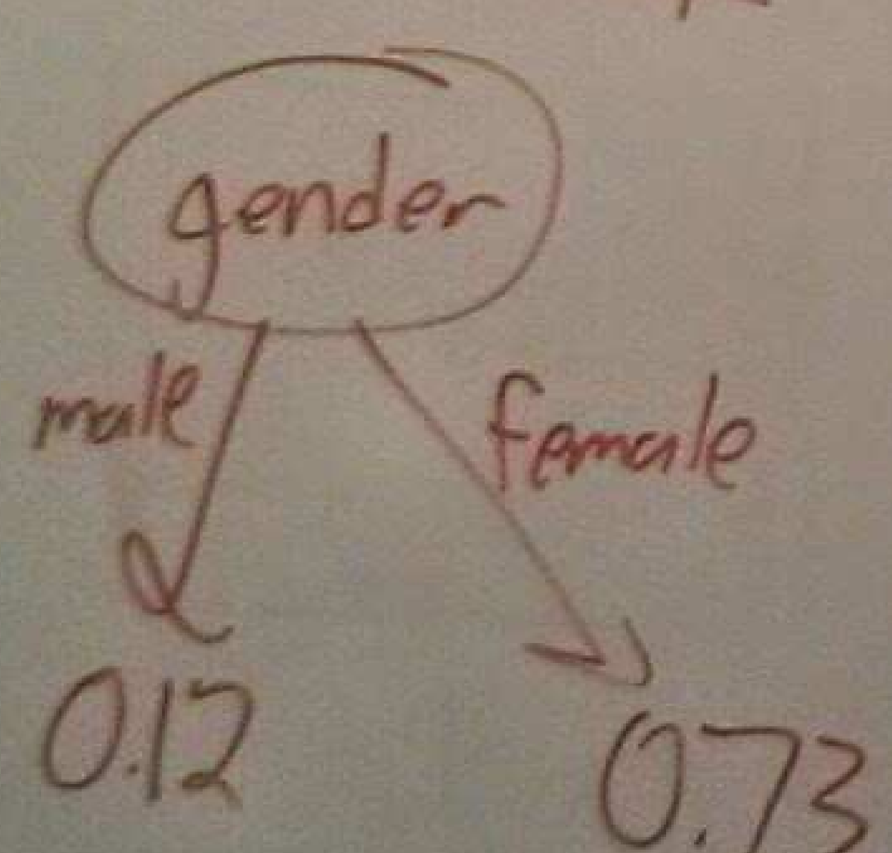
Why does bootstrap select ≈ 63% for large 'N'?

$$p(\text{'i' selected at least once}) = 1 - p(\text{not selected 'N' times})$$
$$= 1 - \left(1 - \frac{1}{n}\right)^n$$
$$\approx 1 - 1/e$$
$$\approx 63\%$$

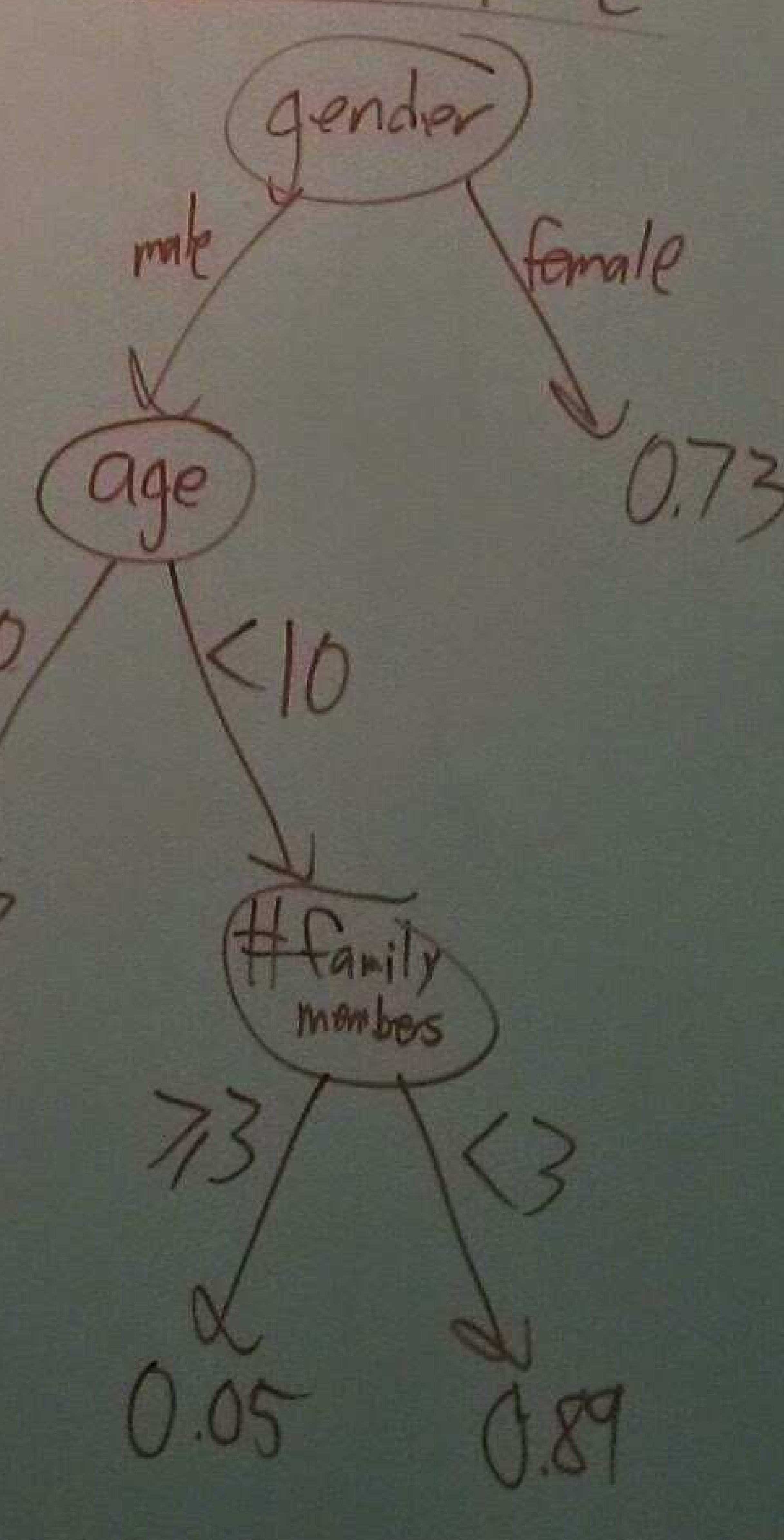
Example of decision tree

$$
X = \begin{array}{ccc}
\text{Gender} & \text{Age} & \text{\# family members} \\
\end{array}
\left[\begin{array}{ccc}
\text{Male} & 33 & 5 \\
\text{Female} & 10 & 1 \\
\vdots & \vdots & \vdots
\end{array}\right]
\qquad
y = \left[\begin{array}{c}
\text{'died'} \\
\text{'lived'} \\
\vdots
\end{array}\right]
$$

Decision Stump



Decision tree



Random Forests

$X = \begin{bmatrix} \quad \end{bmatrix} \xrightarrow{\text{Bootstrap}} X_1 = \begin{bmatrix} \quad \end{bmatrix} \rightarrow$ fit Decision tree to $\{X_1, y_1\}$

$X_2 = \begin{bmatrix} \quad \end{bmatrix} \rightarrow$ fit Decision tree to $\{X_2, y_2\}$

$\xrightarrow{\text{Bootstrap}} X_3 = \begin{bmatrix} \quad \end{bmatrix} \rightarrow$ fit Decision tree to $\{X_3, y_3\}$

each decision node is "random decision stump"
- picks random features
- use infogain
- you can prune with "left-out" samples

$X = \begin{bmatrix} X_1 \\ \text{-----} \end{bmatrix}$ } $\approx 63\%$ of original data

Hidden values
Expectation Maximization
Mixture Models

Hidden Values

- Learning when some values are unobserved, missing, hidden, latent.

Example of decision tree

$$X = \begin{array}{ccc} \text{Gender} & \text{Age} & \text{\# family members} \\ \hline \text{Male} & 33 & 5 \\ \text{Female} & 10 & 1 \\ \text{Female} & ? & 2 \\ \text{Male} & 22 & 0 \end{array} \qquad y = \begin{bmatrix} \text{'died'} \\ \text{'lived'} \\ \text{'died'} \\ ? \end{bmatrix}$$

Semi-Supervised Learning

Idea: getting labels is expensive,
getting unlabeled data is cheap.

- Can we can train on $\{X_L, y_L\}$ and $\{X_U\}$

$$X_L = \begin{bmatrix} \quad \end{bmatrix} \qquad y_L = \begin{bmatrix} \quad \end{bmatrix}$$

$$X_U = \begin{bmatrix} \quad \end{bmatrix} \qquad y_U = \begin{bmatrix} ? \end{bmatrix}$$

"Missing at Random" (MAR)

- the fact that it is missing does not depend on the missing value.
- E.g., digit classification

$\boxed{2} \longrightarrow$ "2"

$\boxed{3} \longrightarrow$ "3"

- missing random pixels: MAR
- hide the labels of all the "2" examples. (not MAR)
- hide the top half of every digit: MAR

$\boxed{\overset{?}{2}}$

- If not MAR, you need to model WHY data is missing.

Approach #1

1. Imputation: replace "?" w/ most likely value.
2. Fit model with "imputed" values.

"hard-EM"

idden values

pectation Maximization

xture Models

Probabilistic Approach

Notation: X: observed variables

H: hidden variables

$$P(X) = \sum_h P(X, H=h)$$

(integral if H is a continuous r.v.)

E.g. SSL:

$$p(\bar{y}_L, X_L, X_U) = \prod_{i=1}^{N} p(y_i, x_i) \prod_{j=1}^{T} \left(\sum_{y_j} p(y_j, x_j) \right)$$

Problem:

Assume $-\log P(X, H)$ is "nice"

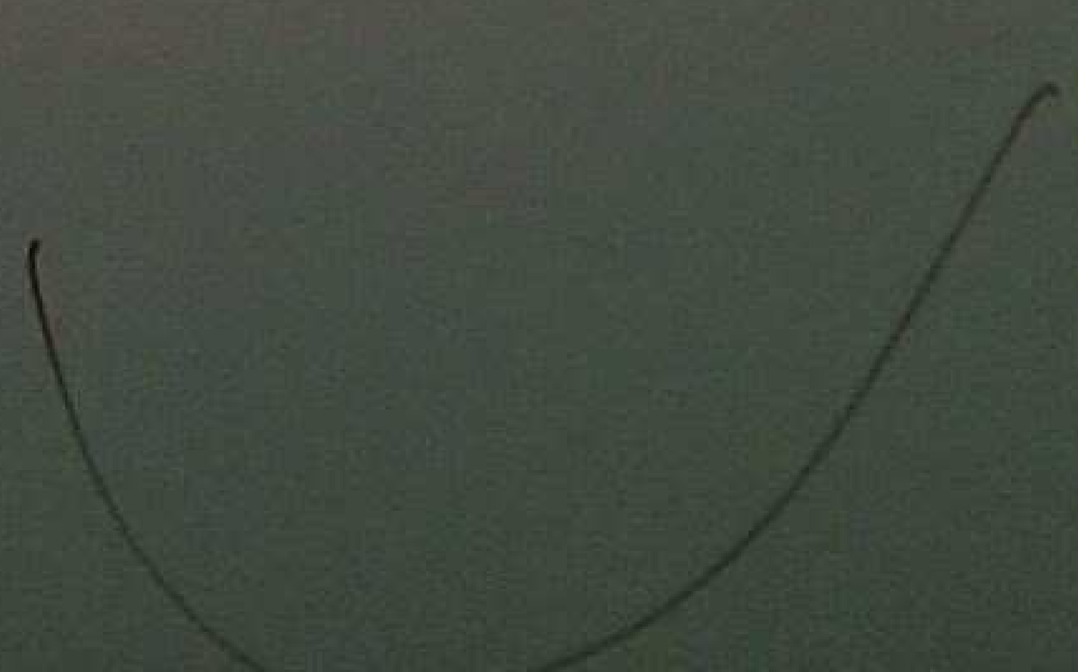
(closed-form, convex)

$$-\log(p(X)) = \underbrace{-\log\left(\sum_{h} p(X, H=h)\right)}_{\text{not convex}}$$

"problem is $\sum$ inside log"

$\log(1 + \exp(w^T x))$

$\log(\exp(1) + \exp(w^T x))$ "convex"



Expectation Maximization

- local optimizer when ($\log p(X, H)$) is "nice", with parameters θ.

$$E_x[f(x)] = \sum_x p(x) f(x)$$

$$E_{x|y}[f(x)] = \sum_x p(x|y) f(x)$$

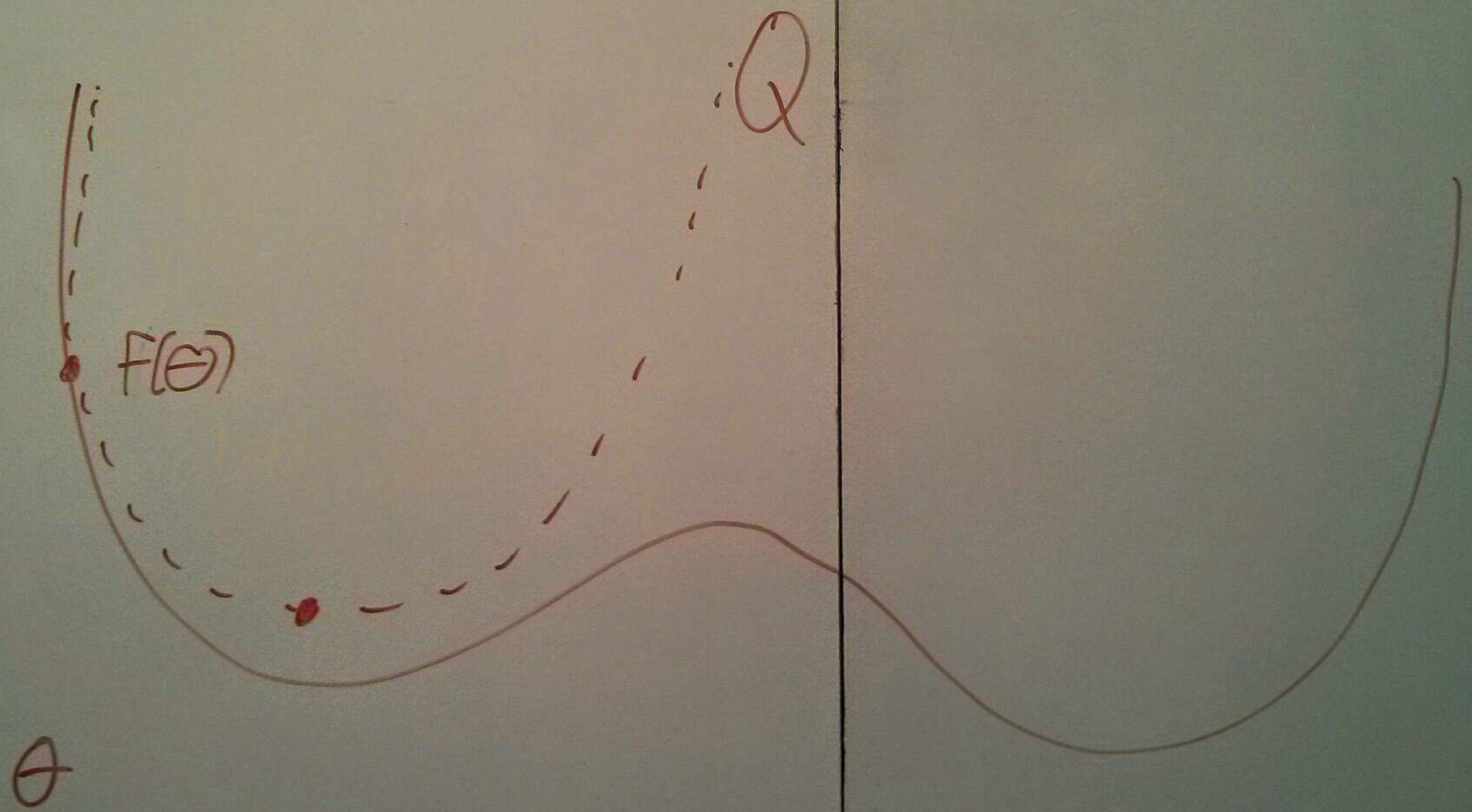
Problem: $\max_\theta p(X|\theta)$

Iterations: θ^t

θ^0

"E"-step: Define $\underline{Q}(\theta|\theta^t) = E_{H|X,\theta^t}[\log p(X, H|\theta)]$

"M"-step: $\theta^{t+1} = \arg\max_\theta Q(\theta|\theta^t)$ $= \sum_h \underbrace{p(h|X,\theta^t)}_{\alpha_h^t} \underbrace{\log p(X, h|\theta)}_{\text{"nice"}}$



Q
F(θ)
θ

r Maximization

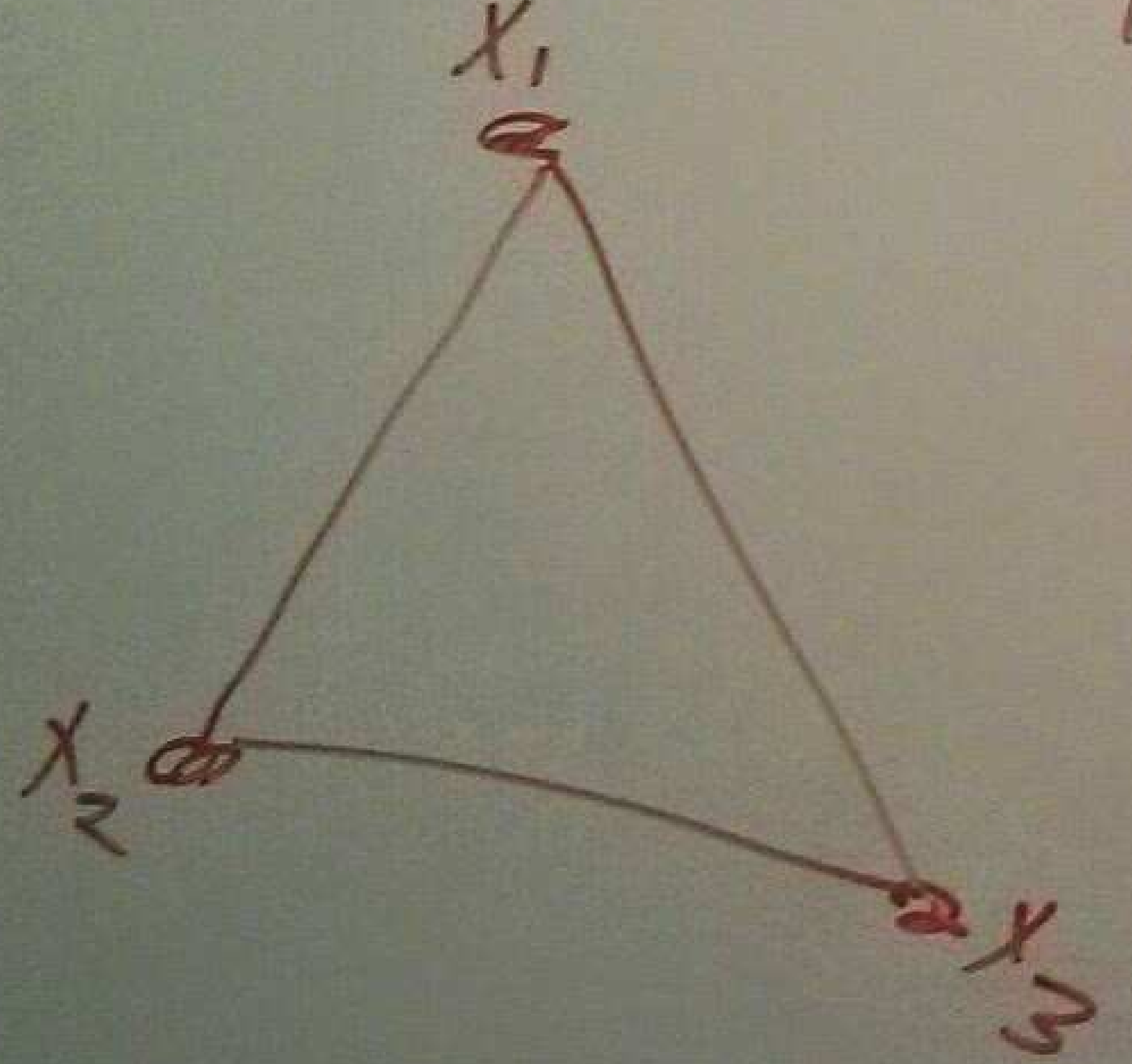
Models

Theorem:

$$\log p(X|\theta^{t+1}) - \log p(X|\theta^{t}) \geq Q(\theta^{t+1}|\theta^{t}) - Q(\theta^{t}|\theta^{t})$$

"Convex Combination"

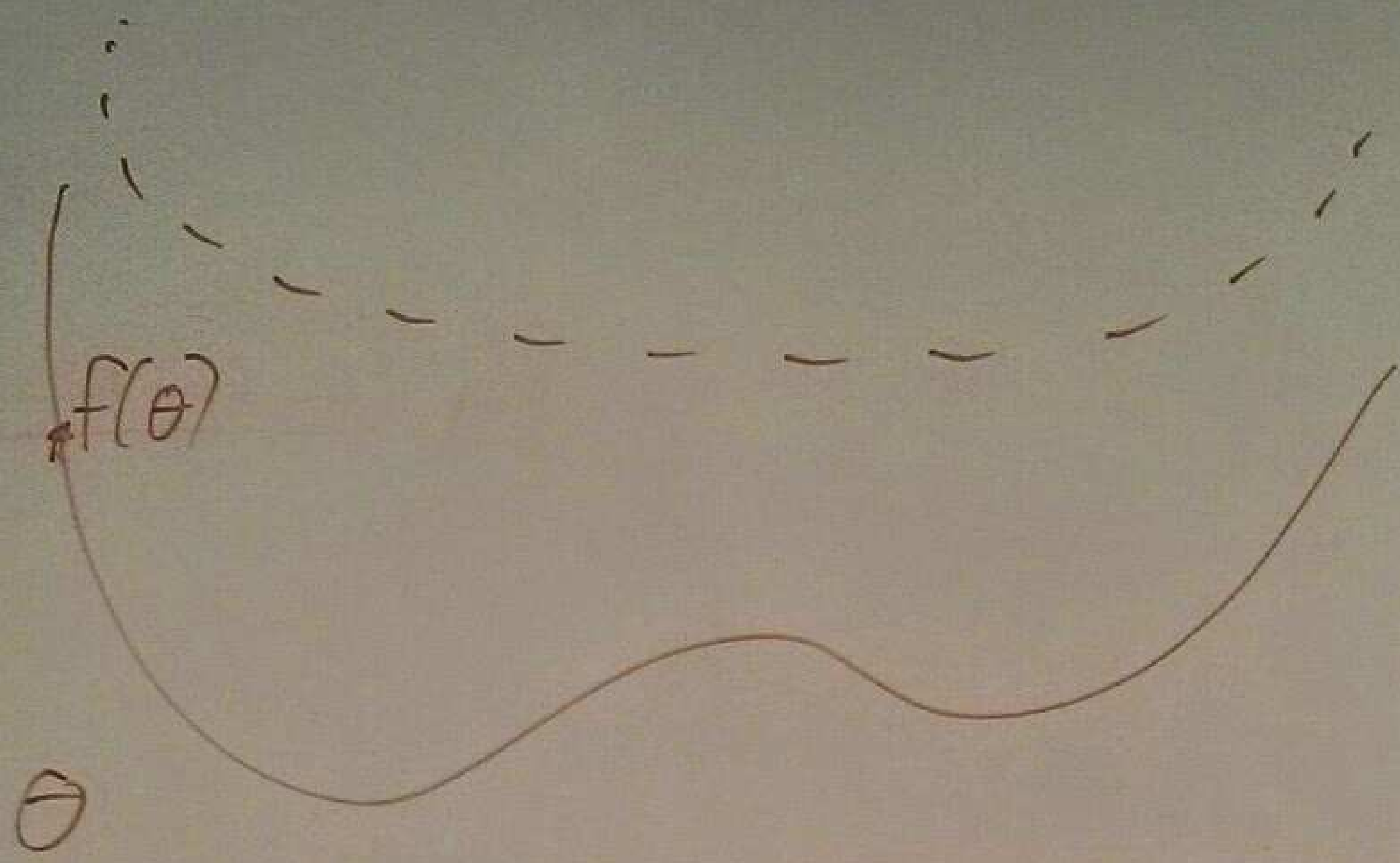
$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n,$$



$$\alpha_i \geq 0 \quad \sum \alpha_i = 1$$

If f is convex,

$$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i)$$



−lo

Problem: $\max_{\theta} \; p(X|\theta)$

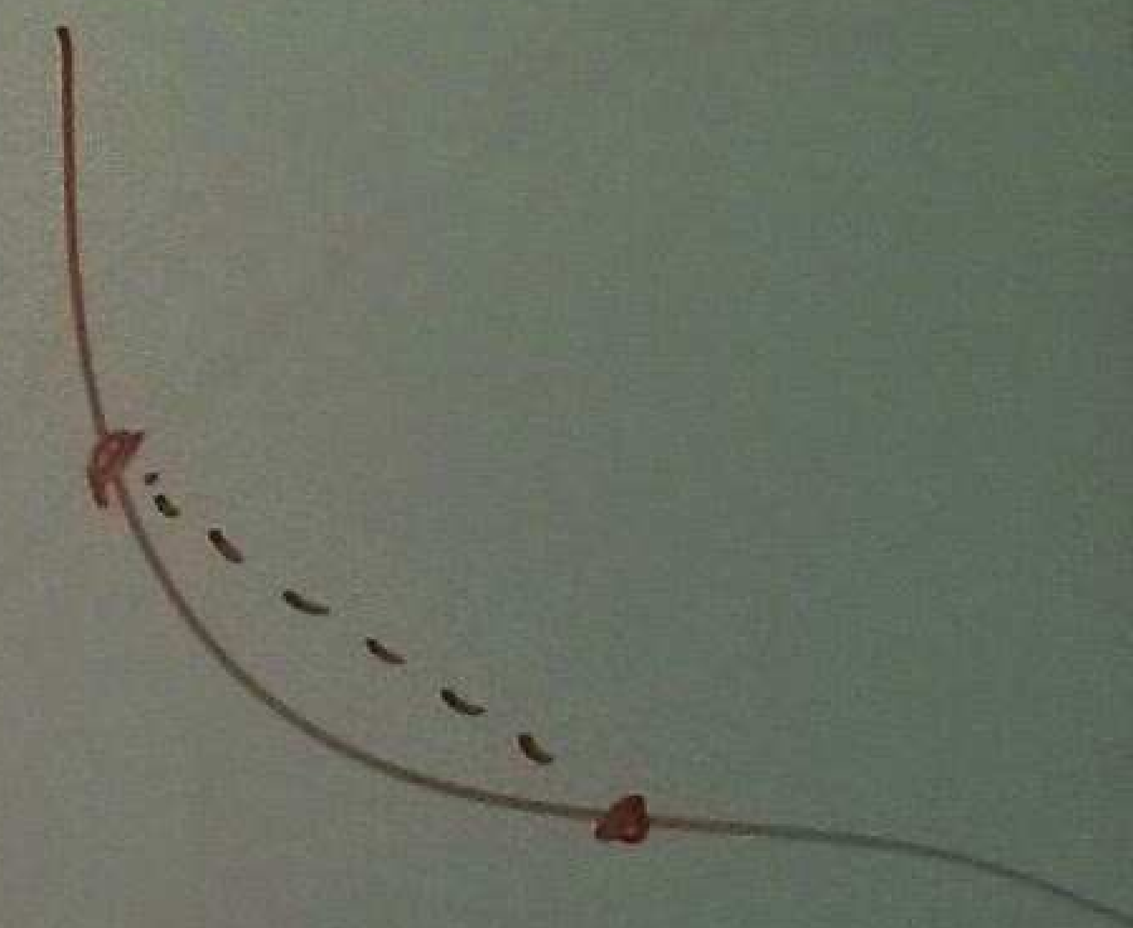
Iterations: θ^t

"E"-step: Define $\underline{Q}(\theta|\theta^t) = E_{H|X,\theta^t}\left[\log p(X,H|\theta)\right]$

"M"-step: $\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t) \quad = \sum_h \underbrace{p(h|X,\theta^t)}_{\alpha_h^t} \underbrace{\log p(X,h|\theta)}_{\text{"nice"}}$

$$-\log p(X|\theta) = -\log \sum_h p(X,h|\theta)$$

$$= -\log\left(\sum_h \alpha_h \frac{p(X,h|\theta)}{\alpha_h}\right)$$



$$\le -\sum_h \alpha_h \log \frac{p(X,h|\theta)}{\alpha_h}$$

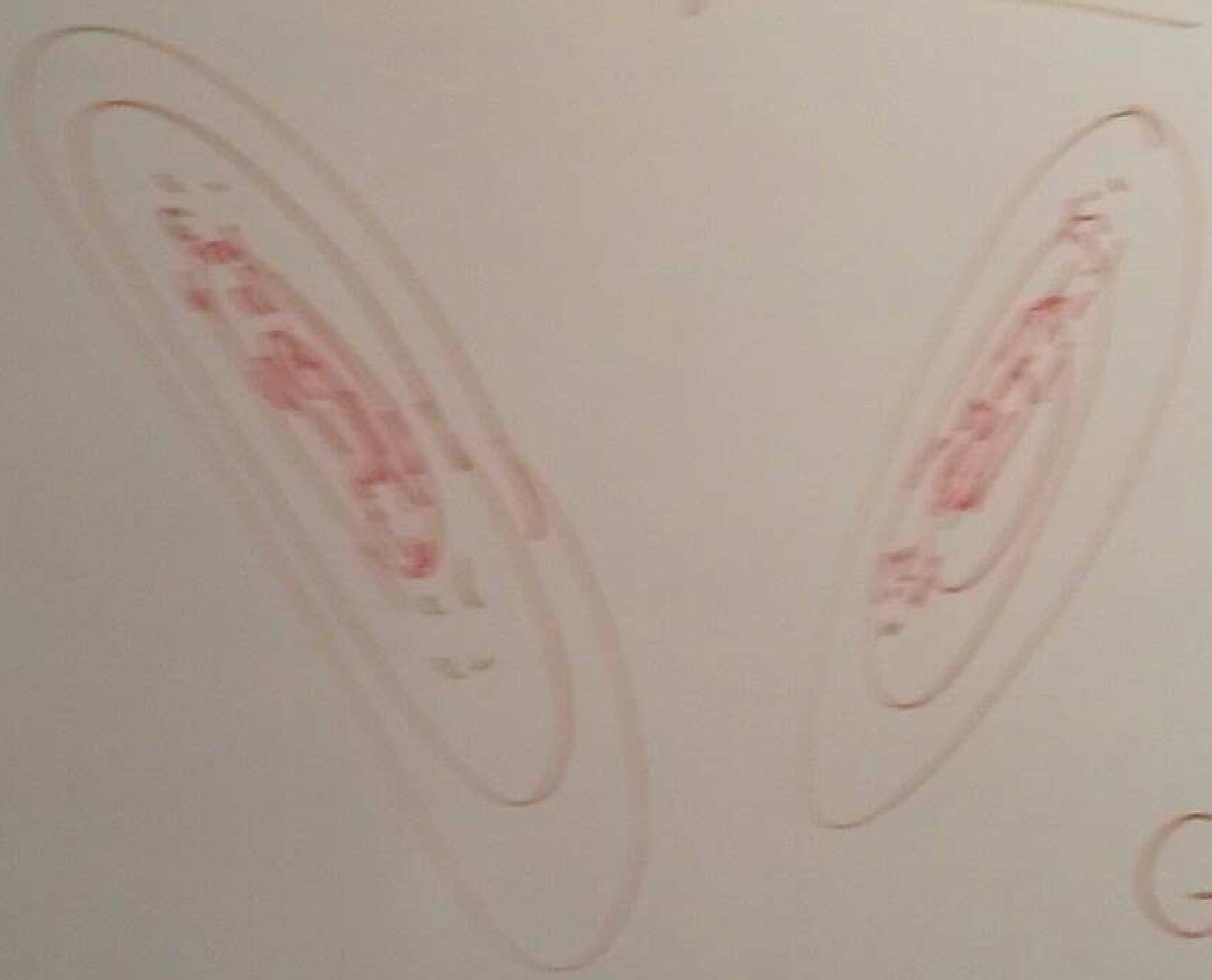
$$= -\sum_h \alpha_h \log p(X,h|\theta) + \underbrace{\sum_h \alpha_h \log \alpha_h}_{-H(\vec{\alpha})}$$

$$= -Q(\theta|\theta^t) + \text{const.}$$

xture Models

Mixture Models

Recall fitting Gaussians:



GDA:

Want to fit probabilistic model
- don't have labels
- but it would be easy if we did
- motivates "mixture" model

Let $z_i \in \{1, 2, \ldots, k\}$.

$$p_k(x_i \mid \theta) = p(x_i \mid z_i = k, \theta)$$

"Gaussian"

z_i is a "latent" variable.

$$p(x_i \mid \theta) = \sum_{k=1}^{K} p(x_i, z_i = k \mid \theta)$$

$$= \sum_{k=1}^{K} \underbrace{p(x_i \mid z_i = k, \theta)}_{\text{Gaussian for cluster 'k'}} \underbrace{p(z_i = k \mid \theta)}_{\text{probability that data point belongs to mixture}}$$

k-means: hard-EM with Gauss with $\Sigma_k = I$