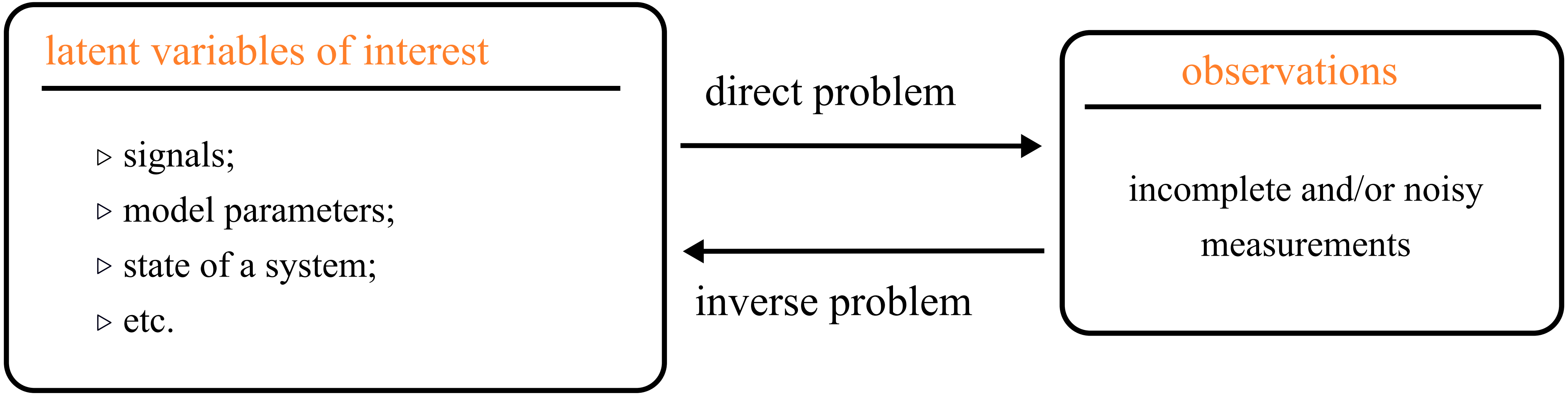
Bayesian Methods for Machine Learning

Lecture 3 (part 2) - Inference in latent variable models

Simon Leglaive

CentraleSupélec, 2020-2021



latent variables of interest

- ▷ signals;
- ▷ model parameters;
- ▷ state of a system;
- ▷ etc.

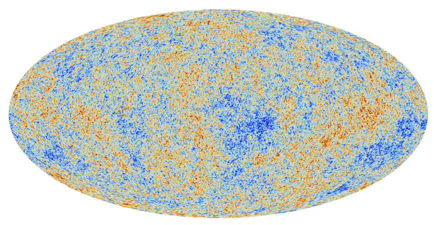
direct problem

inverse problem

observations

incomplete and/or noisy measurements

Astrophysics

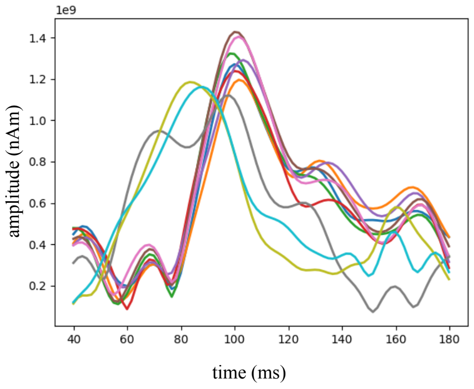


Credit: ESA - Planck Collaboration

Medical imaging

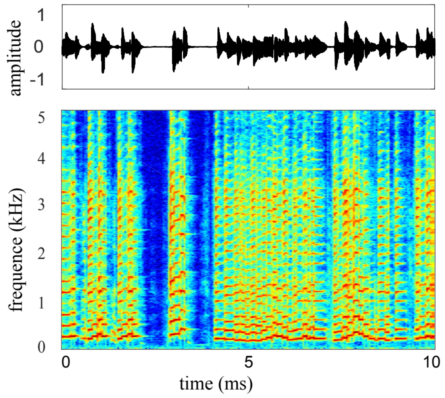
Credit: Wikipedia

Biomedical signal processing



Credit: MNE Python

Audio signal processing



Generative latent-variable model

Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$ denote the observed and latent random variables, respectively.

Developing a probabilistic model consists in defining the joint distribution of the observed and latent variables, also called complete-data likelihood:

$$p(\mathbf{x}, \mathbf{z}; \theta) = p(\mathbf{x}|\mathbf{z}; \theta)p(\mathbf{z}; \theta),$$

where θ is a set of unknown deterministic parameters.

Generative latent-variable model

Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$ denote the observed and latent random variables, respectively.

Developing a probabilistic model consists in defining the joint distribution of the observed and latent variables, also called complete-data likelihood:

$$p(\mathbf{x}, \mathbf{z}; \theta) = p(\mathbf{x}|\mathbf{z}; \theta)p(\mathbf{z}; \theta),$$

where θ is a set of unknown deterministic parameters.

Remarks:

- $p(\mathbf{z}; \theta)$ is the prior over the latent variables.
- $p(\mathbf{x}|\mathbf{z}; \theta)$ is the likelihood.
- to simplify notations we use θ to denote both the parameters of the prior and likelihood, but these two distributions usually depend on disjoint sets of parameters.
- In a full Bayesian setting, there are no deterministic parameters to be estimated. Parameters are also treated as latent variables and θ denotes known and manually-fixed hyperparameters.

The two central problems in statistical inference for latent variable models

Computing the posterior distribution of the latent variables

$$p(\mathbf{z}|\mathbf{x};\theta) = \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{p(\mathbf{x};\theta)} = \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{\int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)d\mathbf{z}}.$$

where $p(\mathbf{x};\theta) = \int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)d\mathbf{z}$ is the marginal likelihood, also called evidence.

- Computing this posterior is the inference step.
- The posterior summarizes our knowledge on latent variables of interest, once we have observed the data.
- For numerous probabilistic models, the posterior is analytically intractable, e.g. because its normalizing constant – the marginal likelihood – cannot be computed analytically.

Estimating the model parameters by maximum (marginal) likelihood

$$\hat{\theta} = \arg\max_{\theta} \; p(\mathbf{x}; \theta) = \arg\max_{\theta} \int p(\mathbf{x}|\mathbf{z}; \theta) p(\mathbf{z}; \theta) d\mathbf{z}.$$

Quite often, directly solving the optimization problem associated with this ML estimation procedure is difficult, if not impossible when the marginal likelihood cannot be computed analytically.

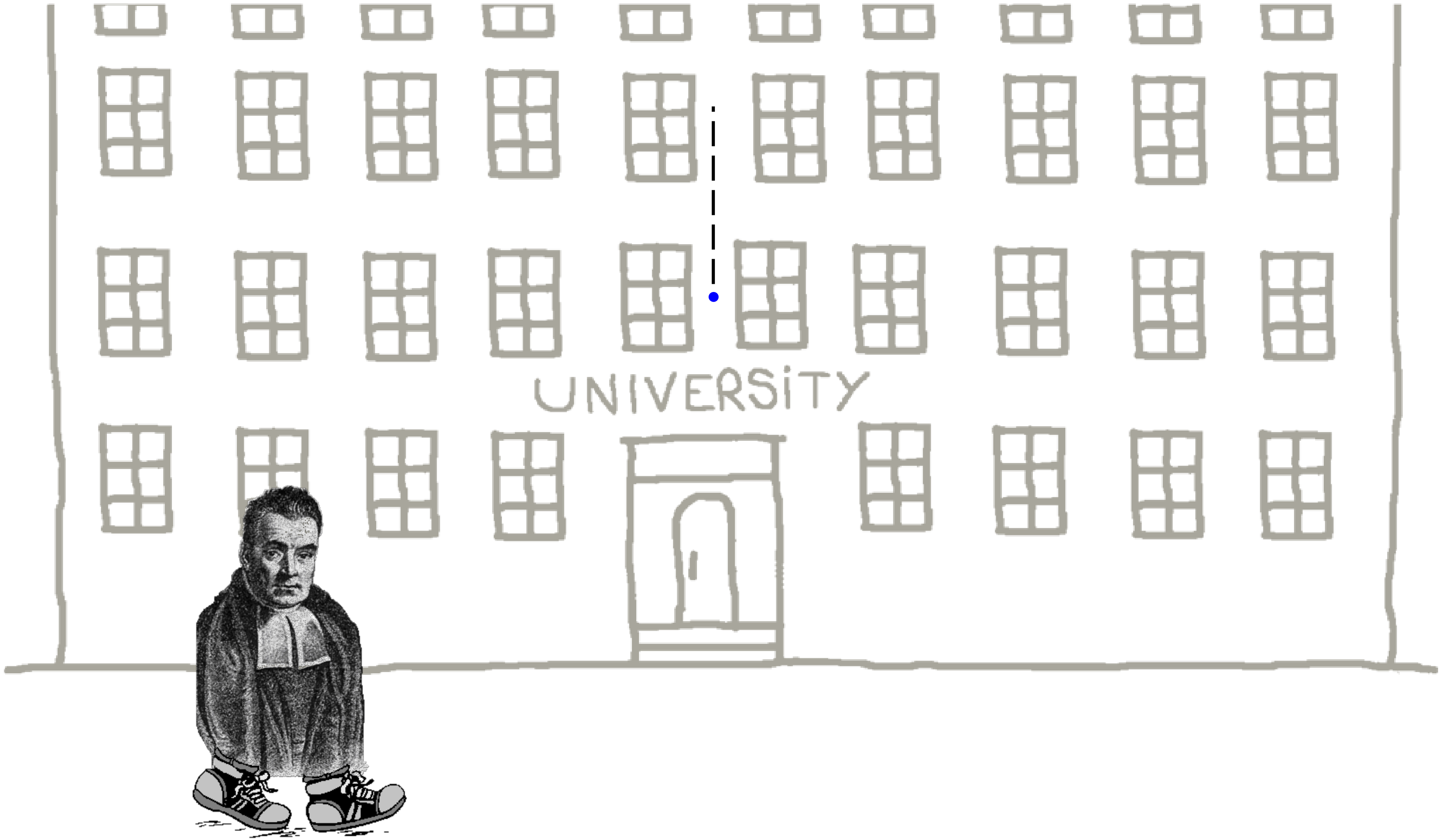
Today, we will focus on models where directly maximizing the likelihood is difficult, but the posterior distribution of the latent variables can be computed analytically.

These two problems (parameters estimation and posterior inference) are actually strongly related:

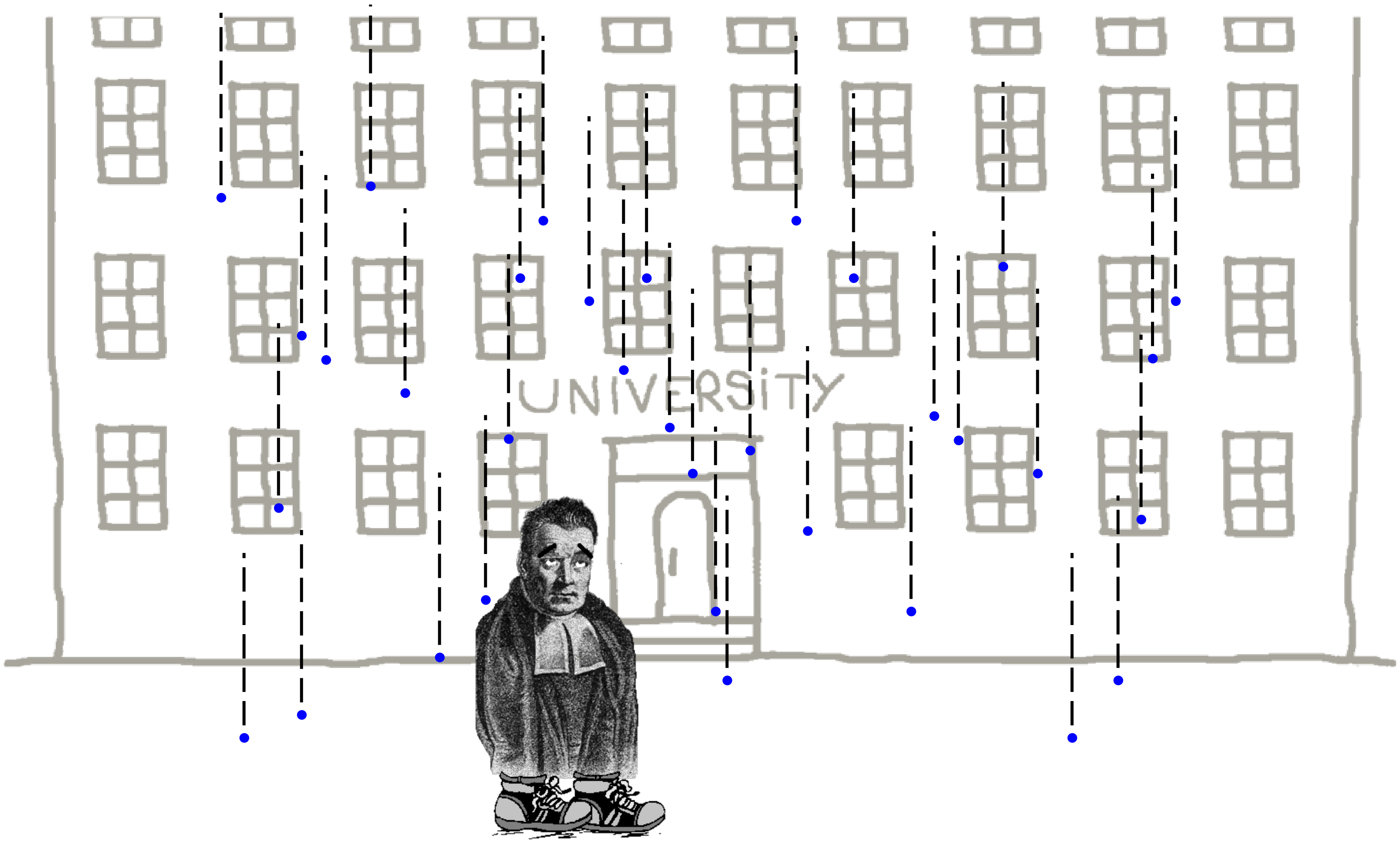
1. the posterior distribution of the latent variables depends on the model parameters.
2. as we are going to see, when direct maximization of the marginal likelihood is impossible, we usually consider the maximization of a lower bound, which precisely depends on the posterior distribution of the latent variables.

Illustrative example

The following example and drawings are adapted from a tutorial on Bayesian Learning for Signal Processing given by Antoine Deleforge at the LVA/ICA 2015 Summer School.



UNIVERSiTY


UNIVERSiTY

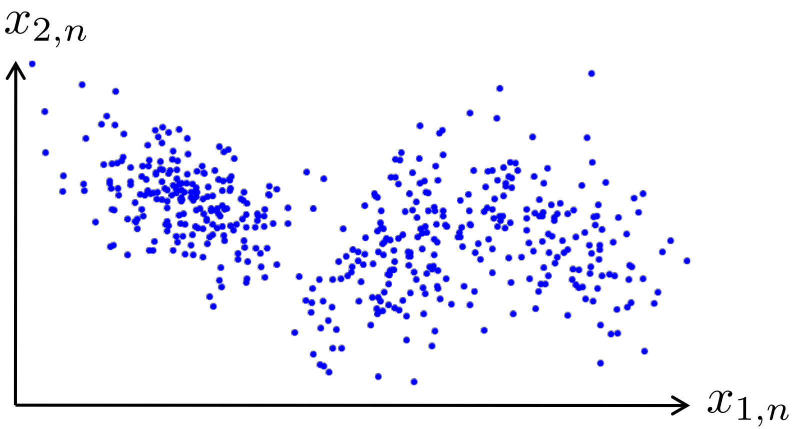
UNIVERSITY

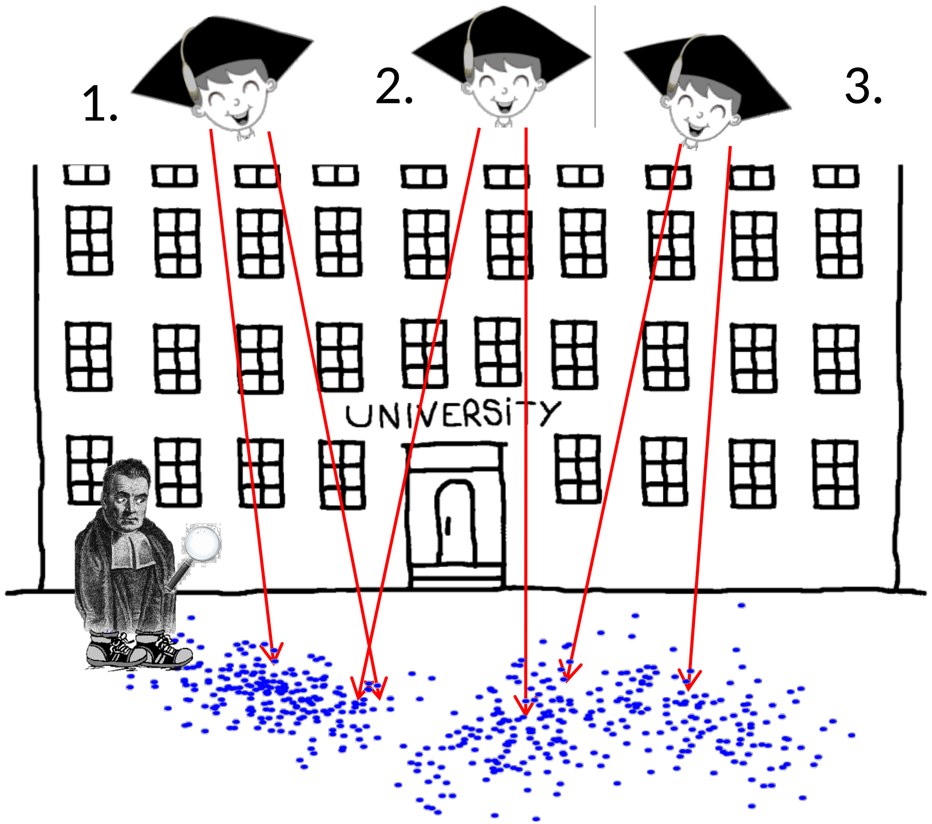
UNIVERSITY
They are on the roof!

Modeling

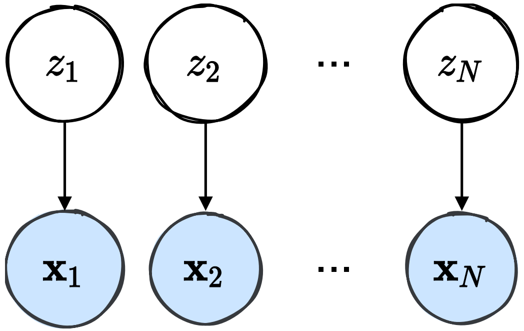
Observed variables: $\{\mathbf{x}_n \in \mathbb{R}^2\}_{n=1}^N$.

Hidden variables: $\{z_n \in \{1, 2, 3\}\}_{n=1}^N$.

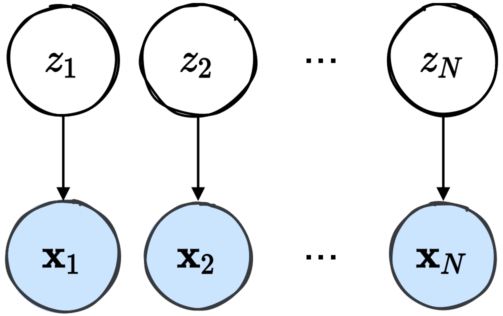


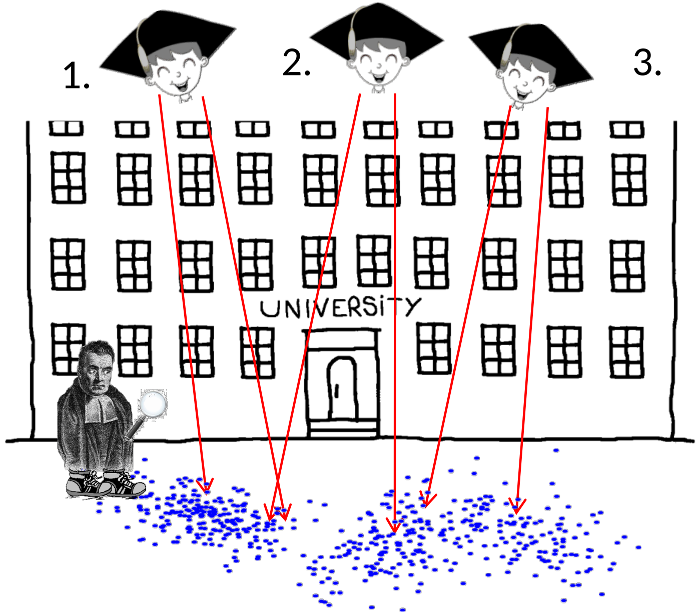


Bayesian network



$$p(\{\mathbf{x}_n, z_n\}_{n=1}^N; \theta) = \prod_{n=1}^N p(\mathbf{x}_n | z_n; \theta) p(z_n; \theta).$$





Prior $p(z_n = k; \theta) = \pi_k, \qquad \sum_{k=1}^{K} \pi_k = 1, \qquad K = 3$

Likelihood $p(\mathbf{x}_n | z_n = k; \theta) = \mathcal{N}\left(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$

Parameters $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}.$

Marginal likelihood

$$\begin{aligned} p(\{\mathbf{x}_n\}_{n=1}^{N}; \theta) &= \prod_{n=1}^{N} p(\mathbf{x}_n; \theta) \\ &= \prod_{n=1}^{N} \sum_{k=1}^{K} p(\mathbf{x}_n | z_n = k; \theta) p(z_n = k; \theta) \\ &= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \end{aligned}$$

Observations are independent and identically drawn from a Gaussian mixture model (GMM) with $K = 3$ components.

The parameters π_k are called the mixing coefficients, they give the prior probability of picking the k-th component to generate a data point $\mathbf{x}_n$.

Inference

Posterior distribution

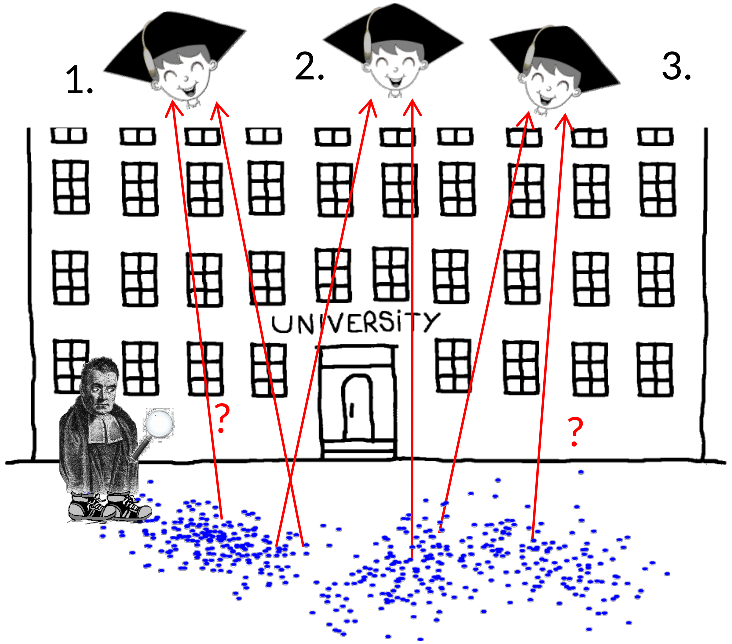
$$p(\{z_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N; \theta) = \prod_{n=1}^N p(z_n | \mathbf{x}_n; \theta),$$

where

$$p(z_n = k | \mathbf{x}_n; \theta) = \frac{p(\mathbf{x}_n | z_n = k; \theta) p(z_n = k; \theta)}{p(\mathbf{x}_n; \theta)}$$

$$= \frac{p(\mathbf{x}_n | z_n = k; \theta) p(z_n = k; \theta)}{\sum_{j=1}^K p(\mathbf{x}_n | z_n = j; \theta) p(z_n = j; \theta)}$$

$$= \frac{\pi_k p(\mathbf{x}_n | z_n = k; \theta)}{\sum_{j=1}^K \pi_j p(\mathbf{x}_n | z_n = j; \theta)}.$$



The posterior probabilities $p(z_n = k | \mathbf{x}_n; \theta)$ are also known as the responsabilities.

The argmax of the responsability assigns the observation to a component, i.e. it clusters the data.

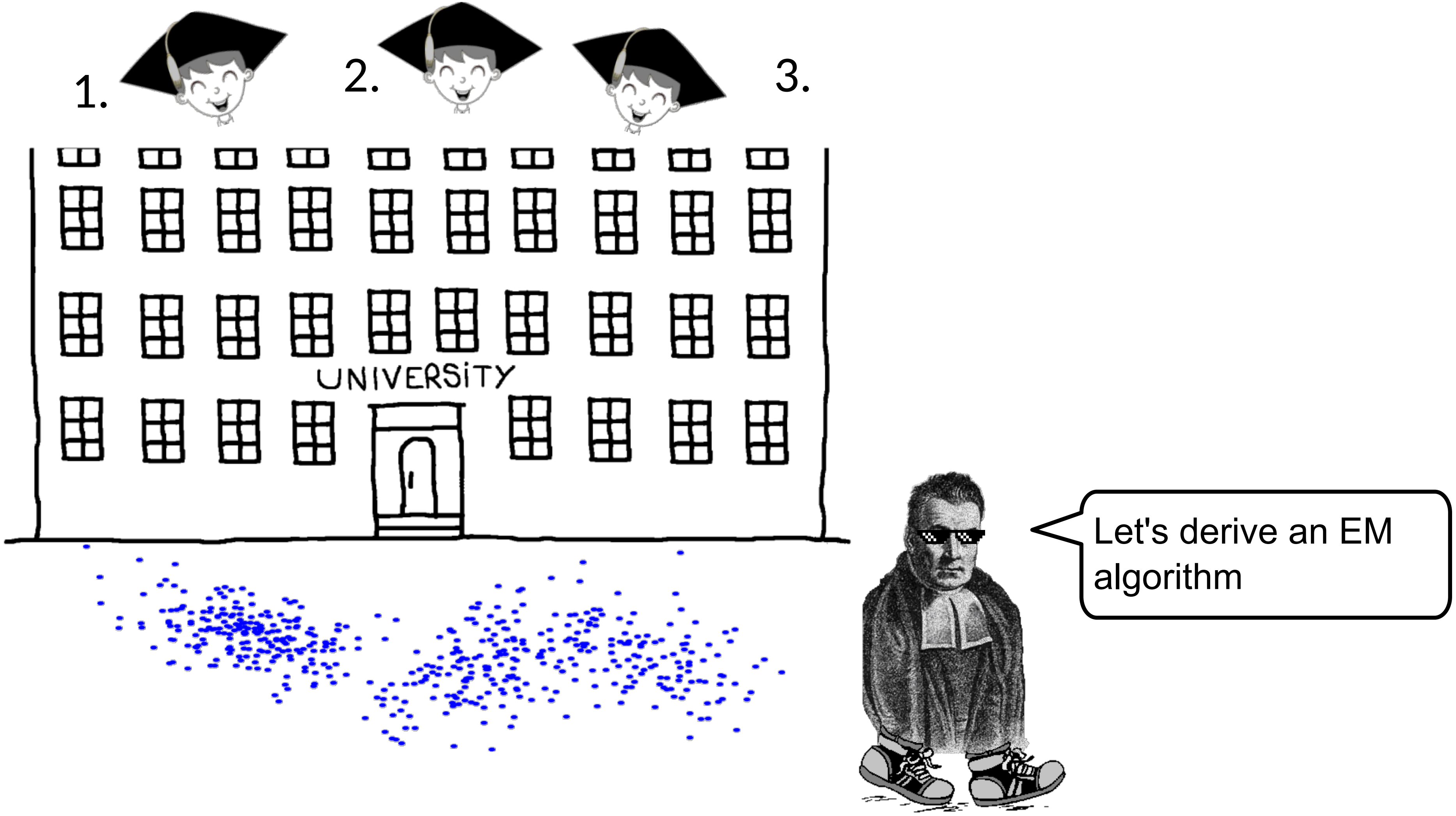
Parameters estimation

The posterior distribution can be computed analytically, but it depends on the unknown model parameters $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

Ideally, we would like to estimate them by maximizing the log-marginal likelihood:

$$\begin{aligned}
\mathcal{L}(\theta) &= \ln p(\{\mathbf{x}_n\}_{n=1}^{N}; \theta) \\
&= \ln \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \\
&= \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right).
\end{aligned}$$

Due to the presence of the sum over k inside the logarithm, the maximum marginal likelihood solution for the parameters does not have a closed-form analytical solution.


1.
2.
3.
UNIVERSITY
Let's derive an EM algorithm

The expectation-maximization algorithm

The expectation-maximization (EM) algorithm is a general technique introduced by Dempster et al. in 1977 for maximum likelihood parameters estimation in probabilistic models having latent variables.

Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$ denote the observed and latent random variables, respectively, which are assumed to be continuous, although the discussion is identical in the discrete setting.

We assume that direct optimization of the marginal likelihood $p(\mathbf{x}; \theta)$ is difficult, while optimization of the complete-data likelihood function $p(\mathbf{x}, \mathbf{z}; \theta)$ is much simpler.

The evidence lower bound

We first introduce a distribution over the latent variables whose probability density function is denoted by $q(\mathbf{z})$.

For any distribution $q(\mathbf{z})$, the following decomposition of the log-marginal likelihood holds:

$$\ln p(\mathbf{x};\theta) = \mathcal{L}(q(\mathbf{z}),\theta) + D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)),$$

where $\mathcal{L}(q(\mathbf{z}),\theta)$ is called the evidence lower bound (ELBO), and it is defined by

$$\mathcal{L}(q(\mathbf{z}),\theta) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q(\mathbf{z})].$$

The Kullback-Leibler (KL) divergence is defined by:

$$D_{\mathrm{KL}}(q \parallel p) = \mathbb{E}_q[\ln(q) - \ln(p)],$$

and it satisfies $D_{\mathrm{KL}}(q \parallel p) \geq 0$ with equality if and only if $q = p$.

Proof: $\ln p(\mathbf{x};\theta) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x};\theta)] = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln p(\mathbf{z}|\mathbf{x};\theta) - \ln q(\mathbf{z}) + \ln q(\mathbf{z})]$

$$\ln p(\mathbf{x};\theta) = \mathcal{L}(q(\mathbf{z}),\theta) + D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta))$$

As the KL divergence is always non-negative, we have:

$$\ln p(\mathbf{x};\theta) \geq \mathcal{L}(q(\mathbf{z}),\theta),$$

with equality if and only if $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta)$.

The ELBO is indeed a lower bound of the log-marginal likelihood, which is tight if $q(\mathbf{z})$ matches the true posterior.

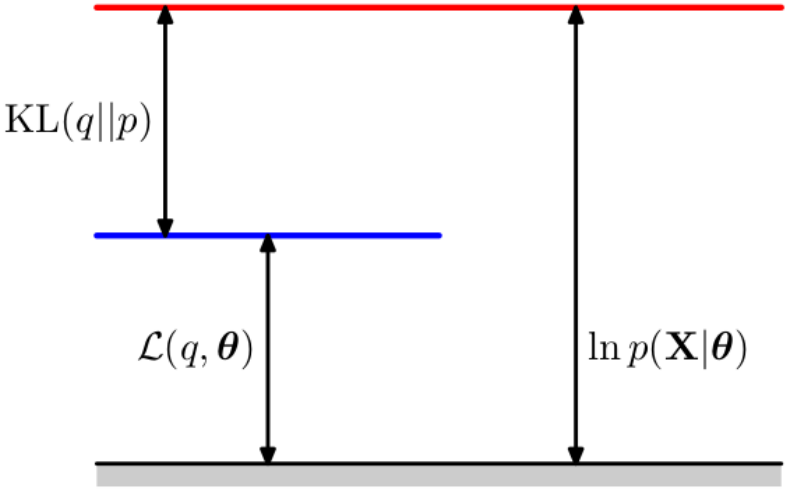

$\mathrm{KL}(q||p)$
$\mathcal{L}(q,\boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

Image credit: Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

EM algorithm

The EM algorithm is an iterative algorithm which alternates between optimizing the ELBO with respect to q in the E-Step and with repspect to θ in the M-step.

We first initialize θ_0, then we iterate for $t \geq 0$

- E-Step: $q_{t+1}(\mathbf{z}) = \arg\max_{q} \mathcal{L}(q(\mathbf{z}), \theta_t)$
- M-Step: $\theta_{t+1} = \arg\max_{\theta} \mathcal{L}(q_{t+1}(\mathbf{z}), \theta)$

E-Step

We recall the decomposition of the log-marginal likelihood:

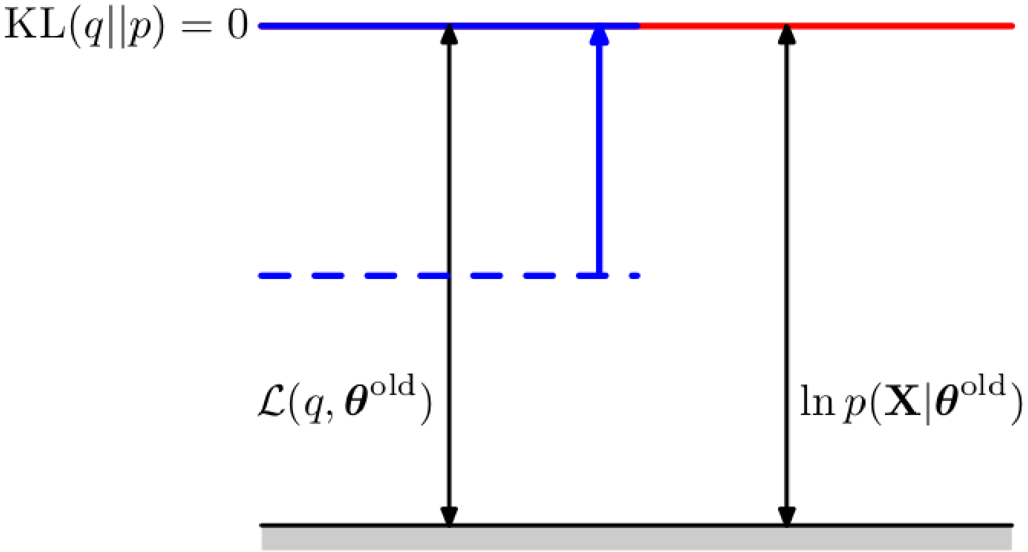
$$\ln p(\mathbf{x};\theta) = \mathcal{L}(q(\mathbf{z}),\theta) + D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta)).$$

The solution of the E-step is given by:

$$\begin{aligned} q_{t+1}(\mathbf{z}) &= \arg\max_q \mathcal{L}(q(\mathbf{z}),\theta_t) \\ &= \arg\min_q D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_t)) \\ &= p(\mathbf{z}|\mathbf{x};\theta_t). \end{aligned}$$

After the E-Step, we have $D_{\mathrm{KL}}(q_{t+1}(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_t)) = 0$, and the ELBO is equal to the log-marginal likelihood (i.e. the lower-bound is tight):

$$\ln p(\mathbf{x};\theta_t) = \mathcal{L}(q_{t+1}(\mathbf{z}), \theta_t).$$


$\mathrm{KL}(q||p) = 0$
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{old}})$

Therefore, maximizing the lower-bound with respect to the model parameters in the M-step will necessarily increase the log-marginal likelihood.

Image credit: Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

M-Step

- We recall the expression of the ELBO:

$$\mathcal{L}(q(\mathbf{z}), \theta) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}, \mathbf{z}; \theta) - \ln q(\mathbf{z})],$$

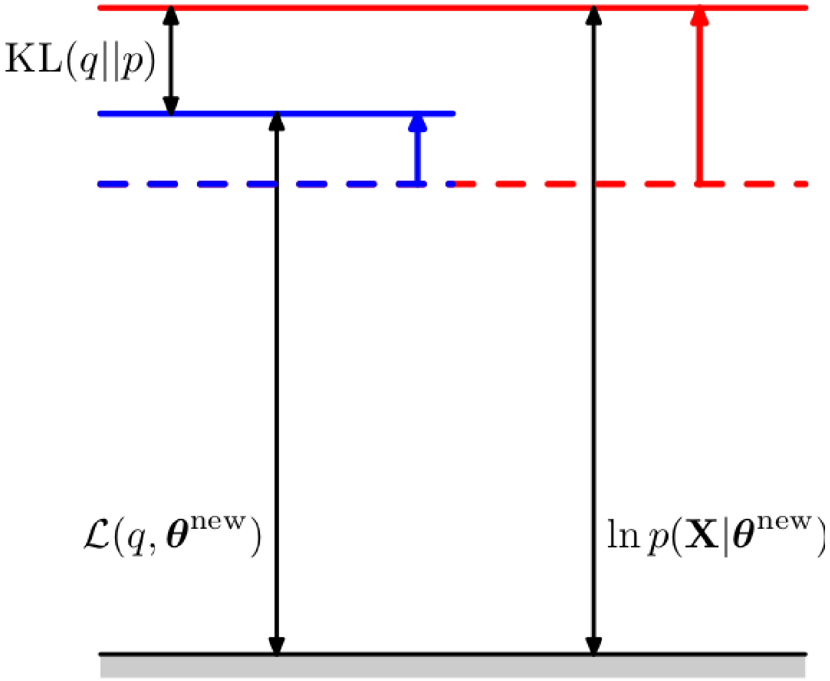
- The solution of the M-step is given by:

$$\begin{aligned}\theta_{t+1} &= \arg\max_{\theta} \mathcal{L}(q_{t+1}(\mathbf{z}), \theta) \\ &= \arg\max_{\theta} \mathcal{L}\big(p(\mathbf{z}|\mathbf{x}; \theta_t), \theta\big) \\ &= \arg\max_{\theta} \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta) - \ln p(\mathbf{z}|\mathbf{x}; \theta_t)] \\ &= \arg\max_{\theta} \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)] + cst(\theta),\end{aligned}$$

where the constant is the differential entropy of $p(\mathbf{z}|\mathbf{x}; \theta_t)$ which is independent of θ.

After the M-step, because $q_{t+1}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_t)$ has been held fixed for computing the new model parameters θ_{t+1}, the KL divergence $D_{\text{KL}}(q_{t+1}(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_{t+1}))$ will be non zero.

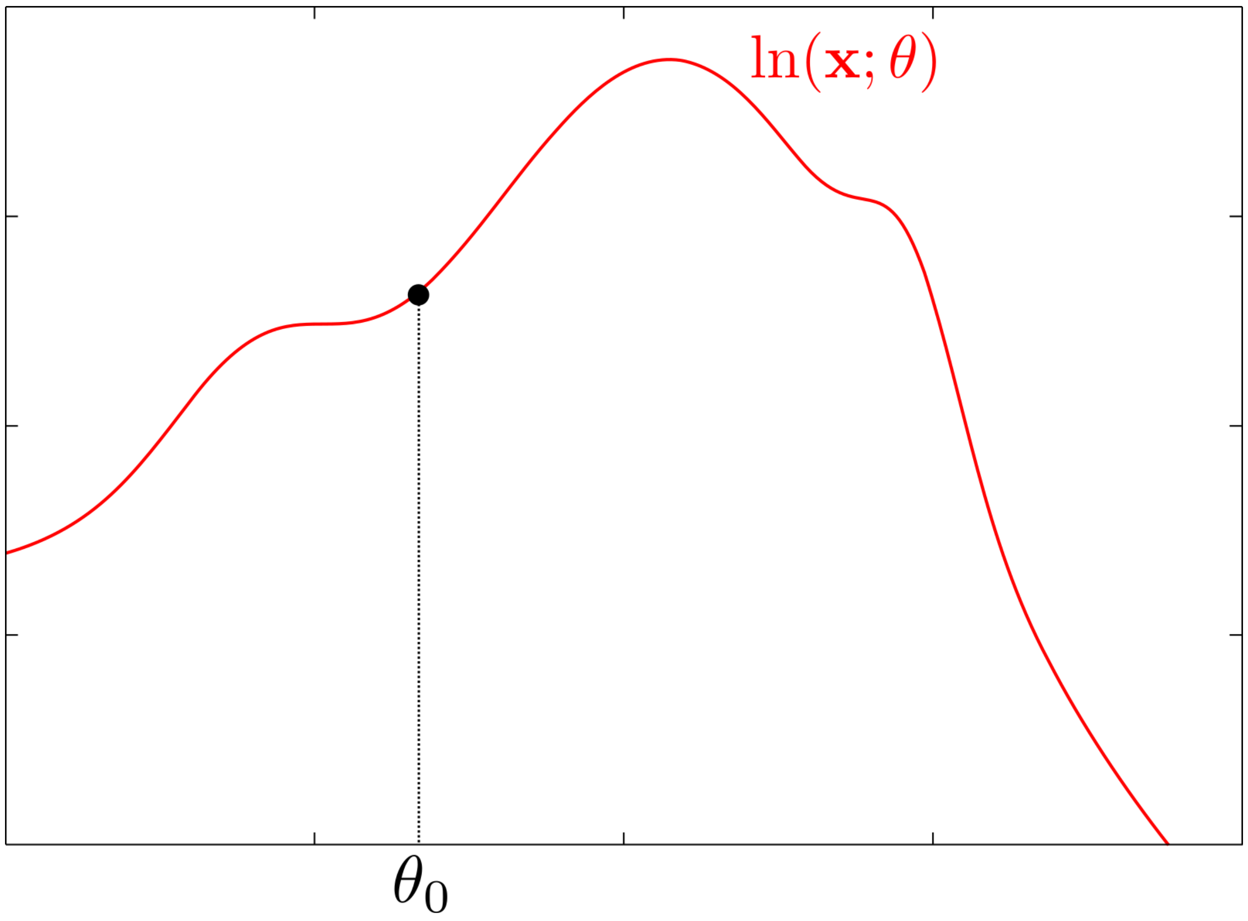
The increase in the log-marginal likelihood function is therefore greater than the increase in the ELBO, as shown below.


$\text{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta}^{\text{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{new}})$

We recall the decomposition of the log-marginal likelihood $\ln p(\mathbf{x};\theta) = \mathcal{L}(q(\mathbf{z}),\theta) + D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta))$.

Image credit: Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

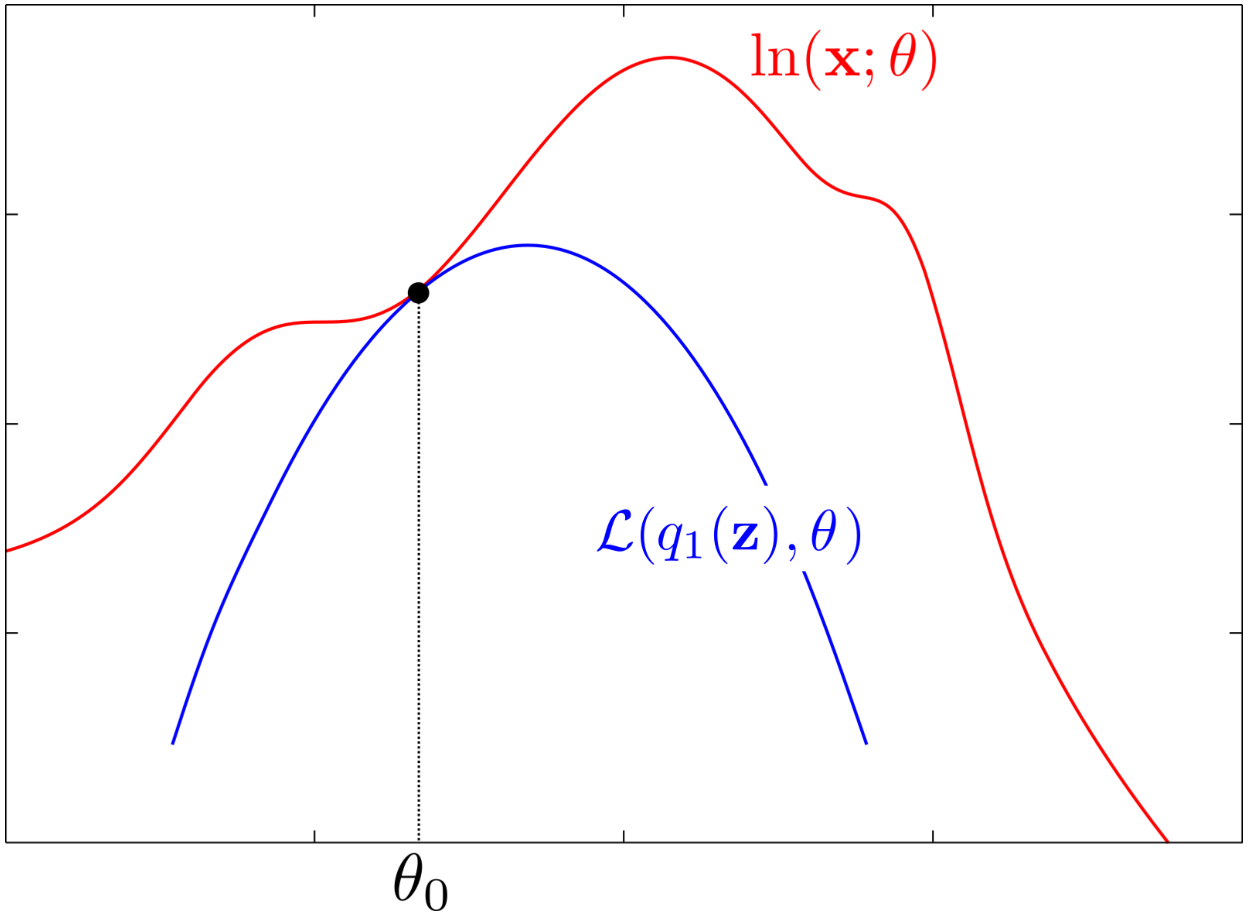
Initialize θ_0.



Iteration $t = 1$:

- E-Step:

$$q_1(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_0)$$



We have $D_{\mathrm{KL}}(q_1(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_0)) = 0$ such that $\ln p(\mathbf{x};\theta_0) = \mathcal{L}(q_1(\mathbf{z}),\theta_0)$.

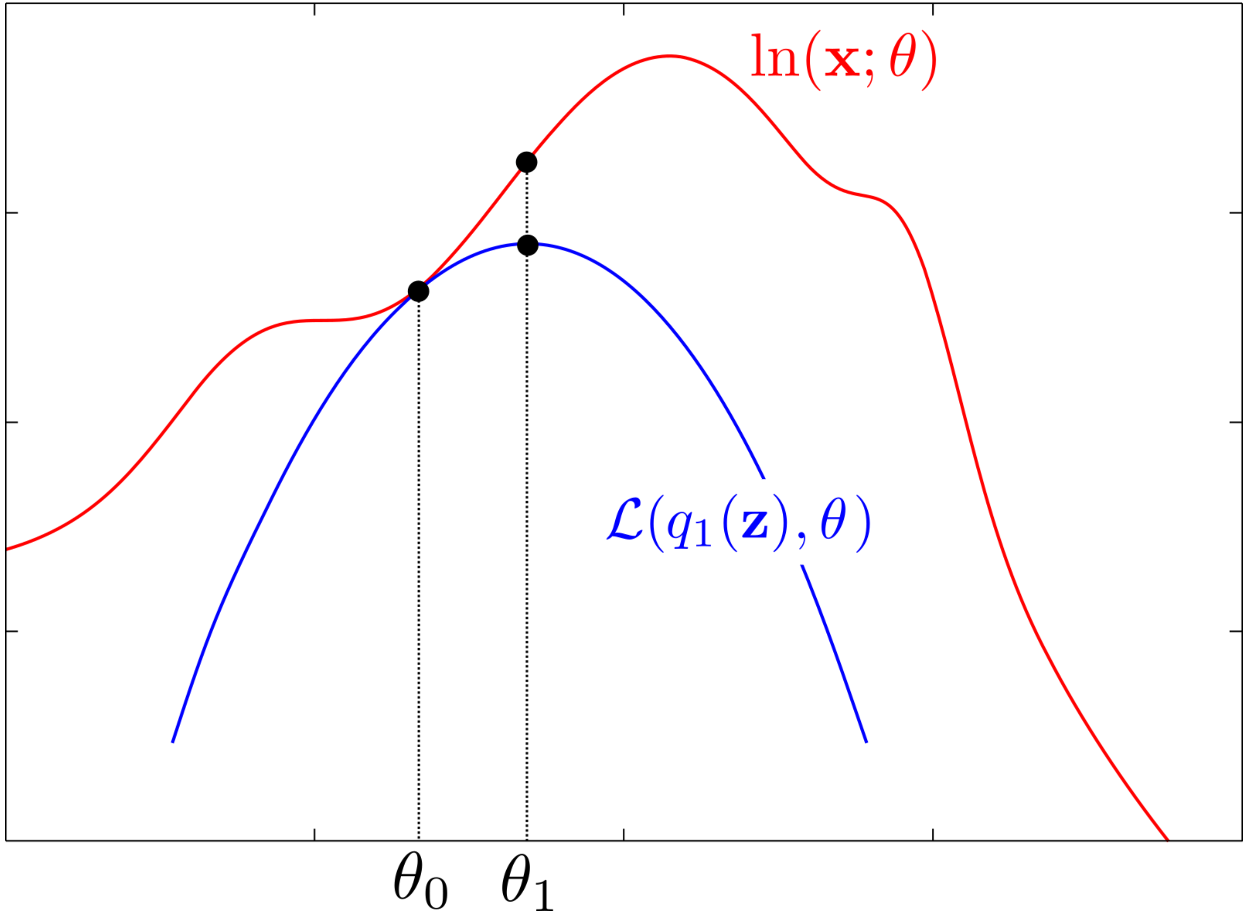
Iteration $t = 1$:

- E-Step:

$$q_1(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_0)$$

- M-Step:

$$\theta_1 = \arg\max_\theta \mathcal{L}(q_1(\mathbf{z}),\theta)$$



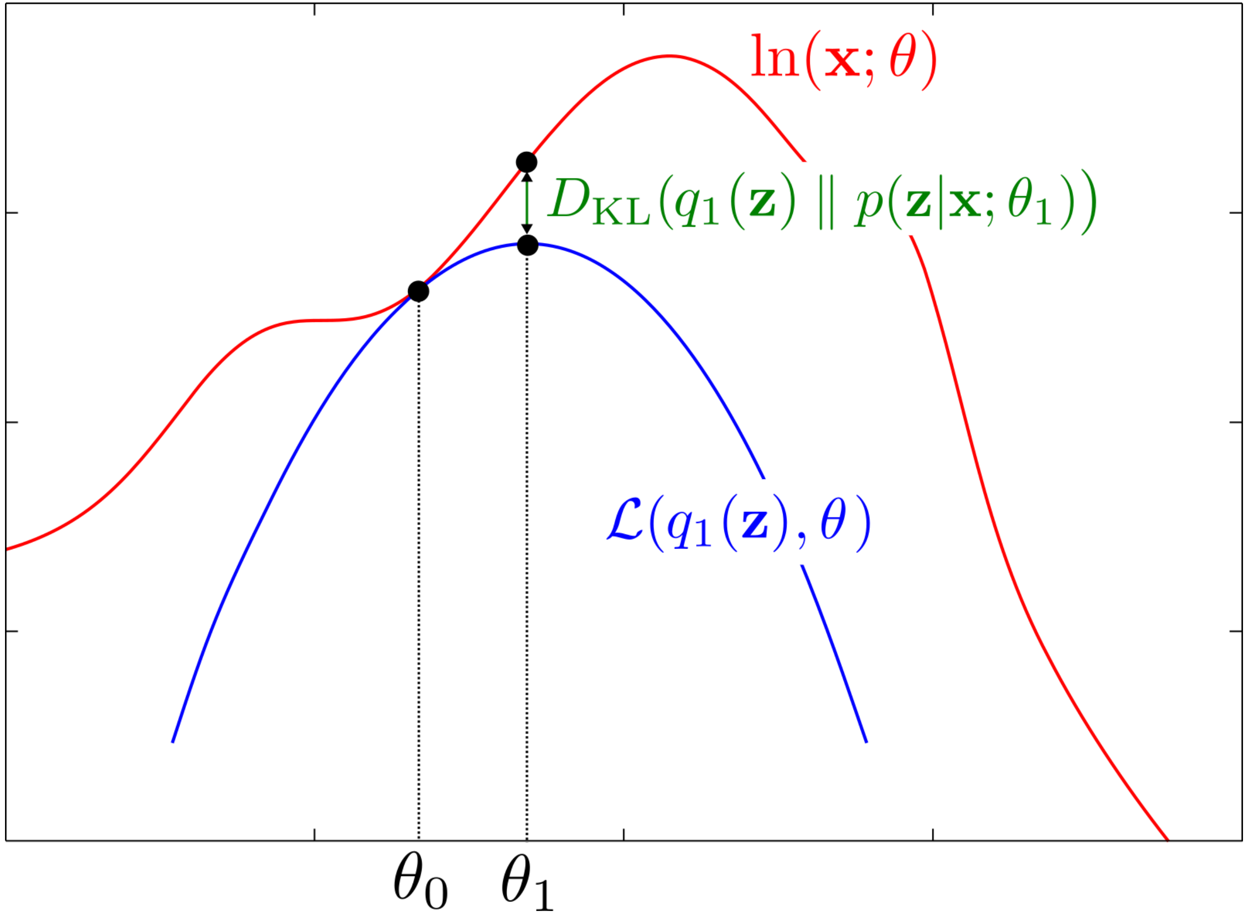
Iteration $t = 1$:

- E-Step:

$$q_1(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_0)$$

- M-Step:

$$\theta_1 = \arg\max_{\theta} \mathcal{L}(q_1(\mathbf{z}), \theta)$$

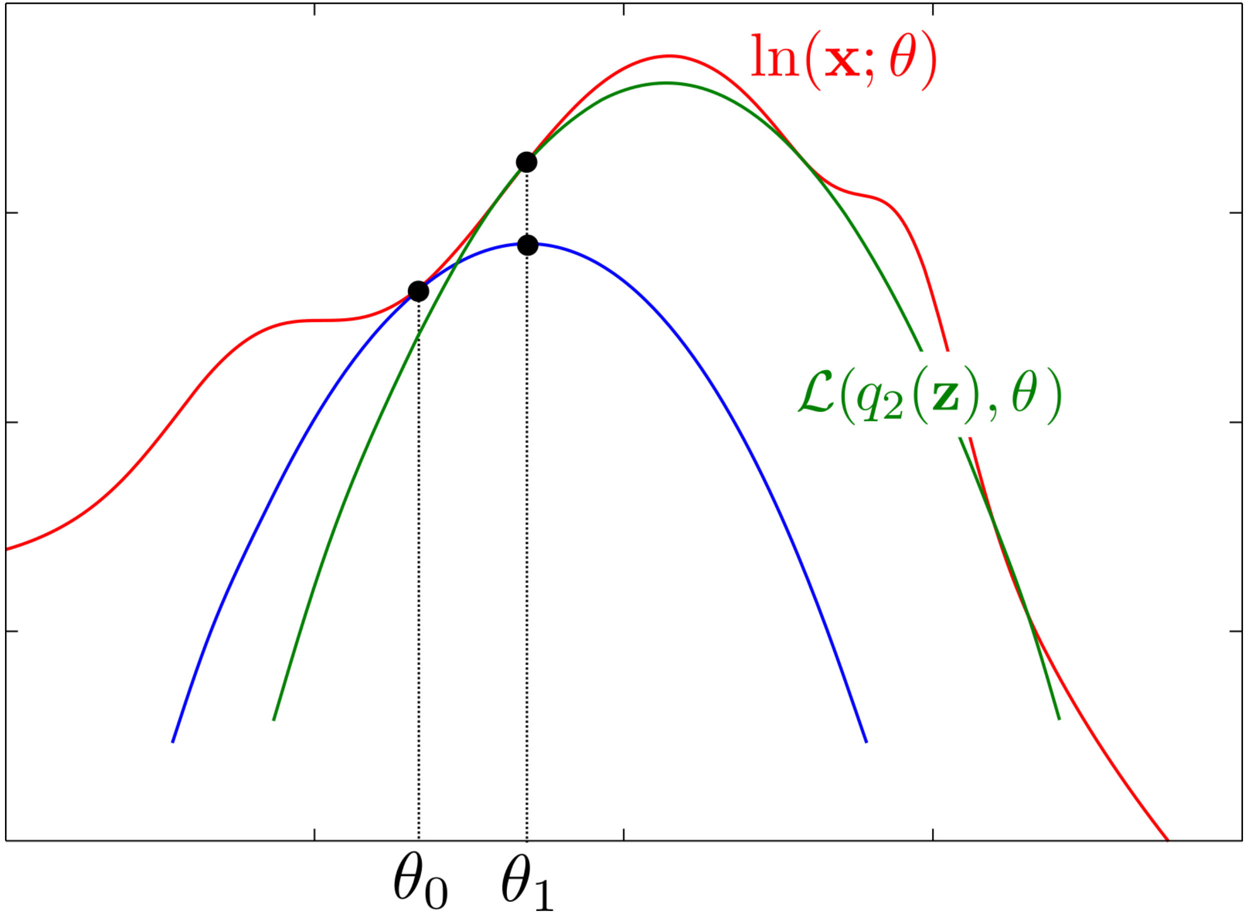

$\ln(\mathbf{x};\theta)$
$D_{\mathrm{KL}}(q_1(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_1))$
$\mathcal{L}(q_1(\mathbf{z}), \theta)$
θ_0
θ_1

We have $D_{\mathrm{KL}}(q_1(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_1)) \neq 0$.

Iteration $t = 2$:

- E-Step:

$$q_2(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_1)$$



We have $D_{\mathrm{KL}}(q_2(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta_1)) = 0$ such that $\ln p(\mathbf{x};\theta_1) = \mathcal{L}(q_2(\mathbf{z}),\theta_1)$.

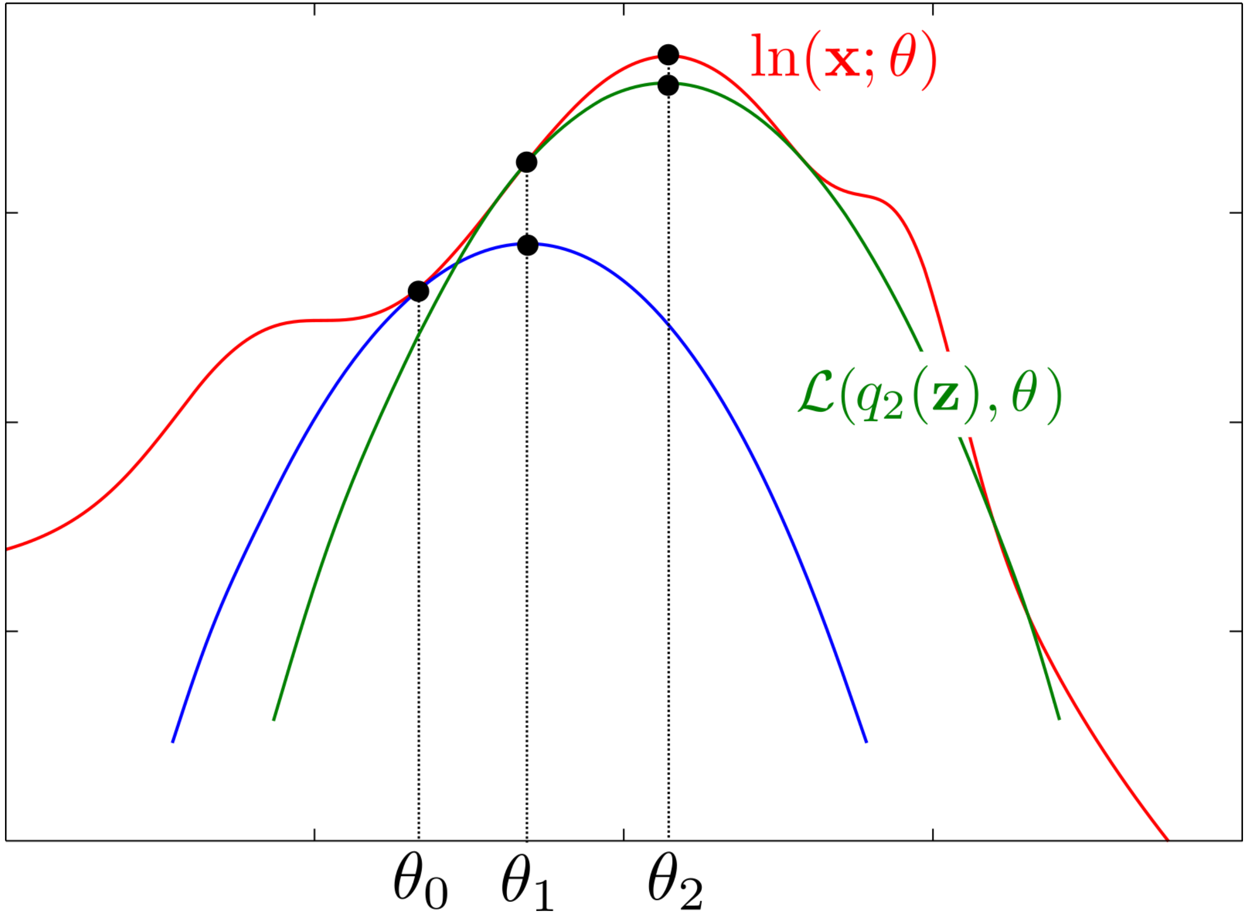
Iteration $t = 2$:

- E-Step:

$$q_2(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_1)$$

- M-Step:

$$\theta_2 = \arg\max_{\theta} \mathcal{L}(q_2(\mathbf{z}), \theta)$$



Iteration $t = 2$:

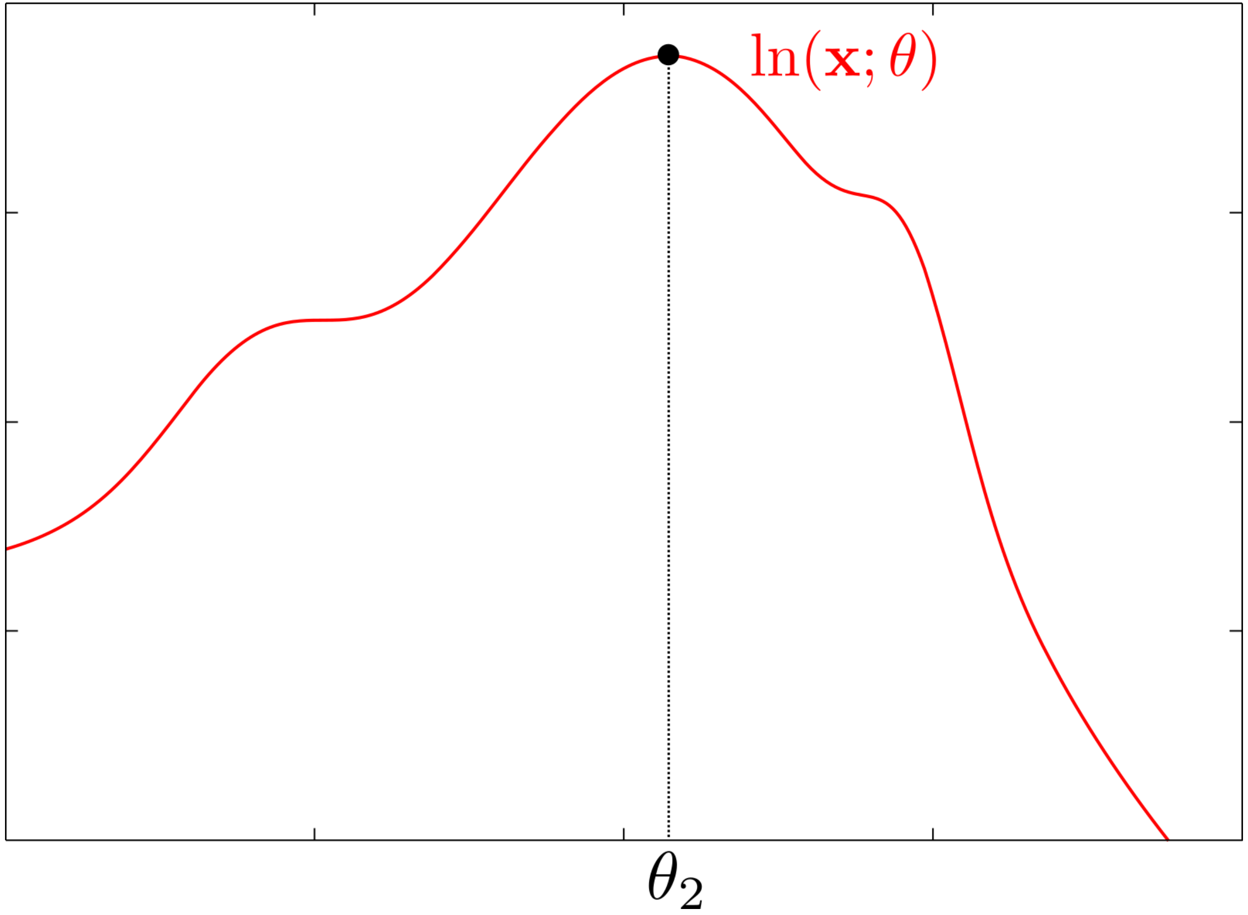
- E-Step:

$$q_2(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_1)$$

- M-Step:

$$\theta_2 = \arg\max_{\theta} \mathcal{L}(q_2(\mathbf{z}), \theta)$$

- We reached a stationary point.



Properties of the EM algorithm

- The log-marginal likelihood is monotonically increasing.

Properties of the EM algorithm

- The log-marginal likelihood is monotonically increasing.

Proof:

Using the fact that $\ln p(\mathbf{x};\theta) = \mathcal{L}(q(\mathbf{z}),\theta) + D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\theta))$ and $q_{t+1}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta_t)$ we deduce:

$$\mathcal{L}(q_{t+1}(\mathbf{z}),\theta_t) = \ln p(\mathbf{x};\theta_t),$$

$$\mathcal{L}(q_{t+1}(\mathbf{z}),\theta_{t+1}) \leq \ln p(\mathbf{x};\theta_{t+1}).$$

Moreover, by definintion of the M-step:

$$\mathcal{L}(q_{t+1}(\mathbf{z}),\theta_{t+1}) \geq \mathcal{L}(q_{t+1}(\mathbf{z}),\theta_t).$$

Putting all together we have:

$$\ln p(\mathbf{x};\theta_{t+1}) \geq \mathcal{L}(q_{t+1}(\mathbf{z}),\theta_{t+1}) \geq \mathcal{L}(q_{t+1}(\mathbf{z}),\theta_t) = \ln p(\mathbf{x};\theta_t).$$

Properties of the EM algorithm

- The log-marginal likelihood is monotonically increasing.
- The algorithm converges to a stationary point of the log-marginal likelihood.
- As the problem is generally not convex, the algorithm generally converges to a local optimum which strongly depends on the initialization.

EM algorithm summary

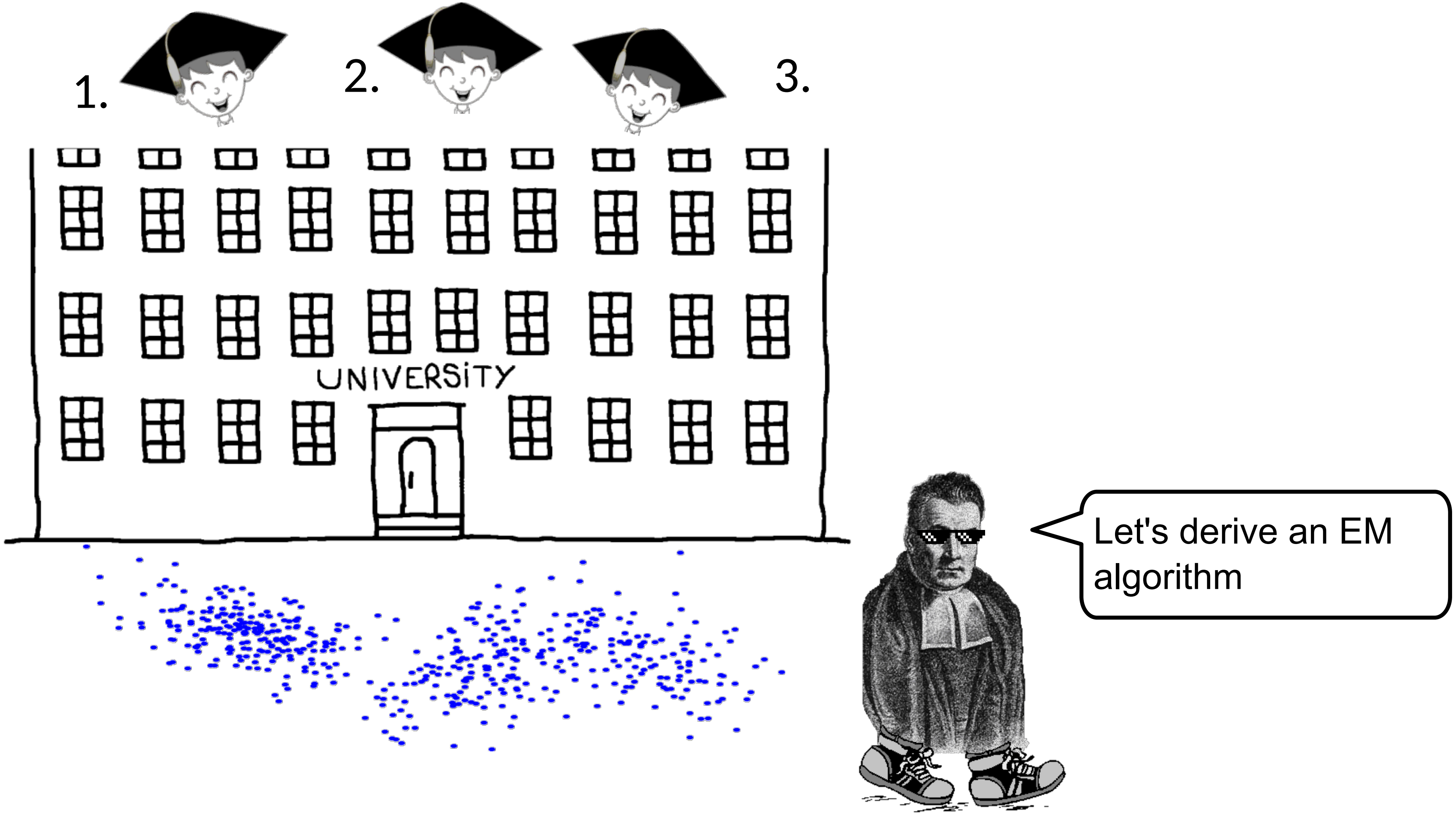
The EM algorithm can be reformulated in the space of the model parameters only.

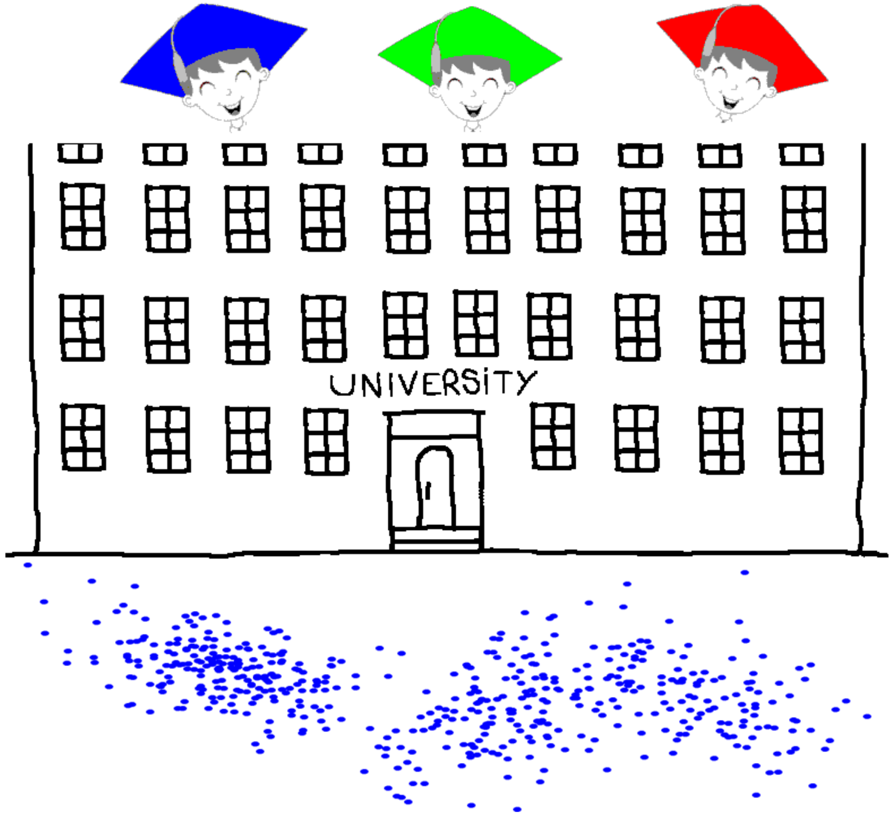
Given an initialization θ_0 of the model parameters, iterate for $t = 0 : T - 1$:

- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$;
- M-Step: $\theta_{t+1} = \arg\max_{\theta} Q(\theta, \theta_t)$.

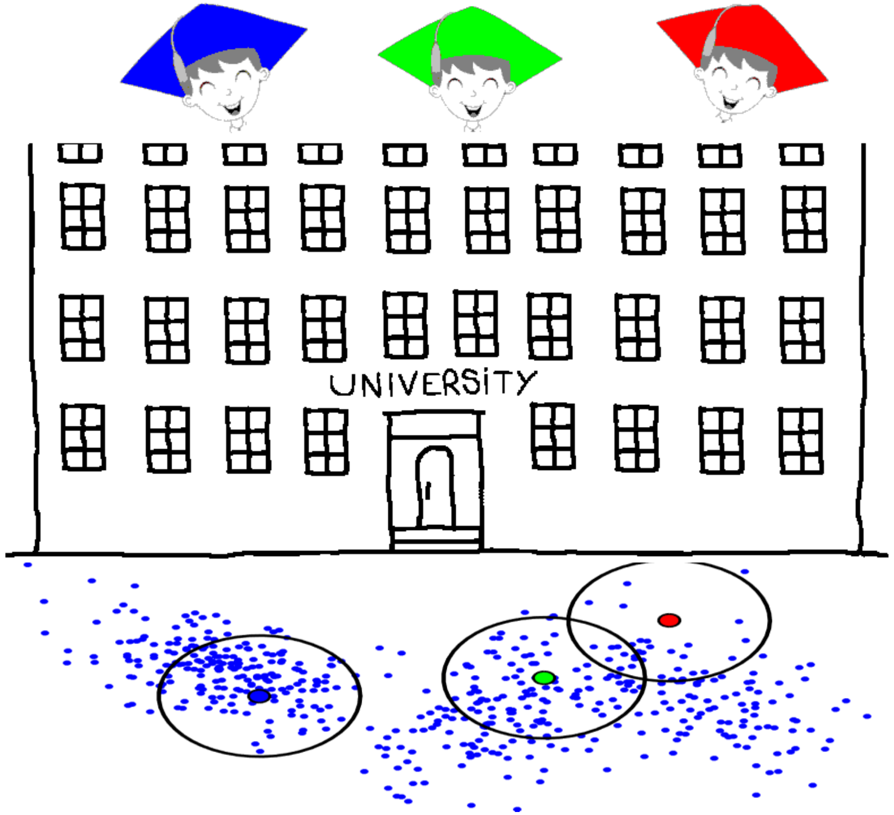
This is the recipe you should remember and use to derive an EM algorithm.

Back to the adventures of Thomas Bayes

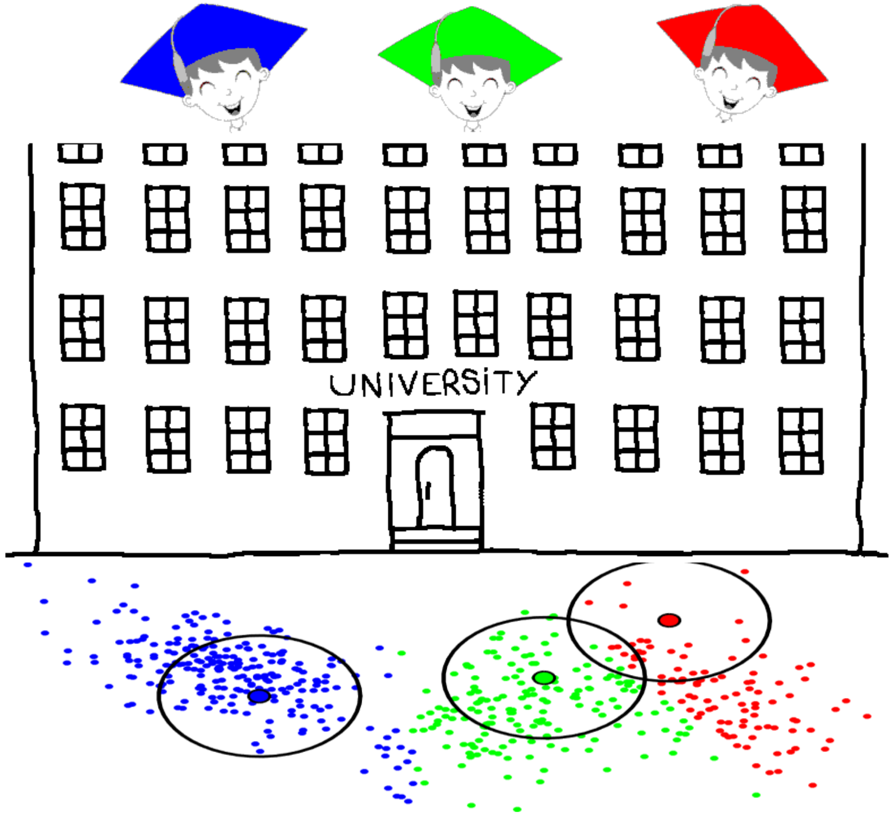




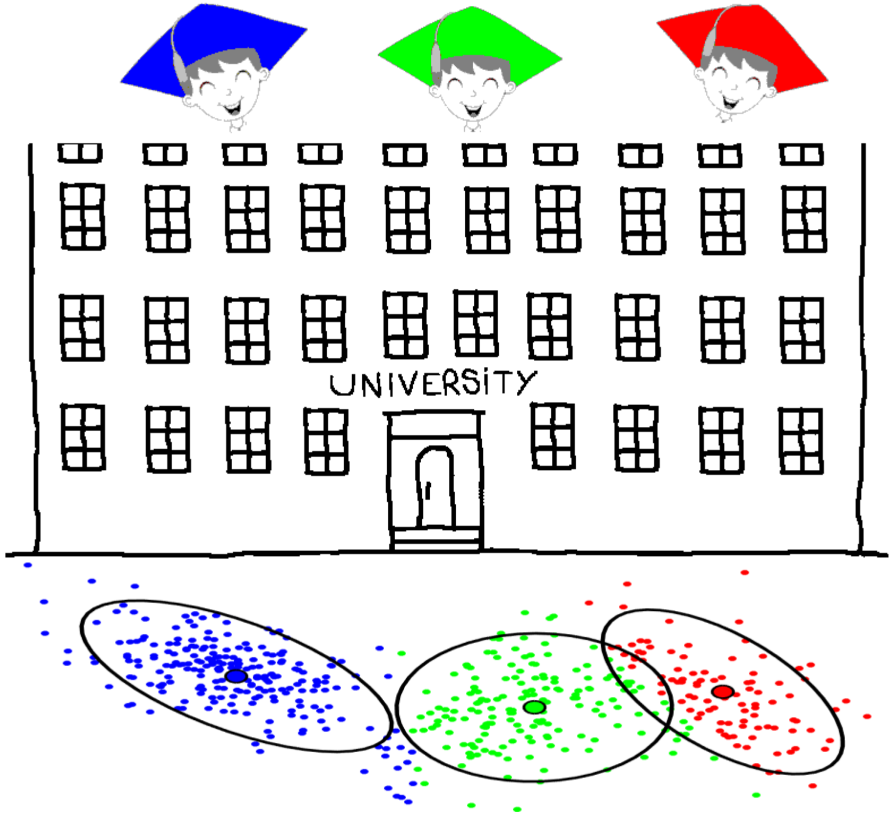
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_{\theta} Q(\theta, \theta_t)$
- Convergence



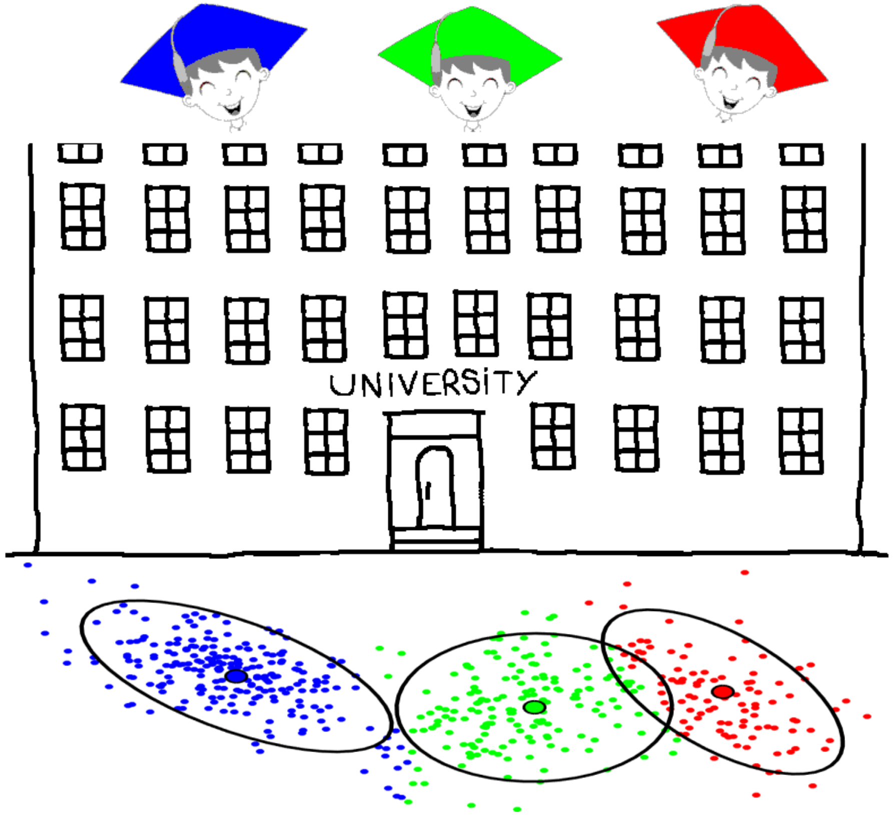
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



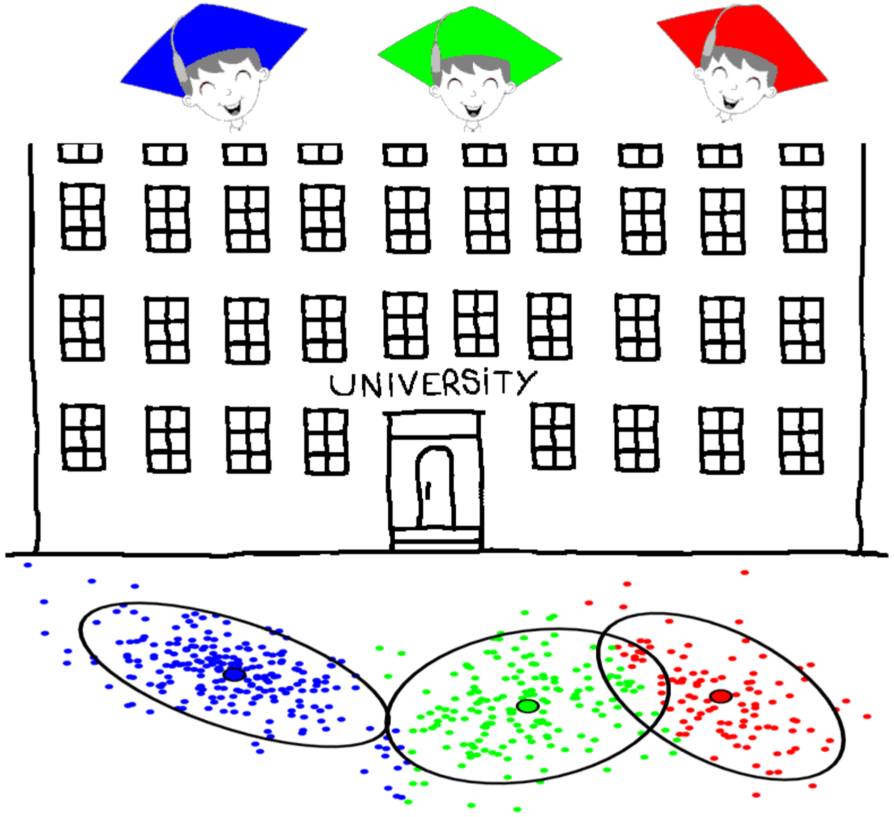
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



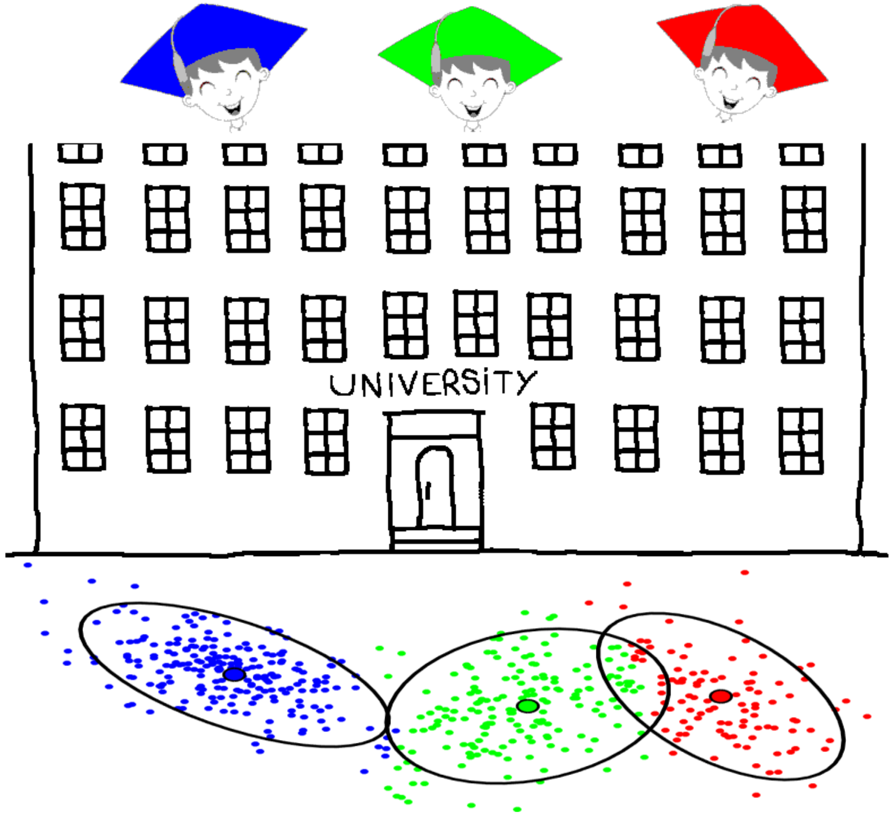
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



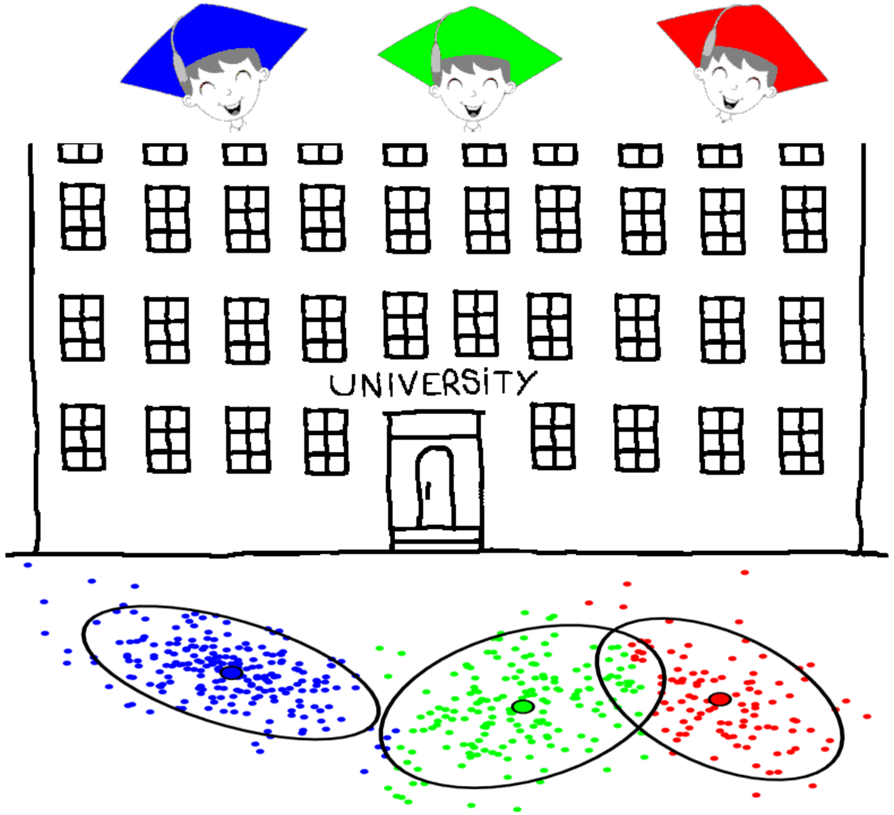
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \ Q(\theta, \theta_t)$
- Convergence



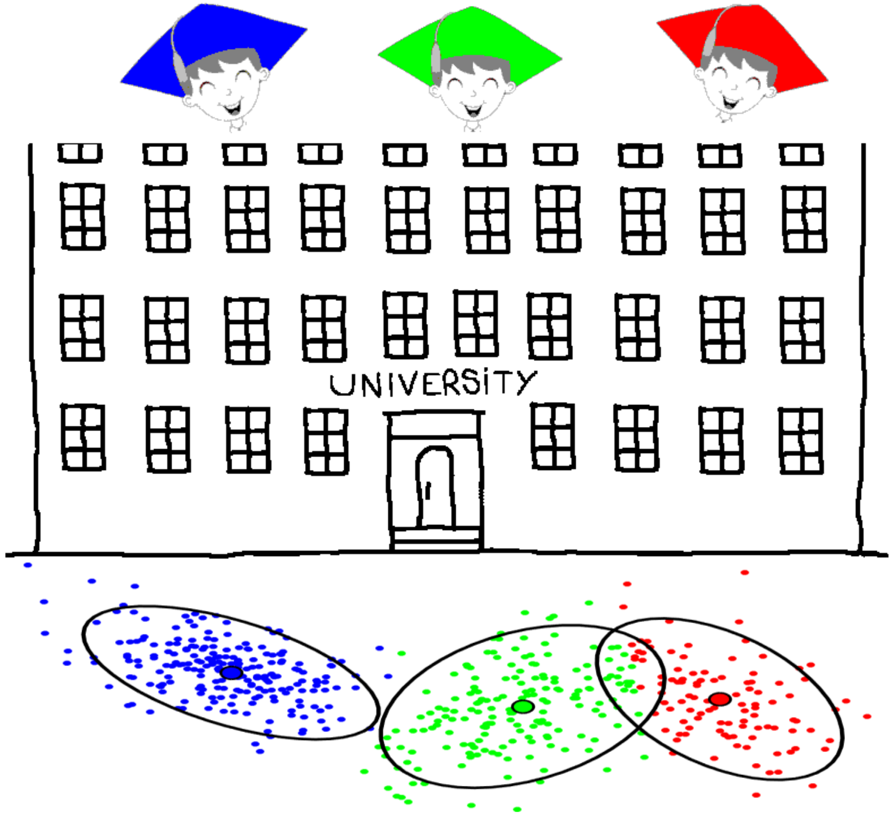
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



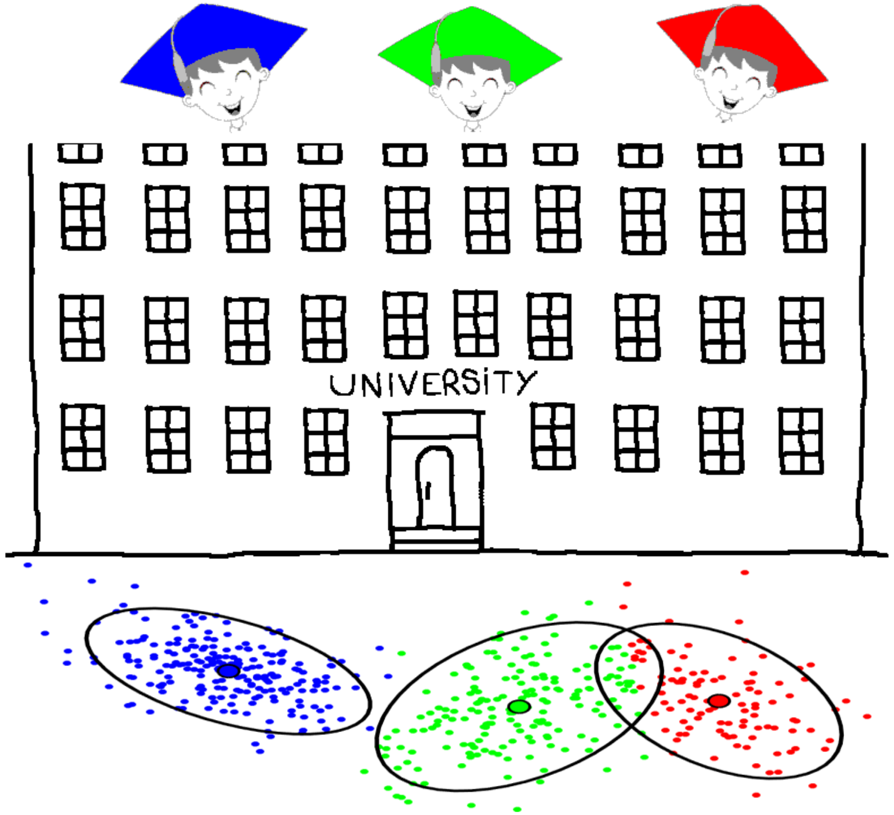
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



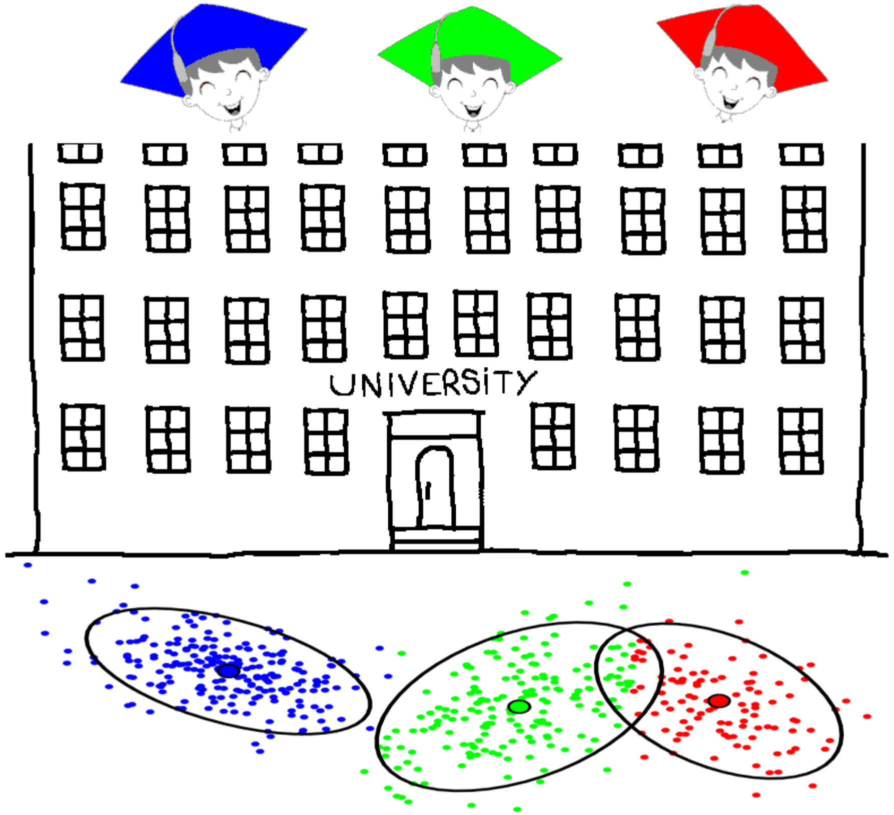
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \ Q(\theta, \theta_t)$
- Convergence



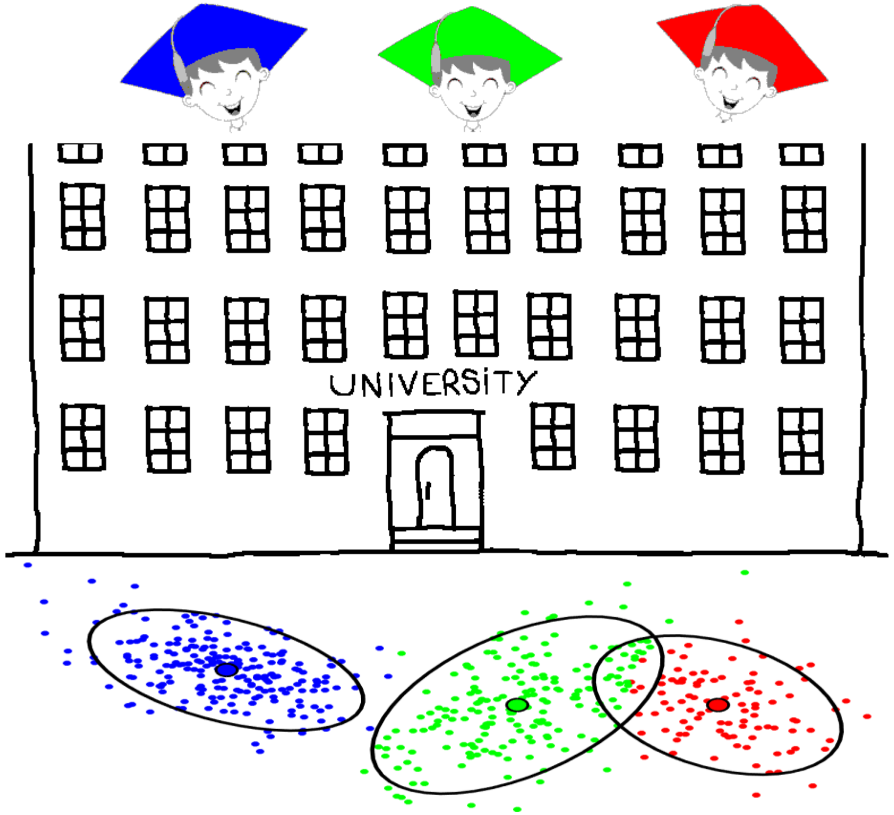
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



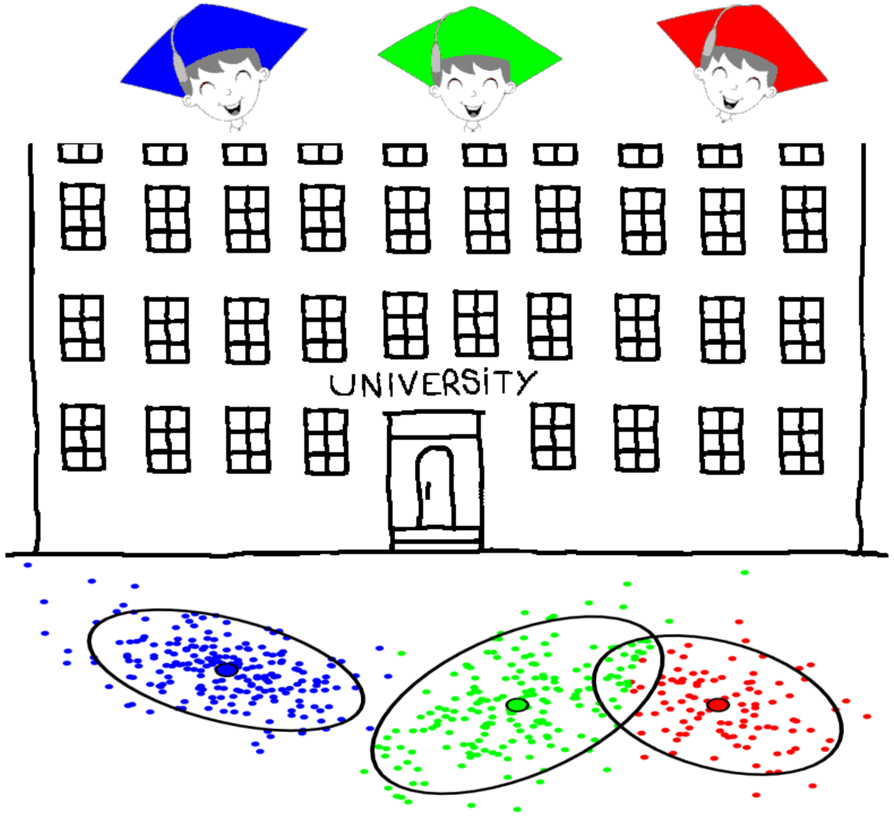
- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



- Initialization: Random "guess" for θ_0
- E-Step: $Q(\theta, \theta_t) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\ln p(\mathbf{x}, \mathbf{z}; \theta)]$
- M-Step: $\theta_{t+1} = \arg\max_\theta \; Q(\theta, \theta_t)$
- Convergence



Minus 10 points for Mr. Green, minus 5 points for the others!

Bayesian model selection

How to choose the number of clusters?

- Let $\mathcal{M}_M$ denote the model associated with $M \in \{1, \ldots, K\}$ clusters, whose parameters are denoted by $\theta_M = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{M}$.

  Our uncertainty is expressed through a prior distribution $p(\mathcal{M}_M)$.

- Given observed data $\mathbf{x}$, we wish to evaluate the posterior:

$$p(\mathcal{M}_M|\mathbf{x}) \propto p(\mathbf{x}|\mathcal{M}_M)p(\mathcal{M}_M).$$

- Assume equal probability for all models $p(\mathcal{M}_1) = \ldots = p(\mathcal{M}_K) = 1/K$.

  The interesting term is the marginal likelihood $p(\mathbf{x}|\mathcal{M}_M)$ (or $p(\mathbf{x}; \theta_M)$) which expresses the "preference" shown by the data for model $\mathcal{M}_M$.

Bayes factor

The Bayes factor is defined by

$$B_{ij} = \frac{p(\mathbf{x}|\mathcal{M}_i)}{p(\mathbf{x}|\mathcal{M}_j)}.$$

$B_{ij} > 1$ means that there is more evidence for model $\mathcal{M}_i$ than $\mathcal{M}_j$.

Bayes factor

The Bayes factor is defined by

$$B_{ij} = \frac{p(\mathbf{x}|\mathcal{M}_i)}{p(\mathbf{x}|\mathcal{M}_j)}.$$

$B_{ij} > 1$ means that there is more evidence for model $\mathcal{M}_i$ than $\mathcal{M}_j$.

In latent variable models, it may be difficult to compute the marginal likelihood and therefore the Bayes factor. We may rather use other criteria such as:

- Akaike Information Criterion
- Bayesian Information Criterion