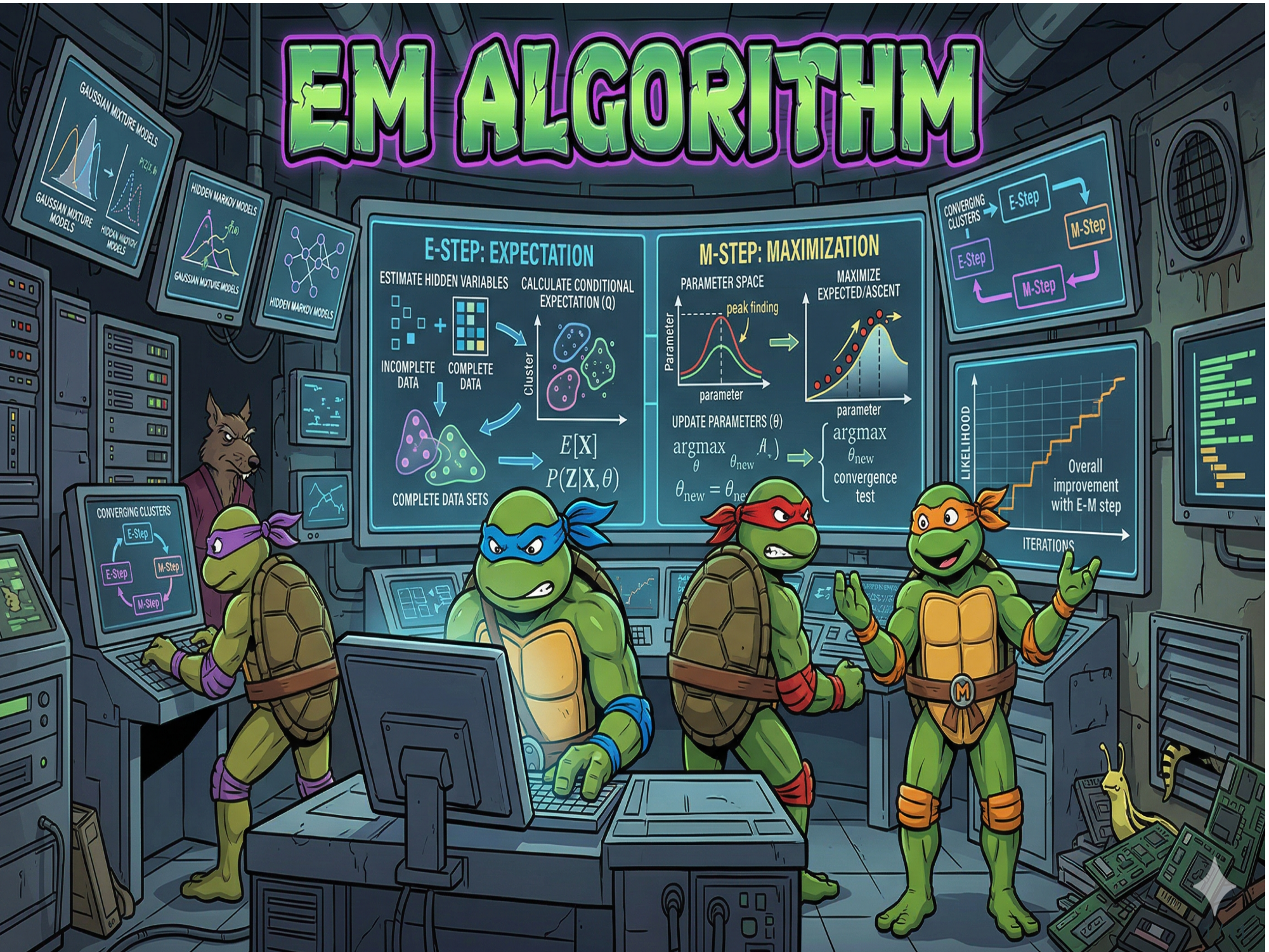
EM ALGORITHM
GAUSSIAN MIXTURE MODELS
HIDDEN MARKOV MODELS
E-STEP: EXPECTATION
ESTIMATE HIDDEN VARIABLES
CALCULATE CONDITIONAL EXPECTATION (Q)
INCOMPLETE DATA
COMPLETE DATA
Cluster
COMPLETE DATA SETS
$E[X]$
$P(Z|X,\theta)$
M-STEP: MAXIMIZATION
PARAMETER SPACE
peak finding
Parameter
parameter
MAXIMIZE EXPECTED/ASCENT
parameter
UPDATE PARAMETERS (θ)
$\theta_{new} = \theta_{ne}$
argmax
θ_{new}
convergence test
CONVERGING CLUSTERS
E-Step
M-Step
E-Step
M-Step
LIKELIHOOD
Overall improvement with E-M step
ITERATIONS
CONVERGING CLUSTERS
E-Step
M-Step
E-Step
M-Step

Finding posterior modes

- In many applications, we might be more interested in searching for the posterior mode, rather than characterizing the posterior distribution, e.g., mixture models for clustering/classification, model selection in regression, etc.
- Posterior mode is also referred to as the maximum a posteriori probability (MAP) estimate.
- This essentially reduces to a generic optimization problem that finds

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|y) = \arg\max_{\theta} p(y|\theta)p(\theta)$$

  and many standard optimization tools can be used to find $\hat{\theta}_{MAP}$.
- When there are many parameters, we may want to find the posterior mode for a subset of them, i.e., let $\theta = (\gamma, \phi)$, we may want to find

$$\hat{\phi}_{MAP} = \arg\max_{\phi} p(\phi|y) = \arg\max_{\phi} \int p(\phi, \gamma|y) d\gamma$$

- In this lecture, we will focus on a particular type of method that finds this type of marginal maxima, called the EM algorithm.

Expectation-Maximization (EM) algorithm

- The idea is to treat ϕ as the parameter and γ as missing data.
- The complete data likelihood is $L(\phi|\gamma, y) = p(y, \gamma|\phi)$
- The observed data likelihood is $L(\phi|y) = p(y|\phi) = \int p(y, \gamma|\phi) d\gamma$
- Notice here we consider (y, γ) as data thus the word *likelihood*.
- The MAP estimate is

$$\begin{aligned}\hat{\phi}_{MAP} &= \arg\max_{\phi} \log p(\phi|y) \\ &= \arg\max_{\phi} \left(\log L(\phi|y) + \log p(\phi)\right) \\ &= \arg\max_{\phi} \left(\log \int L(\phi, \gamma|y) d\gamma + \log p(\phi)\right)\end{aligned}$$

  which can be difficult to find due to the integral/sum within the first log term.
- However, given γ, it might be easy to find

$$\arg\max_{\phi} \log p(\phi, \gamma|y) = \arg\max_{\phi} \left(\log L(\phi|\gamma, y) + \log p(\phi)\right)$$

- But what value of γ should we use?

EM algorithm: implementation

Conceptually, EM algorithm iterates between:

1. impute (functions of) missing data γ or $h(\gamma)$ by $\mathbb{E}(\gamma)$ or $\mathbb{E}(h(\gamma))$, i.e., their expectation under $p(\gamma|\phi, y)$.
2. find the maximizer for $p(\phi|\gamma, y)$.

- Notice the two steps are essentially using the same conditional probabilities for a block Gibbs sampler, but instead of sampling, conditional mean and mode are used in the two steps.
- Unlike Gibbs sampling, where we look for a converged posterior distribution, here the procedures are repeated so that the parameters do not change any more (also referred to as convergence achieved, but notice the difference).

EM algorithm: implementation

EM algorithm

1. Start with some initial value ϕ^{cur}.
2. E-step: Find the form of the function

$$Q(\phi|\phi^{cur}) = \mathbb{E}_{\gamma|\phi^{cur},y}(\log p(\phi,\gamma|y)|\phi^{cur},y) = \int \log p(\phi,\gamma|y)p(\gamma|\phi^{cur},y)d\gamma$$

3. M-step: Set the new ϕ^{cur} to be

$$\arg\max_{\phi} Q(\phi|\phi^{cur})$$

4. Repeat 2 and 3 until convergence.

- It can be shown that after each step $\log p(\phi|y)$ increases monotonically.
- Thus EM algorithm finds a local maxima for $\log p(\phi|y)$

Silly example: normal model

- Consider the normal model $y_i|\mu,\sigma^2 \sim N(\mu,\sigma^2)$ for $i = 1,...,n$ with priors $\mu \sim N(\mu_0,\tau^2)$ and $p(\sigma^2) \propto \sigma^{-2}$.
- Consider the task of finding the MAP estimate for μ. We treat σ^2 as missing data.
- The complete data likelihood is

$$p(y,\sigma^2|\mu) = p(y|\mu,\sigma^2)p(\sigma^2) \propto (\frac{1}{\sigma^2})^{n/2+1}\exp(-\frac{1}{2\sigma^2}\sum_i(y_i-\mu)^2)$$

- The observed data likelihood is

$$p(y|\mu) \propto \int (\frac{1}{\sigma^2})^{n/2+1}\exp(-\frac{1}{2\sigma^2}\sum_i(y_i-\mu)^2)d\sigma^2$$

- The EM algorithm works with the posterior density

$$\begin{aligned}\log p(\sigma^2,\mu|y) &= \log p(y,\sigma^2|\mu) + \log p(\mu) + const \\ &= -(n/2+1)\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i(y_i-\mu)^2 - \\ &\quad \frac{1}{2\tau^2}(\mu-\mu_0)^2 + const\end{aligned}$$

Example: normal model

E-step: find the form of the function

$$Q(\mu|\mu^{cur}) = \mathbb{E}_{\sigma^2|\mu^{cur},y}(\log p(\sigma^2, \mu|y)|\mu^{cur}, y)$$
$$= -(n/2+1)\mathbb{E}(\log(\sigma^2)) - \mathbb{E}(\frac{1}{2\sigma^2})\sum_i (y_i - \mu)^2 - \frac{1}{2\tau^2}(\mu - \mu_0)^2 + const$$

which consists of finding $\mathbb{E}(\log \sigma^2)$ and $\mathbb{E}(1/\sigma^2)$.

- Notice that $\sigma^2|\mu, y \sim$ InvGamma$(n/2, \sum_i (y_i - \mu)^2/2)$. So both these expectations can be evaluated, but let us hold on this thought for now.

Example: normal model

M-step: find the $\hat{\mu} = \arg\max_\mu Q(\mu|\mu^{cur})$. Taking the first order derivative and setting it to 0,

$$\frac{d}{d\mu}Q(\mu|\mu^{cur}) = -\mathbb{E}(\frac{1}{\sigma^2})n\mu + \mathbb{E}(\frac{1}{\sigma^2})\sum_i y_i - \frac{1}{\tau^2}\mu + \frac{\mu_0}{\tau^2},$$

the new $\hat{\mu}$ is

$$\hat{\mu} = \frac{\sum_i y_i \mathbb{E}(1/\sigma^2) + \mu_0/\tau^2}{n\mathbb{E}(1/\sigma^2) + 1/\tau^2}$$

- We only need to evaluate $\mathbb{E}(1/\sigma^2)$ in order to compute $\hat{\mu}$,
- Since $1/\sigma^2|\mu, y \sim \text{Gamma}(n/2, \sum_i (y_i - \mu)^2/2)$, we have

$$\mathbb{E}(1/\sigma^2|\mu^{cur}, y) = \frac{n}{\sum_i (y_i - \mu^{cur})^2}.$$

- Plugging in the expectation, the M-step updates μ^{cur} with $\hat{\mu}$:

$$\hat{\mu} = \frac{\frac{\sum_i y_i}{\sum_i (y_i - \mu^{cur})^2/n} + \mu_0/\tau^2}{\frac{n}{\sum_i (y_i - \mu^{cur})^2/n} + 1/\tau^2}$$

The EM algorithm then proceeds by iteratively update μ using the formula above.

Example: normal mixture model

- Consider the normal mixture model $p(y_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega}) = \sum_{k=1}^{K} \omega_k N(y_i; \mu_k, \sigma_k^2)$.
- Consider the missing data $z_{ik} \in \{0, 1\}^K$ such that $z_{ik} = 1$ if y_i belongs to group k and 0 otherwise and the following augmented model

$$(z_{i1}, ..., z_{iK}) \sim Multi(1, \boldsymbol{\omega})$$

and

$$y_i|z_{i1}, ..., z_{iK} \sim \prod_{k=1}^{K} N(y_i; \mu_k, \sigma_k^2)^{z_{ik}}$$

- This is equivalent to the categorical missing data representation we looked at before, but makes it easier for us to design an EM algorithm.
- The complete data likelihood is

$$p(y, z|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \omega_k^{z_{ik}} N(y_i; \mu_k, \sigma_k)^{z_{ik}}$$

- The observed data likelihood is

$$p(y|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \omega_k N(y_i; \mu_k, \sigma_k)$$

Example: normal mixture model

- E-step: Let $\phi = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega})$, find the form of the function

$$Q(\phi|\phi^{cur}) = \mathbb{E}_{z|\phi^{cur},y}(\log p(\phi, z|y)|\phi^{cur}, y) = \int \log p(\phi, z|y) p(z|\phi^{cur}, y) dz$$

where

$$\log p(\phi, z|y) = \log \Big(\prod_{i=1}^{n} \prod_{k=1}^{K} \omega_k^{z_{ik}} N(y_i; \mu_k, \sigma_k^2)^{z_{ik}} p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega}) \Big)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \big(\log \omega_k + \log N(y_i; \mu_k, \sigma_k^2) \big) + \log p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega})$$

- The only term that involves the missing data is z_{ik}, so

$$Q(\phi|\phi^{cur}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}(z_{ik}|\phi^{cur}, y) \big(\log \omega_k + \log N(y_i; \mu_k, \sigma_k^2) \big) + \log p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega})$$

where the expectation is available in closed form due to the Multinomial conjugacy,

$$\mathbb{E}(z_{ik}|\phi^{cur}, y) = \frac{\omega_k^{cur} N(y_i; \mu_k^{cur}, (\sigma_k^{cur})^2)}{\sum_{k'} \omega_{k'}^{cur} N(y_i; \mu_{k'}^{cur}, (\sigma_{k'}^{cur})^2)}$$

Example: normal mixture model

- M-step: Set the new ϕ^{cur} to be

$$\hat{\phi} = \arg\max_{\phi} Q(\phi|\phi^{cur})$$

- If we let $p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\omega}) \propto 1$ for simplicity, and denote $\hat{z}_{ik} = \mathbb{E}(z_{ik}|\phi^{cur}, y)$,

$$Q(\phi|\phi^{cur}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik}\big(\log\omega_k + \log N(y_i; \mu_k, \sigma_k^2)\big)$$

- The optimization task becomes finding

$$\arg\max_{(\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\sigma}^2)} \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik}\big(\log\omega_k - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_i-\mu_k)^2\big)$$

- By setting the first order derivative to 0, we can get the following updates, where the last line is due to the fact that for any value of $\hat{\sigma}$, the same $\hat{\mu}$ maximizes the sum above.

$$\hat{\omega}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ik}}{n} \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ik}y_i}{\sum_{i=1}^{n}\hat{z}_{ik}}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n}\hat{z}_{ik}(y_i-\hat{\mu}_k)^2}{\sum_{i=1}^{n}\hat{z}_{ik}}$$