CSE/ISyE 6740
Computational Data Analysis

EM Algorithm and Convex Functions

09/08/2025

Kai Wang, Assistant Professor in Computational Science and Engineering
kwang692@gatech.edu

Georgia Tech

Outline

- **Unsupervised Learning**
  - Density estimation
    - Gaussian mixture models (GMMs)
      - Probability density function of mixture of Gaussians
      - Expectation-Maximization (EM) algorithm
      - Mathematical meaning of the EM algorithm
      - Convex/concave function and Jensen inequality
      - Expectation step as a lower bound
      - EM algorithm review
- **Supervised Learning (Wednesday)!**

Georgia Tech

EM (Expectation-Maximization) Algorithm

- Associate each data and each component with a τ_k^i
- Initialized $(\pi_k, \mu_k, \Sigma_k), k = 1,2,\dots,K$
- Iterate the following two steps until convergence:
  - **Expectation step (E-step)**: update τ_k^i given the current (π_k, μ_k, Σ_k)

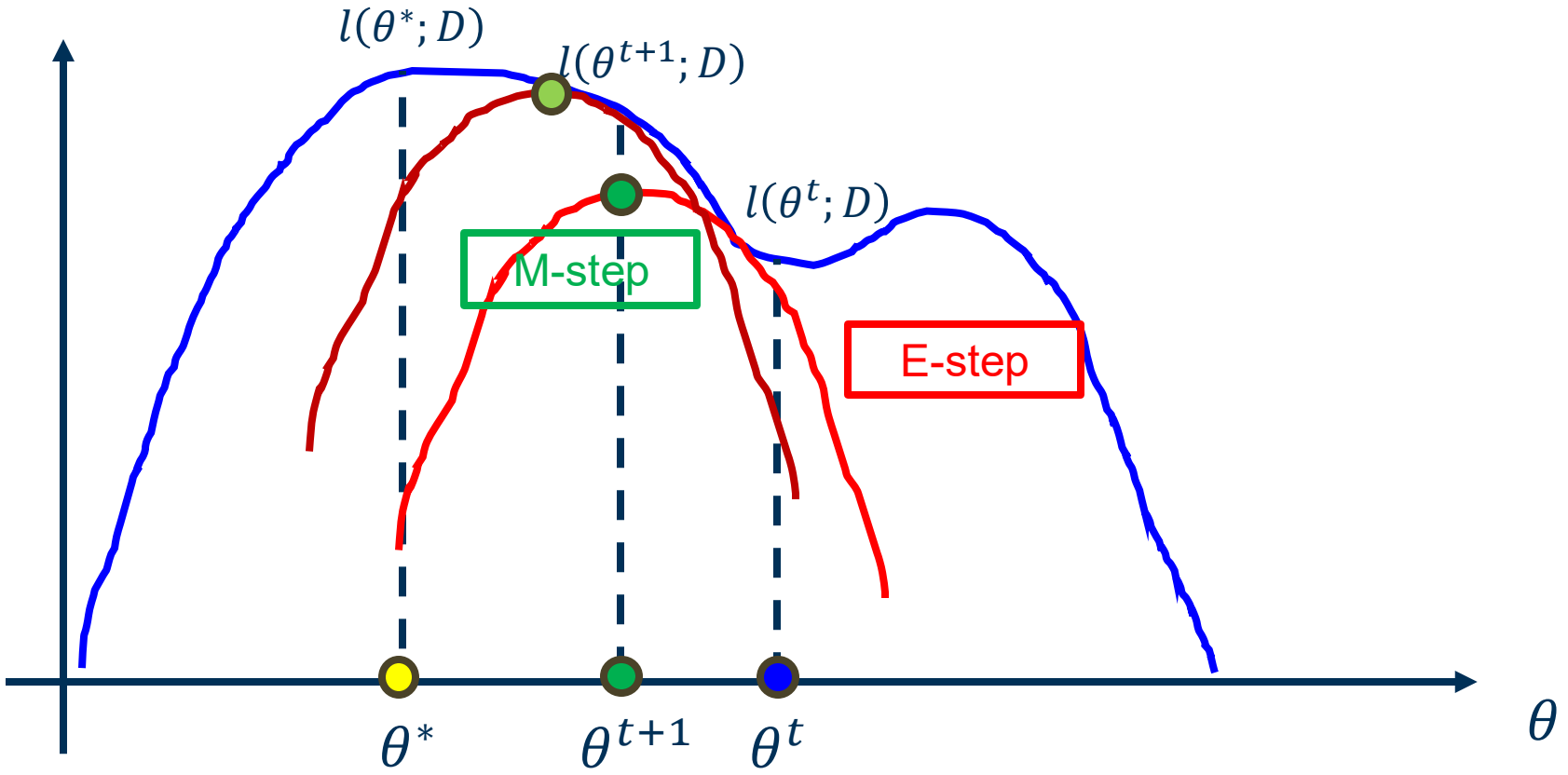
$$\tau_k^i = p(z^i = k \mid D, \mu, \Sigma) = \frac{\pi_k \, \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \, \mathcal{N}(x|\mu_{k'}, \Sigma_{k'})}, \qquad \forall k \in \{1,2,\dots,K\}; i \in \{1,2,\dots,n\}$$

  - **Maximization step (M-step)**: update (π_k, μ_k, Σ_k) given τ_k^i

$$\pi_k = \frac{\sum_i \tau_k^i}{n}, \qquad \mu_k = \frac{\sum_i \tau_k^i x^i}{\sum_i \tau_k^i}$$

$$\Sigma_k = \frac{\sum_i \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^\top}{\sum_i \tau_k^i}, \qquad \forall k \in \{1,2,\dots,K\}$$

EM Graphically



Details of EM

- We intend to learn the parameters that maximizes the log-likelihood of the data

$$l(\theta; D) = \log \prod_{i=1}^{n} \left(\sum_{Z^i=1}^{K} p(x^i, z^i|\theta) \right)$$

- **Expectation step (E-step)**: what do we take expectation over?

$$l(\theta; D) \geq l(\theta, D; q) = \mathbb{E}_{z^1, z^2, \ldots, z^n \sim q} \left[\log \prod_{i=1}^{n} p(x^i, z^i|\theta) \right] + H(q)$$

Why? **Jensen inequality!!**

- **Maximization step (M-step)**: how to maximize?

$$\theta^{t+1} = argmax_\theta l(\theta, D; q)$$

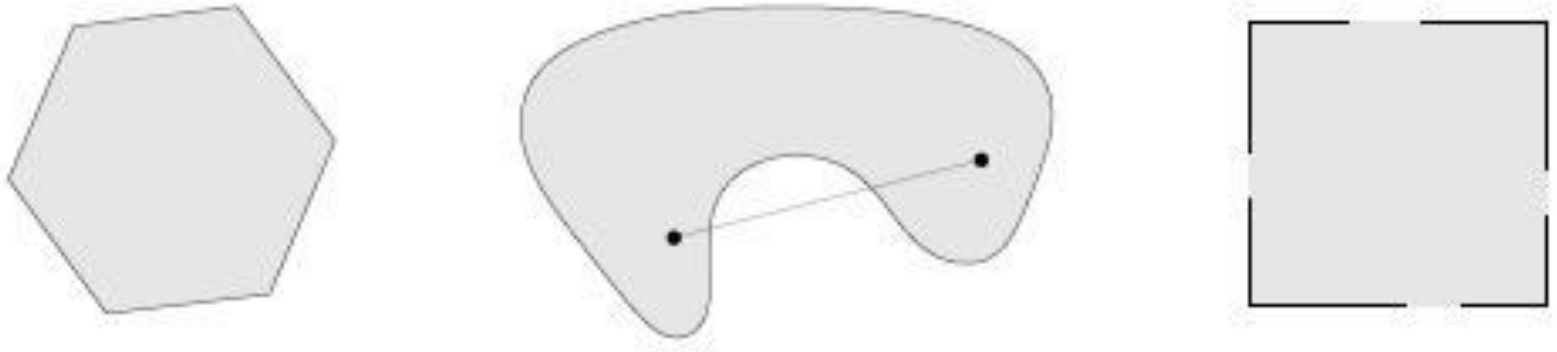
Convex / Concave Function & Jensen Inequality

Convex Sets

- **Definition**: A set A is convex, if for every $0 \leq \alpha \leq 1$ it satisfies:
  - $\forall x, y \in A \Rightarrow \alpha x + (1 - \alpha) y \in A$

- The line segment between any two points is also in the set.

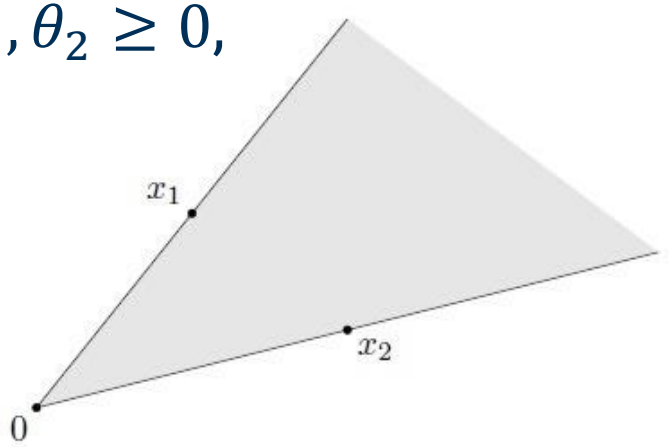
- Examples of convex and non-convex sets



Common Convex Sets

- **Cones**: a set C is a convex cone, if for any $\mathrm{x}_1, \mathrm{x}_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have:

$$\theta_1 x_1 + \theta_2 x_2 \in C$$

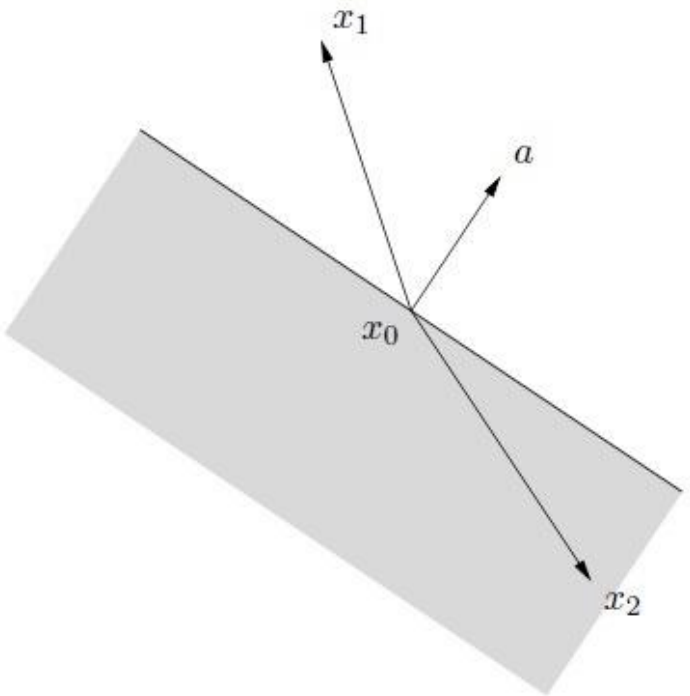


- **Hyperplanes and halfspaces**:
  - A hyperplane is

$$\{x | a^{\mathsf{T}}(x - x_0) = 0, a \neq 0\}$$

  - A halfspace is

$$\{x | a^{\mathsf{T}}(x - x_0) \leq 0, a \neq 0\}$$



Common Convex Sets

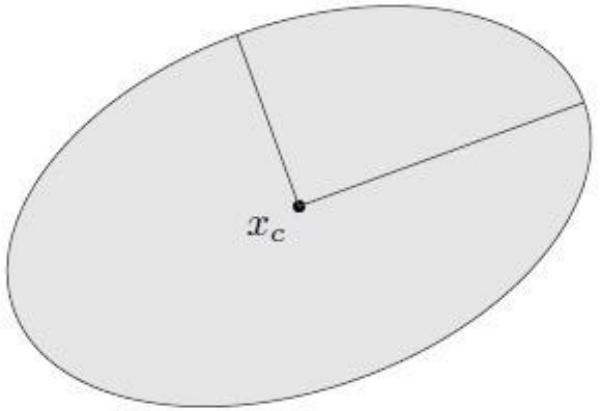
- **Euclidean balls**: A Euclidean ball has the form

$$B(x_C, r) = \{x \mid \|x - x_c\|_2 \leq r\}$$

- **Ellipsoids**:

$$E = \{x \mid (x - x_c)^\top P^{-1}(x - x_c) \leq 1\}$$

  - The eigenvectors and eigenvalues determine the direction and shape of the semi-axes

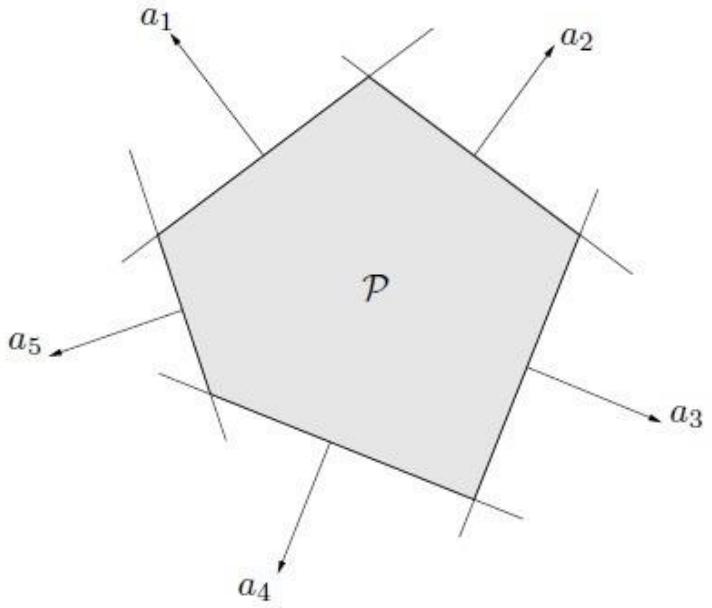


Common Convex Sets

- **Polyhedra**: intersection of a finite set of halfspaces/hyperplances

$$P = \{x \mid a_j^\top x \leq b_j, j = 1,2,\dots,m, \qquad c_j^\top x = d_j, j = 1,2,\dots,p\}$$

- It is defined by as the solution set of a finite number of linear equalities and inequalities



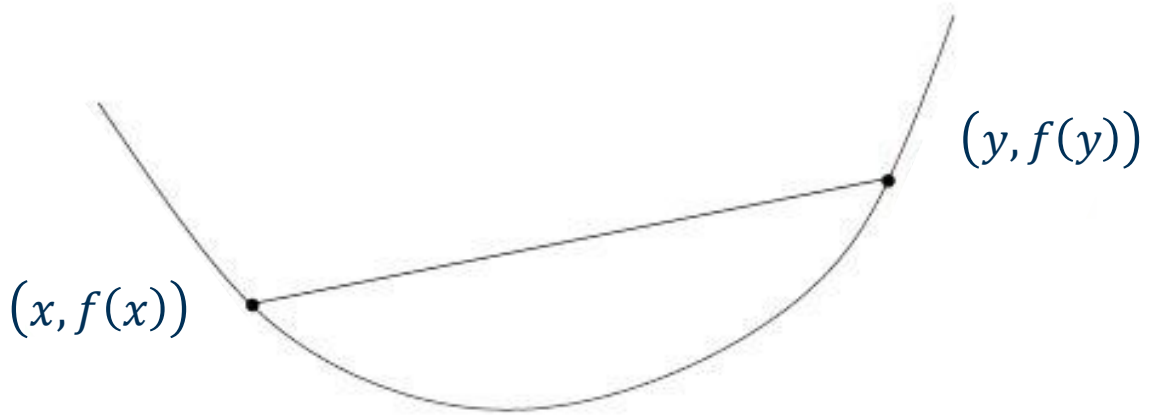
Georgia Tech

Convex Functions

- **Definition**: A function $f: \mathbb{R}^n \to \mathbb{R}$ is **convex** if the domain $\mathbf{dom}\ f \subset \mathbb{R}^n$ is a convex set, and if for all $x, y \in \mathbf{dom}\ f$ and $0 \leq \theta \leq 1$, we have:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies **above** the graph of f

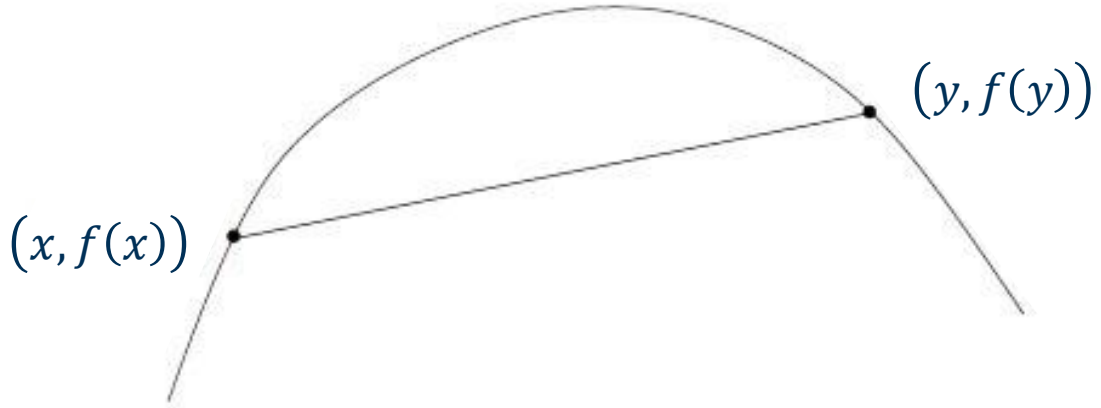


Concave Functions

- **Definition**: A function $f: \mathbb{R}^n \to \mathbb{R}$ is **concave** if the domain $\mathbf{dom}\ f \subset \mathbb{R}^n$ is a convex set, and if for all $x, y \in \mathbf{dom}\ f$ and $0 \leq \theta \leq 1$, we have:

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y)$$

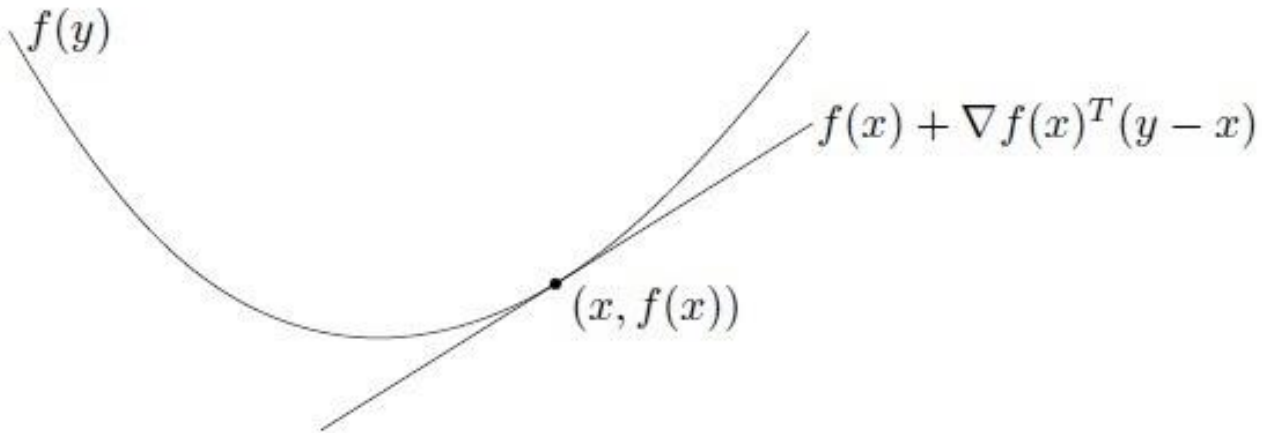
- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies **below** the graph of f



First-order Conditions

- If f is differentiable, another way to characterize it is the first-order condition:
  - f is convex iff
    - $\mathbf{dom}\ f$ is convex
    - $f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y - x)$
  - Holds for all $x, y \in \mathbf{dom}\ f$

- Geometrically, this means that the tangent line of f at point x lies below the function:



Second-order Conditions

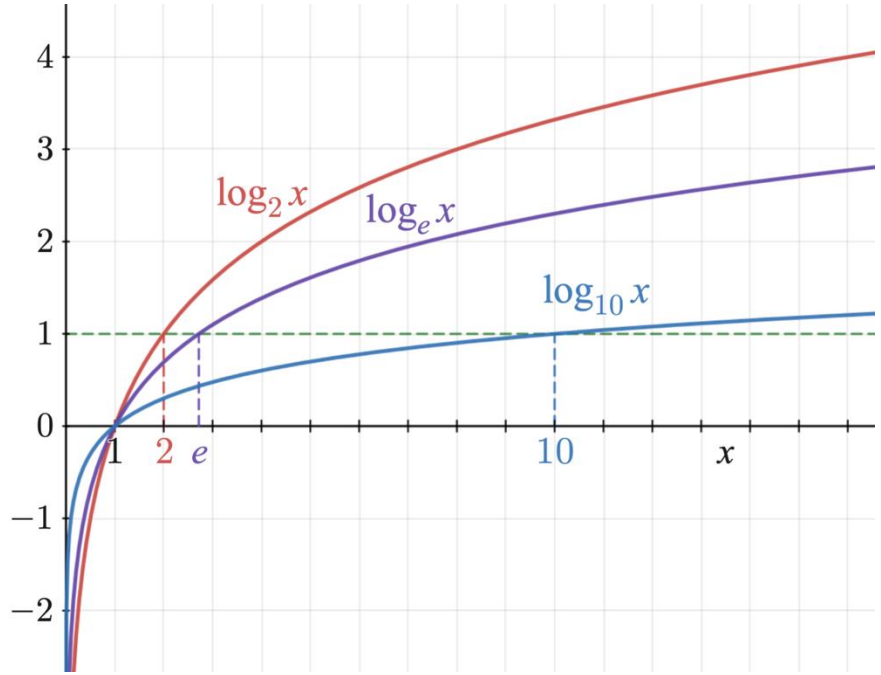
- If f is twice differentiable, the second-order condition is:
  - f is convex iff
    - $\mathbf{dom}\ f$ is convex
    - $\nabla^2 f(x) \succcurlyeq 0.$ ($A \succcurlyeq 0$, i.e., A is positive semidefinite PSD iff $y^\top A y \geq 0\ \forall y$)
  - Holds for all $x \in \mathbf{dom}\ f$
  - In other words, the Hessian is positive semidefinite.

- Geometrically, the graph of the function has positive (upward) curvature at every point.

- E.g., $f(x) = \frac{1}{2} x^\top A x$, where $A \succcurlyeq 0$, is convex in x

Examples

- **Exponential**: e^{ax} for every $a \in \mathbb{R}$
- **Powers**: x^a is convex on $\mathbb{R}_{++}$ when $a \geq 1$ or $a \leq 0$; x^a is concave for $0 \leq a \leq 1$
- **Powers of absolute value**: $|x|^p$ for $p \geq 1$
- **Logarithm**: $\log x$ is concave on $\mathbb{R}_{++}$
- **Negative entropy**: $x \log x$ is convex
- **Norms**: all norms are convex (nonnegative; homogeneous; triangular inequality)
- **Max function**: $f(x) = \max\{x_1, x_2, \ldots, x_n\}$ is convex
- **Log-determinant**: $f(X) = \log(\det X)$ is convex for all positive definite matricies.



Georgia Tech

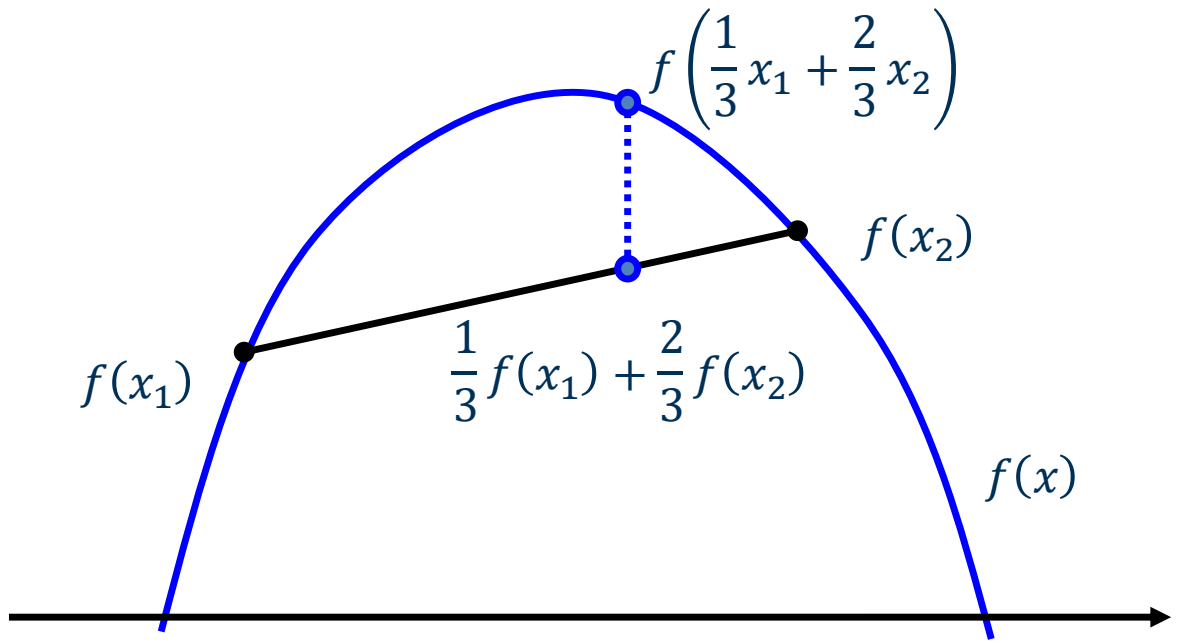
Jensen's Inequality

- For concave function $f(x)$

$$f(\sum_i a_i x_i) \geq \sum_i a_i f(x_i), \text{ where } \sum_i a_i = 1, a_i \geq 0$$

- Most general case: if x is a random variable, and f is concave in x, then:

$$f(\mathbb{E}_x[x]) \geq \mathbb{E}_x[f(x)]$$



Back to EM

- We intend to learn the parameters that maximizes the log-likelihood of the data

$$l(\theta; D) = \log \prod_{i=1}^{n} \left(\sum_{Z^i=1}^{K} p(x^i, z^i|\theta) \right) = \sum_{i=1}^{n} \log \left(\sum_{Z^i=1}^{K} p(x^i, z^i|\theta) \right) = \sum_{i=1}^{n} l(\theta; x_i)$$

- **Expectation step (E-step)**: what do we take expectation over?

$$l(\theta; D) \geq l(\theta, D; q) = \mathbb{E}_{z^1, z^2, \ldots, z^n \sim q} \left[\log \prod_{i=1}^{n} p(x^i, z^i|\theta) \right] + H(q)$$

To show this, we use Jensen inequality on $l(\theta; x_i)$

- **Maximization step (M-step)**: how to maximize?

$$\theta^{t+1} = argmax_\theta l(\theta, D; q)$$

Expectation Step as a Lower Bound

Back to EM

$$l(\theta; x) = \log \sum_{z=1}^{K} p(x, z|\theta)$$

$= \log \sum_{z=1}^{K} q(z) \frac{p(x,z|\theta)}{q(z)}$ (arbitrary distribution $q(z)$)

$\geq \sum_{z=1}^{K} q(z) \log \frac{p(x,z|\theta)}{q(z)}$ (Jensen's inequality: $f(\sum_i a_i x_i) \geq \sum_i a_i f(x_i)$)

$$= \sum_{z=1}^{K} q(z) \log p(x, z|\theta) - q(z) \log q(z)$$

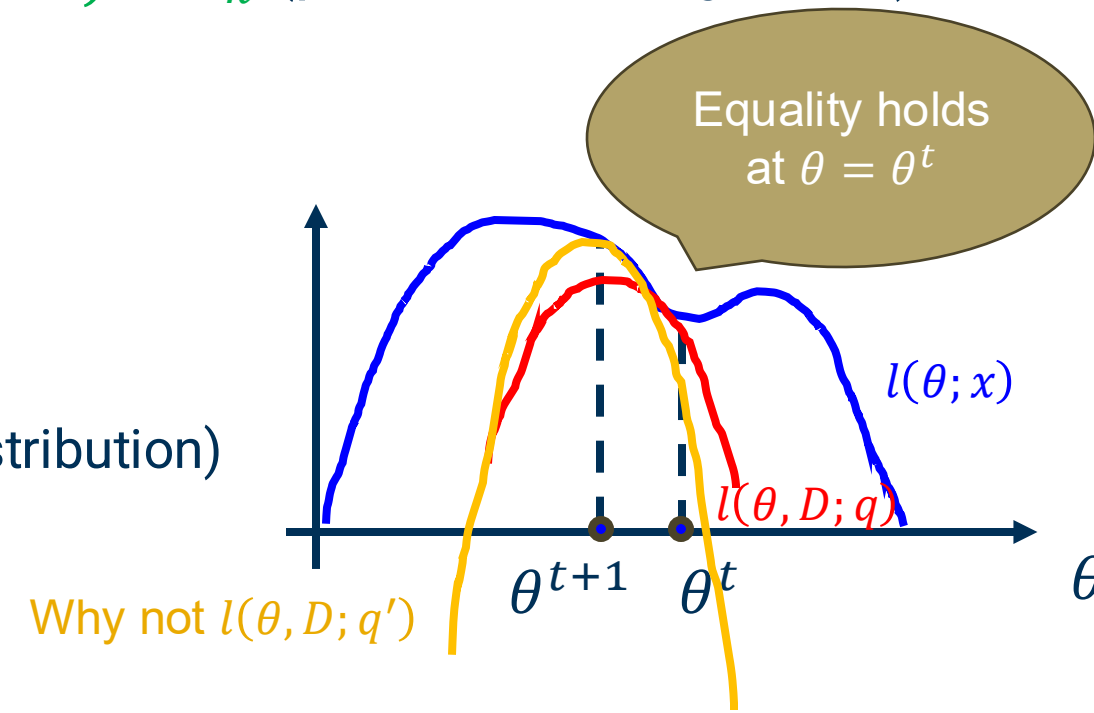
$$= \mathbb{E}_{z \sim q(z)}[\log p(x, z|\theta)] + H(q)$$

$$l(\theta; D) = \sum_{i=1}^{n} l(\theta; x^i) \geq \sum_{i=1}^{n} \left(\mathbb{E}_{z^i \sim q(z^i)}[\log p(x^i, z^i|\theta)] + H(q^i) \right) = \mathbb{E}_{q(z^1, z^2, \ldots, z^n)} \left[\log \prod_{i=1}^{n} p(x^i, z^i|\theta) \right] + H(q) = l(\theta, D; q)$$

This concludes the proof!

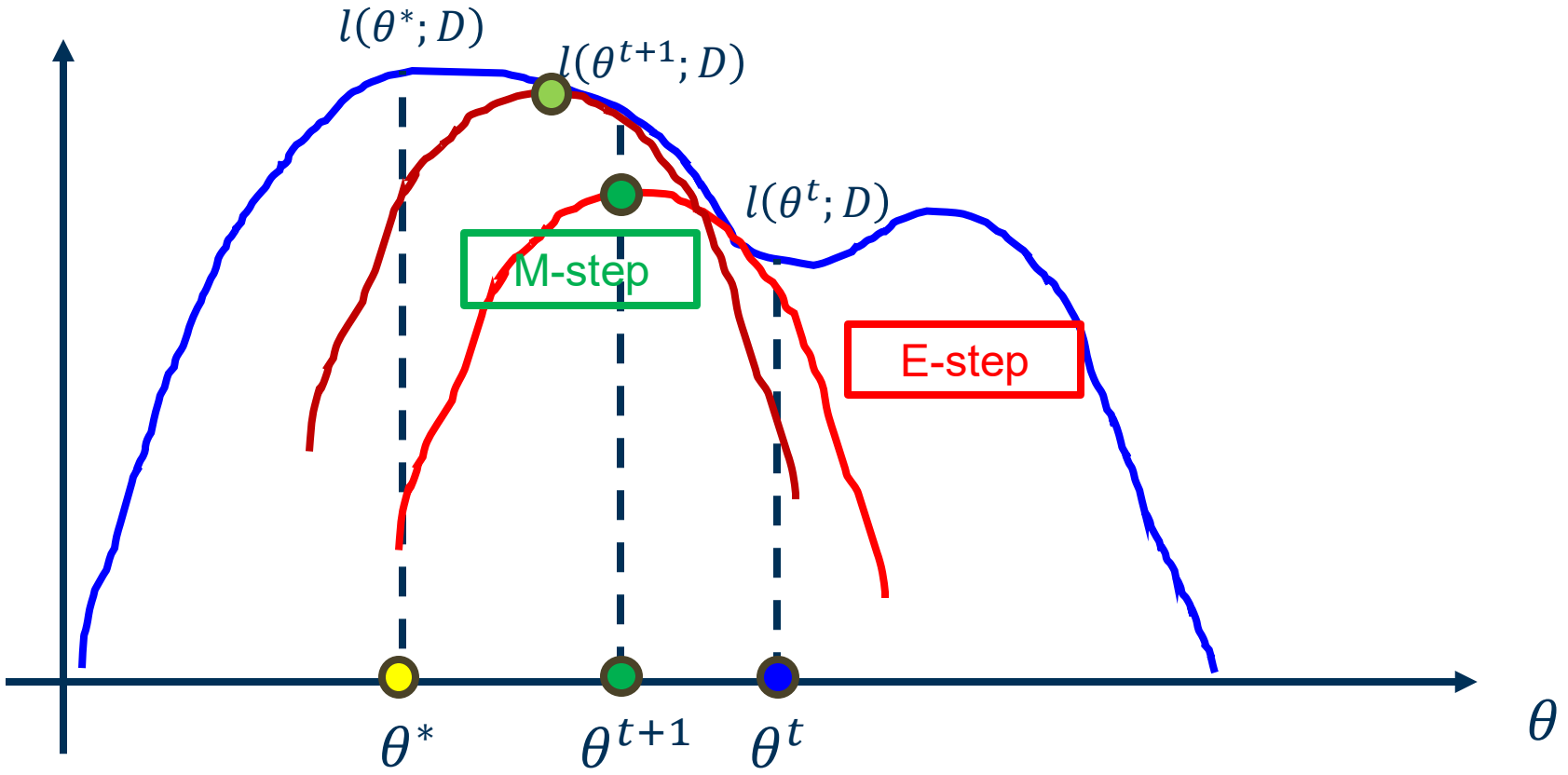
What Distribution q to Pick?

- **E-step**: why do we pick $q(z^i = k) = p(z^i = k|x^i, \theta) = \tau_k^i$ (posterior of z given x)?
  - Make Jensen inequality's equality holds!
  - $\frac{p(x,z|\theta)}{q(z)} = \frac{p(z,x|\theta)}{p(z|x,\theta)} = p(x|\theta)$. (identical for every z)
  - $\sum_{z=1}^{K} q(z) = \sum_{z=1}^{K} p(z|x,\theta) = 1$ (it is a probability distribution)



- Sanity check:
  - $l(\theta; x) = \log \sum_{z=1}^{K} q(z) \frac{p(x,z|\theta)}{q(z)} = \log \sum_{z=1}^{K} q(z) p(x|\theta) = \log p(x|\theta) = \sum_{z=1}^{K} q(z) \log p(x|\theta) = \sum_{z=1}^{K} q(z) \log \frac{p(x,z|\theta)}{q(z)} = l(\theta, x; q)$

EM Graphically



Questions?

- Why is $l(\theta, D; q)$ a lower bound?
  - Jensen's inequality: $l(\theta; D) \geq l(\theta, D; q) = \mathbb{E}_{q(z^1, z^2, \ldots, z^n)}\left[\log \prod_{i=1}^{n} p\left(x^i, z^i | \theta\right)\right] + H(q)$

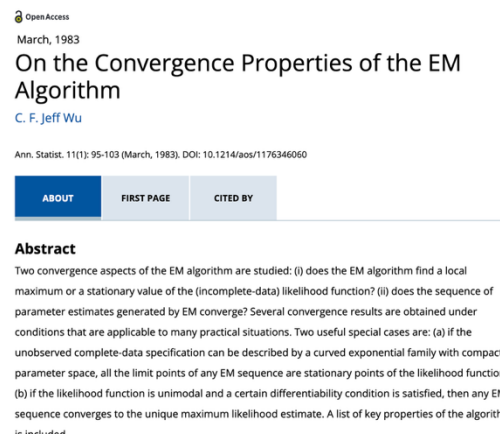
- Why choosing $q(z^1, z^2, \ldots, z^n) = \prod_{i=1}^{n} p\left(z | x^i, \theta^t\right)$ given parameter θ^t?
  - Equality condition of Jensen's inequality
  - Pick the best lower bound

- Why will EM converge?
  - Expectation step (E-step): find q to maximize the lower bound at θ^t

$$l(\theta^t; D) \geq l(\theta^t, D; q^t) \coloneqq \max_q \mathbb{E}_{q(z^1, z^2, \ldots, z^n)}\left[\log \prod_{i=1}^{n} p\left(x^i, z^i | \theta^t\right)\right] + H(q)$$

  - Maximization step (M-step): find θ to maximize the lower bound

$$\theta^{t+1} = argmax_\theta l(\theta^t, D;\ q^t)$$


Open Access

March, 1983

On the Convergence Properties of the EM Algorithm

C. F. Jeff Wu

Ann. Statist. 11(1): 95-103 (March, 1983). DOI: 10.1214/aos/1176346060

ABOUT | FIRST PAGE | CITED BY

Abstract

Two convergence aspects of the EM algorithm are studied: (i) does the EM algorithm find a local maximum or a stationary value of the (incomplete-data) likelihood function? (ii) does the sequence of parameter estimates generated by EM converge? Several convergence results are obtained under conditions that are applicable to many practical situations. Two useful special cases are: (a) if the unobserved complete-data specification can be described by a curved exponential family with compac parameter space, all the limit points of any EM sequence are stationary points of the likelihood functio (b) if the likelihood function is unimodal and a certain differentiability condition is satisfied, then any EI sequence converges to the unique maximum likelihood estimate. A list of key properties of the algorit

Revisit EM

Details of EM

- We intend to learn the parameters that maximizes the log-likelihood of the data

$$l(\theta; D) = \log \prod_{i=1}^{n} \left(\sum_{Z^i=1}^{K} p(x^i, z^i|\theta) \right)$$

- **Expectation step (E-step)**: what do we take expectation over?

$$l(\theta; D) \geq l(\theta, D; q) = \mathbb{E}_{z^1, z^2, \ldots, z^n \sim q(z^1, z^2, \ldots, z^n)} \left[\log \prod_{i=1}^{n} p(x^i, z^i|\theta) \right] + H(q)$$

- **Maximization step (M-step)**: how to maximize?

$$\theta^{t+1} = argmax_\theta l(\theta, D; q)$$

E-step: Compute the Expectation

$$l(\theta; D, q) = \mathbb{E}_{z^1, z^2, \ldots, z^n \sim q}\left[\log \prod_{i=1}^{n} p(x^i, z^i|\theta)\right] + H(q)$$

$$= \mathbb{E}_{z^1, z^2, \ldots, z^n \sim q}\left[\sum_{i=1}^{n}\left[\log p(x^i, z^i|\theta)\right]\right] + H(q)$$

$$= \mathbb{E}_{z^i \sim p(z^i|\, x^i)\ \forall i} \sum_{i=1}^{n}\left[\log p(x^i, z^i|\theta)\right] + H(q)$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z^i \sim p(z^i|\, x^i)}\left[\log\left(\pi_{z^i}\, \mathcal{N}(x|\mu_{z^i}, \Sigma_{z^i})\right)\right] + H(q)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_k^i \log\left(\pi_k\, \mathcal{N}(x|\mu_k, \Sigma_k)\right) + H(q)$$

- Expand log of Gaussian $\log \mathcal{N}(x|\mu_{z^i}, \Sigma_{z^i})$

$$l(\theta; D, \tau) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_k^i \left[\log \pi_k - \frac{1}{2}(x^i - \mu_k)^\top \Sigma_k^{-1}(x^i - \mu_k) - \frac{1}{2}\log|\Sigma_k| - c\right] + H(q)$$

M-step: Maximize $l(\theta; D, q)$

- $l(\theta; D, q) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_k^i \left[\log \pi_k - \frac{1}{2}(x^i - \mu_k)^\top \Sigma_k^{-1}(x^i - \mu_k) - \frac{1}{2}\log|\Sigma_k| - c \right] + H(q)$

- For instance, we want to find π_k, and $\sum_{k=1}^{K} \pi_k = 1$
  - Form Lagrangian

$$L = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_k^i [\log \pi_k - \text{other terms}] + \lambda\left(1 - \sum_{i=1}^{K} \pi_k\right)$$

  - Take partial derivative and set to 0

$$\frac{\partial L}{\partial \pi_k} = \left(\sum_{i=1}^{n} \frac{\tau_k^i}{\pi_k}\right) - \lambda = 0, \qquad \Rightarrow \pi_k = \frac{1}{\lambda}\sum_{i=1}^{n} \tau_k^i, \qquad \Rightarrow \lambda = n$$

EM (Expectation-Maximization) Algorithm

- Associate each data and each component with a τ_k^i
- Initialized $(\pi_k, \mu_k, \Sigma_k), k = 1,2,\dots,K$
- Iterate the following two steps until convergence:
  - **Expectation step (E-step)**: update τ_k^i given the current (π_k, μ_k, Σ_k)

$$\tau_k^i = p(z^i = k \mid D, \mu, \Sigma) = \frac{\pi_k \, \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \, \mathcal{N}(x|\mu_{k'}, \Sigma_{k'})}, \qquad \forall k \in \{1,2,\dots,K\}; i \in \{1,2,\dots,n\}$$

  - **Maximization step (M-step)**: update (π_k, μ_k, Σ_k) given τ_k^i

$$\pi_k = \frac{\sum_i \tau_k^i}{n}, \qquad \mu_k = \frac{\sum_i \tau_k^i x^i}{\sum_i \tau_k^i}$$

$$\Sigma_k = \frac{\sum_i \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^\top}{\sum_i \tau_k^i}, \qquad \forall k \in \{1,2,\dots,K\}$$

General Applicability of EM Algorithm

- Applicable to other models with latent (or missing) variables

- Expectation maximization applied to a coin toss example (python example)
  - Assume you have a sequence of coin flip observations from two coins, but you don't know from which coin each of the observations is from
  - The EM algorithm starts by initializing a random prior
  - Then it calculates the expected log probability distribution over the observations, and based on the log probability updates the prior

```
In [1]:
# N.B. each coin label in `labels` corresponds to the sequence

labels = ['B', 'A', 'A', 'B', 'A']

flip_seqs = ['HTHHHTTHHHHTHHHTTHHHHHTHHHH',
             'HHTHHHHHTTTTTTTHHTT',
             'HTHHHHTTHHTTTTTTTHTTTTHHTTT',
             'HTHTTHHTTHHHHHTHHHHHHHTHHHHHTHHHHHHTTHHHHH',
             'THHHTHHHTHTTTTTTTTTT']
```

```
In [12]:
weight_A

Out[12]:
[0.9993617897653697,
 0.04041543659201761,
 0.0001453833718729461,
 0.999992580222675,
 0.0007623605427559703]

In [13]:
weight_B

Out[13]:
[0.0006382102346302874,
 0.9595845634079825,
 0.9998546166281271,
 7.419777324936436e-06,
 0.999237639457244]
```

Beyond GMM and EM

Beyond Gaussian Mixture Models and EM

- Do we need the distribution to be Gaussian?
- Do we need the expectation step to be exact?
- Do we need the maximization step to be exact?
- Convergence?