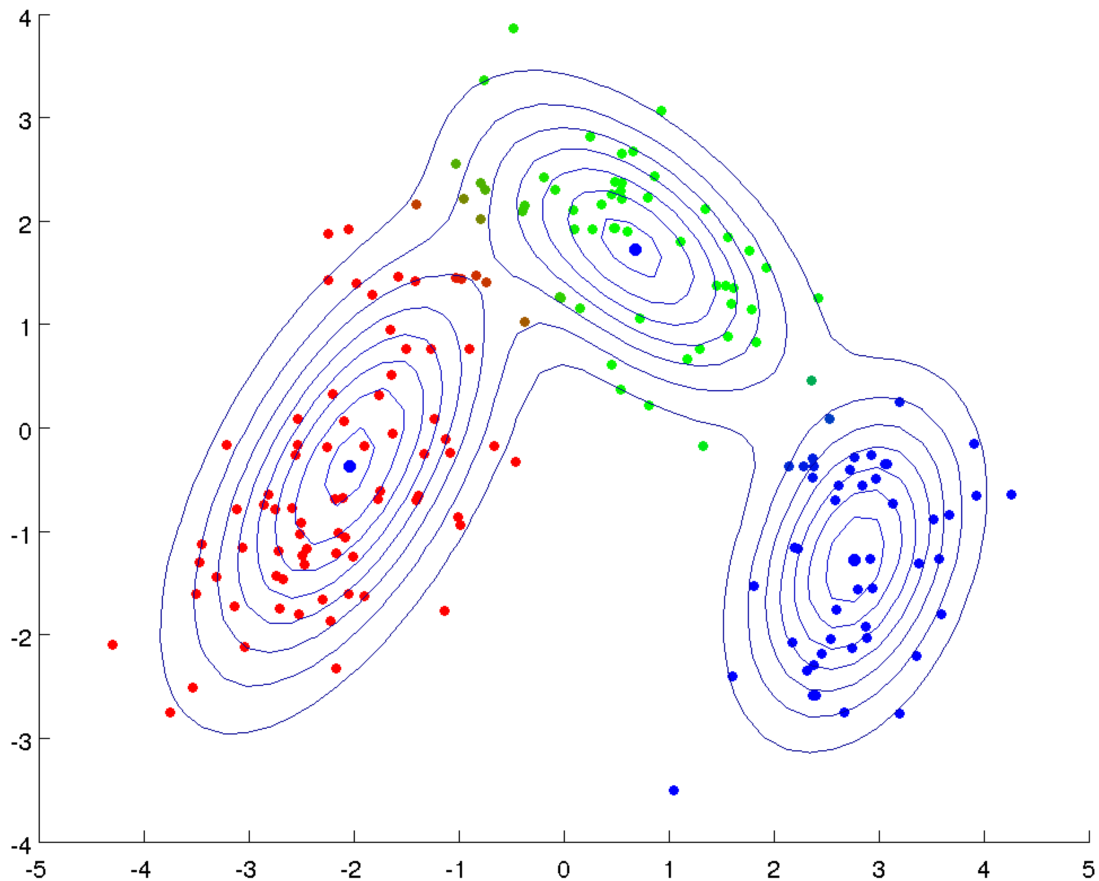
EM Algorithm

Le Song

Machine Learning
CSE/ISYE 6740, Fall 2019

Wine dataset

- First run PCA to reduce the dimension to 2
- Clear cluster structure
- Can we fit 3 Gaussians?



Gaussian mixture model

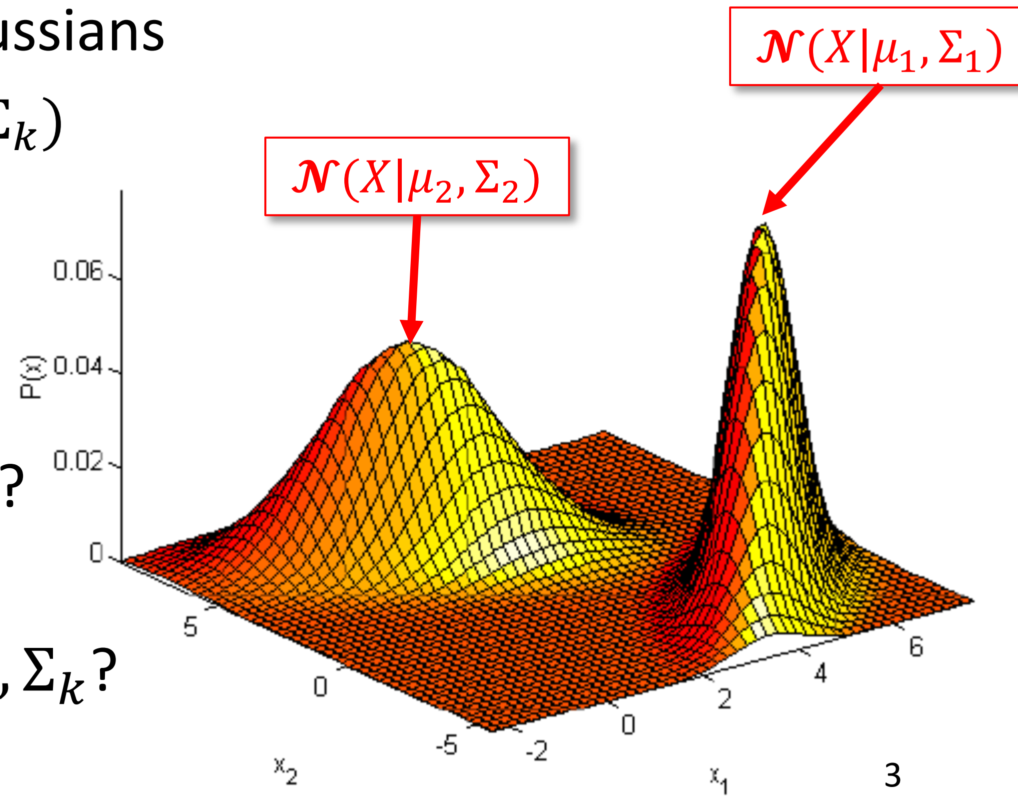
- A density model $p(X)$ may be multi-modal: model it as a mixture of uni-modal distributions (e.g. Gaussians)

$$\mathcal{N}(X|\mu_k, \Sigma_k) := \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} exp\left(-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right)$$

- Consider a mixture of K Gaussians
  - $p(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\mu_k, \Sigma_k)$

mixing proportion

mixture Component



- Parametric or noparametric?

- How to learn $\pi_k \in (0,1), \mu_k, \Sigma_k$?

EM algorithm

- Associate each data and each component with a τ_k^i
- Initialize (π_k, μ_k, Σ_k), $k = 1 \ldots K$
- Iterate the following two steps till convergence:
  - Expectation step (E-step): update τ_k^i given current (π_k, μ_k, Σ_k)

$$\tau_k^i = p\left(z_k^i = 1 \middle| D, \mu, \Sigma\right) = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x_i | \mu_{k'}, \Sigma_{k'})}$$

$$(k = 1 \ldots K, i = 1 \ldots m)$$

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

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

$$\Sigma_k = \frac{\sum_i \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^T}{\sum_i \tau_k^i}$$

$$(k = 1 \ldots K)$$

Expectation-Maximization Iterations

- Mixture of two Gaussian components, $K = 2$
- Use τ_1^i as the proportion of red, and τ_2^i proportion of blue

tau1 + tau2 = 1

- Draw only one contour for each Gaussian component

Intialization

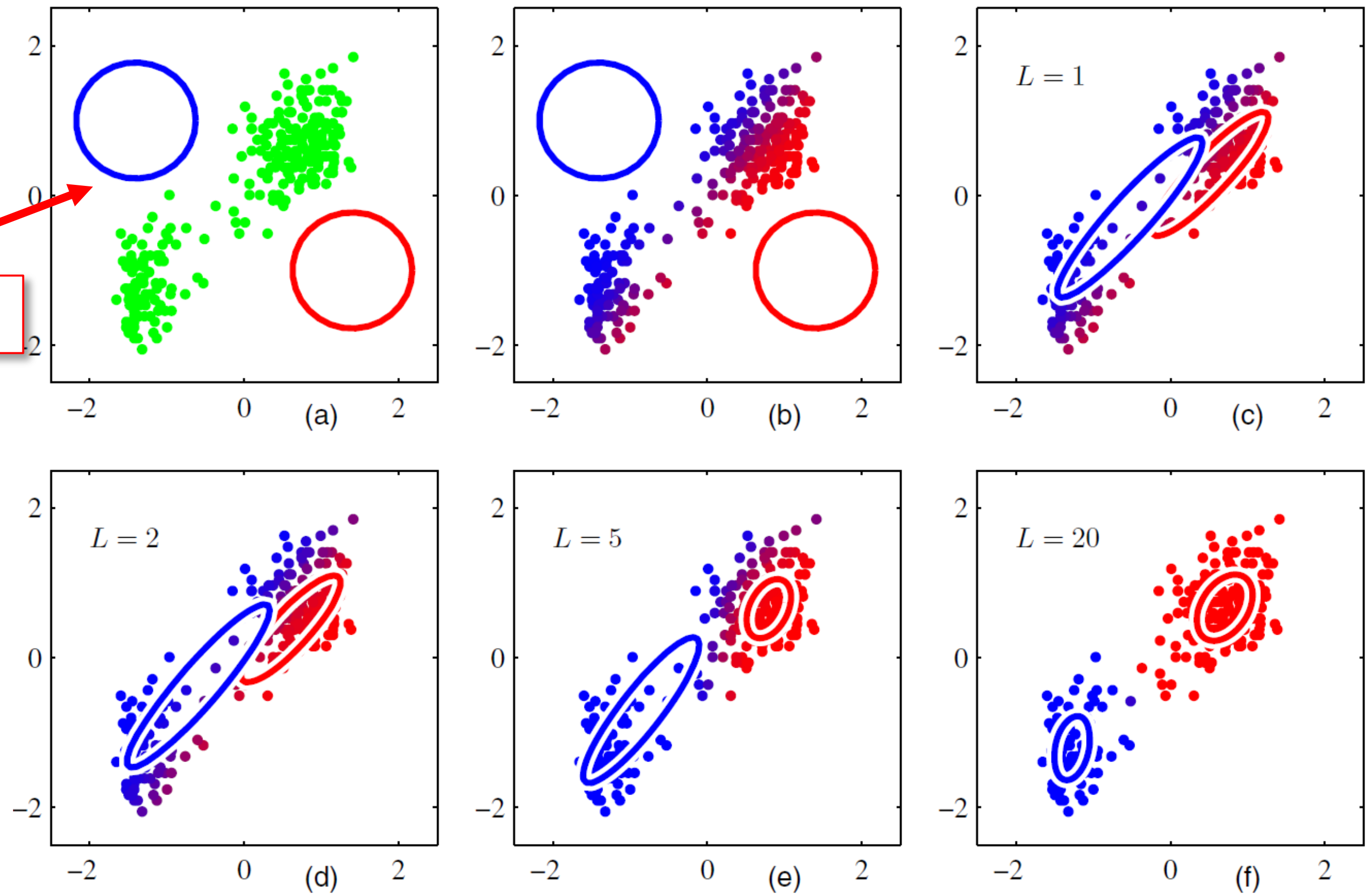
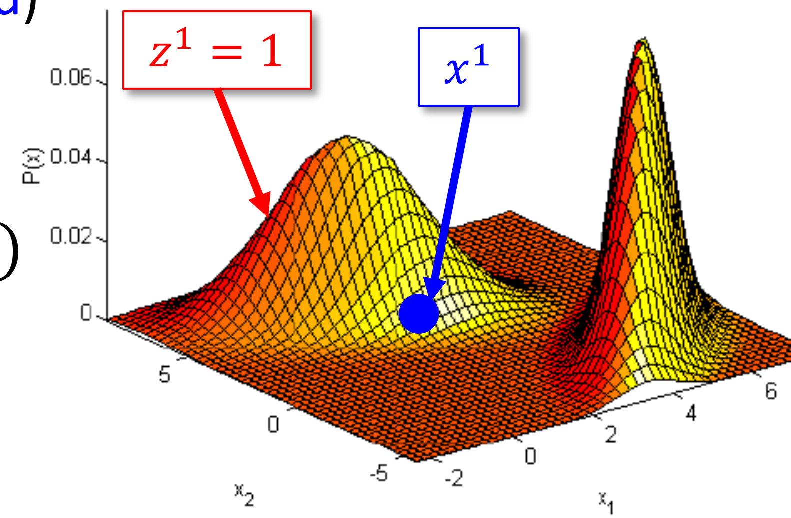


Image a generative process for data points

- For each data point x^i:
  - Randomly choose a mixture component, $z^i = \{1,2,\dots K\}$, with probability $p(z^i) = \pi_{z^i}$ (hidden)
  - Then sample the actual value of x^i from a Gaussian distribution $p(x^i|z^i) = \mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})$ (observed)



- Joint distribution over $p(x,z)$

$$p(x,z) = p(x|z)p(z) = \pi_z \mathcal{N}(x|\mu_z, \Sigma_z)$$

- Marginal distribution $p(x)$

$$p(x) = \sum_{z=1}^{K} p(x,z) = \sum_{z=1}^{K} \pi_z \mathcal{N}(x|\mu_z, \Sigma_z)$$

Learning the parameters

- How to learn, given a dataset $D = \{x^1, x^2 \ldots, x^m\}$?
- Maximum likelihood learning (let $\theta = (\pi_k, \mu_k, \Sigma_k)$, $k = 1 \ldots K$)

$$\theta^* = \mathrm{argmax}_\theta \, l(\theta; D) = \log \prod_{i=1}^{m} p(x^i|\theta)$$

- Use our generative process

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

$$= \log \prod_{i=1}^{m} \left(\sum_{z^i=1}^{K} p(x^i|z^i, \theta) p(z^i|\theta) \right)$$

$$= \log \prod_{i=1}^{m} \left(\sum_{z^i=1}^{K} \pi_{z^i} \mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i}) \right)$$

Details of EM

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

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

Nonconvex Difficult!

no simplification

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

so we simplify

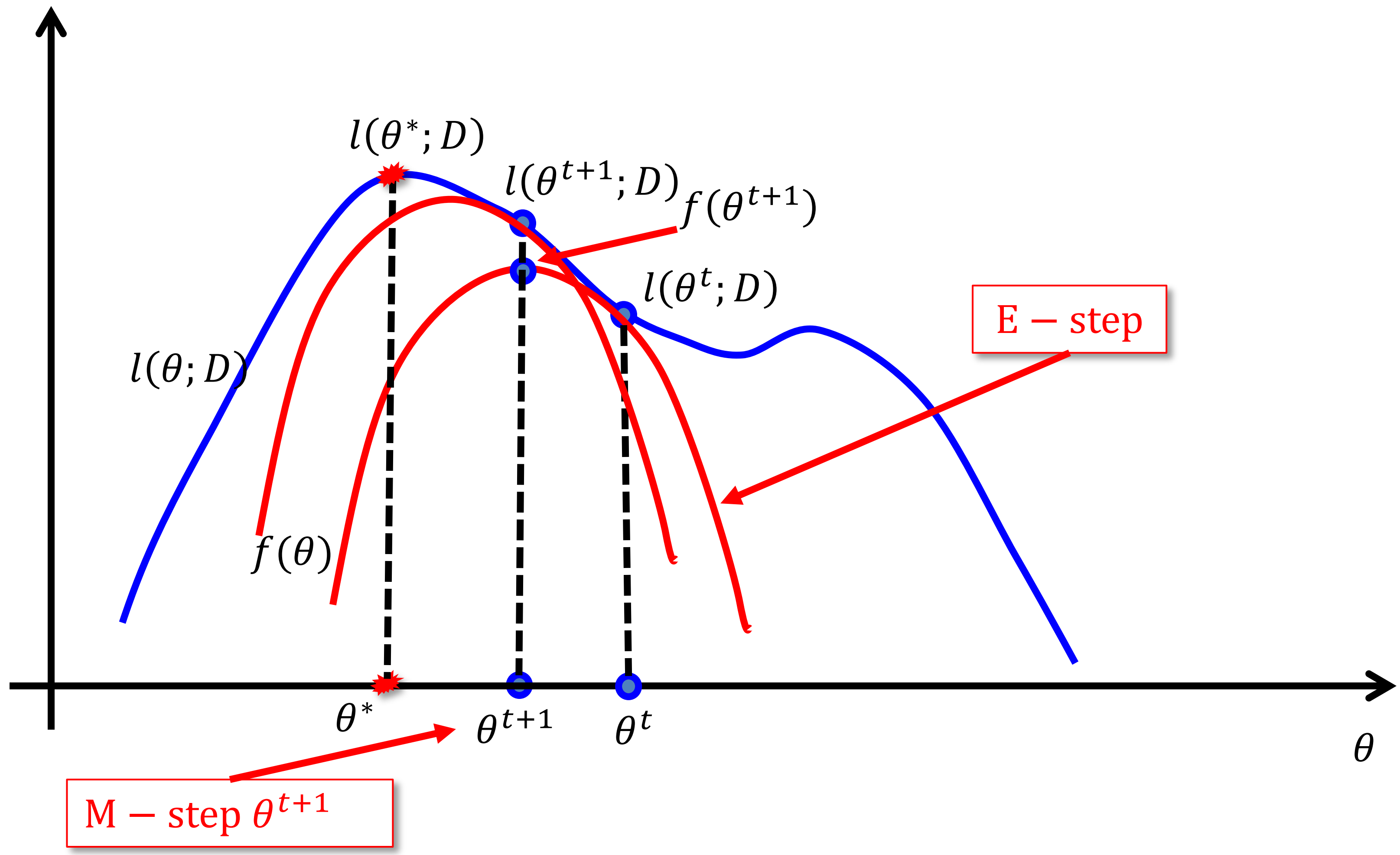
$$l(\theta; D) \geq f(\theta) = E_{q(z^1, z^2, \ldots, z^m)} [\log \prod_{i=1}^{m} p(x^i, z^i | \theta)]$$

original target function is hard to find its optimization. so change to an easier function

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

$$\theta^{t+1} = argmax_\theta \; f(\theta)$$

EM graphically



E-step: compute the expectation

- Precise computation is model dependent

- In some models (eg. mxiture of Gaussians)

$$q(z^1, z^2, .., z^m | \theta^t) = \prod_{i=1}^{m} p(z^i | x^i, \theta^t)$$

Then

$$f(\theta) = E_{q(z^1, z^2, .., z^m | \theta^t)} \left[\log \prod_{i=1}^{m} p(x^i, z^i | \theta) \right]$$

$$= E_{\prod_{i=1}^{m} p(z^i | x^i, \theta^t)} \left[\sum_{i=1}^{m} \log p(x^i, z^i | \theta) \right]$$

$$= \sum_{i=1}^{m} E_{p(z^i | x^i, \theta^t)} [\log p(x^i, z^i | \theta)]$$

E-step: mixture of Gaussians

likelihood Prior

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)} = \frac{P(x,z)}{\sum_{z'} P(x,z')}$$

posterior normalization constant

Prior: $p(z^i) = \pi_{z^i}$

Likelihood: $p(x^i|z^i) = \mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})$

Completely likelihood: $p(x^i, z^i) = \pi_{z^i}\mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})$

Posterior: $\tau_k^i = p(z^i = k|x^i) = \dfrac{\pi_k \mathcal{N}(x^i|\mu_k, \Sigma_k)}{\sum_{k'=1..K} \pi_{k'} \mathcal{N}(x^i|\mu_{k'}, \Sigma_{k'})}$

E-step: mixture of Gaussians (cont.)

Expand $\log \pi_{z^i} \mathcal{N}\left(x^i \middle| \mu_{z^i}, \Sigma_{z^i}\right)$ and use τ_k^i

$$f(\theta) = \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}\left[\log p\left(x^i, z^i \middle| \theta\right)\right]$$

$$= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}\left[\log \pi_{z^i} \mathcal{N}\left(x^i \middle| \mu_{z^i}, \Sigma_{z^i}\right)\right]$$

$$= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}\left[\log \pi_{z^i} - \left(x^i - \mu_{z^i}\right)^{\top} \Sigma_{z^i}^{-1}\left(x^i - \mu_{z^i}\right) - \frac{1}{2}\log\left|\Sigma_{z^i}\right| + c\right]$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{K} \tau_k^i \left[\log \pi_k - \left(x^i - \mu_k\right)^{\top} \Sigma_k^{-1}\left(x^i - \mu_k\right) - \frac{1}{2}\log\left|\Sigma_{z^i}\right| + c\right]$$

M-step: mixture of Gaussians

$$f(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} \tau_k^i \left[\log \pi_k - \left(x^i - \mu_k\right)^{\top} \Sigma_k^{-1} \left(x^i - \mu_k\right) - \frac{1}{2} \log \left|\Sigma_{z^i}\right| + c \right]$$

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

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

  - Take partial derivative and set to 0

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{m} \frac{\tau_k^i}{\pi_k} - \lambda = 0$$

$$\Rightarrow \pi_k = \frac{1}{\lambda} \sum_{i=1}^{m} \tau_k^i$$

$$\Rightarrow \lambda = m$$

EM algorithm

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

$$\tau_k^i = p\left(z_k^i = 1 \middle| D, \mu, \Sigma\right) = \frac{\pi_k \boldsymbol{\mathcal{N}}(x_i | \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \boldsymbol{\mathcal{N}}(x_i | \mu_{k'}, \Sigma_{k'})}$$

$$(k = 1 \dots K, i = 1 \dots m)$$

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

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

$$\Sigma_k = \frac{\sum_i \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^T}{\sum_i \tau_k^i}$$

$$(k = 1 \dots K)$$

EM vs. modified K-means

- The EM algorithm for mixture of Gaussian is like a soft clustering algorithm

- K-means:
  - "E-step", we do hard assignment:
    - $z^i = argmax_k (x^i - \mu_k)\, \Sigma_{\mathrm{k}}^{-1} (x^i - \mu_k)$
  - "M-step", we update the means and covariance of cluster using maximum likelihood estimate:
    - $\mu_k = \frac{\sum_i \delta(z^i,k) x^i}{\sum_i \delta(z^i,k)}$
    - $\Sigma_k = \frac{\sum_i \delta(z^i,k)(x^i - \mu_k)\,(x^i - \mu_k)^T}{\sum_i \delta(z^i,k)}$

  $\delta(z^i, k) = 1\ if\ z^i = k$; otherwise 0.

General applicability of EM algorithm

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

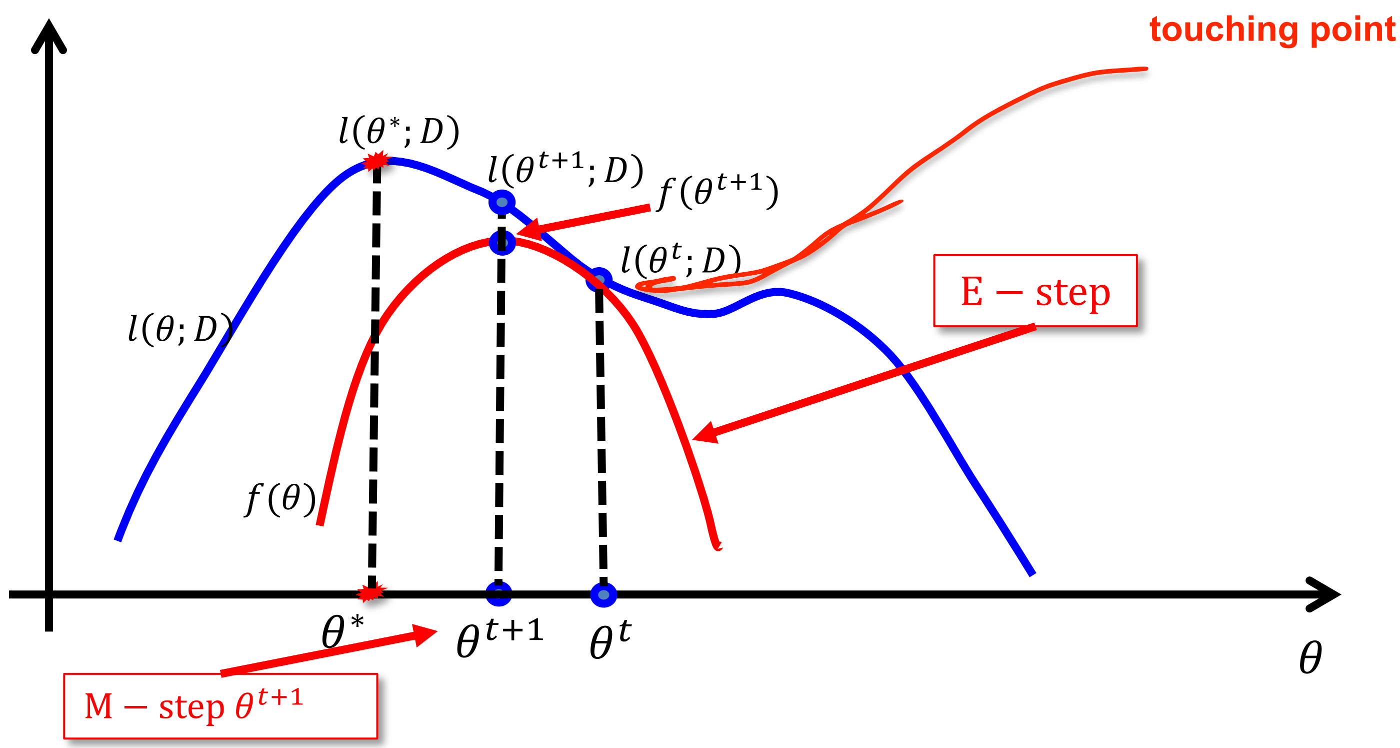
- Example (coin_toss_em.m):
    - Expectation maximization applied to a coin toss example
    - Assume you have five observations of 10 coin flips from two coins but you don't know from which coin each of the observations is from
    - The EM algorithm starts by initializing a random prior
    - Then it calculates the expected log probability distribution over the observations, and based on the log probability updates the prior

Questions?

- Why is $f(\theta)$ a lower bound?

$$l(\theta; D) \geq f(\theta) = E_{q(z^1, z^2, \ldots, z^m)}[\log \prod_{i=1}^{m} p(x^i, z^i | \theta)]$$

- Why will EM converge?

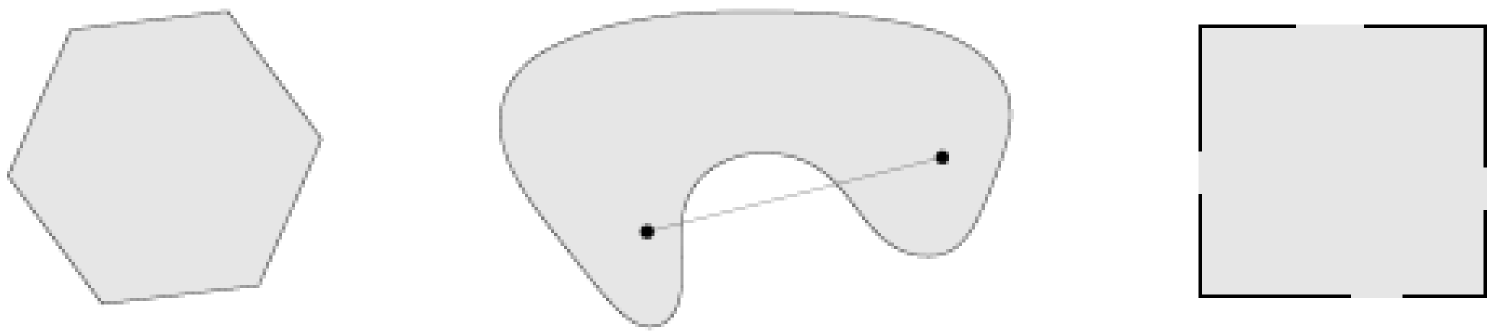


Convex Sets

- 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.

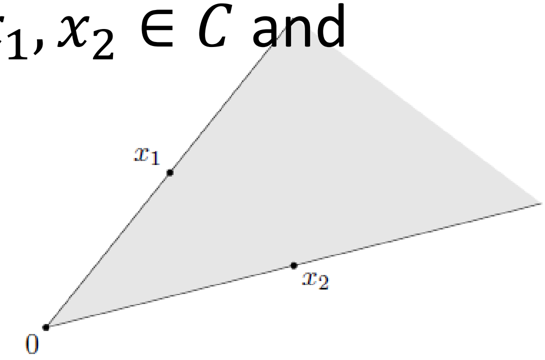
- Examples of convex and non-convex sets



Common Convex Sets

- Cones: A set C is a convex cone, if for any $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have

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

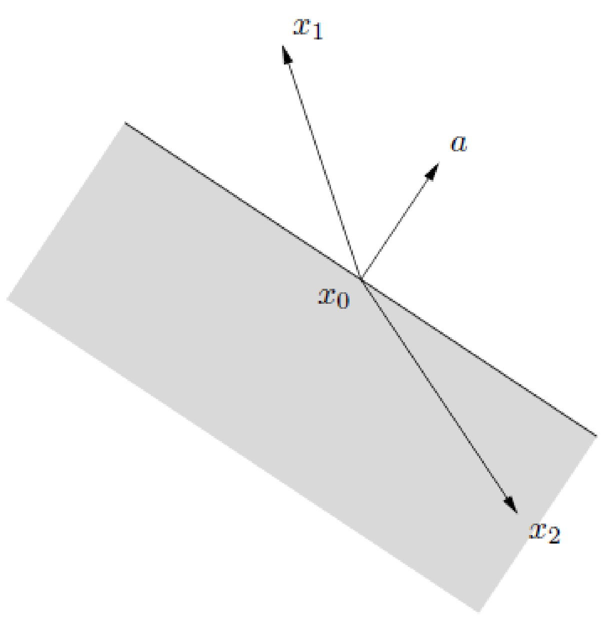


- Hyperplanes and halfspaces:
  A set is hyperplane if

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

  A halfspace is

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



Common Convex Sets

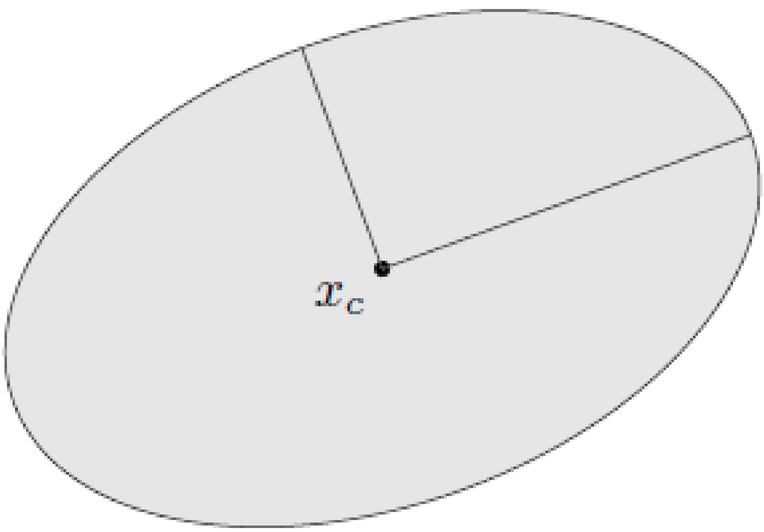
- Euclidean balls: A Euclidean ball has the form

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

- Ellipsoids:

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

  - The eigen-vectors and eigen-values determine the direction and shape of the semi-axes

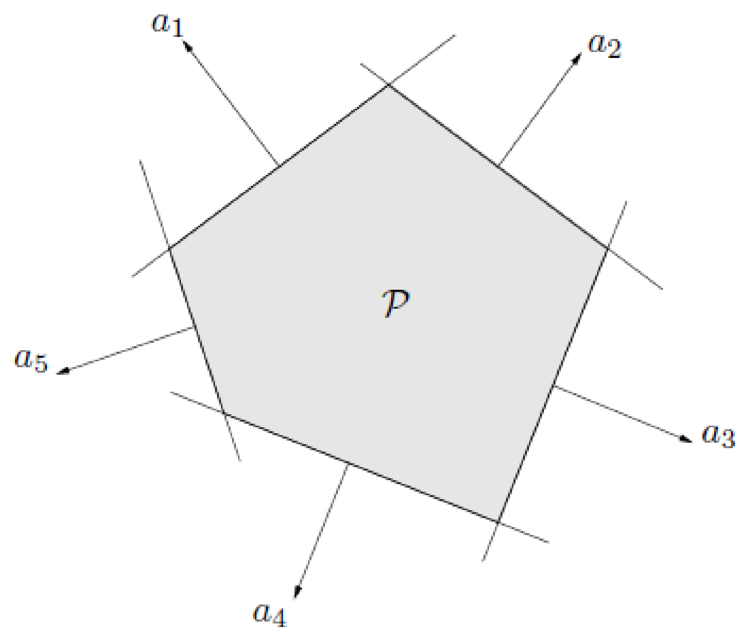


Common Convex Sets

- Polyhedra: Intersection of a *finite* set of halfspaces/hyperplanes

$$P = \{x | a_j^\top x \le b_j, j = 1, \dots, m, c_j^\top x = d_j, j = 1, \dots, p\}$$

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



Operations that Preserve Convex Sets

- Intersections: In fact, *every* closed convex set S is the intersection of all halfspaces that contain it:

$$S = \bigcap \{H | H \text{ is halfspace}, S \subset H\}$$

- Linear combination:

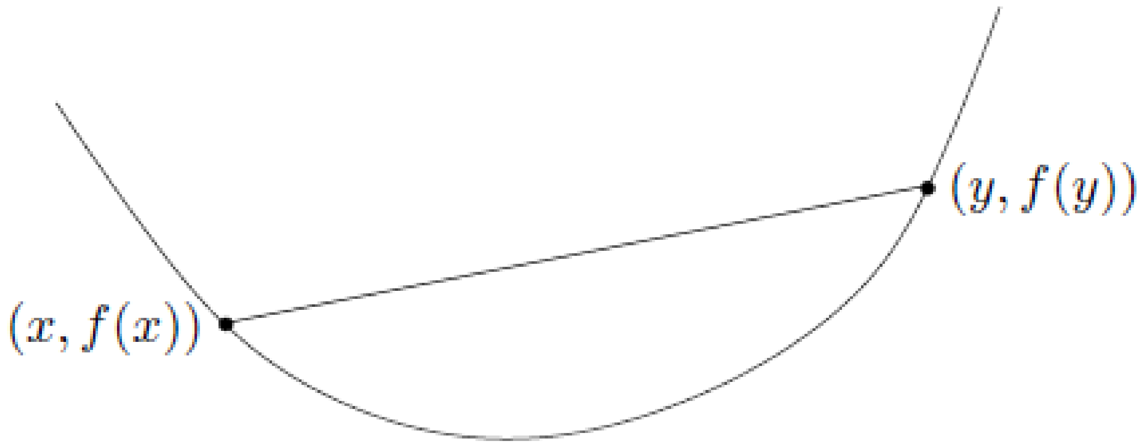
$$\alpha S = \{\alpha x | x \in S\},\ S + \alpha = \{x + a | x \in S\}$$

- Projection/Concatenation

Convex Functions

- Definition: A function $f: R^n \to R$ is convex if the domain $\mathbf{dom} f$ 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

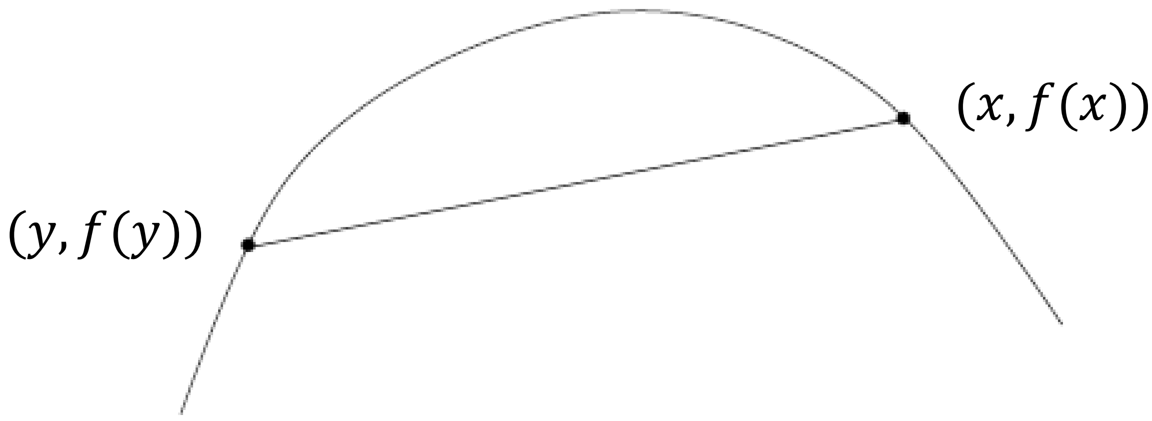


Concave Functions

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

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

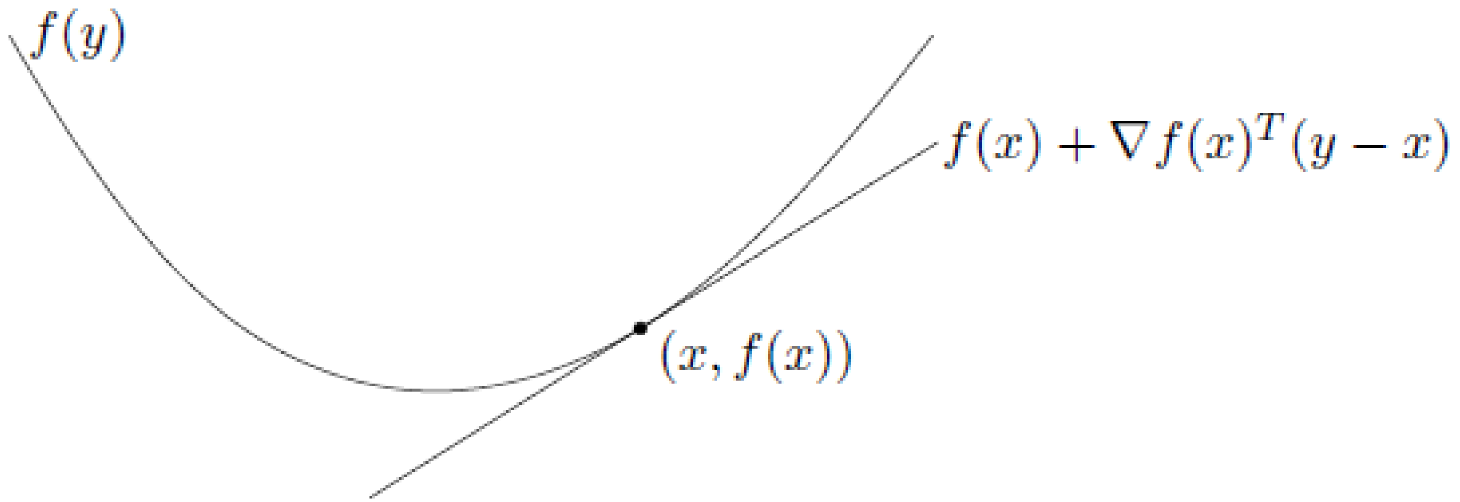
- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies below the graph of f



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 and
$$f(y) \geq f(x) + \nabla f(x)^{\top}(y - x)$$
holds for all $x, y \in \mathbf{dom} f$.

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



Second-order Conditions

- If f is twice differential, the second-order condition is: f is convex iff $\mathbf{dom} f$ is convex and for all $x \in \mathbf{dom} f$
$$\nabla^2 f(x) \succcurlyeq 0$$
positive semidefinite (symmetric and all eigenvalue nonnegative)

- That is the Hessian is positive semidefinite.

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

- Eg. $\mathrm{f(x)} = x^\top A x$, for A positive semidefinite

Used in SVM

Examples

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

Used in EM

Used in
multivariate Gaussian fit

Operations that Preserve Convexity

- Nonnegative weighted sums: If $f_1, \ldots, f_m$ are convex, and $w_1, \ldots, w_m \geq 0$, then
$$f = w_1 f_1 + \cdots + w_m f_m$$
is convex

- Composition with an affine mapping: suppose f is convex, then
$$g(x) = f(Ax + b)$$
with $\mathbf{dom}\, g = \{x | Ax + b \in \mathbf{dom}\, f\}$ is convex

- Pointwise maximum and supremum: If f_1 and f_2 are convex, then $f(x) = \max\{f_1, f_2\}$ is also convex. It easily extends to multiple functions.

Operations that Preserve Convexity

- Composition: If h is convex and nondecreasing, and g is convex, then $f(x) = h(g(x))$ is convex
  - The second derivative of f is
    $$f''(x) = h''\big(g(x)\big)g'(x)^2 + h'\big(g(x)\big)g''(x)$$
    for f to be convex, f'' should be nonnegative

- Log-sum-exp: $f(x) = \log(e^{x_1} + \cdots + e^{x_n})$

Used in
multiclass classification

Theory underlying EM

- Recall that in MLE, we intend to learn the model parameter that would have maximize the likelihood of the data.
  - $l(\theta; D) = \log \sum_z p(x, z|\theta) = \log \sum_z p(x| z, \theta) P(z|\theta)$

- But we are iterating these:
  - Expectation step (E-step)
    - $f(\theta) = E_{q(z)}[\log p(x, z|\theta)], where\ q(z) = P(z|x, \theta^t)$
  - Maximization step (M-step)
    - $\theta^{t+1} = argmax_\theta\ \ f(\theta)$

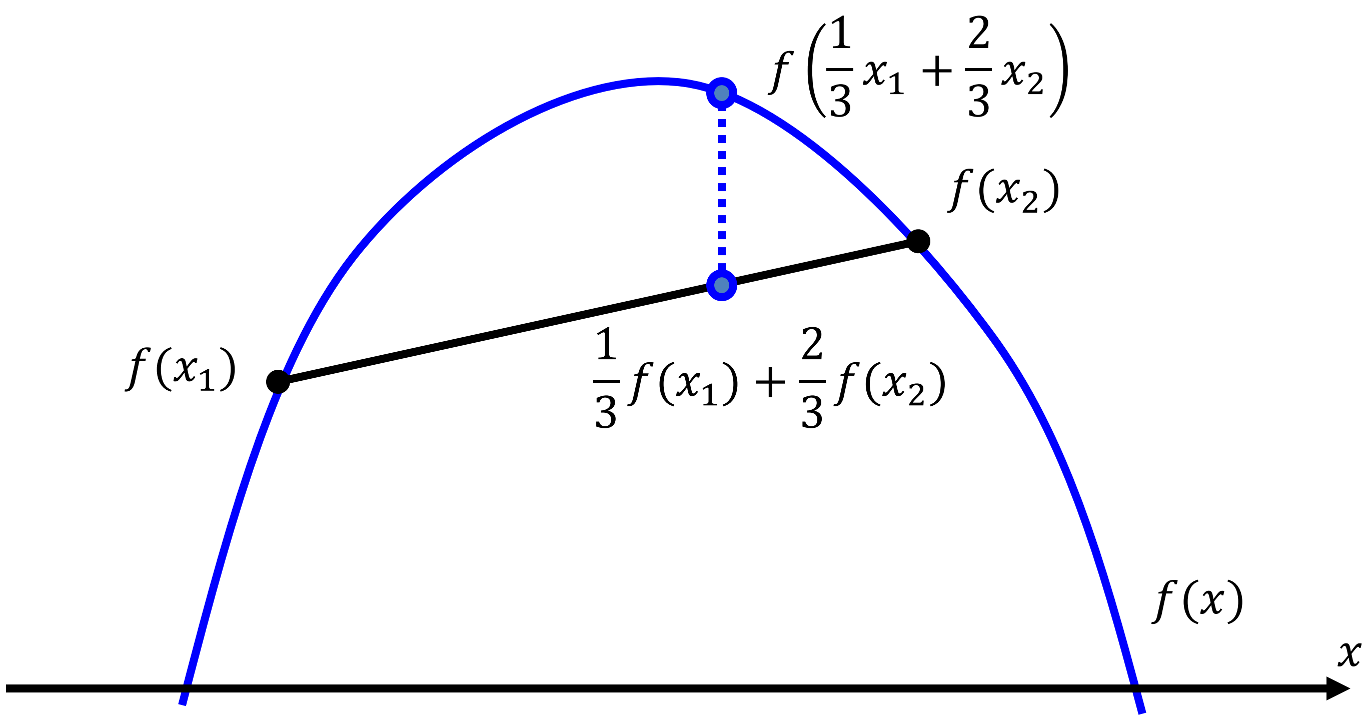
- Does maximizing this surrogate yield a maximizer of the likelihood?

Jensen's inequality

- For concave function $f(x)$
  - $f(\sum_i \alpha_i x_i) \geq \sum_i \alpha_i f(x_i), where\ \sum_i \alpha_i = 1, \alpha_i \geq 0$

- Most general case: If x is a random variable, and f is concave,

$$f(\mathbf{E}x) \geq \mathbf{E}f(x)$$



Lower bound of log-likelihood

- Log-likelihood $l(x;\theta) = \log \sum_z p(x,z|\theta)$

$$= \log \sum_z q(z) \frac{p(x,z|\theta)}{q(z)} \quad \text{(arbitrary } q(z)\text{)}$$

$$\geq \sum_z q(z) \log \frac{p(x,z|\theta)}{q(z)} \quad (Jensen's\ inequality\ f\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i f(x_i))$$

$$= \sum_z q(z) \log p(x,z|\theta) - \sum_z q(z) \log q(z)$$

$$= E_{q(z)}[\log p(x,z|\theta)] + H_{q(z)}$$

$\boxed{\text{What } q \text{ to use?}}$

What attains equality?

- $q(z) = p(z|x, \theta^t)$: posterior of z given x attains the equality at θ^t

- Let $F(q, \theta) = \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(Z|x)} \leq l(x; \theta) = \log \sum_z p(x, z|\theta)$

- $F(p(z|x, \theta^t), \theta^t) = \sum_z p(z|x, \theta^t) \log \frac{p(x, z|\theta^t)}{p(z|x, \theta^t)}$
- $= \sum_z p(z|x, \theta^t) \log p(x|\theta^t)$
- $= \log p(x|\theta^t)$
- $= \log \sum_z p(x, z|\theta^t)$