ESE 650: Learning in Robotics
Lecture 3

Lecturer:

Nikolay Atanasov: atanasov@seas.upenn.edu

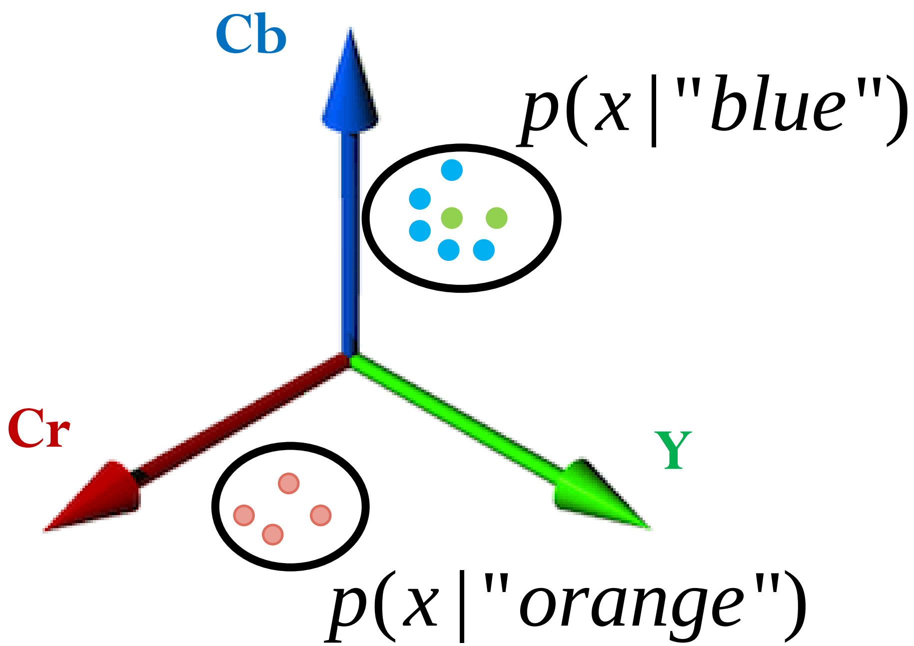
Teaching Assistants:

Jinwook Huh: jinwookh@seas.upenn.edu

Heejin Jeong: heejinj@seas.upenn.edu

Kelsey Saulnier: saulnier@seas.upenn.edu

- Learn a probabilistic model $p(w \mid x)$ of the color classes w given training color-space data $D = \{(x_i, w_i)\}$ where
  - Each pixel is a 3-D vector: $x = (Y, Cb, Cr)$
  - Discrete color labels: $w \in \{1, \ldots, N\}$



Maximum likelihood estimation

- Model $\mathrm{p}(x \mid w = \alpha)$ for each color class α as the pdf of a **Gaussian distribution**

- Assume that, given the color class, the pixel realizations are independent!

- Choose all pixels $D_\alpha := \{(x_i, w_i) | w_i = \alpha\} \subseteq D$ from class α (e.g., red) and use MLE to determine the most likely parameters (μ, Σ) for the Gaussian distribution representing this color class

(data likelihood)

$$\mu^*, \Sigma^* = \arg\max_{\mu,\Sigma} \prod_{x \in D_\alpha} \phi(x; \mu, \Sigma)$$

$$= \arg\max_{\mu,\Sigma} \sum_{x \in D_\alpha} \log \phi(x; \mu, \Sigma) = \arg\max_{\mu,\Sigma} J(\mu, \Sigma)$$

Matrix Calculus (numerator layout)

1. $$\frac{d}{dX_{ij}} X = e_i e_j^T$$

2. $$\frac{d}{dx} Ax = A$$

3. $$\frac{d}{dx} x^T Ax = x^T \left(A + A^T \right)$$

4. $$\frac{d}{dx} M^{-1}(x) = -M^{-1}(x) \frac{dM(x)}{dx} M^{-1}(x)$$

5. $$\frac{d}{dX} tr\left(AX^{-1}B \right) = -X^{-1}BAX^{-1}$$

6. $$\frac{d}{dX} \log\det X = X^{-1}$$

Note that:

$$\exp(X) := \sum_{k=0}^{\infty} \frac{X^k}{k!}$$

Maximum likelihood mean

$$J(\mu,\Sigma) = \sum_{x\in D_\alpha} \log \phi\left(x\,;\mu,\Sigma\right) =$$

$$= -\frac{n\,|\,D_\alpha\,|}{2}\log 2\pi - \frac{|\,D_\alpha\,|}{2}\log\det\Sigma - \frac{1}{2}\sum_{x\in D_\alpha}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$0 = \frac{d}{d\mu}J(\mu,\Sigma) = -\frac{1}{2}\sum_{x\in D_\alpha}\frac{d}{d\mu}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$= -\sum_{x\in D_\alpha}(x-\mu)^T\Sigma^{-1} \qquad \Rightarrow \qquad \boxed{\mu^* = \frac{1}{|\,D_\alpha\,|}\sum_{x\in D_\alpha} x}$$

Maximum likelihood covariance

$$J(\mu,\Sigma)=\sum_{x\in D_\alpha}\log\phi\left(x\,;\mu,\Sigma\right)=$$

$$=-\frac{n\,|\,D_\alpha\,|}{2}\log 2\pi-\frac{|\,D_\alpha\,|}{2}\log\det\Sigma-\frac{1}{2}\sum_{x\in D_\alpha}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$0=\frac{d}{d\Sigma}J(\mu,\Sigma)=-\frac{|\,D_\alpha\,|}{2}\frac{d}{d\Sigma}\log\det\Sigma-\frac{1}{2}\sum_{x\in D_\alpha}\frac{d}{d\Sigma}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$=-\frac{|\,D_\alpha\,|}{2}\Sigma^{-1}-\frac{1}{2}\sum_{x\in D_\alpha}\frac{d}{d\Sigma}tr\left(\Sigma^{-1}(x-\mu)(x-\mu)^T\right)$$

$$=-\frac{|\,D_\alpha\,|}{2}\Sigma^{-1}-\frac{1}{2}\sum_{x\in D_\alpha}-\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1}$$

$$\Rightarrow\qquad\boxed{\Sigma^*=\frac{1}{|\,D_\alpha\,|}\sum_{x\in D_\alpha}(x-\mu)(x-\mu)^T}$$

Maximum likelihood estimation

- Model $\mathrm{p}(x \mid w = \alpha)$ for each color class α as the pdf of a **Gaussian mixture** distribution

- Assume that, given the color class, the pixel realizations are independent!

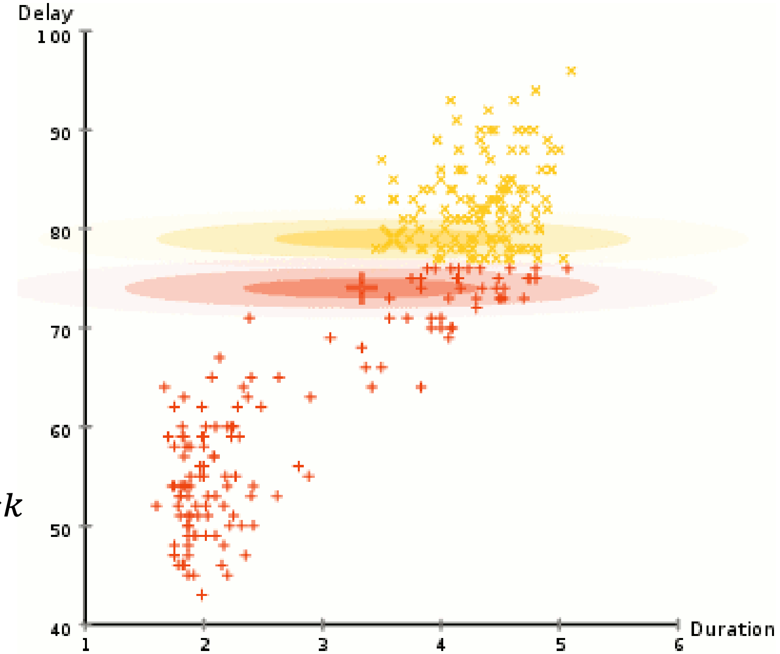
- Choose all pixels $D_\alpha := \{(x_i, w_i) | w_i = \alpha\} \subseteq D$ from class α (e.g., red) and use MLE to determine the most likely parameters $\{\alpha_k, \mu_k, \Sigma_k\}$ for the Gaussian mixture distribution representing this color class

(data likelihood)

$$\left\{\alpha_k^*, \mu_k^*, \Sigma_k^*\right\} = \underset{\{\alpha_k, \mu_k, \Sigma_k\}}{\arg\max} \prod_{x \in D_\alpha} p\left(x; \{\alpha_k, \mu_k, \Sigma_k\}\right)$$

$$= \underset{\{\alpha_k, \mu_k, \Sigma_k\}}{\arg\max} \sum_{x \in D_\alpha} \log p\left(x; \{\alpha_k, \mu_k, \Sigma_k\}\right)$$

- Gaussian mixtures are well suited for modeling clusters of points:
  - each cluster is assigned a Gaussian
  - mean is somewhere in the middle of the cluster
  - covariance measures the cluster spread



- **Generative model**:
  - Draw an integer between 1 and K with probability a_k of drawing k
  - Draw a random vector x from the k-th Gaussian density $\phi(x; \mu_k, \Sigma_k)$

- **Problem**: how do we determine the parameters $\theta := (a_1, \ldots, a_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ that specify the model from which the points are "most likely" to be drawn?

$$\theta^* = \arg\max_{\theta} \prod_{x \in D_\alpha} p(x;\theta) = \arg\max_{\theta} \sum_{x \in D_\alpha} \log\, p(x;\theta)$$

$$\Lambda(X,\theta) := \prod_{x \in D_\alpha} p(x;\theta) \qquad \lambda(X,\theta) := \sum_{x \in D_\alpha} \log\, p(x;\theta)$$

(data likelihood) **(log likelihood)**

Membership probabilities

(data likelihood) $$\Lambda(X,\theta) := \prod_{x \in D_\alpha} \sum_{k=1}^{K} a_k \phi(x; \mu_k, \Sigma_k)$$

- It is useful to understand the meaning of the terms: $q(k,x) = a_k \phi(x; \mu_k, \Sigma_k)$
- We assume that the event of drawing component k of the generative model is **independent** of the event of drawing a particular data point x out of a component
- $q(k,x)dx$ is the joint probability of drawing component k and data point x in volume dx around it
- **Membership probabilities**: the conditional probability of having selected component k given data point x:

$$r(k \mid x) = \frac{q(k,x)}{\sum_{m=1}^{K} q(m,x)} \qquad \sum_{k=1}^{K} r(k \mid x) = 1$$

Local maxima of $\lambda(X,\theta) := \sum_{x \in D_\alpha} \log \sum_{k=1}^{K} a_k \phi\left(x; \mu_k, \Sigma_k\right)$

$$\frac{d}{d\mu}\phi\left(x;\mu,\Sigma\right) = \phi\left(x;\mu,\Sigma\right)\frac{d}{d\mu}\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$= \phi\left(x;\mu,\Sigma\right)\Sigma^{-1}(\mu - x)$$

$$\frac{d}{d\mu_k}\lambda(X,\theta) = \sum_x \frac{a_k}{\sum_j a_j \phi\left(x;\mu_j,\Sigma_j\right)}\frac{d}{d\mu_k}\phi\left(x;\mu_k,\Sigma_k\right)$$

$$= \sum_x r(k \mid x)\Sigma_k^{-1}(\mu_k - x)$$

Local maxima of $\lambda(X,\theta) := \sum_{x \in D_\alpha} \log \sum_{k=1}^{K} a_k \phi\left(x; \mu_k, \Sigma_k\right)$

$$\frac{d}{d\Sigma}\phi\left(x;\mu,\Sigma\right) = \phi\left(x;\mu,\Sigma\right)\frac{d}{d\Sigma}\log\phi\left(x;\mu,\Sigma\right)$$

$$= \frac{1}{2}\phi\left(x;\mu,\Sigma\right)\Sigma^{-1}\left[(x-\mu)(x-\mu)^T\Sigma^{-1} - 1\right]$$

$$\frac{d}{d\Sigma_k}\lambda\left(X;\theta\right) = \sum_x \frac{a_k}{\sum_j a_j \phi\left(x;\mu_j,\Sigma_j\right)}\frac{d}{d\Sigma_k}\phi\left(x;\mu_k,\Sigma_k\right)$$

$$= \sum_x r(k\mid x)\frac{1}{2}\Sigma_k^{-1}\left[(x-\mu_k)(x-\mu_k)^T\Sigma_k^{-1} - 1\right]$$

Local maxima of $\lambda(X,\theta) := \sum_{x \in D_\alpha} \log \sum_{k=1}^{K} a_k \phi(x; \mu_k, \Sigma_k)$

- The derivative with respect to a_k is a trickier because a_k are restricted to a simplex
- **Trick**: express a_k through a *softmax* function:

$$a_k = \frac{e^{\gamma_k}}{\sum_j e^{\gamma_j}} \qquad \frac{da_k}{d\gamma_j} = \begin{cases} a_k - a_k^2, & \text{if } j = k \\ -a_j a_k, & \text{else} \end{cases}$$

$$\frac{\partial}{\partial \gamma_j} \lambda(X;\theta) = \sum_x \frac{1}{\sum_i a_i \phi(x; \mu_i, \Sigma_i)} \sum_k \frac{da_k}{d\gamma_j} \phi(x; \mu_k, \Sigma_k)$$

$$= \sum_x \left(r(j \mid x) - a_j \right)$$

Local maxima of Λ

- Setting the previous derivatives to zero, we obtain:

$$\frac{d}{d\mu_k}\lambda(X,\theta) = \sum_x r(k\,|\,x)\Sigma^{-1}(\mu_k - x) = 0 \quad \Rightarrow \quad \boxed{\mu_k = \frac{\sum_x r(k\,|\,x)x}{\sum_x r(k\,|\,x)}}$$

$$\frac{d}{d\Sigma_k}\lambda(X;\theta) = \sum_x r(k\,|\,x)\frac{1}{2}\Sigma_k^{-1}\left[(x-\mu_k)(x-\mu_k)^T\Sigma_k^{-1} - 1\right] = 0$$

$$\Rightarrow \boxed{\Sigma_k = \frac{\sum_x r(k\,|\,x)(x-\mu_k)(x-\mu_k)^T}{\sum_x r(k\,|\,x)}}$$

$$\frac{\partial}{\partial\gamma_j}\lambda(X;\theta) = \sum_x\left(r(j\,|\,x) - a_j\right) = 0 \quad \Rightarrow \quad \boxed{a_k = \frac{1}{|D_\alpha|}\sum_x r(k\,|\,x)}$$

Local maxima of Λ

$$\mu_k = \frac{\sum_x r(k \mid x)x}{\sum_x r(k \mid x)} \quad \Sigma_k = \frac{\sum_x r(k \mid x)(x-\mu_k)(x-\mu_k)^T}{\sum_x r(k \mid x)} \quad a_k = \frac{1}{|D_\alpha|}\sum_x r(k \mid x)$$

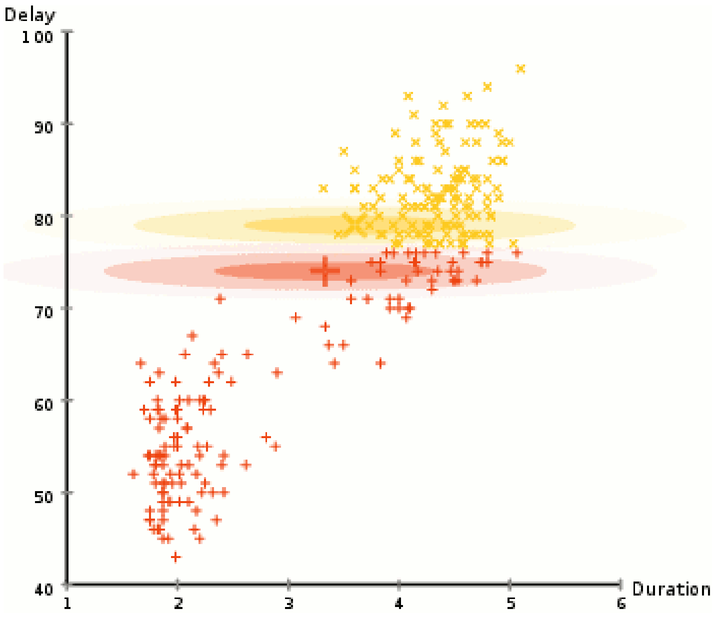
- First two are the sample mean and covariance of the data, weighted by the conditional probability that data point x was generated by mode k.
- The mixture weights are equal to the sample mean of the conditional probabilities $r(k \mid x)$ assuming a uniform distribution over D_α.
- The three equations are couple through $r(k \mid x)$ and hence are **hard to solve directly:**

$$r(k \mid x) = \frac{q(k,x)}{\sum_{m=1}^{K} q(m,x)} \qquad \sum_{k=1}^{K} r(k \mid x) = 1 \qquad q(k,x) = a_k \phi(x; \mu_k, \Sigma_k)$$

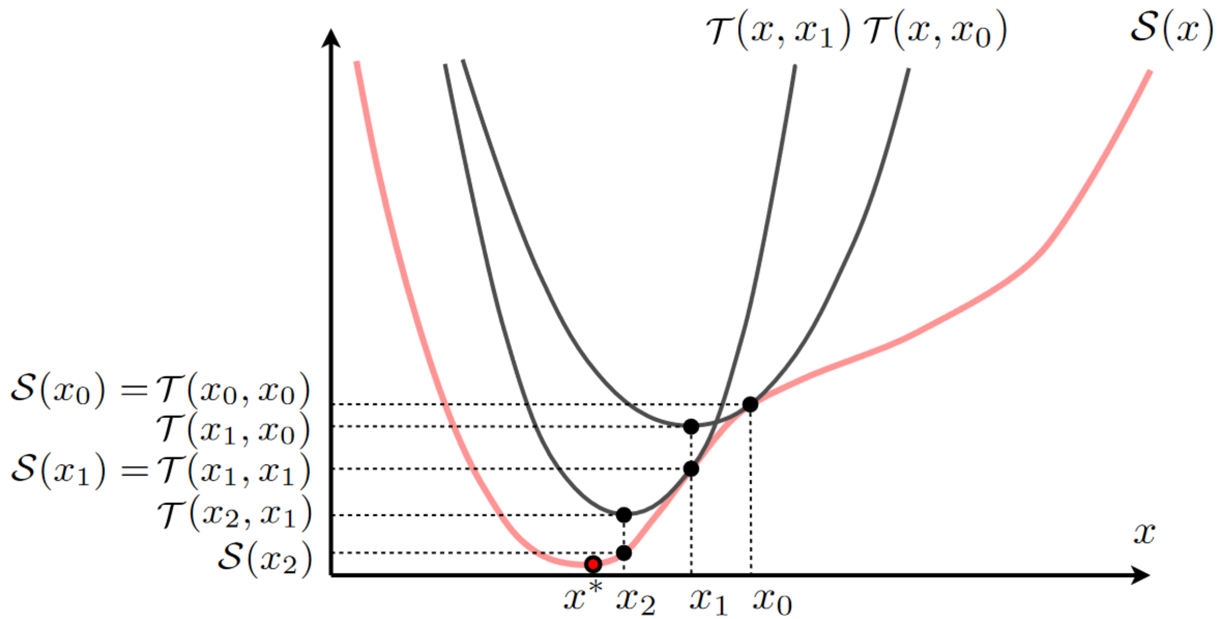
- **Idea:** start with a guess of the parameters θ^0 and iterate between computing $r(k \mid x)$ and updating θ

Expectation Maximization

- Iterative optimization technique based on auxiliary lower bound functions
  - Old idea (late 50's) but formalized by Dempster, Laird and Rubin in 1977
  - Subject of much investigation. See McLachlan & Krishnan book 1997
- Has two steps:
  - Expectation (E)
  - Maximization (M)
- Applicable to a wide range of problems:
  - Fitting mixture models
  - Probabilistic latent semantic analysis: produce concepts related to documents and terms (NLP)
  - Learning parts and structure models (vision)
  - Segmentation of layers in video (vision)



Expectation Maximization



Goal: $\min_x S(x)$

- Iterative approach:
  - Initialize at x_0
  - Construct an auxiliary lower-bound function $\mathcal{T}$ at x_0
  - Optimize the auxiliary function to get x_1

$$\boxed{\begin{aligned} \mathcal{T}(x, x_i) &\leq S(x_i) \\ \mathcal{T}(x_i, x_i) &= S(x_i) \end{aligned}}$$

- Each step gets closer to a **local max** of $\mathcal{S}(x)$:

$$S(x_{i+1}) \geq \max_x \mathcal{T}(x, x_i) \geq \mathcal{T}(x_i, x_i) = S(x_i)$$

- The lower bound function $\mathcal{T}$ need **not** be a paraboloid (Newton's method)

Jensen's Inequality

- **Jensen's inequality:** If f is a convex function and Z is a random variable, then:

$$f\left(\mathbb{E}[Z]\right) \leq \mathbb{E}\left[f\left(Z\right)\right]$$

- **Jensen's inequality (finite form):** If f is a convex function, $\{z_i\}$ are points in its domain and $\{a_i\}$ are positive weights:

$$f\left(\frac{\sum_i a_i z_i}{\sum_i a_i}\right) \leq \frac{\sum_i a_i f(z_i)}{\sum_i a_i}$$

- **Example:**

$$\log\left(\sum_i z_i\right) = \log\left(\sum_i a_i \frac{z_i}{a_i}\right) \geq \sum_i a_i \log\left(\frac{z_i}{a_i}\right)$$

Expectation Maximization

- **E-step**: Starting with an estimate $\theta^{(i)}$ of the parameters of $\lambda(X, \theta)$, construct a lower bound $\mathcal{T}\left(\theta, \theta^{(i)}\right) \leq \lambda(X, \theta)$
- **M-step**: maximize $\mathcal{T}\left(\theta, \theta^{(i)}\right)$ with respect to θ to obtain $\theta^{\{i+1\}}$

- Idea from Jensen's inequality:

$$\lambda(X,\theta) := \sum_{x \in D_\alpha} \log \sum_{k=1}^{K} q(k,x) \geq \sum_{x \in D_\alpha} \sum_{k=1}^{K} r^{(i)}(k \mid x) \log \frac{q(k,x)}{r^{(i)}(k \mid x)} = \mathcal{T}(\theta, \theta^{i})$$

Auxiliary Function

- Introduce a latent variable Z with pdf $r(z|X)$ conditioned on the data X

log likelihood

$$\lambda(X,\theta) := \log p(X;\theta) = \log \int p(X,z;\theta)dz = \log \int r(z\,|\,X)\frac{p(X,z;\theta)}{r(z\,|\,X)}dz$$

Jensen's

$$\geq \int r(z\,|\,X)\log\frac{p(X,z;\theta)}{r(z\,|\,X)}dz = \mathcal{T}(q,\theta)$$

Auxiliary function

- The **auxiliary function** is concave in r for a fixed θ and concave in θ for fixed r (but **not jointly** concave) assuming that $\log \mathrm{p}(\mathrm{z},\mathrm{X};\theta)$ is concave
- Local maxima of $\mathcal{T}(r,\theta)$ are local maxima of $\log p(X;\theta)$!

$$\text{(E step)} \qquad r^{(i)}(z\,|\,X) = \arg\max_{r} \mathcal{T}\left(r,\theta^{(i)}\right) = p(z\,|\,X,\theta^{(i)})$$

$$(M \;\; \text{step}) \qquad \theta^{(i+1)} = \arg\max_{\theta} \mathcal{T}(r^{(i)},\theta)$$

Gaussian Mixture MLE via EM (summary)

- Start with initial guess $\theta^{(i)} \coloneqq \left(\left\{a_k^{(i)}\right\}, \left\{\mu_k^{(i)}\right\}, \left\{\Sigma_k^{(i)}\right\}\right)$ and iterate:

(E step) $$r^{(i)}(k \mid x) = \frac{a_k^{(i)} \phi(x; \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{m=1}^{K} a_m^{(i)} \phi(x; \mu_m^{(i)}, \Sigma_m^{(i)})}$$

(M step) $$\mu_k^{(i+1)} = \frac{\sum_x r^{(i)}(k \mid x) x}{\sum_x r^{(i)}(k \mid x)}$$

$$\Sigma_k^{(i+1)} = \frac{\sum_x r^{(i)}(k \mid x)(x - \mu_k^{(i+1)})(x - \mu_k^{(i+1)})^T}{\sum_x r^{(i)}(k \mid x)}$$

$$a_k^{(i+1)} = \frac{1}{|D_\alpha|} \sum_x r^{(i)}(k \mid x)$$

Gaussian Mixture MLE via EM (comments)

- Sometimes the data is not enough to estimate all these parameters. Instead:
  - Fix the weights: $a_k = 1/K$
  - Fix diagonal $\left(\Sigma_k = diag\{\sigma_{k,1}^2, \dots, \sigma_{k,n}^2\}\right)$ or spherical $\left(\Sigma_k = \sigma_k^2 I\right)$ covariances
  - Estimate a **diagonal covariance**:

$$\Sigma_k^{(i+1)} = \frac{\sum_x r^{(i)}(k \mid x) diag(x - \mu_k^{(i+1)}) diag(x - \mu_k^{(i+1)})}{\sum_x r^{(i)}(k \mid x)}$$

  - Estimate a **spherical covariance**:

$$\sigma_k^{(i+1)} = \sqrt{\frac{1}{n} \frac{\sum_x r^{(i)}(k \mid x) \left\| x - \mu_k^{(i+1)} \right\|^2}{\sum_x r^{(i)}(k \mid x)}}, \qquad x \in \mathbb{R}^n$$

- How should we initialize θ? Use **k-means++**! If $\sigma_k \to 0$, the assignments become hard and the algorithm works like K-means.

E step (details)

why?

$$\text{(E step)} \qquad r^{(i+1)}(z \mid X) = \arg\max_{r} \mathcal{T}\left(r, \theta^{(i)}\right) = p(z \mid X, \theta^{(i)})$$

- **Kullback-Leibler (KL) divergence** from pdf p to pdf q is:

$$\boxed{d_{\mathcal{KL}}\left(p \,\|\, q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx}$$

$$\lambda(X, \theta) := \log p(X; \theta) \geq \mathcal{T}(r, \theta) = \int r(z \mid X) \log \frac{p(z \mid X; \theta) p(X; \theta)}{r(z \mid X)} dz$$

$$= -d_{\mathcal{KL}}\left(r(\cdot \mid X) \,\|\, p(\cdot \mid X; \theta)\right) + \log p(X; \theta)$$

- When maximizing the lower bound $\mathcal{T}(r, \theta)$ with respect to r, we are maximizing the similarity between $r(\cdot \mid X)$ and the conditional pdf $p(\cdot \mid X; \theta)$
- Choosing the optimal $r^*(\cdot \mid X) \equiv p(\cdot \mid X; \theta)$, makes the lower bound $\mathcal{T}(r^*, \theta)$ **tight**, i.e., it touches the log-likelihood function:

$$\mathcal{T}(r^*, \theta) = \mathcal{T}(p(\cdot \mid X; \theta), \theta) = \int p(z \mid X; \theta) \log \frac{p(z \mid X; \theta) p(X; \theta)}{p(z \mid X; \theta)} dz = \log p(X; \theta) = \lambda(X, \theta)$$

M step (details)

$$(M \ \text{step}) \qquad \max_\theta \mathcal{T}(r^{(i)}, \theta) = \max_\theta \int r^{(i)}(z \mid X) \log \frac{p(z, X; \theta)}{r^{(i)}(z \mid X)} dz$$

- **Differential Entropy** of pdf p is:

$$\boxed{H(p) = -\int p(x) \log p(x) dx}$$

$$\mathcal{T}(r^{(i)}, \theta) = \underbrace{H(r^{(i)}(\cdot \mid X))}_{\text{Entropy of } r^{(i)} \text{ does not depend on } \boldsymbol{\theta}} + \underbrace{\int r^{(i)}(z \mid X) \log p(z, X; \theta) dz}_{\text{weighted MLE}}$$

Entropy of $r^{(i)}$ does not depend on $\boldsymbol{\theta}$

weighted MLE where labeled examples $\{(x_i, z_i)\}$ are weighted by $r^{(i)}(z_i|x_i)$