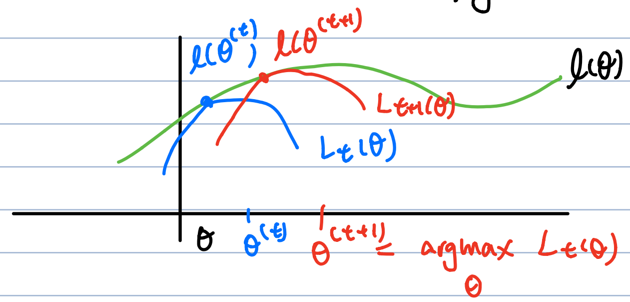
## Overview

- General Algo
- EM Algorithm, relate to MLE
- Factor Analysis

## Picture of General Algorithm




Property

1. $L_t(\theta) \le \ell(\theta)$ (lowerbound)
2. $L_t(\theta^{(t)}) = \ell(\theta^{(t)})$ (tight)

Hope: $L_t(\theta)$ easy to optimize

## Rough Algo

(E-STEP) 1. Find $L_t(\theta)$ given $\theta^t$

(M-STEP) 2. $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\ L_t(\theta)$

We go term $\log \sum_z P(x^{(i)}, z; \theta)$ by term

$$\log \sum_z P(x^{(i)}, z; \theta) = \log \sum_z \underbrace{\frac{Q^{(i)}(z)}{Q^{(i)}(z)}}_{=1} P(x^{(i)}, z; \theta)$$

Any prob distribution for $Q$

Pick $Q^{(i)}(z)$ s.t. $\sum_z Q^{(i)}(z) = 1$ & $Q^{(i)}(z) \ge 0$ $\forall z$. (*)

$$= \log \mathbb{E}_{z \sim Q^{(i)}} \left[ \frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)} \right]$$

Recall: $\log(E[X]) \ge E[\log(X)]$ (Jensons)

$$\ge \mathbb{E}_{z \sim Q^{(i)}} \left[ \log \frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)} \right]$$ (found lowerbound)

\# Note this holds for any $Q$ that satisfy (*)

Property 1

$$= \sum_z Q^{(i)}(z) \log \frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)}$$

How to Pick $Q^{(i)}(z)$ that satisfy Property 2?

$\ell(\theta^{(t)}) = L_t(\theta^{(t)})$ "tight"

want $\log \sum_z P(x, z; \theta) = \underbrace{\sum_z Q^{(i)}(z) \log \frac{P(x^{(i)}, z, \theta)}{Q^{(i)}(z)}}_{\bullet}$ (7)

TAKE $Q^{(i)}(z) = P(z \mid x^{(i)}; \theta)$

$$\frac{P(x^{(i)}, z; \theta)}{P(z \mid x^{(i)}; \theta)} = P(x^{(i)}; \theta) \quad \text{does not depend on } z$$

$$= \log P(x^{(i)}; \theta) \underbrace{\sum_z Q^{(i)}(z)}_{\to 1} = \log P(x^{(i)}; \theta)$$

#Note $Q^{(i)}$ varies for every point.

We call $\text{ELBO}(x, Q, \theta) = \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)}$

we've shown $\ell(\theta) \geq \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, Q^{(i)}, \theta)$

$$\ell(\theta^{(t)}) = \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, Q^{(i)}, \theta^{(t)})$$

Warm Up: Mixture of Gaussian

$$P(x^{(i)}, z^{(i)}) = P(x^{(i)} \mid z^{(i)}) P(z^{(i)})$$

"In Clusters" $z^{(i)} \sim \text{Multinomial}(\phi)$. $\phi_i \geq 0$, $\sum_i \phi_i = 1$

"cluster Means" $x^{(i)} \mid z^{(i)} = j \sim N(\mu_j, \sigma_j^2)$

$z^{(i)}$: Latent variable

What is EM here?

$$Q^{(i)}_{j} = P(z^{(i)} = j \mid x; \theta) \sim P(x^{(i)} \mid z^{(i)} = j; \theta)$$

"GMM" in terms of

$$\frac{P(x^{(i)} \mid z^{(i)} = j; \theta)\, P(z^{(i)} = j)}{P(x; \theta)}$$

(via Bayes rule)

M-step.

$$\max_{\phi, \mu, \sigma^2} \sum_{i=1}^{n} \overbrace{\sum_{z^{(i)}} Q^{(i)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}}^{f_i(\theta)}$$

$$w_j^{(i)} = Q^{(i)}_{j}$$

$$f_i(\theta) = \sum_j w_j^{(i)} \log \frac{\frac{1}{2\pi |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j) \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\} \phi_j}{w_j^{(i)}}$$

$$\nabla_{\mu_j} \sum_i f_i(\theta) = \sum_i \nabla_{\mu_j} w_j^{(i)} \cancel{\log \exp} \left\{-\frac{1}{2}(x^{(i)} - \mu_j) \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\}$$

$$= -\frac{1}{2}\sum_i w_j^{(i)} \Sigma_j^{-1}(x^{(i)}-\mu_j) = -\frac{1}{2}\Sigma_j^{-1}\left(\sum_i w_j^{(i)}(x^{(i)}-\mu_j)\right)$$

Assume $\Sigma_j^{-1}$ exists (full rank)

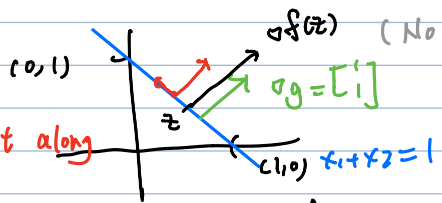
Setting to 0. $\mu_j = \frac{\sum_i w_j^{(i)} x^{(i)}}{\sum_i w_j^{(i)}}$

$\nabla_{\phi_j} f_i(\theta) = \sum_j \nabla_{\phi_j} w_j^{(i)} \log \phi_j$ $\quad$ $\phi_j$ is constrained s.t. $\sum \phi_j = 1$

$\Rightarrow \nabla_{\phi_j} f_i(\theta) = \sum_j \nabla_{\phi_j}\left(w_j^{(i)} \log \phi_j + \lambda\left(\sum_{m=1}^{k} \phi_j - 1\right)\right)$ $\quad$ Lagrangian

Detour: Want to find critical points of

$f: \mathbb{R}^2 \to \mathbb{R}$, s.t. $x_1 + x_2 = 1$ $\quad$ $(g(x) = 1,\ g(x) = x_1 + x_2)$




Some component along the line

$\Rightarrow$ Not a critical point

If z is a critical point then $\nabla f(z)$ is parallel to $\nabla g(z)$. $\quad \nabla f(z) = -\lambda \nabla g(z)$

$L(x, \lambda) = f(x) + \lambda(g(x) - 1)$

$\nabla_x L = \nabla f(x) + \lambda \nabla g(x) = 0$ $\rightarrow$ Parallel Condition

$\nabla_x L = g(x) - 1 = 0$ $\rightarrow$ Constraint is Satisfied

$$\Rightarrow \nabla_{\phi_j} f_i(\theta) = \sum_{i=1}^{n}\left(\nabla_{\phi_j} w_j^{(i)} \log \phi_j + \lambda\left(\sum_{m=1}^{k} \phi_m - 1\right)\right)$$

$$= \sum_{i=1}^{n} \frac{w_j^{(i)}}{\phi_j} + \lambda = 0 \quad \Rightarrow \phi_j = \frac{-1}{\lambda}\sum_{i=1}^{n} w_j^{(i)}$$

Since $\phi_j = 1$ $\Rightarrow 1 = \sum_j \phi_j = -\frac{1}{\lambda}\sum_{i=1}^{n} \overset{1}{\sum_j w_j^{(i)}} = -\frac{n}{\lambda}$ $\Rightarrow \lambda = -n$

$w_j^{(i)} = Q_j^{(i)}$

$$\Rightarrow \phi_j = \frac{1}{n}\sum_{i=1}^{n} w_j^{(i)}$$

Message: EM recovers GMM Algorithm

NB: z is descrete, can replace with sum of integrals

# Factor Analysis

when "$n \ll d$" ( $n \gg d$, GMMs )

How do this happen?

Place senses all over the campus, record temp.
But only records for 30 days ( $n < d$ )

**Key idea**: Assume there is some latent variable that is not complex and explains behavior

1st. lets see the problem w/ fitting a single Gaussian

Given: $x^{(1)} \dots x^{(\overset{\text{small}}{n})} \in \mathbb{R}^{d}$ large

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

n rank 1 vecs

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} (x^{(i)}-\mu)(x^{(i)}-\mu)^T \in \mathbb{R}^{d\times d}$$

**Rank $(\Sigma) \le n < d \Rightarrow$ Not full rank**

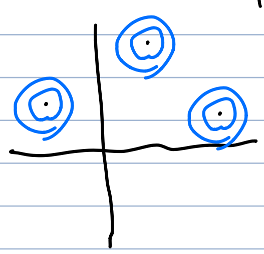
$$P(x;\mu,\Sigma) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\{(x-\mu)^T \Sigma^{-1} (x-\mu)\}$$

zero

Not defined

## Building Block 1.

Suppose independent & identical r.v.



Covariance are circles

$$\Sigma = \sigma^2 I$$

scalar

$$\min_{\mu,\Sigma} \sum_{i=1}^{n} (x-\mu)^T \Sigma^{-1} (x-\mu) + \log|\Sigma|$$

$$= \sigma^{-2} \sum_{i=1}^{n} (x^{(i)}-\mu)^T (x-\mu) + \log \sigma^{2d}$$

let $z = \sigma^2$

$$= \sigma^{-2} \sum_{i=1}^{n} (x^{(i)}-\mu)^T (x-\mu) + d \log \sigma^2$$

$$= z^{-1} \sum_{i=1}^{n} (x^{(i)}-\mu)^T (x-\mu) + d \log z$$

$$= z^{-1} \underbrace{\sum_{i=1}^{n} \| x^{(i)}-\mu \|^2}_{c} + d \log z$$

$$\nabla_z = -z^{-2} c + n\frac{d}{z} \qquad \Rightarrow -c = dz$$

$$\sigma^2 = \frac{c}{nd} \qquad z = \frac{c}{nd}$$