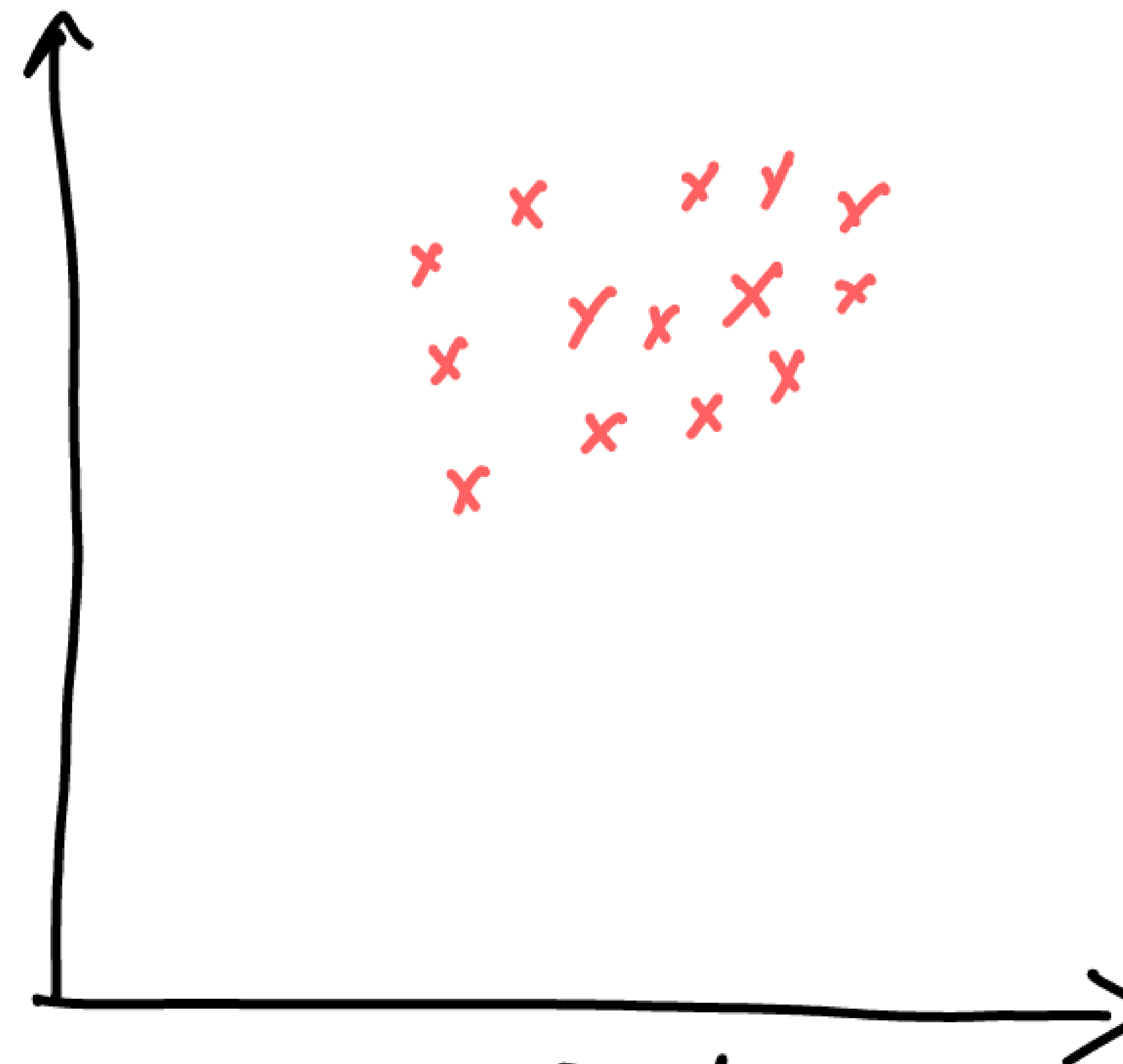
$$p(\theta|X) = \frac{p(X|\theta)\,p(\theta)}{p(X)}$$



X = Data $= \{x_1, x_2 \dots x_n\}$

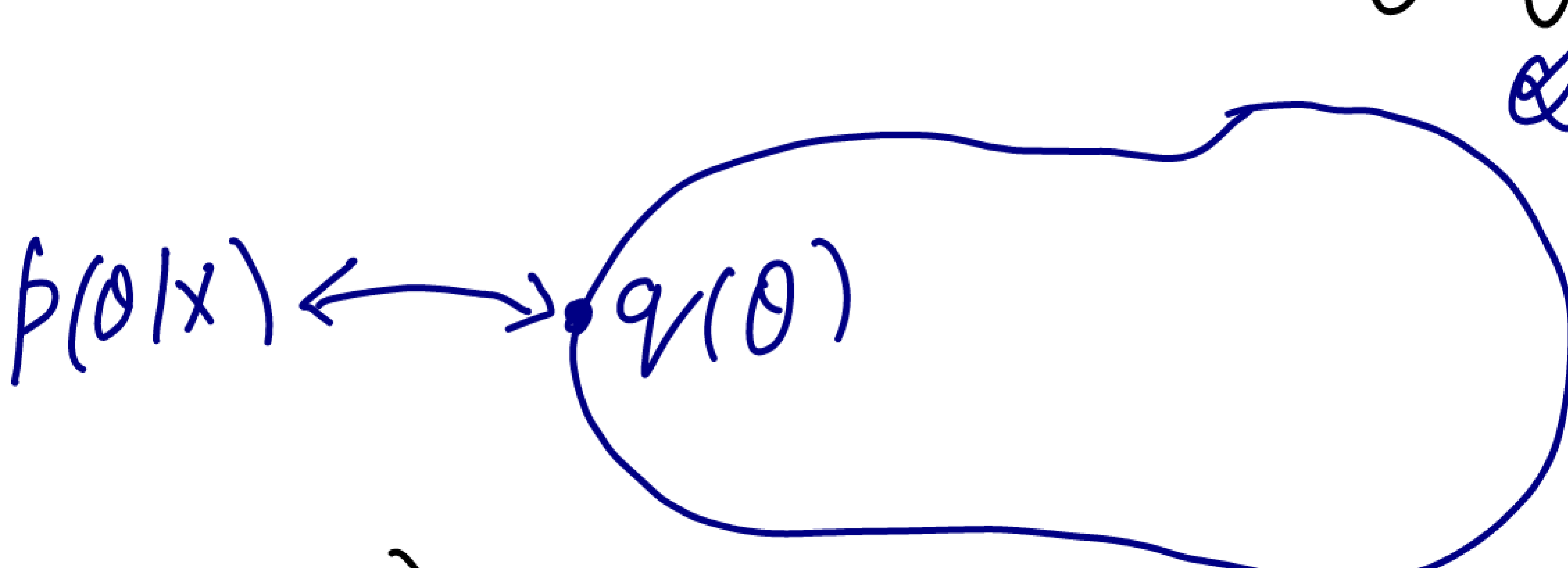
where

$$p(X) = \int_\theta p(X|\theta)\,p(\theta)\,d\theta$$

$$= \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_m} p(x|\theta_1,\theta_2\cdots\theta_m)\,d\theta_1\ldots d\theta_m$$

multiple integral intractable to compute.

To approximate the posterior $p(\theta|X)$ we have to find another posterior from family of dist^n $\mathcal{Q}$.



$$KL\left(q(\theta)\,||\,p(\theta|X)\right) = \int q(\theta)\log\frac{q(\theta)}{p(\theta|X)}\,d\theta$$

$$= \int q(\theta)\left(\log q(\theta) - \log p(\theta|X)\right)d\theta$$

$$= \int q(\theta)\log q(\theta)\,d\theta - \int q(\theta)\log p(\theta|X)\,d\theta$$

$$= \mathbb{E}_q\left[\log q(\theta)\right] - \mathbb{E}_q\left[\log p(\theta|X)\right]$$

$$\because\ p(\theta|x) = \frac{p(\theta,x)}{p(x)}$$

$$= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(x,\theta) - \log p(x)]$$

$$KL\left(q(\theta) \| p(\theta|x)\right) = \mathbb{E}_q\left[\log q(\theta) - \log p(x,\theta)\right] + \log p(x)$$

$$\underbrace{\mathbb{E}_q\left[\log p(x,\theta) - \log q(\theta)\right]}_{\text{Evidence lower bound}} = \log p(x) - KL\left(q(\theta) \| p(\theta|x)\right)$$

$$ELBO(q) = \mathbb{E}_q\left[\log \frac{p(x,\theta)}{q(\theta)}\right]$$

we need to maximize ELBO to find parameter of Approximate posterior

Let's us unpack ELBO

$$ELBO(q) = \mathbb{E}_q\left[\log p(X,\theta) - \log q(\theta)\right]$$

$$= \mathbb{E}_q\left[\log p(x,\theta)\right] - \mathbb{E}_q\left[\log q(\theta)\right]$$

$$= \mathbb{E}_q\left[\log\left(p(x|\theta)\, p(\theta)\right)\right] - \mathbb{E}_q\left[\log q(\theta)\right]$$

$$ELBO(q) = \mathbb{E}_q\left[\log p(x|\theta)\right] + \mathbb{E}_q\left[\log p(\theta) - \log q(\theta)\right]$$

$$= \mathbb{E}_q\left[\log p(x|\theta)\right] + \int q(\theta) \log \frac{p(\theta)}{q(\theta)}\, d\theta$$

$$ELBO(q) = \underbrace{\mathbb{E}_q\left[\log p(x|\theta)\right]}_{\text{log liklihood}} - \underbrace{KL\left(q(\theta) \| p(\theta)\right)}_{\text{KL-div}}$$

In our case of 2-dim data $\left(\begin{matrix}\text{dim}\\ K=2\end{matrix}\right)$

$\theta = (\mu, \Sigma)$

$q(\theta) \longrightarrow$ parameterize by variational parameter μ_q, Σ_q we have to find this

$p(\theta) \longrightarrow$ parametrize by prior parameter obtain from data

$$\mu_p = \left(\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \; \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$\Sigma_p = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(x_i - \mu_x)}_{2\times 1} \underbrace{(y_i - \mu_y)^T}_{1\times 2}$$

Now log liklihood:

$$\text{log-liklihood} = \log p(X|\theta) \quad \theta = (\mu_q, \Sigma_q)$$

$$= \log\left(\prod_{i=1}^{n} p(x_i|\theta)\right) \quad \text{data } X \text{ in iid}$$

$$= \log\left(\prod_{i=1}^{n} p(x_i \mid \mu_q, \Sigma_q)\right)$$

$$= \log\left(\prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma_q|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_q)^T \Sigma_q^{-1} (x-\mu_q)\right\}\right)$$

$$\log p(X|\theta) = \sum_{i=1}^{n}\left[\log\left(\frac{1}{(2\pi)^{K/2} |\Sigma_q|^{1/2}}\right) - \frac{1}{2}(X_i - \mu_q)^T \Sigma_q^{-1} (X_i - \mu_q)\right]$$

$$\log p(X|\theta) = n \log\left(\frac{1}{(2\pi)^{K/2} \, |\Sigma_q|^{1/2}}\right) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_q)^T \Sigma_q^{-1}(x_i-\mu_q)$$

$$\boxed{\log p(X|\theta) = n\left[-\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_q|\right] - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_q)^T \Sigma_q^{-1}(x_i-\mu_q)} \quad \text{— } A$$

where $x_i \in X = \underbrace{\{x_1, x_2, \dots x_n\}}_{\text{Data}}$

KL Divergence :

$$KL(q(\theta)\,\|\,p(\theta)) \qquad q(\theta) = \mathcal{N}(\mu_q, \Sigma_q)$$

$$p(\theta) = \mathcal{N}(\mu_p, \Sigma_p)$$

$$\boxed{KL(q\|p) = \frac{1}{2}\left[-\log\frac{|\Sigma_q|}{|\Sigma_p|} - K + (\mu_p-\mu_q)^T \Sigma_p^{-1}(\mu_p-\mu_q) + \mathrm{tr}\{\Sigma_p^{-1}\Sigma_q\}\right]} \quad \text{— } B$$

To update variational distribution $q(\theta)$ parameter we need to find gradient of ELBO w.r.t parameters

$$\mu_{q\_new} = \mu_{q_{old}} + \eta \frac{\partial ELBO}{\partial \mu_q}$$

$$\Sigma_{q_{new}} = \Sigma_{q_{old}} + \eta \frac{\partial ELBO}{\partial \Sigma_q}$$

Applying Gradient Ascent

<u>Now</u>

$$\frac{\partial ELBO}{\partial \mu_q} = \frac{\partial \log p(X|\theta)}{\partial \mu_q} - \frac{\partial KL(q(\theta)\|p(\theta))}{\partial \mu_q}$$

from (A)

let

$$\frac{\partial \log p(X|\theta)}{\partial \mu_q} = -\frac{1}{2}\sum_{i=1}^{n}(-1)\,2\,(x_i-\mu_q)\Sigma_q^{-1}$$

$$= \sum_{i=1}^{n}(x_i-\mu_q)\Sigma_q^{-1}$$

from (B)

$$\frac{\partial KL(q(\theta)\|p(\theta))}{\partial \mu_q} = -\Sigma_p^{-1}(\mu_p-\mu_q)$$

$$\boxed{\frac{\partial ELBO}{\partial \mu_q} = \sum_{i=1}^{n}(x_i-\mu_q)\Sigma_q^{-1} + \Sigma_p^{-1}(\mu_p-\mu_q)}$$

Now $$\frac{\partial ELBO}{\partial \Sigma_q} = \frac{\partial \log p(X|\theta)}{\partial \Sigma_q} - \frac{\partial KL\left(q(\theta)\,\|\,p(\theta)\right)}{\partial \Sigma_q}$$

let

$$\frac{\partial \log p(X|\theta)}{\partial \Sigma_q} = \frac{\partial}{\partial \Sigma_q}\left[n\left\{-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_q|\right\} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_q)^T\Sigma_q^{-1}(x_i-\mu_q)\right]$$

$$= -\frac{n}{2}\frac{1}{|\Sigma_q|}\frac{\partial|\Sigma_q|}{\partial \Sigma_q} - \frac{1}{2}\sum_{i=1}^{w}\left\{-\Sigma_q^{-1}(x_i-\mu_q)(x_i-\mu_q)^T\Sigma_q^{-1}\right\}$$

$$\frac{\partial |X|}{\partial X} = |X|(X^{-1})^T$$

$$\frac{\partial a^T X^{-1} b}{\partial X} = -X^{-1}ab^TX^{-1}$$

if x is symmetrical matrix

$$= -\frac{n}{2}\frac{|\Sigma_q|(\Sigma_q^{-1})^T}{|\Sigma_q|} + \frac{1}{2}\sum_{i=1}^{w}\left[\Sigma_q^{-1}(x_i-\mu_q)^2\Sigma_q^{-1}\right]$$

$$\frac{\partial \log p(X|\theta)}{\partial \Sigma_q} = -0.5\,n\left(\Sigma_q^{-1}\right)^T + 0.5\sum_{i=1}^{n}\left[\Sigma_q^{-1}(x_i-\mu_q)^2\Sigma_q^{-1}\right]$$

$$\boxed{\frac{\partial \log p(X|\theta)}{\partial \Sigma_q} = -0.5\left[n\left(\Sigma_q^{-1}\right)^T - \sum_{i=1}^{n}\left[\Sigma_q^{-1}(x_i-\mu_q)^2\Sigma_q^{-1}\right]\right]}$$

$$\frac{\partial KL(q(\theta)\|p(\theta))}{\partial \Sigma_q} = \frac{\partial}{\partial \Sigma_q}\left\{ \frac{1}{2}\left\{ -\log\frac{|\Sigma_q|}{|\Sigma_p|} - k + (\mu_p-\mu_q)^T \Sigma_p^{-1}(\mu_p-\mu_q) + tr(\Sigma_p^{-1}\Sigma_q) \right] \right.$$

$$= \frac{1}{2}\left[-\underbrace{\frac{\partial}{\partial \Sigma_q}\left\{\log\frac{|\Sigma_q|}{|\Sigma_p|}\right\}} + \underbrace{\frac{\partial}{\partial \Sigma_q}\left\{tr(\Sigma_p^{-1}\Sigma_q)\right\}} \right]$$

$$\frac{\partial}{\partial \Sigma_q}\left[\log|\Sigma_q| - \log|\Sigma_p|\right]$$

$$= \frac{1}{|\Sigma_q|}\frac{\partial}{\partial \Sigma_q}|\Sigma_q|$$

$$= \frac{1}{|\Sigma_q|}|\Sigma_q|(\Sigma_q^{-1})$$

$$= (\Sigma_q^{-1})^T$$

$$\boxed{\because \frac{\partial |X|}{\partial X} = |X|(X^{-1})^T}$$

$$\frac{\partial}{\partial \Sigma_q}\left[tr\{\Sigma_p^{-1}\Sigma_q\}\right]$$

$$= (\Sigma_p^{-1})^T$$

$$\boxed{\because \frac{\partial\, Tr(AX)}{\partial X} = A^T}$$

$$\frac{\partial ELBO}{\partial \Sigma_q} = \frac{\partial \operatorname{Log} p(X|\theta)}{\partial \Sigma_q} - \frac{\partial KL(q(\theta)\,||\,p(\theta))}{\partial \Sigma_q}$$

$$\boxed{\begin{aligned}\frac{\partial ELBO}{\partial \Sigma_q} &= -0.5\left[n\left(\Sigma_q^{-1}\right)^T - \sum_{i=1}^{n}\left(\Sigma_q^{-1}(x_i-\mu_q)^2\Sigma_q^{-1}\right)\right] \\ &\quad -\frac{1}{2}\left[-\Sigma_q^{-1}+\left(\Sigma_p^{-1}\right)^T\right]\end{aligned}}$$