Density fitting:

$\rho_\theta(x) \in \mathbb{R}^d$ density.

α_θ as a measure

$$\frac{d\alpha_\theta}{dx} = \rho_\theta$$

- ρ_θ
- Algorith$_\theta$ create sample

Max likelihood:

$\{x_i\}_i$ are indep. identically dist

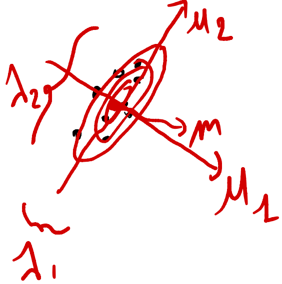
i.i.d $\alpha_\theta // \rho_\theta$

$$\max_\theta \mathbb{P}(\{x_i\}) = \prod_{i=1}^{n} \underbrace{\mathbb{P}(x_i)}_{\rho_\theta(x_i)}$$

$$\min_\theta \mathcal{L}(\theta) = -\log\left(\prod \dots\right)$$

$$= \underbrace{-\frac{1}{n}\sum_i}_{\text{“}\int\text{”}} \log(\rho_\theta(x_i))$$

Gaussian case: $\theta = (m \in \mathbb{R}^d, \underbrace{\Sigma}_{\succeq 0} \in \mathbb{R}^{d\times d})$



$U = (u_1 | u_2) \qquad u_i \in \text{eigenvectors}(\Sigma)$

$$\rho_\theta(x) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\Big(-\frac{1}{2} \langle \underbrace{\Sigma^{-1}}_{\text{"precis°"}}(x-m), x-m \rangle\Big)$$

$$\min_{m,\Sigma} \sum_i \frac{1}{2} \underbrace{\log\det(\Sigma)}_{\text{convex}} + \frac{1}{2} \underbrace{\langle \Sigma^{-1}(m-x_i), m-x \rangle}_{\text{convex}}$$

$\dfrac{x^2}{\sigma}$ convex

$$\nabla \mathcal{E}\begin{pmatrix} m \\ \Sigma \end{pmatrix} = 0 \qquad m^* = \frac{1}{n} \sum_i x_i$$

$$\Sigma^* = \frac{1}{n} \sum_i \underbrace{(x_i - m)(x_i - m)^T}_{\text{rank } 1}$$

Convexo - MLE / IT : $n \to +\infty$

$(x_i)_i \sim \text{iid } \hat{\mu}$

$$\min_\theta -\frac{1}{n} \sum \log(\rho_\theta(x_i)) = \sum \log\left(\frac{1/n}{\rho_\theta(x_i)}\right) 1/n \quad (+ \log(n))$$

$$\| \\ \mathcal{E}(\theta)$$

$$\tilde{\mathcal{L}}(\theta) \xrightarrow{n\to+\infty} \underbrace{\int \log\left(\frac{d\hat{\mu}}{d\mu_\theta}(x)\right) d\hat{\mu}(x)}$$

$$\triangleq KL(\hat{\mu}|\mu_\theta)$$

test DATA ref MODEL

Prop: $KL(\mu|\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu \geqslant 0$

$= 0 \iff \mu = \nu$

KL "distance like"

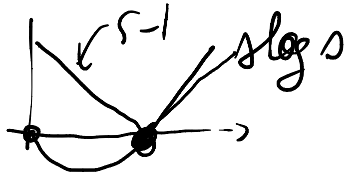
Def: φ-divergence // Ciszar div

$$D_\varphi(\mu|\nu) = \int \varphi\left(\frac{d\mu}{d\nu}\right) d\nu$$

KL using $\varphi(s) = s\log(s)$

Hyp: φ convex + $\varphi(1) = 0$

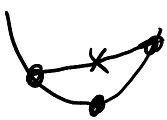
$\varphi(s) = |s-1|$



$D_\varphi(\mu|\nu) = \int |\mu-\nu|$

Jensen - Shanan

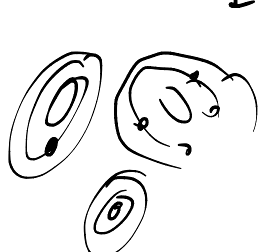
Proof: $D_\varphi(\mu|\nu) = \int \varphi\left(\frac{d\mu}{d\nu}(x)\right) d\nu(x) \geq 0$



JENSEN: φ cvx $\rightarrow$ $\varphi\left(\underbrace{\int \frac{d\mu}{\cancel{d\nu}}(x)\, \cancel{d\nu(x)}}_{}\right) \leq \int \varphi\left(\frac{d\mu}{d\nu}\right) d\nu$

$\varphi\left(\underset{\substack{\parallel \\ 1}}{\int d\mu}\right) = \varphi(1) = 0$

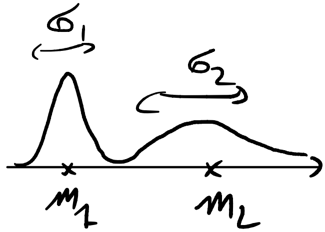
Mixture models:



$\pi = (\pi_k)_{k=1}^{K} \longrightarrow$ draw k according π

$\hookrightarrow x_i \sim p_{\theta_k}(x)$

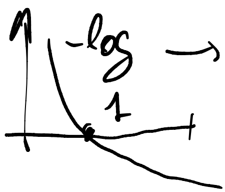
$\rho_\theta(x) = \sum_k \pi_k \, \rho_{\theta_k}(x)$



$\mathcal{L}(\theta) = -\sum_i \log \underbrace{\sum_k \pi_k \, \rho_{\theta_k}(x_i)}$ non cvx

EM Expectation Maximization

$-\log \to$ cvx



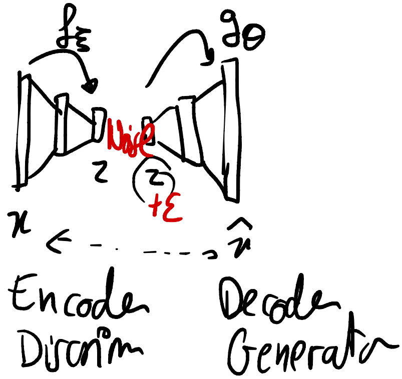
PDF: $\rho_\theta \to f(\cdot|\theta)$

k-means: $\to$ hard cluster $p_{ik} \in \{0,1\}$

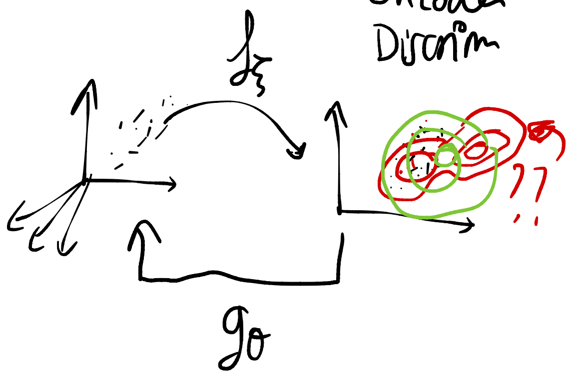
$\to \Sigma \propto Id$

$\Sigma = \sigma Id, \ \sigma \to 0$ EM $\to$ k-means

$\boxed{\text{V}}$AE : $\boxed{\text{AE}}$



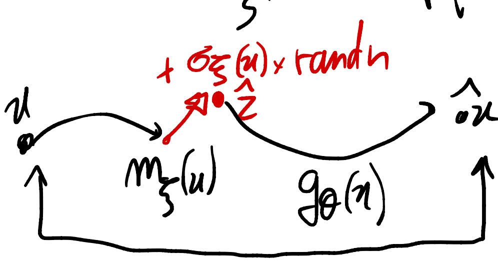
Non-linear PCA



?? ξ, θ param

$$AE(\theta,\xi) = \frac{1}{n}\sum_i \|x_i - g_\theta(f_\xi(x_i))\|^2$$

$$f_\xi(x) = \begin{pmatrix} m_\xi(x) \\ \sigma_\xi(x) = \exp(s_\xi(x)) > 0 \end{pmatrix} \begin{matrix} \leftarrow \text{mean} \\ \leftarrow \text{noise level} \end{matrix}$$



$$VAE(\theta,\xi) = \mathbb{E}_{\text{noise}} \sum_i \|x_i - \hat{x}_i\|^2 + KL(\mathcal{N}_i, \mathcal{N}(0,1))$$

$$\mathcal{N}_i = \mathcal{N}(m_\xi(x_i), \sigma_\xi(x_i))$$

$$KL(\mathcal{N}_i, \mathcal{N}(0,1)) = \sigma^2(x_i) + \|m_\xi(x_i)\|^2 + \log(\sigma_i^2)$$

Norm. Flow : $\rho_\theta(x)$.

$\rho_0(x)$
$N(0, Id)$

$x \sim$ rand. $\qquad g_\theta(x) = T_k T_{k-1} \cdots T_1(x)$.

(*) $\min_\theta \frac{1}{n} \sum \log(\rho_\theta(x_i))$

$$\rho_0(x) \xrightarrow{T = T_1} \rho_1(y) \qquad y = Tx$$

$$\rho_1(y) = \rho_0(\underbrace{T^{-1}y}_{x}) \times \frac{1}{|\det \partial T(T^{-1}y)|}$$

$$\min_{\theta_1 - \theta_k} \frac{1}{n} \sum_i \sum_k \log|\det(\partial T(x_i^{(k)}))| + \frac{1}{2}\|x_i^{(0)}\|^2$$

$$T_\theta : x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longrightarrow \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \exp(A_\theta(x_1)) + B_\theta(x_1) \end{bmatrix}$$

Real-Valued Non-Volume Preserving Flows

R-NVP

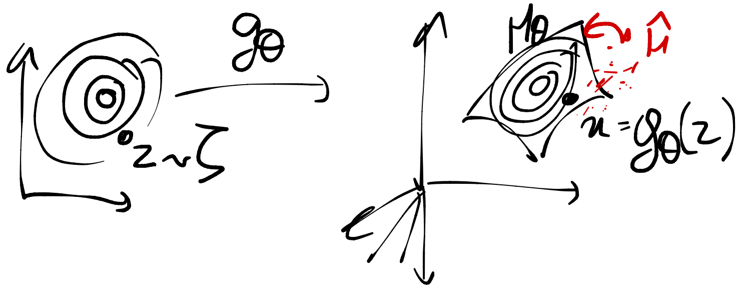
$$y \to x = \begin{bmatrix} x_1 = y_1 \\ x_2 = (y_2 - B_\theta(y_1)) \cdot \exp(-A_\theta(y_1)) \end{bmatrix}$$

$$\frac{\partial T}{\partial x} = \left[\begin{array}{c|c} Id & 0 \\ \hline \cdots & \mathrm{diag}(\exp A) \end{array}\right]$$

$$\det \frac{\partial T}{\partial x} = \prod_d \exp(A(x_1)_d)$$

$$\min \sum_i \sum_k \sum_d A(x_i^{(k)})_d + \| x_i^{(0)} \|^2 = NF(\theta)$$

GAN:



$$\inf_\theta D_\varphi(\hat{\mu} | \mu_\theta) \qquad \varphi(s) = s\log(s)$$

$$JS(\mu|\nu) = J(\nu|\mu) = KL\left(\mu \,\middle|\, \frac{\mu+\nu}{2}\right) + KL\left(\nu \,\middle|\, \frac{\mu+\nu}{2}\right)$$

Prop: sym, satisfies triang. ineq. (a TV).

$0 \leq JS \leq 1$

$$JS = D_\varphi \qquad \varphi(s) = s \log\left(\frac{2s}{s+1}\right) + \log\left(\frac{2}{s+1}\right)$$

Legendre transform: $\varphi^*(t) = \sup_s st - \varphi(s)$

$$t = \varphi'(s)$$

$$(\varphi^*)' = [\varphi']^{-1} \qquad s = (\varphi')^{-1}(t)$$

Prop: if φ is convex, $(\varphi^*)^* = \varphi$

$$\varphi(s) = \sup_t st - \varphi^*(t).$$

Prop / Computation

$$D_\varphi(\mu|\nu) = \int \varphi\left(\frac{d\mu}{d\nu}(x)\right) d\nu(x)$$

$$= \int \left[\sup_{t(x)} t(x) \times \frac{d\mu}{d\nu}(x) - \varphi^*(t(x))\right] d\nu(x)$$

$$= \sup_{t(x)} \int \underbrace{t(x)}\, \underbrace{d\mu(x)}_{\hat{\mu}} - \int \varphi^*\big(\underbrace{t(x)}\big)\, \underbrace{d\nu(x)}_{\mu_\theta}$$

$x = g_\theta(z)$

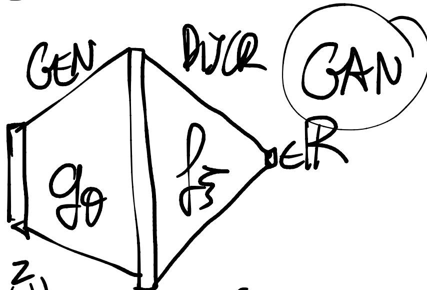
$$\min_\theta D_\varphi(\hat{\mu} \mid \mu_\theta)$$

$$\| $$

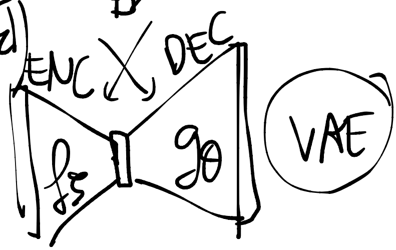
$$\min_\theta \max_{\boxed{t(x)}} \underbrace{\int t(x)\, d\hat{\mu}(x)}_{\frac{1}{n}\sum_i t(x_i)} - \int \varphi^*\big(t(g_\theta(z))\big)\, d\zeta(z)$$

GAN move

$t(x) = f_\xi(x)$



$$\min_\theta \max_\xi \frac{1}{n}\sum_i f_\xi(x_i) - \mathbb{E}_{z\sim N}\Big(\varphi^*\big(f_\xi(g_\theta(z))\big)\Big)$$



in practice $\longrightarrow$ sample $\mathbb{E}_N$

SGD $\longrightarrow$ k_1 step desc θ

k_2 step ascent ξ

$k_1 \gg k_2 = 1$

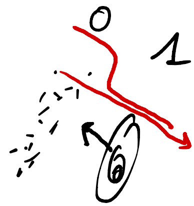
KL $\quad \varphi(s) = s\log s \qquad \varphi^*(t) = e^{t-1}$

JS $\quad \varphi(s) = s\log\left(\frac{2s}{s+1}\right) + \log\left(\frac{2}{s+1}\right)$

$$\varphi^*(t) = \log(t) + \log(1-t)$$

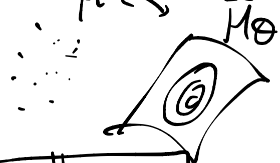
$$0 \le t \le 1$$

$\Longrightarrow$ discr. $f_\xi(x) \in [0,1]$



WGAN :

$\hat{\mu} \leftarrow W_1(\hat{\mu}, \mu_\Theta)$



$\boxed{\|\nabla f_\xi(x)\| \le 1} \longrightarrow |\text{weight}| \le 1$