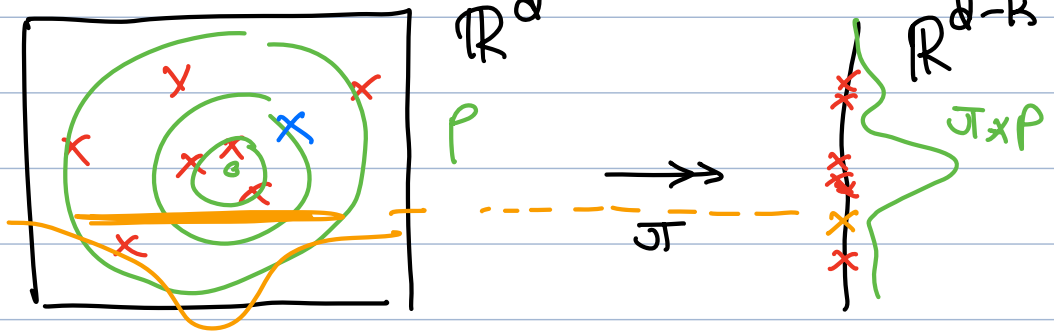
HOW TO FAKE ANYTHING

TALK ON VARIATIONAL AUTOENCODERS
FRIDAY, APRIL 19

I. INTRO + GOAL STATEMENT

MAIN GOAL: GIVEN DATA $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$,

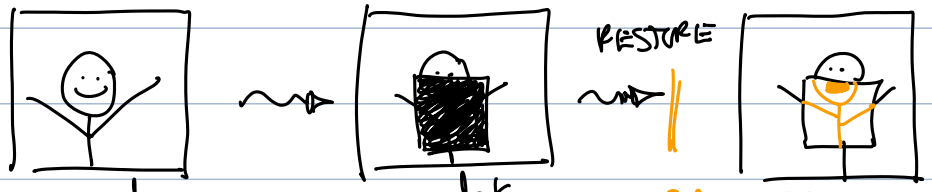
WANT TO FIND + BE ABLE TO USE PDF $x_i \sim p$.



USE? TO SAMPLE, FROM p (AND RELATED DISTS)!

1. SAMPLE FROM p = GENERATE NEW DATA
2. SAMPLE FROM CONDITIONAL $p(-|c)$ =

INPAINT / RESTORE CORRUPTED DATA



$\mathbb{R}^d$ (ORIGINAL)

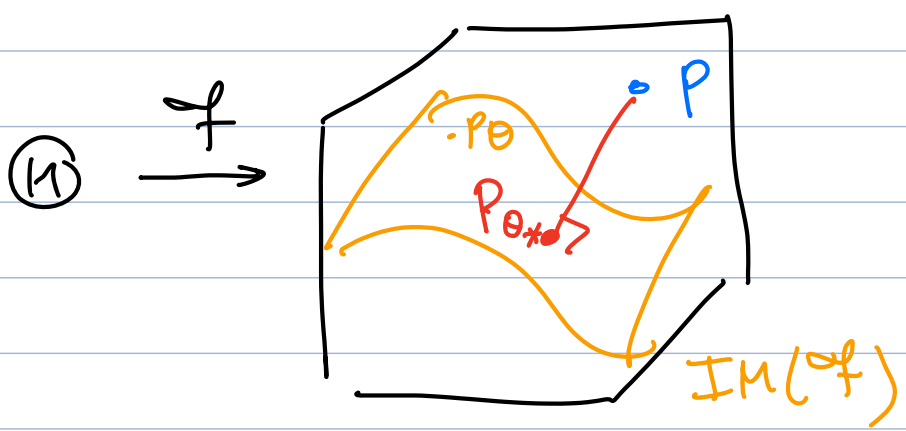
$\mathbb{R}^{d-k}$ CORRUPTED IMAGE IN PROJECTION $\mathbb{R}^{d-k}$

SAMPLE FROM CONDITIONAL $p(x \mid \text{CORR})$

APPROACHES:

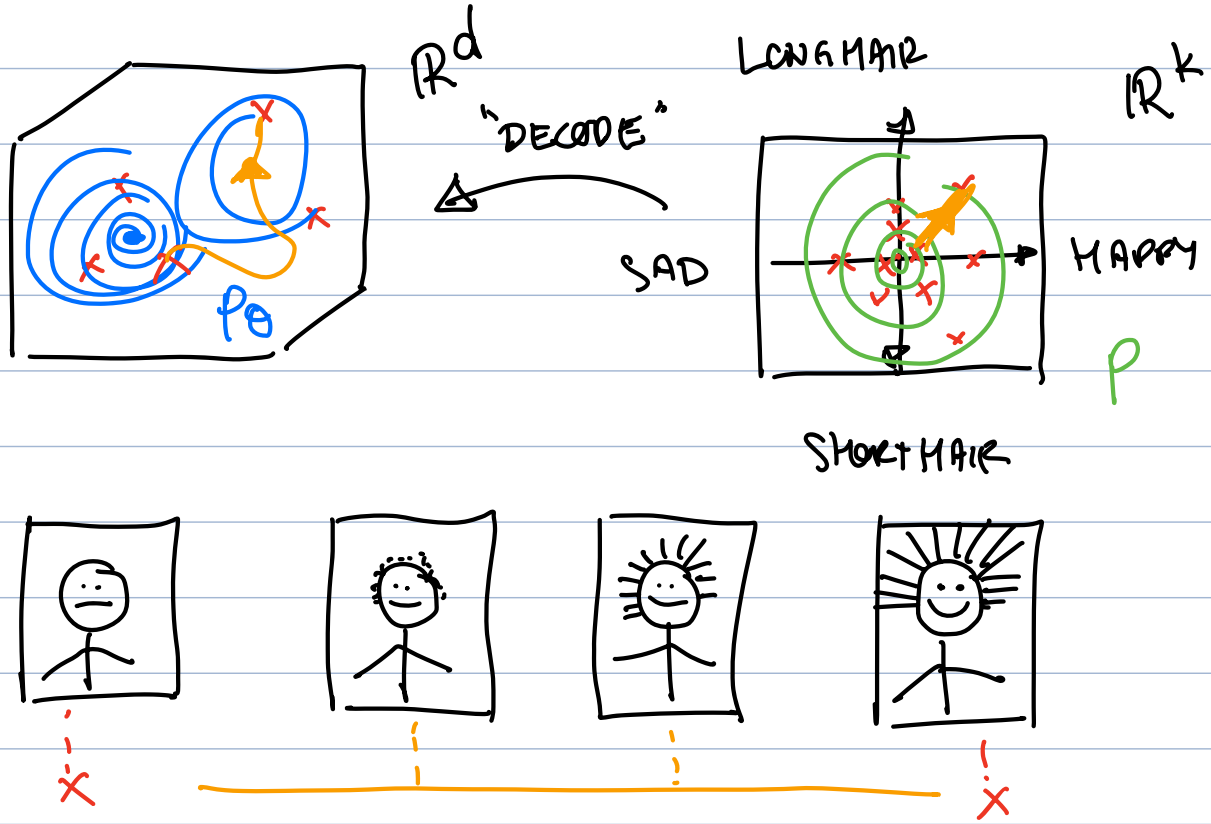
	PROS	CONS
① NON-PARAMETRIC:		
(a) HISTOGRAM	EASY!	
(b) DENSITY ESTIMATION USING	GENERAL!	HARD TO SAMPLE
{ KERNELS: $K_\lambda * P_{empirical}$	ASYMPTOTICALLY PRECISE	(NEED TO LEARN A FLOW TO A KNOWN DIST)
{ ORTHOG FUNCTIONS		
{ ...		
② PARAMETRIC:	EAGER TO SAMPLE	FUNDAMENTALLY IMPRECISE

PICK FAMILY (H) OF NICE DISTRIBUTIONS AND FIND APPROX P_{θ^*}:



ADDING LATENT VARIABLES

USE (3): INTERPOLATION B/T DATA IN A LEARNED FEATURE SPACE



I.E. WE POSIT THE EXISTENCE OF A LATENT SPACE $\mathbb{R}^k$

LATENT VAR $= z$

ST $$p_\theta(x) = \int_{\mathbb{R}^k} p_\theta(x|z)\, p(z)\, dz$$

$p_\theta(x|z)$: LIKELIHOOD — CHOOSE THIS TO BE GAUSSIAN

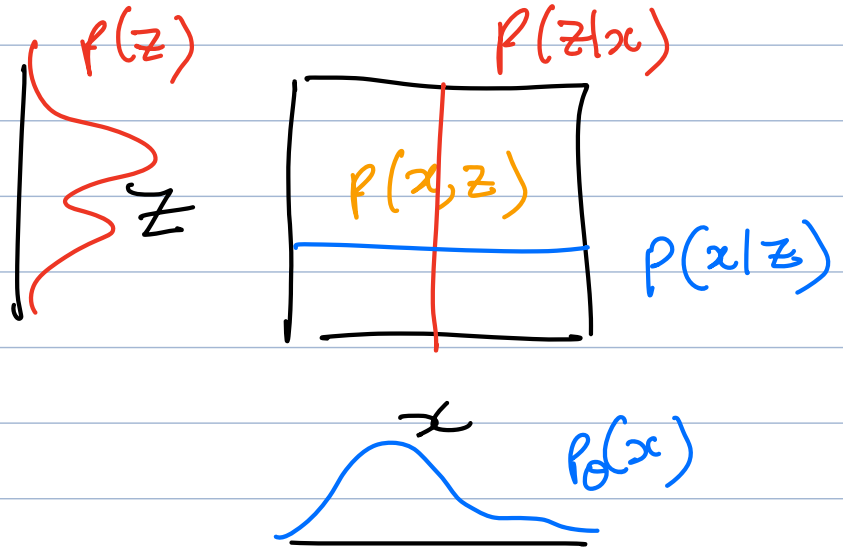
$p(z)$: CHOOSE / FIX THIS TO BE EASY DIST: $\sim N(0, I)$

$N(f(z), cI)$

PRIOR

IE $\theta = (f, c) \in \mathcal{F} \times \mathbb{R}_{>0}$

= (14)



$$p(x|z)p(z) = p(x,z) = p(z|x)p(x)$$

$$\therefore p(x) := \frac{p_\theta(x|z)}{p(z|x)} p(z)\, dz$$

POSTERIOR:

APPROXIMATE IT BY ... GAUSSIAN:

$$q_\phi(z|x) \sim N(g(x), h(x)\mathbb{1})$$

II. WHICH OBJECTIVE FUNCTION?

WE ∴ ARE AFTER MLE ESTIMATE

$$\theta^* = \underset{\theta,\phi}{\text{ARGMAX}} \prod_{i=1}^{n} p_\theta(x_i)$$

$$= \underset{\theta}{\text{ARGMAX}} \sum_{i=1}^{n} \log p_\theta(x_i)$$

$$= \underset{\theta}{\text{ARGMAX}} \sum_{i=1}^{n} \log \int_{R^k} p_\theta(x_i|z)p(z)dz$$

REPLACE BY:

$$(\theta^*, \phi^*) = \underset{\theta,\phi}{\text{ARGMAX}} \sum_{i=1}^{n} \log \int \frac{p_\theta(x_i|z)}{q_\phi(z|x_i)} p(z)\, q_\phi(z|x_i)dz$$

$$= \underset{\theta,\phi}{\text{ARGMAX}} \sum_{i=1}^{n} \log \mathbb{E}_{z\sim q_\phi(x_i)}\left[\frac{p_\theta(x_i|z)}{q_\phi(z|x_i)} p(z)\right]$$

LET'S LOOK AT:

$$\log \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\frac{p_\theta(x_i|z)}{q_\phi(z|x_i)} p(z)\right]$$

JENSEN

$$\geq \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log \frac{p_\theta(x_i|z)}{q_\phi(z|x_i)} p(z)\right]$$

① $\mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log p_\theta(x_i,z)\right] + \mathcal{H}(q_\phi(\cdot|x_i))$

≥ 0

$\text{ELBO} =: \mathcal{L}_i(\theta,\phi)$

② $= \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log \frac{p_\theta(z|x_i)}{q_\phi(z|x_i)} + \log p_\theta(x_i)\right]$

$= -D_{KL}\left(q_\phi(\cdot|x_i)\,||\,p_\theta(\cdot|x_i)\right) + \log p_\theta(x_i)$

≥ 0

THE START OF OUR WHOLE JOURNEY!

$\therefore \log p_\theta(x_i) = D_{KL}\left(q_\phi(\cdot|x_i)\,||\,p_\theta(\cdot|x_i)\right) + \mathcal{L}_i(\theta,\phi)$

≥ 0

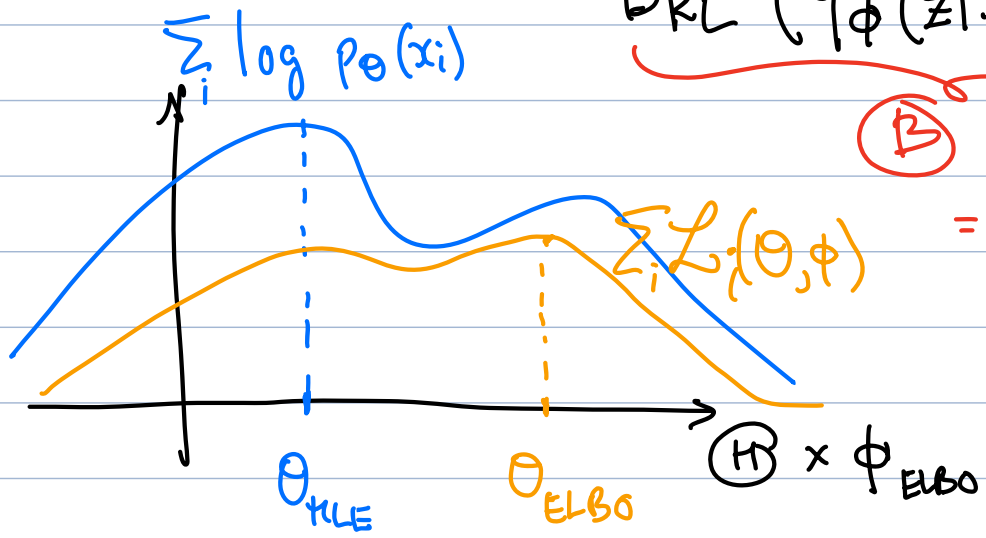
RECALL OUR GOAL: MAXIMIZE $\sum_{i=1}^{n} \log p_\theta(x_i)$

WE NOW REPLACE IT BY THE DIFFERENT BUT RELATED

NEW GOAL: MAXIMIZE

$$\mathcal{L}(\theta,\phi) = \sum_{i=1}^{n} \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log \frac{p_\theta(x_i|z)}{q_\phi(z|x_i)}\, p(z)\right]$$

$$= \sum_{i=1}^{n} \left\{ \underbrace{\mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log p_\theta(x_i|z)\right]}_{\text{(A)}} - \underbrace{D_{KL}\left(q_\phi(z|x_i)\,||\,p(z)\right)}_{\text{(B)}} \right\}$$



$= \mathcal{H}(q_\phi) + \mathbb{E}_{q_\phi}[\log p(z)]$

Q: WHAT DOES THIS NEW GOAL MEAN?

A: (1) LOOK AT ELBO ABOVE: TAKE DATA PT x_i AND

(i) SAMPLE $z \sim q_\phi(\cdot|x_i)$.

z IS CODE DESCRIBING x_i

(SO CALL q_ϕ THE ENCODER)

(ii) TERM (A) IS THE LOG-LIKELIHOOD OF THE OBSERVED x_i, GIVEN THE CODE

z THAT WE SAMPLED.

THIS TERM IS BIG WHEN $p(x_i|z)$ IS BIG ...

$p_\theta(x_i|z)$ TRIES TO RECONSTRUCT x_i FROM CODE

(SO CALL $p_\theta(x|z)$ THE DECODER)

∴ = RECONSTRUCTION ERROR !

(iii) TERM (B): PRESSURE ON D_{KL} TO BE SMALL ... IE FOR q_ϕ TO LOOK LIKE $p(z)$

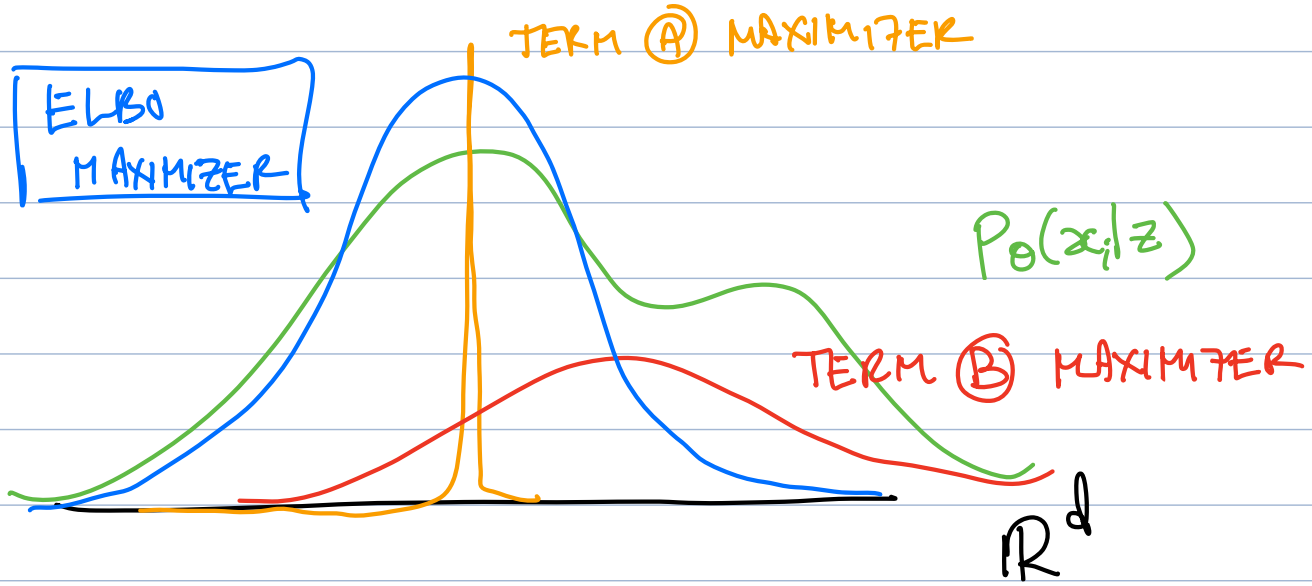
... IE FOR CODES z FOR x_i TO LOOK STD GAUSSIAN. I.E

PREVENTS CODE q_ϕ FROM SPIKING

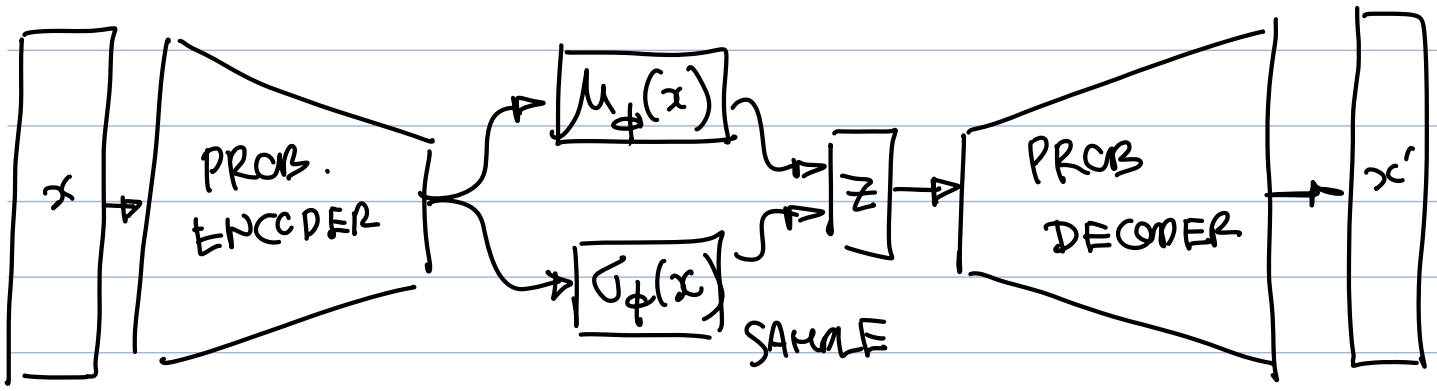
TO REMEMBER JUST THE IDENTITY MAP,

AND ∴ LEARN A MORE INTERESTING MAP

∴ REGULARIZATION TERM !


TERM (A) MAXIMIZER
ELBO
MAXIMIZER
$p_\theta(x_i|z)$
TERM (B) MAXIMIZER
$\mathbb{R}^d$

III. TRAINING



$$q_\phi(z|x) \sim N(\mu_\phi(x), \sigma_\phi(x)) \qquad p_\theta(x|z) \sim N(\mu_\theta(z), \sigma_\theta)$$

RECALL OUR OBJECTIVE: $\mathcal{L}_i(\theta, \phi)$

MAXIMIZE $$\sum_{i=1}^{N} \left\{ \mathbb{E}_{z \sim q_\phi(z|x_i)} \left[\log \left(p_\theta(x_i|z) p(z) \right) \right] + \mathcal{H}\left(q_\phi(z|x_i)\right) \right\}$$

NAIVE ALGORITHM (ONE ITERATION)

FOR EACH x_i (OR MINI-BATCH)

① CALCULATE

$$\nabla_\theta \mathcal{L}_i(\theta, \phi)$$

② CALCULATE

$$\nabla_\phi \mathcal{L}_i(\theta, \phi)$$

③ TAKE STEP $\quad \theta \longmapsto \theta + \alpha \nabla_\theta \mathcal{L}_i$

$$\phi \longmapsto \phi + \alpha \nabla_\phi \mathcal{L}_i$$

HOW TO CALCULATE ① AND ②?

IDEA #1: JUST DO IT. FOLLOW THE APPROACH:

$$\nabla_\theta \mathbb{E}_{q_\phi}[r(\theta, \phi)] = \mathbb{E}_{q_\phi}[\nabla_\theta r(\theta, \phi)]$$

$$\nabla_\phi \mathbb{E}_{q_\phi}[r(\theta, \phi)] = \nabla_\phi \int r(\theta, \phi)\, q_\phi \, dz$$

$$= \int (\nabla_\phi r(\theta, \phi))\, q_\phi \, dz$$

$$+ \int r(\theta, \phi) \nabla_\phi q_\phi \quad q_\phi \, dz$$

q_ϕ

$$= \mathbb{E}_{q_\phi}\left[\nabla_\phi r(\theta,\phi) + r(\theta,\phi)\nabla_\phi \log q_\phi\right]$$

$$r(\phi) = -\log q_\phi$$

FOR US, THIS IS:

$$\begin{cases} \nabla_\theta \mathcal{L}_i \overset{\circledast}{=} \mathbb{E}_{q_\phi}\left[\nabla_\theta \log p_\theta(x_i|z)\right] \\ \nabla_\phi \mathcal{L}_i \overset{\circledast}{=} \mathbb{E}_{q_\phi}\left[\log p_\theta(x_i|z)\nabla_\phi \log q_\phi \quad (7)\right] \\ \qquad + \mathbb{E}_{q_\phi}\left[-\frac{\nabla_\phi q_\phi}{q_\phi} - \log q_\phi \nabla_\phi \log q_\phi\right] \end{cases}$$

IDEA #2: WAIT... SO EVERYTHING IS AN EXPECTATION OVER AN EASY DIST. q_{ϕ,x_i}!

SO: INSTEAD OF CALCULATING, APPROXIMATE BY SAMPLING!

$$z_1 \text{ — } z_k \sim q_{\phi,x_i}$$

$$\hookrightarrow \overset{\circledast}{\approx} \frac{1}{k}\sum_{j=1}^{k} \nabla_\theta \log p_\theta(x_i|z_j)$$

$$\hookrightarrow \overset{\circledast}{\approx} \text{SIMILAR}$$

IDEA #3: WE'RE DOING THIS OVER AND OVER...

MAYBE WE CAN SAMPLE JUST ONE z $(k-1)$?
(FOR EACH OF $\nabla_\theta \mathcal{L}_i, \nabla_\phi \mathcal{L}_i$)

IDEA #4: ACTUALLY... JUST SAMPLE ONE $\varepsilon \sim N(0, \mathbb{1})$:

THAT'S BECAUSE

$$z \sim q_{\phi, x_i}(z) = N\left(\mu_\phi(x_i), \sigma_\phi(x_i)\right)$$

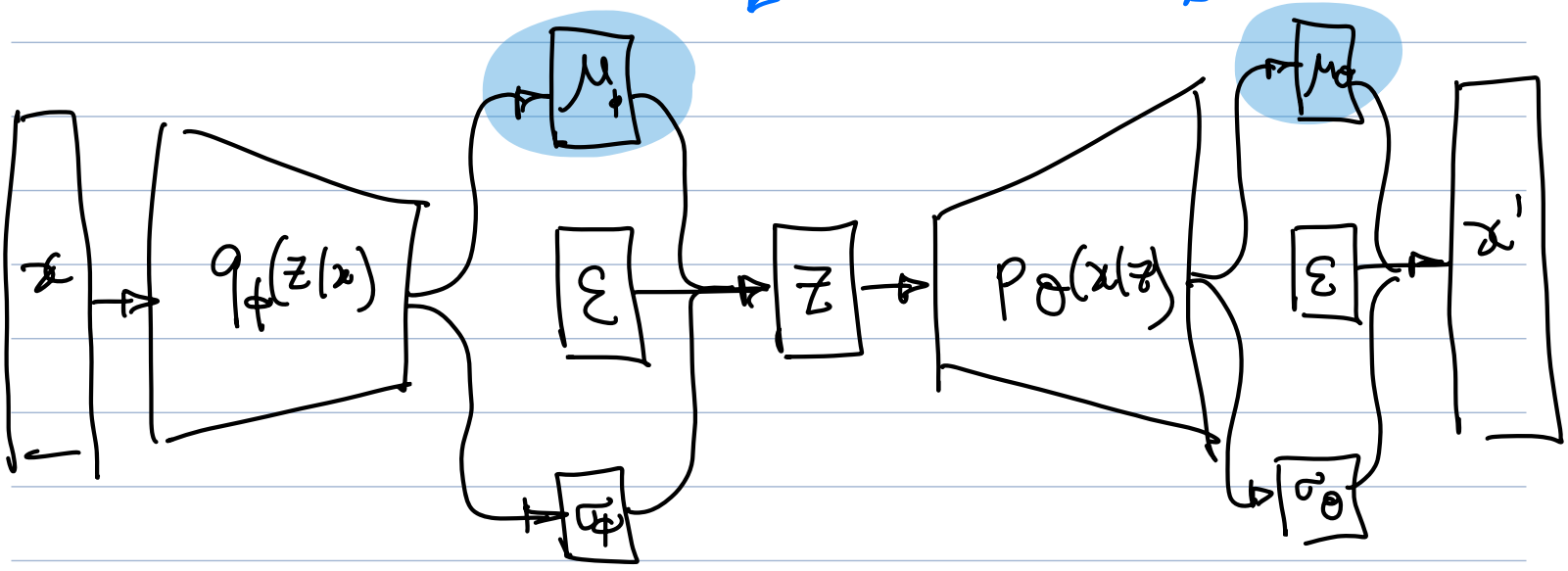
IE.

$$z = \underbrace{\mu_\phi(x_i)}_{\text{DETERMINISTIC (BUILT BY NN'S)}} + \underbrace{\sigma_\phi(x_i)}_{} \underbrace{\varepsilon}_{\sim N(0, \mathbb{1})}$$

... I'LL SOAKE YOU THE FINAL FORMULA.

FINAL SCHEMATIC:

SHADOWS OF A VANILLA AUTOENCODER



FINAL TRAINING ALGO:

FOR EACH x_i : (IE ONE ITERATION)

(1) SAMPLE ONE $\varepsilon_1 \sim N(0,1)$ TO APPROXIMATE $\nabla_\theta \mathcal{L}_i(\theta,\phi)$

(2) SAMPLE A SECOND $\varepsilon_2 \sim N(0,1)$ TO APPROXIMATE $\nabla_\phi \mathcal{L}_i(\theta,\phi)$

(3) STEP $\theta \longmapsto \theta + \alpha \widetilde{\nabla_\theta \mathcal{L}_i}$

$\phi \longmapsto \phi + \alpha \widetilde{\nabla_\phi \mathcal{L}_i}$

IV : GENERATION

TRAINING PRODUCES A DISTRIBUTION THAT APPROXIMATES $p_{\text{EMPIRICAL}}(x)$:

$$x_i \sim \approx \int_{\mathbb{R}^d} \underbrace{p_{\theta^*}(x|z)}_{\text{DECODER}} \underbrace{p(z)}_{\text{ELBO}} dz =: p_{\theta^*}(x)$$

HERE ARE FUN THINGS YOU CAN NOW DO:

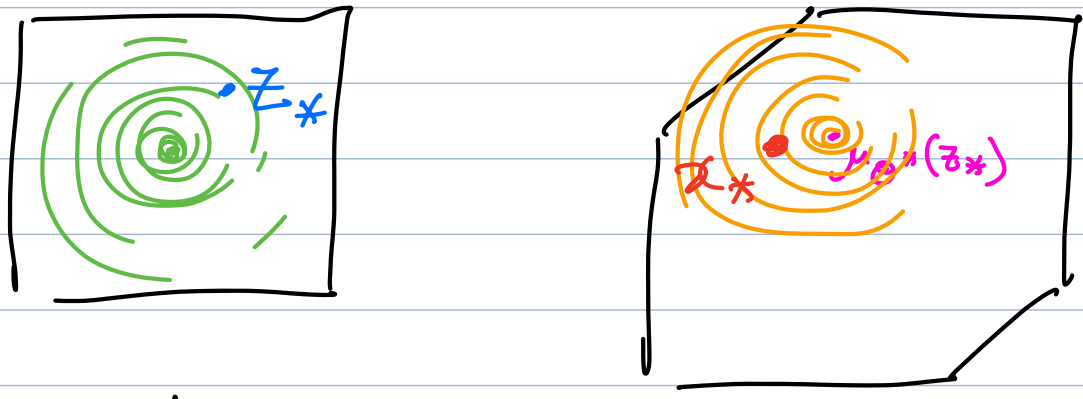
(1) GENERATE NEW DATA:

TO GENERATE "NEW DATA", YOU "SAMPLE" FROM $p_{\theta^*}(x)$ BY:

(1) SAMPLE $z_* \sim p(z) = N(0, \mathbb{1})$

(2) SAMPLE $x_* \sim p_{\theta^*}(x|z_*)$

$$= N(\mu_{\theta^*}(z_*), c_{\theta^*} \mathbb{1})$$



$z \in \mathbb{R}^d$ $\quad$ $x \in \mathbb{R}^n$

THIS x_* SHOULD LOOK LIKE $\{x_i\}_{i=1}^N$!

(2) INTERPOLATE USING LATENT SPACE

GIVEN TWO DATA PCNTS x_0, x_{k+1}

$(\rightarrow z_i := \mu_\phi(x_i))$

[OK JUST TWO LATENT SAMPLES $z_0, z_{k+1} \sim N(0,1)$,]

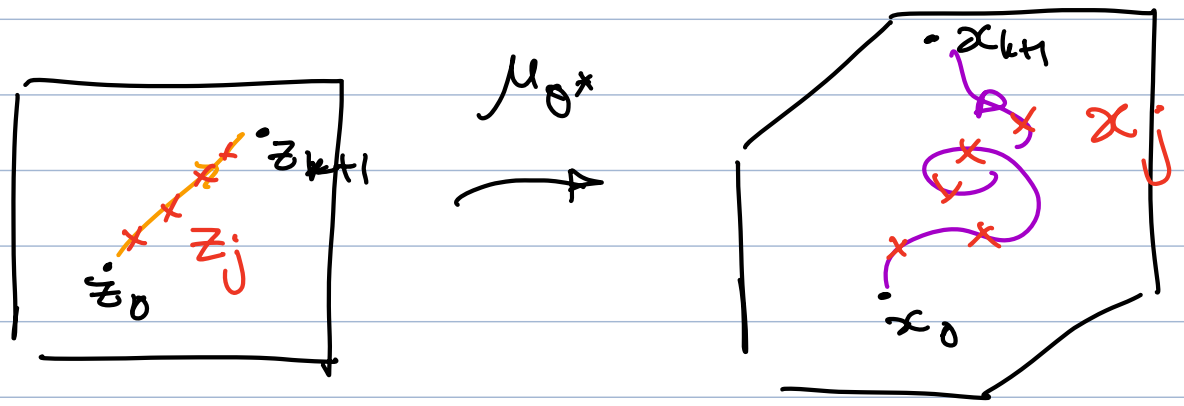
CAN (1) DRAW A STRAIGHT LINE $z_0 \rightsquigarrow z_{k+1}$ IN LATENT SPACE,

(2) PICK EVENLY-SPACED z_j ON THE LINE

(3) YOUR INTERPOLATION WILL BE

$$x_j := \mu_{\theta^*}(z_j)$$

$$x_0 \rightsquigarrow x_{k+1}$$



YOU CAN HOPE THAT THE NN'S LEARNED TO IDENTIFY CERTAIN FEATURES OF THE TRAINING DATA WITH DIRECTIONS IN LATENT SPACE!

... HENCE THE SMILE - FROWN

LONG - SHORT HAIR

CARTOON FROM EARLIER.

THANKS FOR LISTENING!