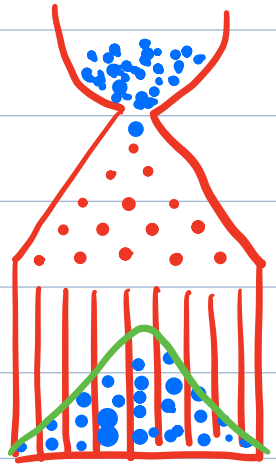
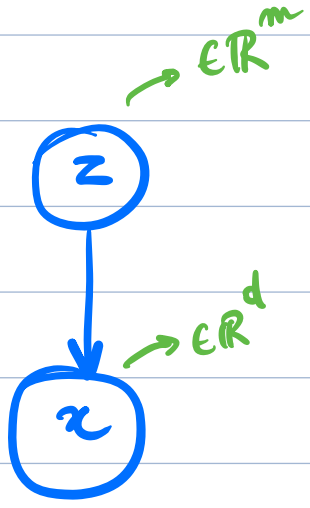
# Deep Latent variable models






$p(z)$

$p_\theta(x|z)$

How to estimate $\theta$?

$$\max_\theta p_\theta(x) = \int p_\theta(x,z)\,dz$$
$$= \int p(z)\, p_\theta(x|z)\,dz$$
$$= \mathbb{E}_{p(z)}\left[p_\theta(x|z)\right]$$
$$\approx \frac{1}{K}\sum_k p_\theta(x|z_k)$$

MC approximation

Poor approximation when $m$ is large due to the curse of dimensionality!

## I. Variational inference

$$p_\theta(x) = \mathbb{E}_{p(z)}\left[p_\theta(x|z)\right]$$

Intractable

$$= \mathbb{E}_{q_\phi(z)}\left[\frac{p(z)}{q_\phi(z)}\, p_\theta(x|z)\right]$$

Importance sampling

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(z)}\left[\frac{p(z)}{q_\phi(z)}\, p_\theta(x|z)\right]$$

Jensen's $\geq$

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

① or $\mathbb{E}_q\left[\log \frac{p(x,z)}{q(z)}\right]$

$$= \mathbb{E}_{q_\phi(z)}[\log p(x|z)] - KL(q_\phi(z)\|p(z)) = \text{ELBO}_{VI}$$

②

$$= \mathbb{E}_{q_\phi(z)}\left[\log \frac{p(z)}{q_\phi(z)}\, p_\theta(x|z)\, \frac{p_\theta(x)}{p_\theta(x)}\right]$$

Reconstruction error

$$= \mathbb{E}_{q_\phi(z)}\left[\log \frac{p_\theta(z|x)}{q_\phi(z)}\, p_\theta(x)\right]$$

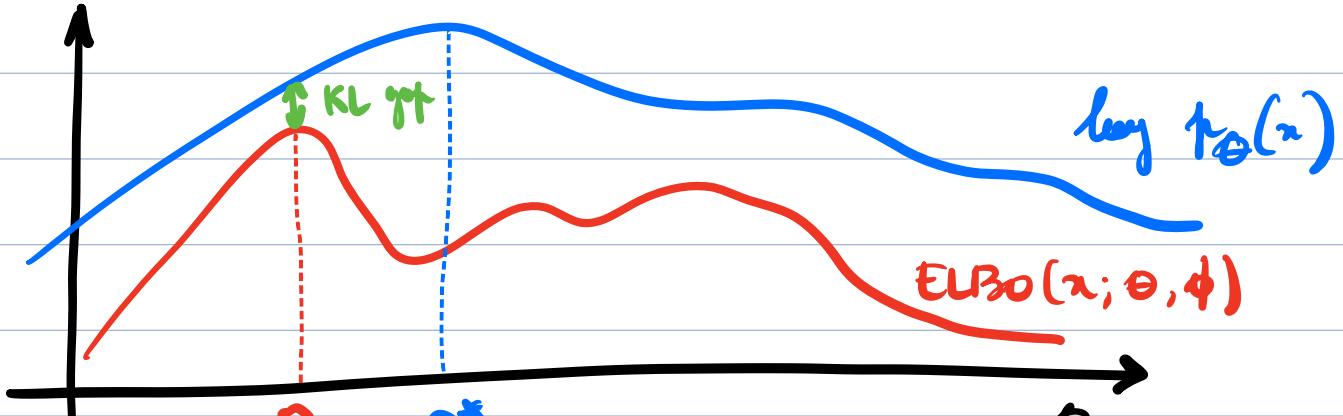
→ When $p(x|z)$ is Gaussian, this results in MSE

$$= \log p_\theta(x) - KL(q_\phi(z)\|p_\phi(z|x))$$

① + ②:

$$\text{ELBO}_{VI}(x;\theta,\phi) = \log p_\theta(x) - \underbrace{KL(q_\phi(z)\|p_\theta(z|x))}_{\geq 0}$$

$\Rightarrow$ Maximizing ELBO, is the same as max $\log p_\theta(x)$ and min $KL(q_\phi(z)\|p_\theta(z|x))$.



We want $KL(q_\phi(z) \| p_\theta(z|x)) \to 0$, otherwise the gap is large and $|\theta^* - \hat{\theta}| \gg 0$.

$\Rightarrow$ Put enough capacity in $q_\phi$.

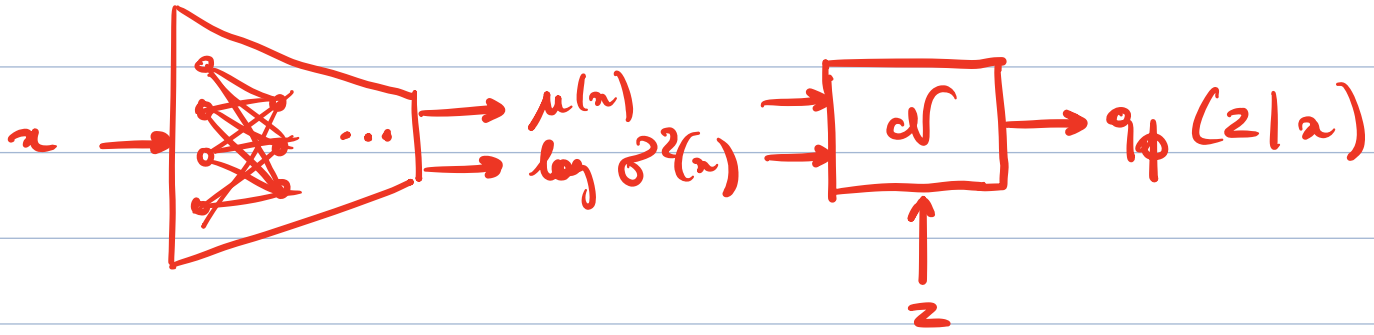
## II. VAEs

Amortize inference for any $x$.

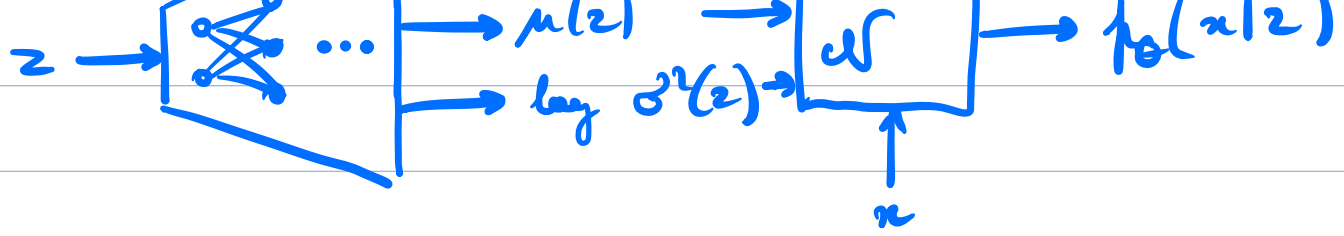
$$q_\phi(z) \longrightarrow q_\phi(z|x)$$

Now a DNN that outputs the parameters of the variational dist.

Encoder




Decoder




Training

$$\max_{\theta,\phi} \mathbb{E}_{p(x)}\left[ELBO(x;\theta,\phi)\right]$$

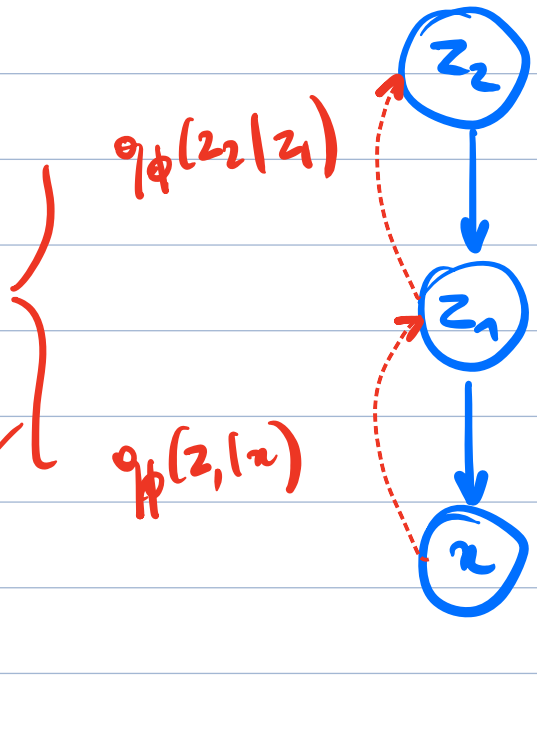
$$= \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x}\left[\log p_\theta(x|z)\right] - KL(q_\phi(z|x)\|p(z))\right]$$

⚠ 1 Issue: posterior collapse when $KL = 0$.
→ x does not bring any info about z
→ z is pure noise

⚠ 2 Tension between $KL(q_\phi(z|x)\|p(z))$ and $KL(q_\phi(z|x)\|p(z|x))$

→ Slides + Code.

## III. Hierarchical VAEs




$$q_\phi(z_1, z_2|x) = q_\phi(z_1|x)\, q_\phi(z_2|z_1)$$

## Training

$$\max_{\phi,\theta} \mathbb{E}_{p(x)}[\mathrm{ELBO}(x)]$$

$p(x|z_1)\,p(z_1|z_2)\,p(z_2)$

$$= \mathbb{E}_{p(x)}\, \mathbb{E}_{q_\phi(z_1,z_2|x)}\left[\log \frac{p(x, z_1, z_2)}{q(z_1, z_2|x)}\right]$$

do it!

$q(z_1|x)\, q(z_2|z_1)$

$$= \mathbb{E}_{p(x)}\, \mathbb{E}_{q_\phi(z_1,z_2|x)}\Big[\log p_\theta(x|z_1) - \mathrm{KL}(q_\phi(z_1|x)\,\|\,p(z_1|z_2)) - \mathrm{KL}(q_\phi(z_2|z_1)\,\|\,p(z_2))\Big]$$

Same as above but with this new term

$\to 0$ when $q_\phi(z_2|z_1)$ has too much capacity

$\Rightarrow q_\phi(z_2|z_1) = p(z_2) = \mathcal{N}$

=> the second latent is not used!

=> Same as regular VAEs!

no info about $x$ in $z_2$

## Top-down VAEs

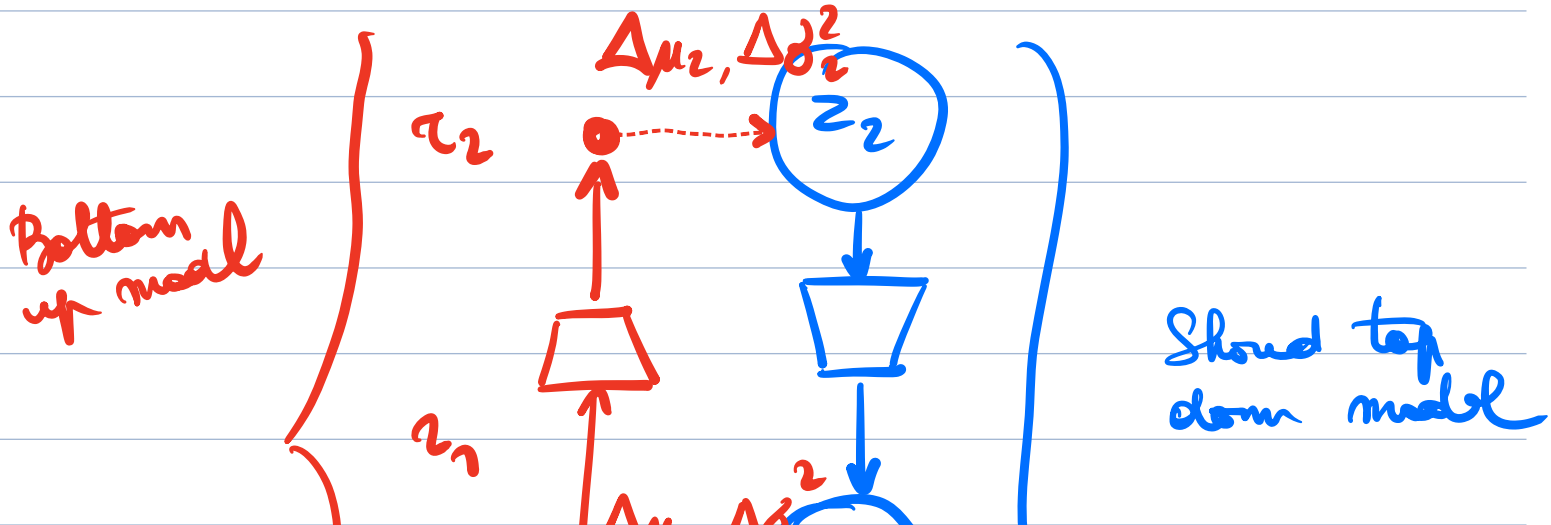
! Swap the directional dependencies between the latents

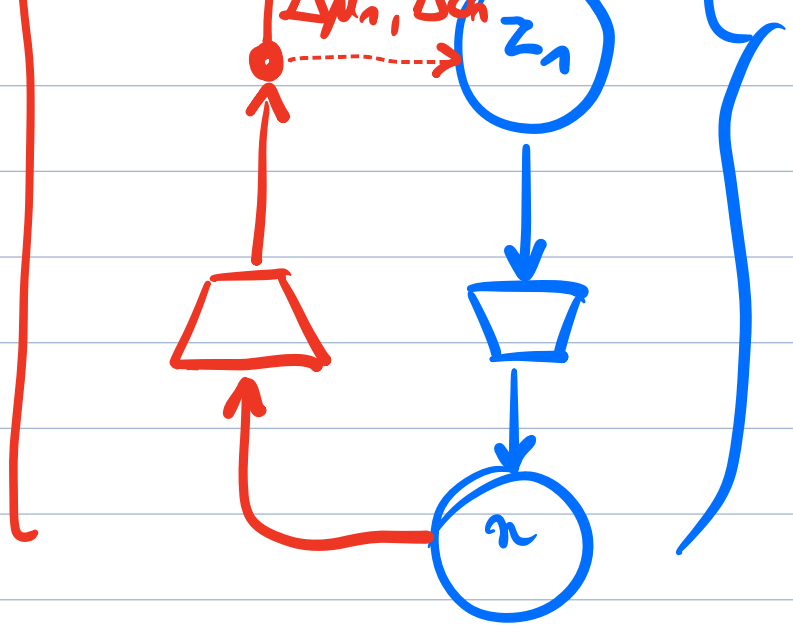
$$q_\phi(z_1, z_2 | x) = q_\phi(z_2 | x)\, q_\phi(z_1 | z_2, x)$$

! Same directed dependencies as in $p_\theta(z_1 | z_2)$

INDUCTIVE BIAS

=> Share a common top-down path.





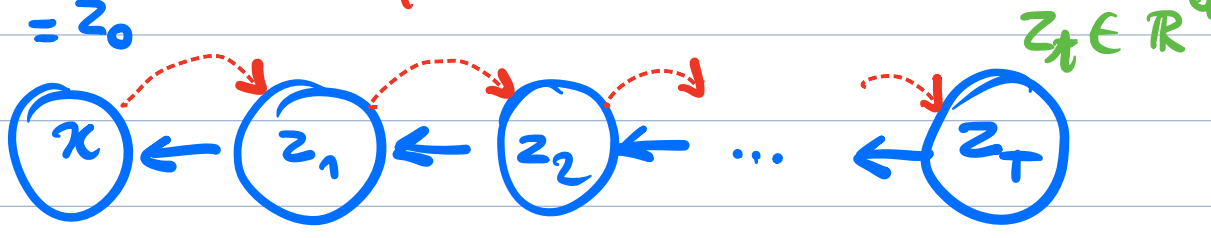

Forces a close connection between $q_\phi$ and $p_\theta$ and helps encode information about $x$ in $z_2$, $z_1$.

→ Slides.

## IV. Deep diffusion probabilistic models

Diffusion →

④ W encoder
↳ $q_\phi(z_t | z_{t-1})$




$\mathbb{R}^d$

$x \in \mathbb{R}$ $\quad p_\theta(z_t \mid z_{t+1})$ ↘ DNN decoder $\quad p(z_T) := \mathcal{N}$

$T = O(1000)$

Big first-order Markov chain

Generation = Reverse diffusion

→ Show diffusion.

$$p_\theta(z_{0:T}) = \left[\prod_{t=0}^{T-1} p_\theta(z_t \mid z_{t+1})\right] p(z_T)$$

$$q_\phi(z_{1:T} \mid z_0) = \prod_{t=1}^{T} q_\phi(z_t \mid z_{t-1})$$

$$q_\phi(z_t \mid z_{t-1}) = \mathcal{N}(z_t \mid \sqrt{1-\beta_t}\, z_{t-1}; \beta_t I)$$

$$\Leftrightarrow \quad z_t := \sqrt{1-\beta_t}\, z_{t-1} + \beta_t \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$$

Training

$$\max_{\phi,\theta} \mathbb{E}_{p(z_0)}[\mathrm{ELBO}(z_0)]$$

$$= \mathbb{E}_{p(z_0)} \mathbb{E}_{q_\phi(z_{1:T} \mid z_0)}\left[\log \frac{p_\theta(z_{0:T})}{q_\phi(z_{1:T} \mid z_0)}\right]$$

$$= \mathbb{E}_{p(z_0)} \mathbb{E}_{q_\phi(z_{1:T}|z_0)} \Big[ \log_{T=1} p_\theta(z_0|z_1) - \sum_{t=1} KL(q_\phi(z_t|z_{t-1}) \| p(z_t|z_{t+1})) - KL(q_\phi(z_T|z_{T-1}) \| p(z_T)) \Big]$$

Ho et al, 2020 : [SKIP] if time is short.

① Since $q$ is linear Gaussian, we have

$$q(z_t|z_0) = \mathcal{N}(z_t | \sqrt{\bar{\alpha}_t} z_0, \ 1 - \bar{\alpha}_t I)$$

where $\alpha_t = 1 - \beta_t$

$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t$

$\Rightarrow$ $z_t$ can be sampled without all intermediate steps !

② $$q(z_{t-1}|z_t, z_0) = \frac{q(z_t|z_{t-1}, z_0)\, q(z_{t-1}|z_0)}{q(z_t|x)}$$

($\mathcal{N}$ ; $\mathcal{N}$ ; $\mathcal{N}$)

$\Downarrow$ $\mathcal{N}$ (conjugate prior)

$$= \mathcal{N}(z_{t-1} | \tilde{\mu}_t(z_t, z_0), \tilde{\beta}_t I)$$

$$\tilde{\beta}_t \to \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$$

$$\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} z_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} z_t$$

Then, the ELBO can be expressed as

$$\mathbb{E}_{p(z_0)} \mathbb{E}_{q(z_{1:T}|z_0)} \Big[ \underbrace{\log p_\theta(z_0|z_1)}_{\mathcal{L}_0} - \sum_{t=1}^{T} \underbrace{KL(q_\phi(z_{t-1}|z_t, z_0) \| p_\theta(z_{t-1}|z_t))}_{\mathcal{L}_t} - \underbrace{KL(q_\phi(z_T|z_0) \| p(z_T))}_{\mathcal{L}_T} \Big]$$

Training

$$\mathbb{E}_{t \sim [1..T]} \mathbb{E}_{p(z_0)} \mathbb{E}_{q(z_t|z_0)} [\mathcal{L}_t]$$

$KL(\mathcal{N} \| \mathcal{N})$ closed form!

→ Update the layers one at a time!
→ much more memory efficient
→ Scale to $T = O(1000)$

→ Slides

$$\mathcal{L}_{t-1} = \mathbb{E}_q \left[ \frac{1}{2\sigma_t^2} \| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t) \|^2 \right]$$

## II. VAE priors

$$\text{ELBO} = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} \left[ \log p_\theta(x|z) + \log \frac{p(z)}{q_\phi(z|x)} \right]$$

Reconstruction $\quad$ $\Omega$

$$\Omega = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} \left[ \log p(z) - \log q_\phi(z|x) \right]$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_\phi(z|x_n)} \left[ \log p(z) - \log q_\phi(z|x_n) \right]$$

$$= \mathbb{E}_{q_\phi(z)} \left[ \log p(z) \right] - \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_\phi(z|x_n)} \left[ - \log q_\phi(z|x_n) \right]$$

$$= CE(q_\phi(z) \,||\, p(z)) + H_{p(z)}[q_\phi(z|x)]$$

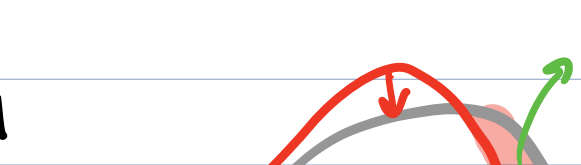
cross-entropy $\quad$ entropy

$$q_\phi(z) = \frac{1}{N} \sum_{n=1}^{N} q_\phi(z|x_n)$$

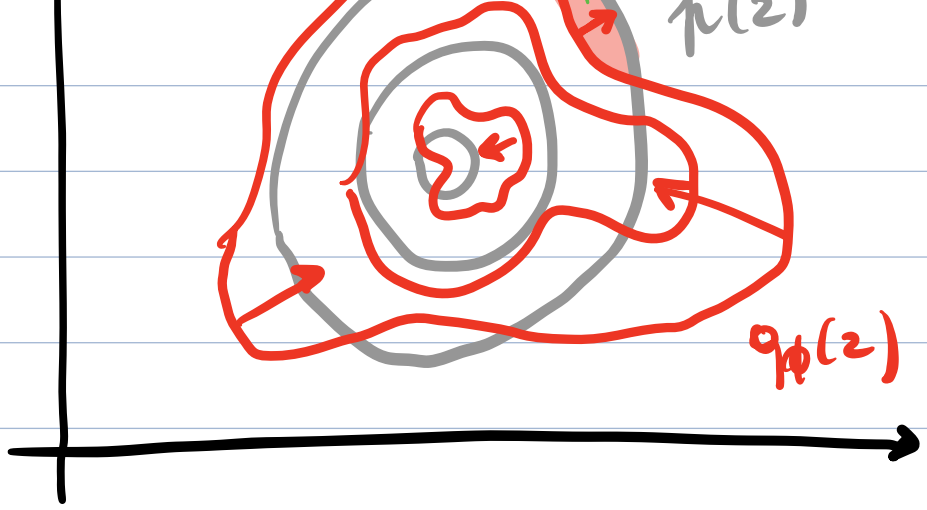
aggregated posterior

Make the aggregated posterior match with the prior.

Make the posterior variance $\to \infty$, but counter-balanced with RE.




Difficult when




the prior is fixed since the decoder does from the encoder to be fixed.

=> Learn the prior with another generative model!

$$p(z) \longrightarrow p_\lambda(z)$$

- MoG
- NFs
- DDPMs
- ...

→ Slides.