Variational Autoencoder (VAE) [2014]

idea: replace deterministic functions $f(x)$ and $g(z)$ with conditional distributions: encoder $p_E(z|x)$, decoder $p_D(x|z)$

$\Rightarrow$ more flexibility, and new training algorithms

- we want to minimize $KL[p^*(x) \| p(x)] \iff \mathbb{E}_{x\sim p^*(x)}[-\log p(x)]$ NLL loss

$$\log p(x) = \log p(x) \cdot \underbrace{\int p_E(z|x)\,dz}_{=1 \text{ (normal.)}} = \int p_E(z|x) \log p(x)\,dz$$

$$= \int p_E(z|x) \log \left[\frac{p(x)\, p_E(z|x)}{p_E(z|x)} \cdot \frac{q(z)}{q(z)} \cdot \frac{p_D(x|z)}{p_D(x|z)} \right] dz \qquad q(z) \text{ desired code distrib.}$$

$$= \int p_E(z|x) \left[\log \frac{p_E(z|x)}{q(z)\, p_D(x|z)/p(x)} - \log \frac{p_E(z|x)}{q(z)} + \log p_D(x|z) \right] dz$$

$$p(x) = \int p_D(x,z)\,dz = \int q(z)\, p_D(x|z)\,dz \qquad p_D(z|x) = \frac{q(z)\, p_D(x|z)}{p(x)}$$

decoder posterior

$$\log p(x) = \underbrace{KL[p_E(z|x) \,\|\, p_D(z|x)]}_{\geq 0} - \underbrace{KL[p_E(z|x) \,\|\, q(z)] + \mathbb{E}_{p_E(z|x)}[\log p_D(x|z)]}_{ELBO(x)}$$

- Observation 1: $\log p(x) \geq ELBO(x) \Rightarrow$ "evidence lower bound"
  $p(x)$

  - maximizing lower bounds also tends to push-up $\log p(x) \Rightarrow$ maximize likelihood

- $KL[p_E(z|x) \,\|\, p_D(z|x)]$ "variational gap" if ≥ 0, measures the amount of inconsistency between encoder and decoder

consistency: joint distributions should be equal

$$p_E(x,z) = p^*(x) \cdot p_E(z|x) \overset{!}{=} p_D(x,z) = q(z) \cdot p_D(x|z)$$

$$p_E(z|x) = \underbrace{\frac{q(z)\, p_D(x|z)}{p^*(x)}}_{p_D^*(z|x)}$$

- observation 2: $-\log p(x) + KL[p_E(z|x) \| p_D(z|x)] = -ELBO(x)$

$\Rightarrow$ minimizing $-ELBO(x)$ has two effects (i.e. minimizing RHS and LHS)

- minimize $-\log p(x) \Rightarrow$ high quality of generative distribution! (minimize NLL)

minimize $KL[p_E \| p_D] \Rightarrow$ improve self-consistency of model

- observation 3: terms in ELBO optimize trade-off between reconstruction quality ($\hat{x} \approx x$) and distribution fidelity ($p(x) \approx p^*(x)$)

Variational Autoencoders (contd.)

- encoder and decoder are not deterministic functions, but conditional distributions

enc. $p_E(z|x)$ dec $p_D(x|z)$ + data $p^*(x)$, desired code / distrib. $q(z)$

$\Rightarrow$ implies two versions of joint distribution of x and z

enc. $p_E(x,z) = p^*(x) \cdot p_E(z|x)$ dec. $p_D(x,z) = q(z)\, p_D(x|z)$

two requirements:

① decoder marginal $p(x) = \int p_D(x,z)\,dz \approx p^*(x)$ should be a good approx

② enc/dec pair must be self-consistent $p_E(x,z) \stackrel{!}{=} p_D(x,z)$

- maximizing ELBO loss enforces ②:

$$\int KL\left[p_E(x,z) \,\|\, p_D(x,z)\right] = \mathbb{E}_{x\sim p^*(x)}\left[\mathbb{E}_{z\sim p_E(z|x)}\left[\log \frac{p_E(z|x)\, p^*(x)}{p_D(x|z)\cdot q(z)}\right]\right]$$

$=$

$$= \mathbb{E}_{x\sim p^*(x)}\Big[\underbrace{\mathbb{E}_{z\sim p_E(z|x)}\Big[\log\frac{p_E(z|x)}{q(z)} - \log p_D(x|z)\Big]}_{-\mathrm{ELBO}(x)} + \underbrace{\log p^*(x)}_{\text{independent of model}}\Big]$$

$\Rightarrow$ drop

$$\hat{E},\hat{D} = \arg\min_{E,D} \mathbb{E}_{x\sim p^*(x)}\big[-\mathrm{ELBO}(x)\big] \Rightarrow \text{ensures self-consistency upon perfect convergence}$$

(in literature, ELBO is usually maximized, conceptually, minimizing $-$ELBO is simpler)

- maximizing the ELBO indirectly maximizes the data likelihood under the model (?)

maximum likelihood principle. TS should by a typical (= high prob.) model outcome

$\Leftrightarrow$ minimize expected negative log-likelihood (NLL)

$$\mathbb{E}_{x\sim p^*(x)}\big[-\log p(x)\big] = \mathbb{E}_{x\sim p^*(x)}\Big[-\log\int p_D(x|z)\,q(z)\,dz\Big]$$

$$= \mathbb{E}_{x\sim p^*(x)}\Big[-\log \underbrace{\int p_E(z|x)\,\frac{p_D(x|z)\,q(z)}{p_E(z|x)}\,dz}_{= \mathbb{E}_{z\sim p_E(z|x)}\left[\frac{p_D(x|z)\,q(z)}{p_E(z|x)}\right]}\Big]$$

[ ]ensen's inequality. if φ is a convex function $\varphi(\mathbb{E}[t]) \leq \mathbb{E}[\varphi(t)]$)

$$\leq \mathbb{E}_{x\sim p^*(x)}\left[\underbrace{\mathbb{E}_{z\sim p_E(z|x)}\left[-\log \frac{p_D(x|z)\,q(z)}{p_E(z|x)}\right]}_{-ELBO(x)}\right]$$

$\Rightarrow$ $\underbrace{\mathbb{E}_{x\sim p^*(x)}\left[-ELBO(x)\right]}_{\text{minimization here tends to minimize this as well}}$ is upper bound for NLL loss

---

Interpretation of ELBO(x) terms: there is a trade-off between two objectives

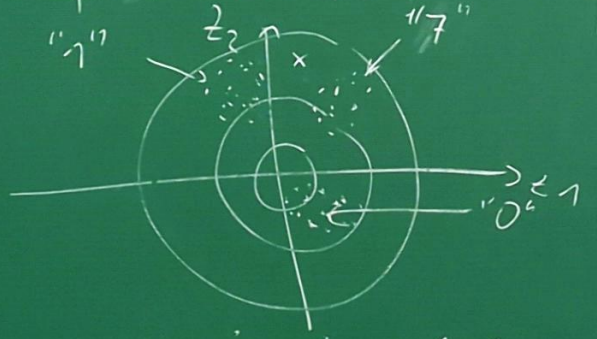
$$-ELBO(x) = \underbrace{KL\left[p_E(z|x)\,\|\,q(z)\right]}_{\text{how close is conditional code distribution } p_E(z|x) \text{ to desired } q(z)} + \underbrace{\mathbb{E}_{z\sim p_E(z|x)}\left[-\log p_D(x|z)\right]}_{\text{reconstruction error: original } x \text{ should be a very probable reconstruction}}$$

minimized if $p_E(z|x) = q(z)$
ie encoder is independent of x ("ignores x")
$\Rightarrow$ output $p(x)$ is unrelated to input, no reconstr.

minimized if encoder & decoder are deterministic with perfect reconstruction:
$p_E(z|x) = \delta(z - f(x))$, $p_D(x|z) = \delta(x - g(z))$, $x = g(f(x))$

typical compromise in practice MNIST (10 classes)

$q(z) = \mathcal{N}(0, \mathbb{I})$



$p_E(z) = \int p^*(x)\, p_E(z|x)\, dx$ has visible clusters

clusters for each digit with visible gaps

$\Rightarrow$ sampling $z \sim q(z)$ can land in a gap

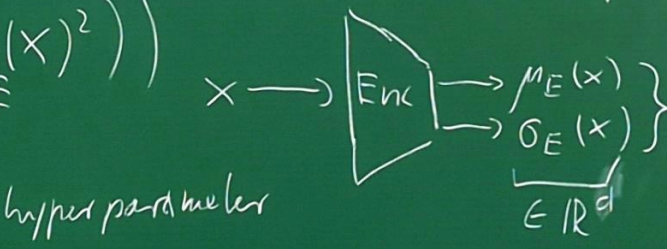
$\Rightarrow$ bad synthetic point $p_D(x|z)$

most common implementation: encoder and decoder are diagonal Gaussians

$$p_E(z|x) = \mathcal{N}(z \mid \mu_E(x), \mathrm{Diag}(\sigma_E(x)^2))$$

$$p_D(x|z) = \mathcal{N}(x \mid \mu_D(z), \beta^2 \mathbb{I})$$

β: hyperparameter



if $q(z) = \mathcal{N}(0, \mathbb{I})$, $KL[p_E \| q]$ can be calculated analytically

$$KL\left[\mathcal{N}(z \mid \mu_E(x), \mathrm{Diag}(\sigma(x)^2)) \,\|\, \mathcal{N}(0, \mathbb{I})\right] = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_j(x)^2 + \sigma_j(x)^2 - \log \sigma_j(x)^2 - 1\right)$$

if $p_D(x|z) = \mathcal{N}(x \mid \mu_D(z), \beta^2 \mathbb{I})$, reconstruction term becomes

$$\mathbb{E}_{z\sim p_E(z|x)}\left[-\log p_D(x|z)\right] = \mathbb{E}_{z\sim p_E(z|x)}\left[\frac{(x-\mu_D(z))^2}{2\beta^2} + \underbrace{\log\left(2\pi\beta^2\right)^{D/2}}_{\text{independent of model} \Rightarrow \text{drop}}\right]$$

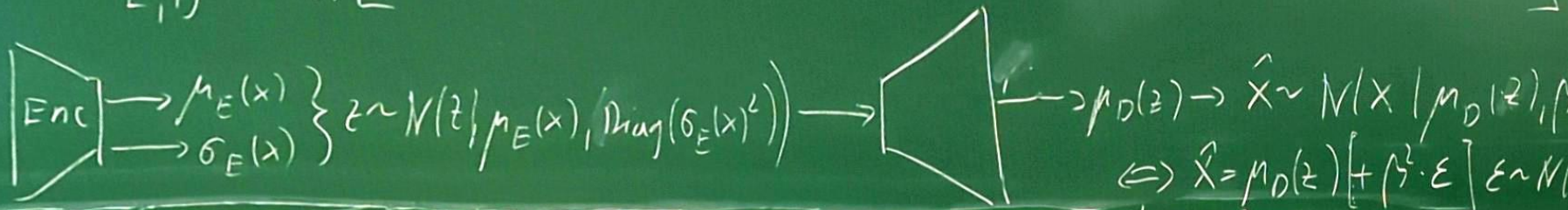
$$\{z_k \sim p_E(z|x)\}_{k=1}^{M} \qquad \approx \frac{1}{M}\sum_{k=1}^{M}\frac{(x-\mu_D(z_k))^2}{2\beta^2}$$

$\beta^2 \gg 1$ downscale squared loss $\Rightarrow$ reconstruction error unimportant

$\beta^2 \ll 1$ upscale — " — $\Rightarrow$ — " — dominates

full loss:

$$\hat{D}, \hat{E} = \arg\min_{E,D} \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{d}\frac{1}{2}\left(\mu_{E,j}(x_i)^2 + \sigma_{E,j}(x_i)^2 - \log\sigma_{E,j}(x_i)^2 - 1\right) + \frac{1}{M}\sum_{k=1}^{M}\frac{(x_i-\mu_D(z_{i,k}))^2}{2\beta^2}\right]$$



Paper: [Kingma & Welling 2013] > 30000 citations

- generation $z \sim q(z)$, $x \sim p_D(x|z) \Leftrightarrow \mu_D(z) + \beta^2 \varepsilon$
- inference if $p(x) = p^*(x)$ and $p_E(x,z) = p_D(x,z)$

$$\Rightarrow p_E(x,z) = p(x)\, p_E(z|x) = q(z)\, p_D(x|z) = p_D(x,z)$$

$$p(x) = \frac{q(z)\, p_D(x|z)}{p_E(z|x)}$$ must give same value for all $z \sim p_E(z|x)$

[measure of self-consistency: variance of $\log q(z) + \log p_D(x|z) - \log p_E(z|x)$]

conditional VAE

- plain VAE learns $p(x)$, conditional VAE learns $p(x|y)$ for some variable y
  ex $y \in \{0, \dots, 9\}$ MNIST label

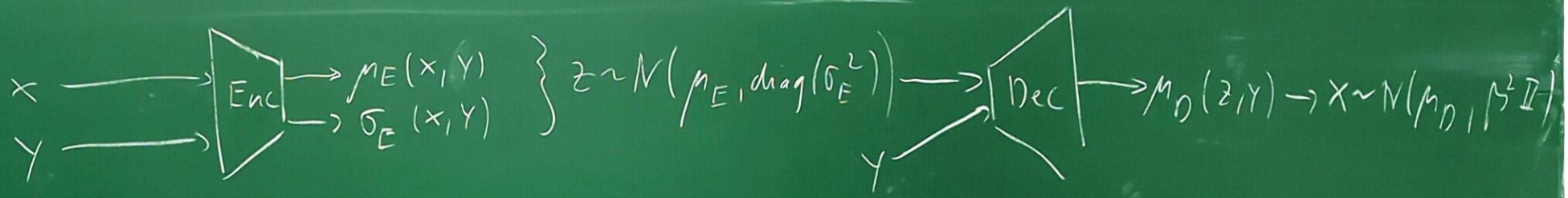
$$p(x|y) = \int q(z)\, p_D(x|y,z)\, dz \;, \Rightarrow p(x) = \sum_y p(x|y)\, p^*(y)$$

hopefully better quality

- model change add Y is an extra condition to enc. & dec.

$$p_E(z \mid x, y) \qquad p_D(x \mid y, z)$$

if enc. & dec. are Gaussian, add Y as input to the networks



application: style transfer $Y = \{$painting, toto, cartoon, drawing$\}$

encode x with true Y (e.g. painting of Newton)

decode with different Y (e.g. toto of Newton)

- generative classification: test x with unknown label

$\Rightarrow$ try encoding with every label and calculate the corresponding prob $p(x \mid y)$

return the label maximizing the probs.

$$\hat{y} = \arg\max_y \left(\int p_E(z \mid x, y)\, p_D(x \mid z, y)\, dz \right)$$