Another look at the ELBO:

$$KL(P||Q) = \int dx\, P(x) \log \frac{P(x)}{Q(x)}$$

"Kullback Leibler divergence" $\longrightarrow$ aka "relative entropy"

not a metric distance, asymmetric $P \leftrightarrow Q$

non negative $KL \geq 0$ but not bounded

0 iff $P = Q$

- $JSD(P,Q) = \frac{KL(P||M) + KL(Q||M)}{2}$ $\quad M = \frac{P+Q}{2}$.

Introduce $p(z|x) = \frac{p_\theta(x|z)\,p(z)}{p(x)}$ true posterior

$$ELBO = \frac{1}{N} \sum_{z \sim q_\phi(z|x)} \log \frac{p(z|x)\,p(x)}{q_\phi(z|x)}$$

$$= \boxed{\log p(x) - KL\left(q_\phi(z|x) \,||\, p(z|x)\right)} \leq \log p(x)$$

So ELBO is saturated when q_ϕ is the true posterior ✓.

confirming our intuition $\longrightarrow$ "variational Bayesian inference"

can use to infer posteriors

$\downarrow$

So what is our generative model?

need $p_\theta(x|z)$ and $q_\phi(z|x)$ (w/ $z \sim p(z)$ fixed prior)

$p_\theta(x|z)$: "probabilistic decoder" $\quad$ $q_\phi(z|x)$: "probabilistic encoder"

$x \longrightarrow z \longrightarrow x$ "variational autoencoder"

Can choose many things for p_θ & q_ϕ (NNs)

e.g. $\begin{cases} q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma_\phi(x)) \\ p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \beta_\theta(z)) \end{cases}$ normal dist'n w/ mean/var DNNs ①

Very simple ansatz...

Another common choice: continuous Bernoulli dist'n

$p(x|z) = N(z) z^x (1-z)^{1-x}$

$x \in [0,1]$

$z \in [0,1]$

Cont. gen. of discrete Bern.

$x = 0$ or 1

$p(x|z) = z$ $x=1$; $1-z$ $x=0$

or BCE for Cont. Bern

$x \log z + (1-x)\log(1-z)$

often $N(z)$ ignored but not strictly correct...

Another look at ELBO loss

$$-\text{ELBO} = \underbrace{-\mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z)}_{} + \underbrace{KL(q_\phi(z|x) \,||\, p(z))}_{}$$

max → min

$$\mathbb{E}_{z\sim q_\phi(z|x)} \frac{||x-\mu_\theta(z)||^2}{2\beta_\theta(z)^2}$$

regulariz

"reconstruction error"

want $x \to z \to x$ to recover x

standard Autoencoder loss

"regularization"

promotes $x \to z \sim p(z)$

regular latent space

"Reparam trick": if q_ϕ & $p(z)$ are Gaussian, can show

$$KL(q_\phi \,|\, p(z)) = \frac{1}{2}\left(-1 + \mu_\phi(x)^2 + \sigma_\phi(x)^2 - \log \sigma_\phi(x)^2\right)$$

So Gaussian VAE objective becomes finally:

$$\min_{\theta,\phi} \sum_{\substack{x\sim \text{data} \\ z\sim q_\phi(z|x)}} \left[\frac{||x-\mu_\theta(z)||^2}{2\beta_\theta(z)^2} + \frac{1}{2}\left(-1 + \mu_\phi(x)^2 + \sigma_\phi(x)^2 - \log \sigma_\phi(x)^2\right)\right]$$

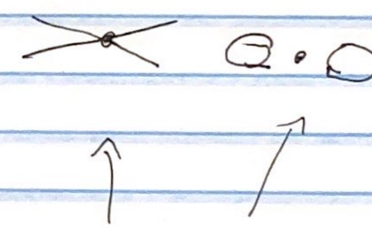
In practice, allowing $\beta_\theta(z) \rightarrow$ unstable traj.

$\rightarrow \beta_\theta = \beta$ const.

"β-VAE"

Like vanilla AE + regularizer term
& probabilistic/generative interpretation

- MNIST example
- radio astronomy example



2102.01007
Bastien, Scaife et al

Trained VAE to generate FR I and FR II radio galaxy images

↳ ~ only ~600 images (!!)
+ augmentation (eg random rotations)
$\rightarrow$ 200k images

Also CVAE
conditioned on label

- SDSS example 2002.10464
Galaxy spectra