1. Introduction to Bayesian Statistics

→ In Bayesian statistics probabilities represent „states of knowledge" or „degrees of belief"

→ Different than frequentist interpretation: Prob.'s express expected frequencies of events

→ Bayesian statistics tells us how to formally update our state of knowledge when new information (a measurement) becomes available

→ Bayesian statistics works with Conditional Probabilities

a) Working with conditional probabilities

X, Y: continuous random variables or discrete events

joint

Conditional: $$P(X|Y) = \frac{P(X,Y)}{P(Y)} \qquad (1)$$

Marginal: $$P(Y) = \int dx \; P(X,Y) \qquad (2)$$

OR $$\sum_{x \in \{x_i\}} P(x_i|Y)$$

b) Bayes Theorem

→ Tells us how to update our state of knowledge as new data becomes available

- Prior state of belief $p(x)$
  $\underline{\underline{x}}$ could be parameters in a physics model

- additional knowledge gained through measurement $\underline{\underline{d}}$

  Likelihood $P(d|x)$

- updated knowledge

  Posterior $p(x|d)$

Bayes Theorem

$$P(x|d) = \frac{P(d|x)\,P(x)}{P(d)}$$

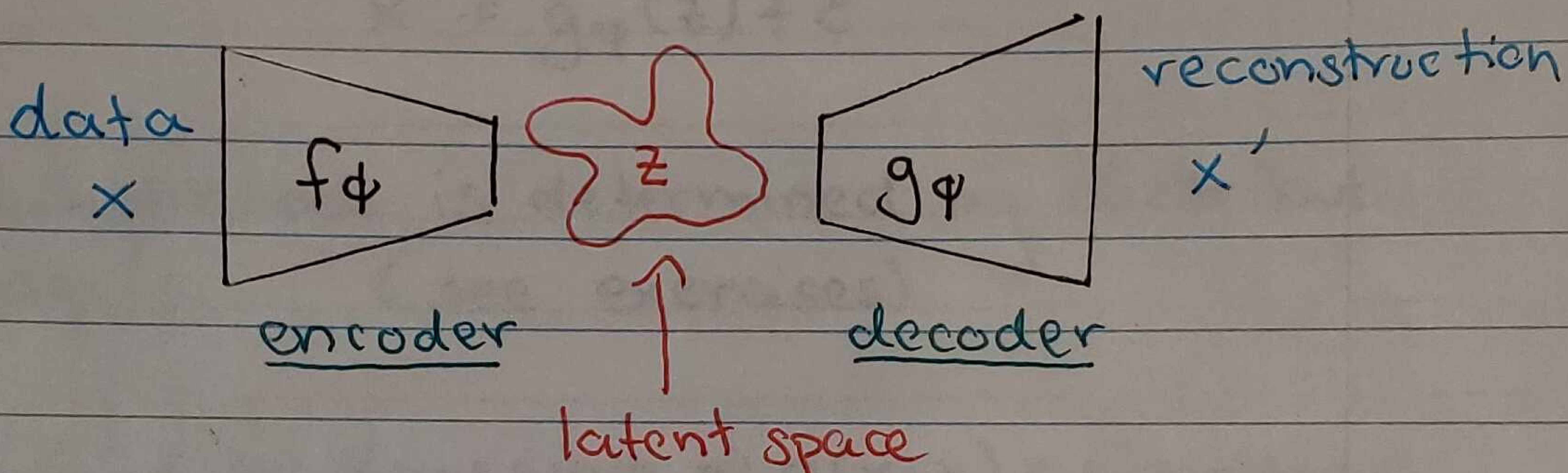
$P(d)$: normalization constant „evidence"

2. Why do I care?

OR

Introduction to Variational AEs

a) recap: Autoencoder



$$x' = g_\varphi(f_\phi(x)) \qquad \underbrace{\mathcal{L}_{AE} = \| x' - g_\varphi(f_\phi(x)) \|_2^2}_{\text{loss function}}$$

b) Probabilistic interpretation

-> Information loss in the compression

=> can't recover x precisely

=> implies probabilistic structure

-> data x follows probability distr. $p(x)$

-> all available data points are drawn from this distribution

$$p(x) = \int dz \, P(X, z) = \int dz \, P(X|Z) \, P(Z)$$

Likelihood $p(x|z)$

→ Information loss in compression implies probabilistic structure

$$x' = g_\psi(z) + \varepsilon$$

→ Likelihood is determined by distribution of ε (see exercises)

→ if $\varepsilon \leftarrow$ Gaussian ⇒ $p(x|z)$ = Gaussian

$$p_\psi(x|z) = \mathcal{G}(x \,|\, \mu = g_\psi(z), \Sigma = \varepsilon)$$

→ Likelihood depends on / is parameterized by network parameters

Prior $p(z)$

→ prior knowledge (before looking at data sample x) is average distribution of z

$$p(z) = \int dx \; p(z|x) \, p(x)$$

c) Derivation of Evidence Lower Bound

$$p(x) = \int dz \; p(x|z)\, p(z)$$

$$\log p(x) = \log \int dz \; p(x|z)\, p(z)$$

$$\log p(x) = \log \underbrace{\int dz \; q(z|x)\, p(x|z) \frac{p(z)}{q(z|x)}}_{\underset{q(z|x)}{E}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]}$$

$$\log p(x) \geq \int dz \; q(z|x) \left[\log p(x|z) - \log \left[\frac{q(z|x)}{p(z)} \right] \right]$$

likelihood term = measure of reconstruction quality

$$\geq \underline{\int dz \; q(z|x) \log p(x|z)}$$

Gaussian likelihood $\log p(x|z) \hat{=} \frac{(x - g(z))^2}{\sigma^2}$

$$- \underline{\int dz \; q(z|x) \log \left[\frac{q(z|x)}{p(z)} \right]}$$

measures similarity between posterior & prior

Kullbach-Leibler Divergence

$$D_{KL}(q \| p) \geq 0$$

→ ELBO

d) <u>In practice</u>

-> $q_\phi(z|x)$ is usually chosen to be a Gaussian with diagonal covariance *
Mean and covariance are parameterized by the encoder network

$$q_\phi(z|x) = \mathcal{G}(z|(\mu,\sigma)=f_\phi(x))$$

* note that this decorrelates latent variables

-> $p(z)$ can be chosen. Typically a standard Normal distribution

$$p(z) = \mathcal{G}(z|\mu=0, \sigma=1)$$

We want to be able to sample from $p(z)$

-> The ELBO is evaluated stochastically

$$ELBO = \underset{z\sim q_\phi(z|x)}{E} \log p_\psi(x|z)$$

$$+ \underset{z\sim q_\phi}{E} \log\left[\frac{q_\phi(z|x)}{p(z)}\right]$$

Training requires derivative $\frac{\partial ELBO}{\partial(\psi,\phi)}$

-> How do we take this derivative?

Reparameterization Trick

sample: $z' \sim G(z' | 0, 1)$

compute: $z = \sigma_\phi z' + \mu_\phi$

e) What is the VAE good for?

① The VAE is a generative model!
- → it allows us to generate artificial data
- → it allows us to approximately sample from $p(x)$

How? $z \sim p(z)$

$x = g_\psi(z)$

② Different to the AE, we obtain a continuous latent space.
This allows us to interpolate between data points.

How? $z_1 = \mu_\phi(x_1)$ $z_2 = \mu_\phi(x_2)$

$$z = z_1 + t \cdot (z_2 - z_1), \quad t \in [0, 1]$$
$$x = g_\psi(z)$$

③ VAE allows us to explore the space of likely reconstructions

$$z \hookleftarrow q_\phi(z|x) \qquad x' = g_\psi(z)$$

„posterior analysis"

④ Take into account physical data properties (mask, noise)

⇒ tomorrow's exercise!