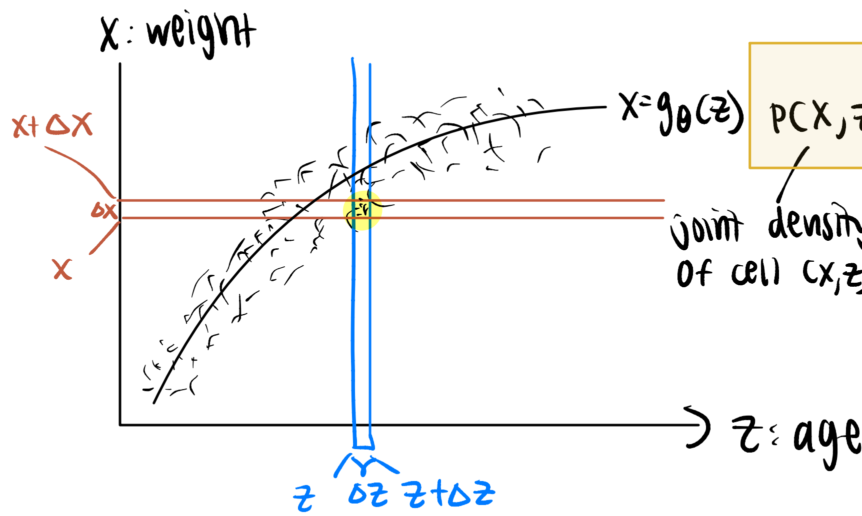
# LECTURE 18

## background on prob. density

$$\text{Density(LA)} = \frac{\text{\# of people in LA}}{\text{size of LA}}$$

$$P(x) = \frac{n(x)}{\Delta x}$$




\# of points in cell (x,z)

$$\boxed{P(x,z) = \frac{n(x,z)/n}{\Delta x\,\Delta z}}$$

→ total # of points

joint density of cell (x,z)

area of cell

Δx, Δz make range infinitesimally small so becomes ≈ continuous

$$P(z) = \frac{n(z)/n}{\Delta z}$$

→ # of people in slice (z)

length for 1D

total # (to normalize/get proportion)

$$\frac{n(z)}{n} = \sum_x \frac{n(x,z)}{n}$$

$$P(z)\cancel{\Delta z} = \sum_x P(x,z)\,\Delta x \cancel{\Delta z}$$

$$\boxed{\begin{aligned} P(z) &= \int P(x,z)\,dx \\ \text{similarly, } P(x) &= \int P(x,z)\,dz \end{aligned}}$$ marginal

$$P(x|z) = \frac{n(x,z)/n(z)}{\Delta x}$$

$$= \frac{n(x,z)/n/\Delta x \Delta z}{n(z)/n/\Delta z} = \frac{P(x,z)}{P(z)}$$

joint

marginal

$$P(x,z) = P(z)\,P(x|z)$$

**Logic:**

count → proportion → density

## Generator

prior
$z \sim p(z) \sim N(0, I_d)$ — $\dim(z)$ (100)

conditional
$p_\theta(x|z) \sim N(g_\theta(z), \sigma^2 I_D)$ — $\dim(x)$ (100 x 100)

$\downarrow$ top-down CNN

posterior: $p_\theta(z|x) \approx q_\phi(z|x)$

joint: $p_\theta(x, z) = p(z)\, p_\theta(x|z) \to$ tractable

marginal: $p_\theta(x) = \int p_\theta(x, z)\, dz \to$ intractable

conditional: $p_\theta(z|x) = \dfrac{p_\theta(x,z)}{p_\theta(x)}$

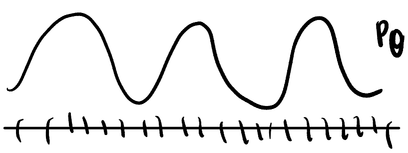
learn $\theta$ by max likelihood
$n$ = # of training examples

$$\max_{\theta,\phi} \frac{1}{n}\sum_{i=1}^{n}\Big[\underbrace{\log p_\theta(x_i)}_{\text{log-likelihood}} - \boxed{D_{KL}\big(q_\phi(z_i|x_i)\,\|\,p_\theta(z_i|x_i)\big)}\Big] = \text{Evidence lower bound (ELBO)}$$

variational lower bound (vary $\phi$)

tractable

$$\boxed{E_{q_\phi(z_i|x_i)}\left[\log \frac{q_\phi(z_i|x_i)}{p_\theta(z_i|x_i)}\right]}$$

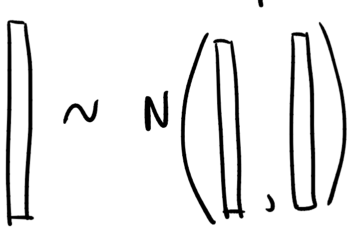


$p_\theta(x)$ — but we don't know $p_\theta(x)$ b/c intractable

$$\text{ELBO} = \frac{1}{n}\sum_{i=1}^{n} E_{q_\phi(z_i|x_i)}\Big[\underbrace{\log p_\theta(x_i)}_{\text{intractable}} - \underbrace{\log q_\phi(z_i|x_i)}_{\text{tractable}} + \underbrace{\log p_\theta(z_i|x_i)}_{\text{intractable}}\Big]$$

$\log p_\theta(x_i, z_i)$
tractable

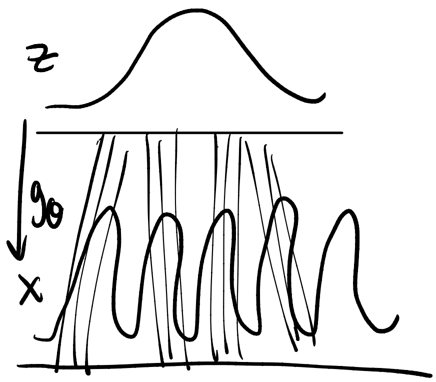
$q_\phi(z|x) \sim N(\mu_\phi(x), \sigma^2_\phi(x))$



elementwise

$$\mathrm{ELBO} = \frac{1}{n}\sum_{i=1}^{n} E_e\left[\frac{1}{2\sigma^2}\left| x_i - g_\theta(\overbrace{\mu_\phi(x_i) + \sigma^2_\phi(x_i)\odot e}^{z_i})\right|^2\right] + \cdots$$

$e \sim N(0, I_d)$
noise term




$g_\theta$ is a mapping

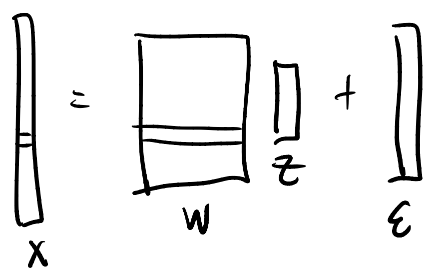
original auto-encoder

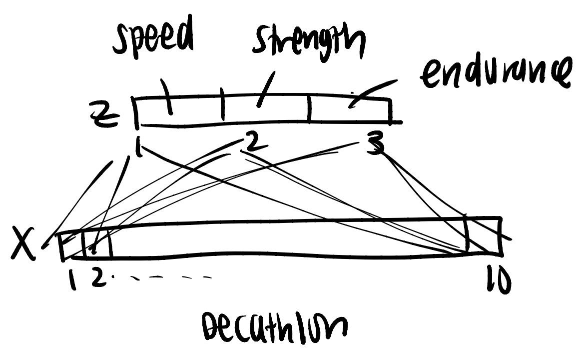
$$\begin{tikzcd} z \arrow[d, "g_\theta(z)"] \\ x \arrow[u, bend left, "\mu_\phi(x)"] \end{tikzcd}$$

want more interpretable model, understand $z$

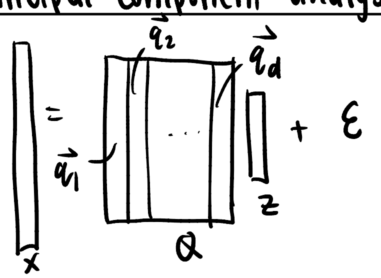
$$\begin{tikzcd} z \arrow[d, "\text{decoder (generator)}"] \\ x \arrow[u, bend left, "\text{(inference) encoder}"] \end{tikzcd}$$

linear $g_\theta$: factor analysis




principal component analysis (PCA)




$(\vec{q}_1, \cdots, \vec{q}_k, \cdots \vec{q}_d)$ : orthogonal basis

## Embedding (representation)

**supervised / prediction:**

output
↑ (decoder)
embedding
↑ (encoder)
input

self-supervised
create own output
(masked words)
↓
in between
supervised /
unsupervised

**unsupervised / generative:**

embedding
(encoder) ↑ ↓ (decoder)
input

auto-encoder
↳ no output
↳ dimension reduction

**relative (relational)**

z' — z — z''
x' — x — x''

(e.g. preserving distance)

**clustering** (figure out categories)

vec (continuous)
↑
word (one-hot)

⟹ opposite direction

z (one-hot)
↑
x (continuous)