Vae

→ Simple dist → complex dist

eg. VAE

Generative Process $z \sim p(z)$ → latent variable → prior, eg. $\mathcal{N}(0, I)$

$x \sim p_\theta(x|z)$ → likelihood, decoder → learn θ → decoder given z what image?

$p(x)$ → complex, eg. image, audio

encoder ← Posterior

$$p_\theta(z|x) = \frac{p_\theta(x|z)\, p_\theta(z)}{p_\theta(x)}$$

likelihood, prior, evidence

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

is intractable

approx posterior

$q_\phi(z|x)$ → need to learn → encoder

Select q → ~~normalizing~~ constant, sample

Goal: Minimize

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

ELBO (evidence lower bound) (Variational Inference)

intractable

$$\log p_\theta(x) = ELBO + KL(\quad)$$

(log likelihood)

$KL \geq 0$

$\log p_\theta(x) \geq ELBO$

ELBO:

$$\mathcal{L}(x; \theta, \phi) = E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - KL\left(q_\phi(z|x) \,\|\, p(z)\right)$$

maximize

Interlude

$$D_{KL}(q_\phi \| p_\theta) = E_{q_\phi}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]$$

Applies to both discrete and cont, use for distillation, generative models etc.
The less likely an event, the more surprising it is, so surprise is 1/P(x)

90% chance no surprise quiz! You can chill.
Surprise quiz!

No quiz? Expected!
less likely → more surprising

$$I(x) = \log \frac{1}{P(x)}$$

mon quiz tues quiz wed quiz

3x surprise

Ind Probabilities multiply → Surprise adds

$$\text{Entropy} = E[I] = H(P) = 90\% \log \frac{1}{90\%} + 10\% \log \frac{1}{10\%}$$

Avg Surprise = bits

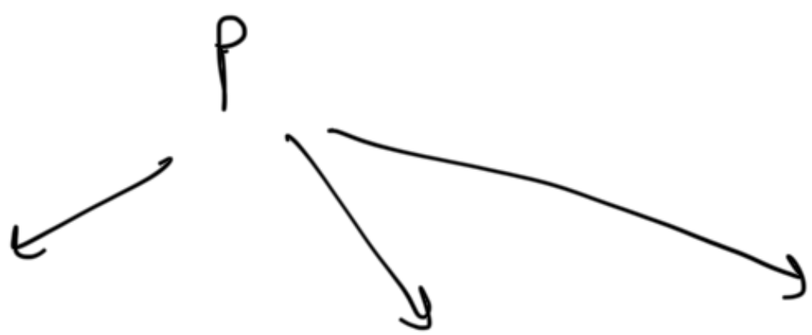
encode outcome $= \sum_i P(x_i) \log_2 \frac{1}{P(x_i)}$

eg: P(Mess food)

$H(P) = 1 \log \frac{1}{1} = 0$

= 0 bits to encode this information

= no surprise



finish syllabus 50%.

last minute

on time 25%.

important topics last minute 25%.

$H(P) = \frac{1}{2} \log \frac{1}{(1/2)} + \frac{1}{4} \log \frac{1}{(1/4)} + \frac{1}{4} \log \left(\frac{1}{1/4}\right)$

$= 1.5$

Encoding: 0 | 10 | 11

Say

suboptimal encoding

10 — Assume 25%

0 — (50%)

11 — (25%)

wrong idea

code optimized for wrong model

Cross Entropy

$H(P, q) = \sum p(x) \log_2 \frac{1}{q(x)}$

$H(P, q)$: P ← real, q ← assumption

$D_{KL}(P \| Q) = H(P, Q) - H(P)$

P: true, Q: approx

$= \left(\frac{1}{2}\right) \log_2 \frac{1}{(1/4)} + \frac{1}{4} \log_2 \left(\frac{1}{1/2}\right)$

$+ \frac{1}{4} \log_2 \frac{1}{(1/4)}$

$= 1 + \frac{1}{4} + \frac{1}{2}$

$= 1.75$

Self Information $I(x) = \log \frac{1}{P(x)}$

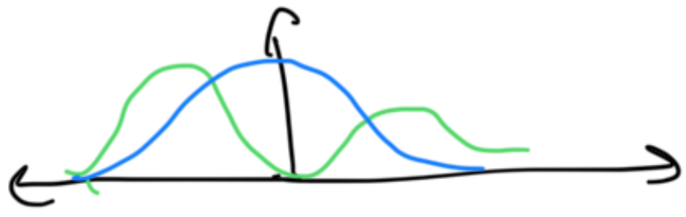
Cross Entropy : Avg surprise under wrong belief

so on average 0.25 bit extra

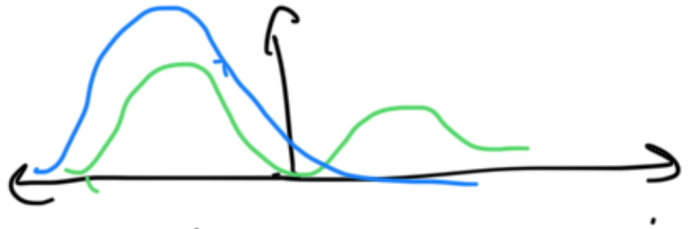
forward KL

$D_{KL}(P||Q)$ → mode covering

$D_{KL}(Q||P)$ → mode seeking (reverse)

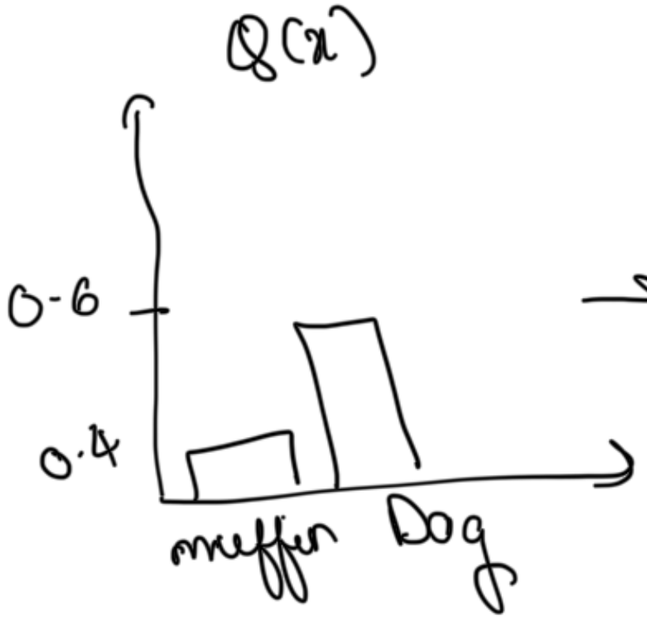


picks one peak

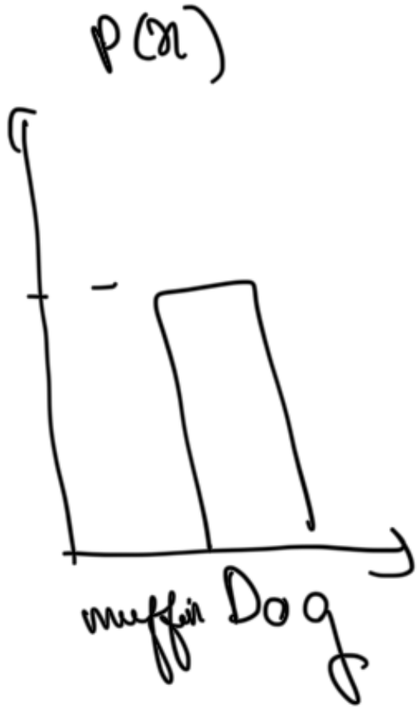


Penalized for assigning mass where P has non

$$E\left[Q(x)\log\frac{Q(x)}{P(x)}\right]$$



→ $D_{KL} \in$



$$\min D(P||Q) = E\left[p(x)\log\frac{p(x)}{Q(x)}\right]$$

ie minimize

$$- \sum_i P(x_i) \log Q(x_i)$$

<u>Problem</u>

Vocab size LLM

$\approx$ 200K token

no closed form solution!

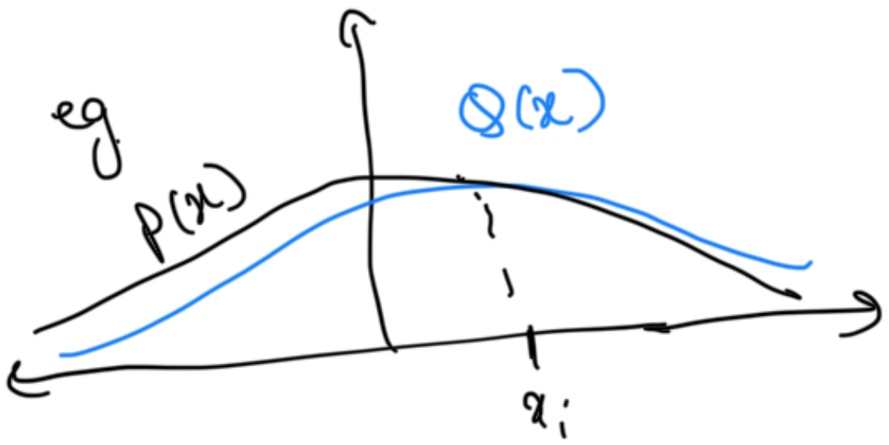
$$E_{x \sim p}\left[\log \frac{P(x)}{Q(x)}\right]$$

Monte Carlo

Sample $x_1, \ldots, x_N \sim P(x)$

$$\hat{D}_{KL}() = \frac{1}{N} \sum^{N} \log \frac{P(x_k)}{Q(x_k)}$$ } Unbiased

D_{KL} ... N ... $\log\frac{P(x_k)}{Q(x_k)}$)



lot of variance

negative if $Q(x_k) > P(x_k)$

$= \frac{1}{N} \sum \frac{1}{2}(\quad)^2$ } lower variance if P, Q are close enough

Bias deviation from true value

Control Variates

simple est → add term ST $E[\text{term}] = 0$

$$= \frac{1}{N} \sum \left[\frac{p(x_k)}{q(x_k)} + \lambda \left(r(x_k) - \mathbb{E}_{x \sim p}[r(x)] \right) \right]$$

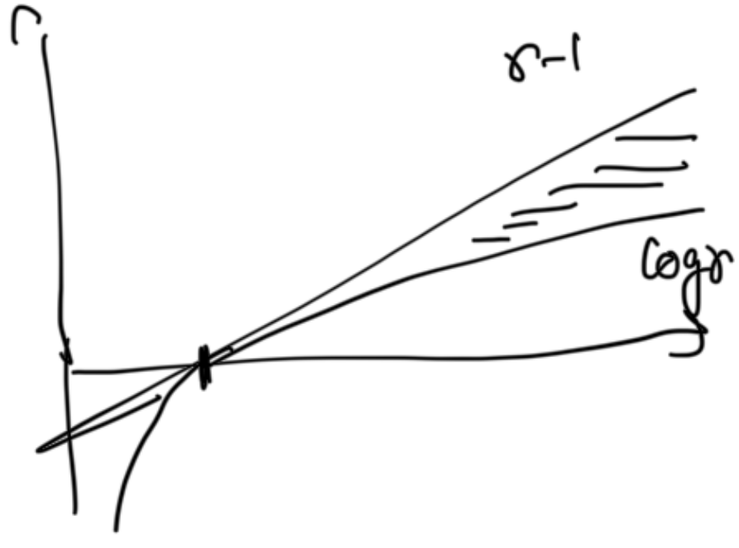
$$r(x) = \frac{q(x)}{p(x)}$$

$$\mathbb{E}\, r(x) = \int p(x)\, r(x)\, dx$$

$$= 1$$

now λ?

$$= \frac{1}{N} \sum \left[-\log r(x_k) + \lambda (r - 1) \right]$$



$\lambda = 1$, estimate remains non negative

$$= \frac{1}{N} \left[r(x_k) - 1 - \log r(x_k) \right]$$

ELBO

$$D_{KL}(q_\phi \| p(\theta) = E_{q_\phi}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]$$

$$= E_{q_\phi}\left[\log q_\phi(z|x)\right] - E_{q_\phi}\left[\log p_\theta(z|x)\right]$$

$$= E_{q_\phi}\left[\log q_\phi(z|x)\right] - E_{q_\phi}\left[\log p(z,x)\right] + \log p_\theta(x) \int q_\phi(z|x)\,dz$$

$$= E_{q_\phi}\left[\log q_\phi(z|x)\right] - E\left[\log p_\theta(z,x)\right] + \underbrace{\log p_\theta(x)}_{\text{marginal log likelihood}}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{component}}$$

$$\log p_\theta(x) = -E_{q_\phi}\left[\log q_\phi(z|x)\right] + E_{q_\phi}\left[\log p_\theta(z,x)\right] + D_{KL}\left(q_\phi \| p_\theta\right)$$

$\log p_\theta(x)$ — can't compute — log evident

component 2 — intractable — ≥ 0 ← always positive

$\log p_\theta(x) \geq$ Evident lower bound

maximize this to minimize $D_{KL}\left(q_\phi \| p_\theta\right)$

$$ELBO = -E_{q_\phi}\left[\log q_\phi(z|x)\right] + E_{q_\phi}\left[\log p_\theta(z,x)\right]$$

$$E[\log p_\theta(x|z)]$$

$$+ E[\log p_\theta(z)]$$

$$= E_{q_\phi}\left[\log p_\theta(\underbrace{\underbrace{x}_{\text{orig}}}_{z})\right] - E_{q_\phi}\left[\log \frac{q_\phi(z|x)}{p_\theta(z)}\right]$$

latent space

Expected Reconstruction error

KL divergence b/w

learnable approx posterior & prior we know

We could also pick $q_\phi(z)$ → not conditioned on x