Foundation of Generative Models

We will discuss neural networks and machine learning in terms of probabilities and distributions instead of input and output.

Terminologies :

$p(x) : X \longrightarrow \mathbb{R}^+$ such that $\begin{cases} \sum_{x \in X} p(x) = 1 \\ \int_X p(x)\, dx = 1 \end{cases}$

$p_\theta(x)$ is a learned model of $p(X)$

Joint distribution $p(x, y)$

Conditional distribution $p(x \mid y)$

From distribution $p(\cdot)$ to instance x:

$x \in$ Discrete Space:

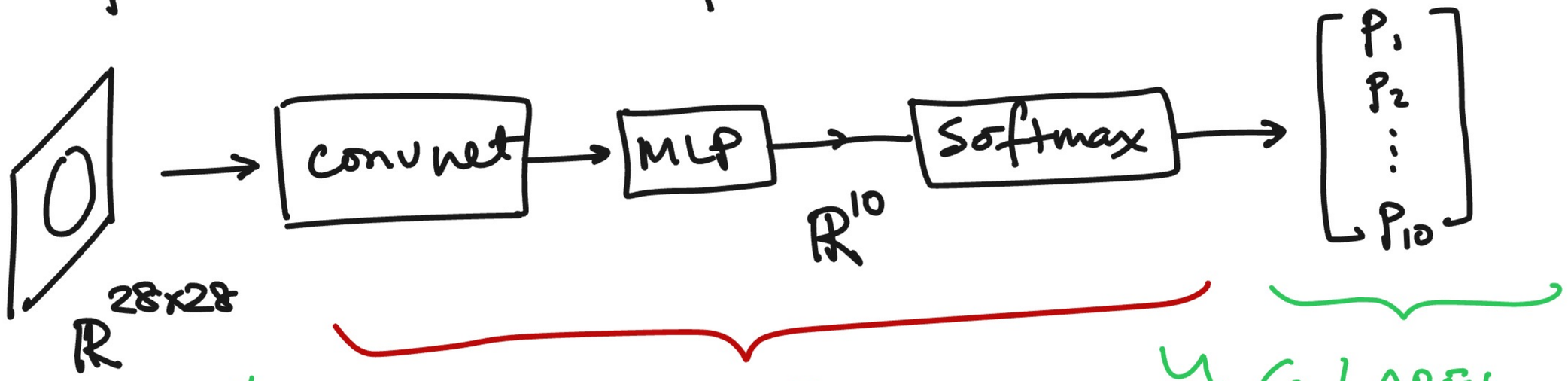
$$\text{Sample}(p) \rightarrow x$$

Pick from discrete space and favour samples x w/ larger $p(x)$.

$x \in \mathbb{R}^d$:

$$x = \mathbb{E}[X|p] = \int_{\mathbb{R}^d} x \cdot p(x)\, dx$$

Classification as a probabilistic model



$x \in$ IMAGE

$p_\theta(y|x)$

$y \in$ LABEL

$= \text{softmax}(\text{logits})$

CLASSIFICATION $=$ DISCRIMINATIVE MODEL

Training data = pairs of random variables

$$(x, y)$$

Learning the conditional distribution

$$P_\theta(y|x) \simeq P_{data}(y|x)$$

GENERATIVE MODEL

Training data $D = \{x_1, x_2 \dots x_N\}$

are samples of a random var $x \in \mathbb{R}^d$.

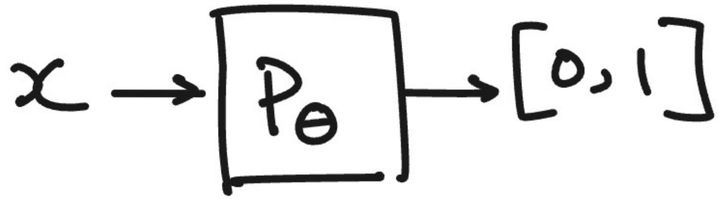
Want to learn $P_\theta(x) \simeq P_{data}(x)$

Given x, we know its likelihood $p(x)$!

(Explicit model)

Why is generative model useful ?

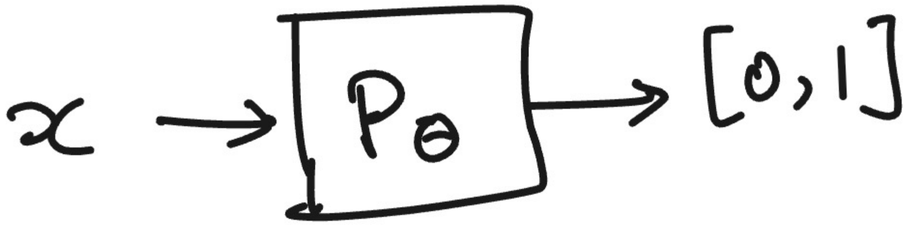
$P_\theta(x)$



Under conditions, we can make new instances of x! (i.e. generate new instances of samples not in the observed data.)

Explicit Density Model vs Implicit Density Model

- We can get the likelihood from model given instance x



- Cannot measure likelihood,
- Generate new "likely" instances directly

Autoregressive LM is an explicit density model

$x = [x_1 \; x_2 \; \cdots \; x_L]$ is a sentence,
x_i are tokens.

What is $p_\theta(x) = ?$

$$p(x_1 x_2 \cdots x_L) = p(x_L \mid x_1 \cdots x_{L-1}) \cdot p(x_1 \cdots x_{L-1})$$

↑

This is cond prob $\quad p(A,B) = p(A|B)\,p(B)$

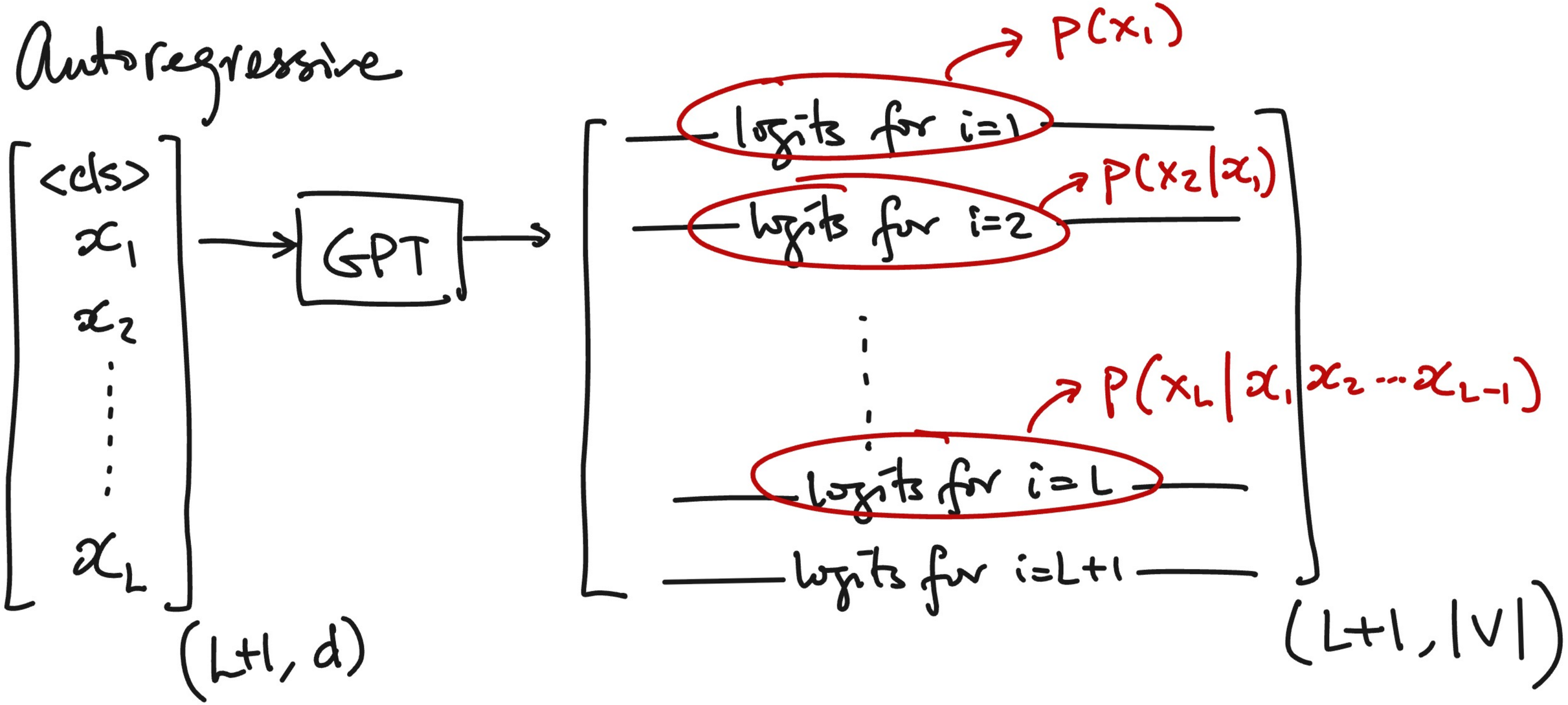
$$p(x_1, x_2 \cdots x_L) = p(x_1, x_2 \cdots x_{L-1}) \cdot p(x_L | x_1 \cdots x_{L-1})$$

$$= p(x_1 \cdots x_{L-2}) \cdot p(x_{L-2} | x_1 \cdots x_{L-3}) \cdot p(x_L | x_1 \cdots x_{L-1})$$

$$= \prod_{i=1}^{L} p(x_i | x[1:i-1])$$

next token

already generated tokens


Autoregressive
<cls>
x_1
x_2
x_L
GPT
$P(x_1)$
logits for i=1
$P(x_2|x_1)$
logits for i=2
$P(x_L|x_1 x_2 \cdots x_{L-1})$
logits for i=L
logits for i=L+1
(L+1, d)
(L+1, |V|)

How to generate from explicit density models ?

$\Rightarrow$ SAMPLING

Relies on finite discrete vocabulary for each x_i

Since $p(x_{i+1} \mid x_1 \dots x_i)$ is known for all $x_{i+1} \in V$,

Sample x_{i+1} by picking most likely next tok.

What about continuous space sampling $\Rightarrow$ Variational Autoencoder...

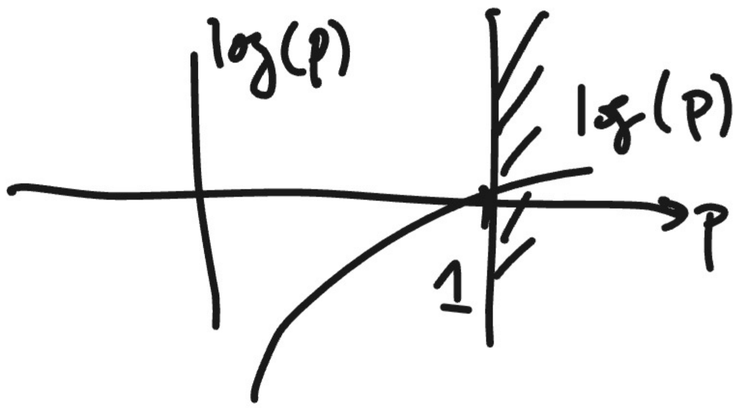
Maximal Likelihood Traing of Explicit Model

$\mathcal{D} = \{ x^{(1)}, x^{(2)} \dots x^{(N)} \}$ is the training data.

$x^{(i)}$ → for LM, each $x^{(i)}$ is a sequence of tokens!

P_θ is trained so $\mathcal{D}$ are likely samples of P_θ.

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P_\theta(x^{(i)})$$

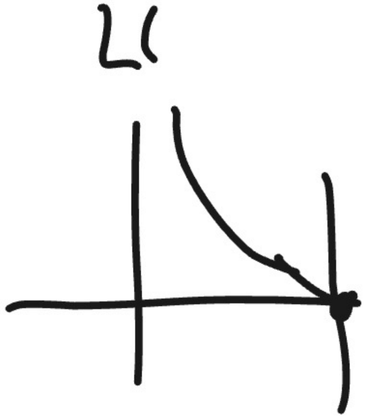


Max. Likelihood Loss Function

$$L_{ML} = -\frac{1}{N}\sum_{i=1}^{N} \log P_\theta(x^{(i)})$$

distance measure between two distributions

Connection to P_{data} via KL-divergence



$$L_{ML} = -\,E\left[\log P_\theta(x) \;\middle|\; P_{data}(x)\right]$$

observing $\log P_\theta(x)$

sample x from P_{data}

About KL Divergence (a quick look)

$$KL(P_{data} \| P_\theta) = E_{x \sim P_{data}}\left[\log\left(\frac{P_{data}(x)}{P_\theta(x)}\right)\right]$$

$$= \underbrace{E_{x \sim P_{data}}\left[\log P_{data}(x)\right]}_{\text{Not dependent on } \theta \Rightarrow \text{constant}} \underbrace{- E_{x \sim P_{data}}\left[\log P_\theta(x)\right]}_{L_{ML}(\theta)}$$

$$\theta^* = \arg\min_\theta KL(P_{data} \| P_\theta)$$

Explicit Density Model w/ Target Distribution

$\mathcal{D} = \{ x^{(1)}, x^{(2)}, \dots \}$ and want p^*

$L_{ME}(\theta)$ $\qquad$ $L_{KL}(\theta) = KL(p^* \| p_\theta)$

$$L(\theta) = L_{ME}(\theta) + \lambda L_{KL}(\theta)$$

This trick will be used in latent variable models.

Latent Variable Models as Explicit Models

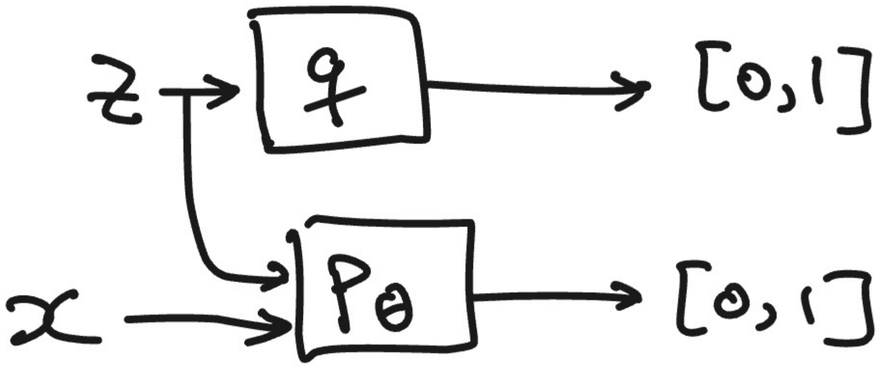
2 stage models

Stage #1 : explicit model on z

$$z \sim q(z)$$

Stage #2 : conditional model

$$x \sim p_\theta(x|z)$$



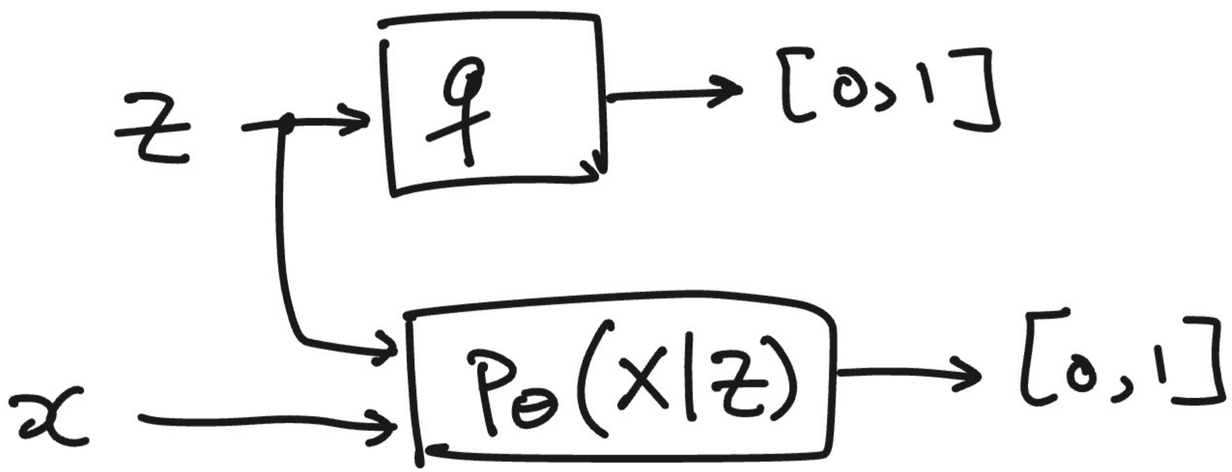
Latent Var. Model

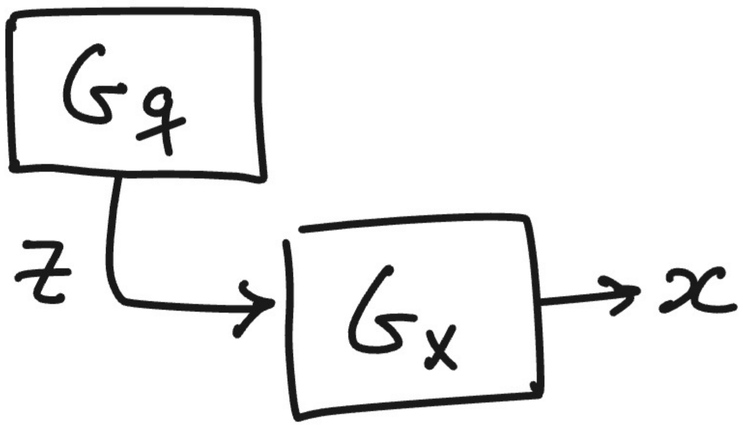
Given x, what is $p(x)$?

$$P_\theta(x) = \int P_\theta(x|z) \cdot q(z)\, dz \qquad \text{continuous}$$

$$= \sum_{z \in Z} P_\theta(x|z)\, q(z) \qquad \text{discrete}$$

More on latent variable explicit models





SAMPLING:

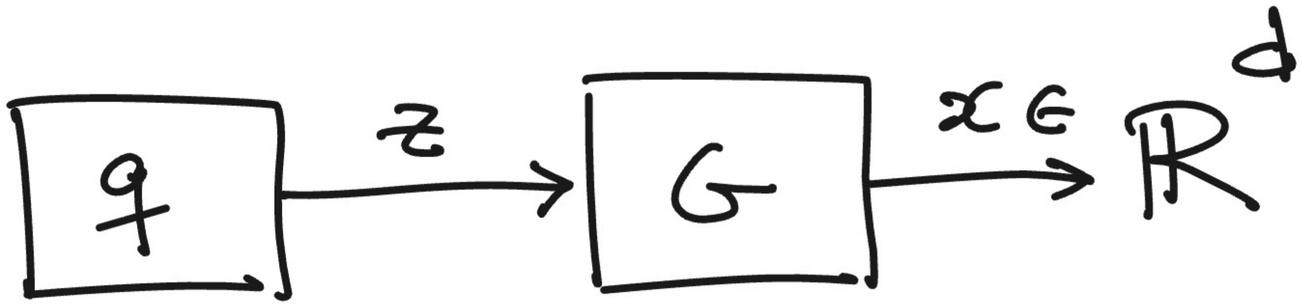
$x \sim P_\theta(X)$

INFERENCE

$P(z|x)$

Latent Variable Implicit Models

LATENT VARIABLE z is a control



for large d

Simple Gaussian

$N(0,1)$ in $\mathbb{R}^k$ for small k

NEXT WEEK:

Variational Autoencoder ← latent variable, explicit model

Diffusion Model ← latent variable, implicit model

IMAGE GENERATION