Lecture 14

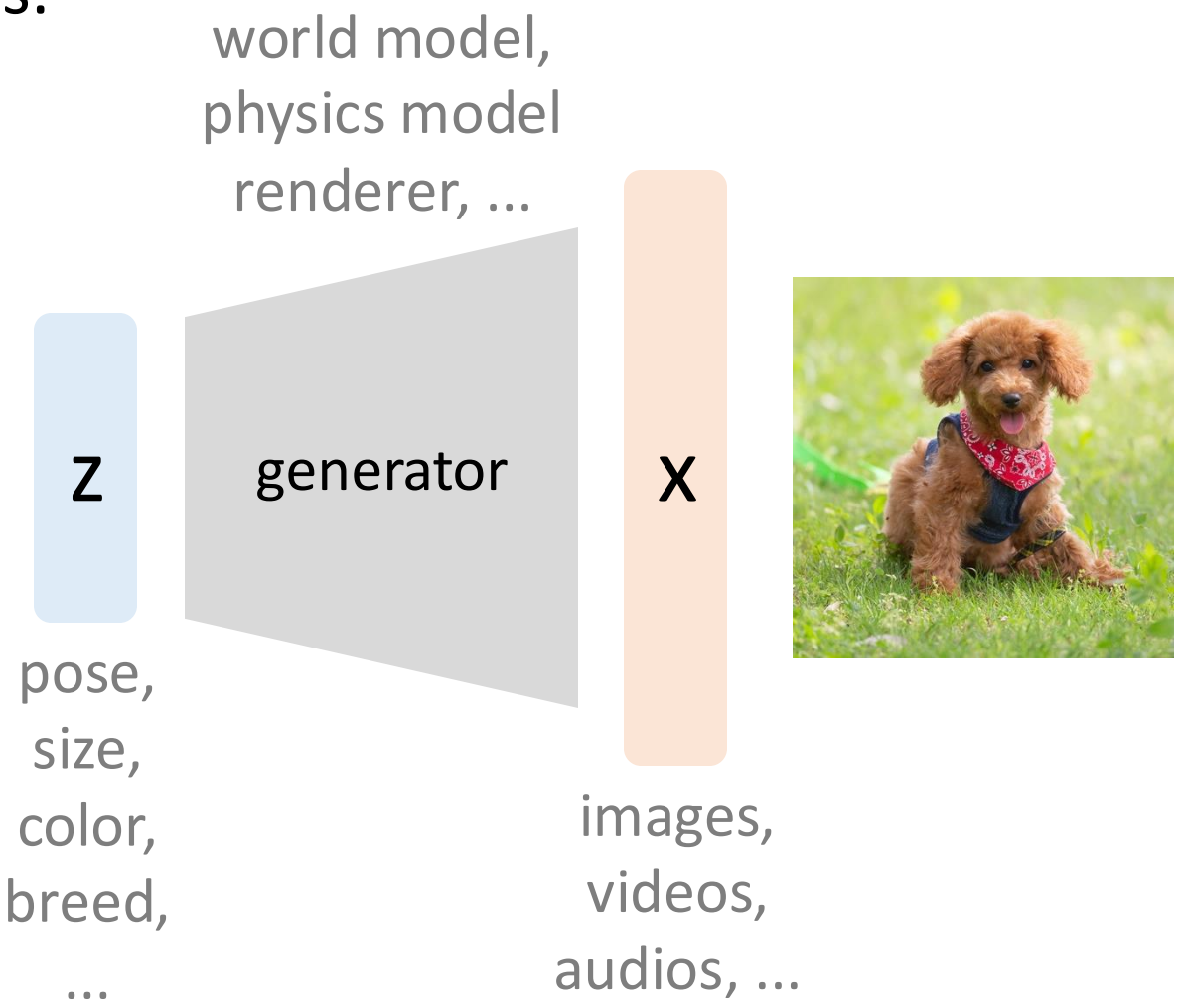
Generative Models: VAE and GAN

Speaker: Kaiming He

Latent Variable Models

Assuming a data generation process:

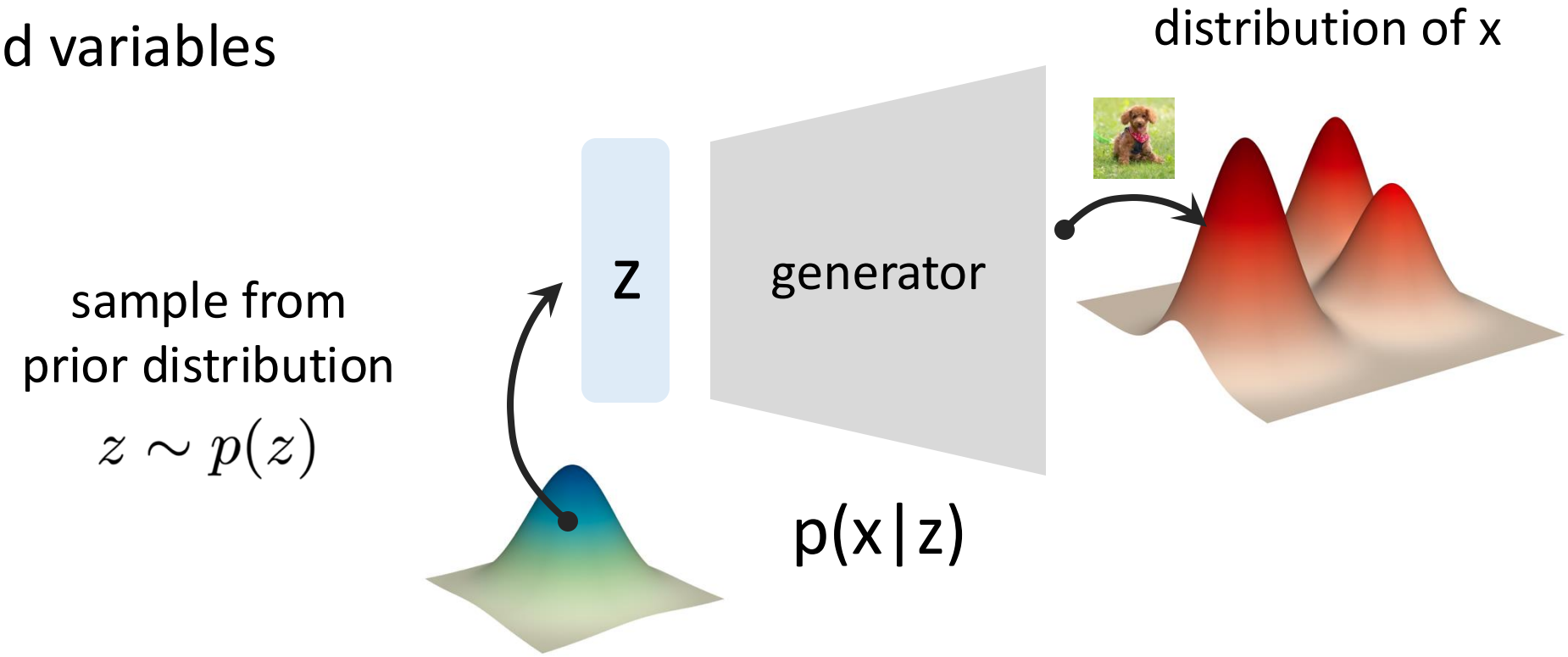
- z - latent variables
- x - observed variables



Latent Variable Models

Assuming a data generation process:

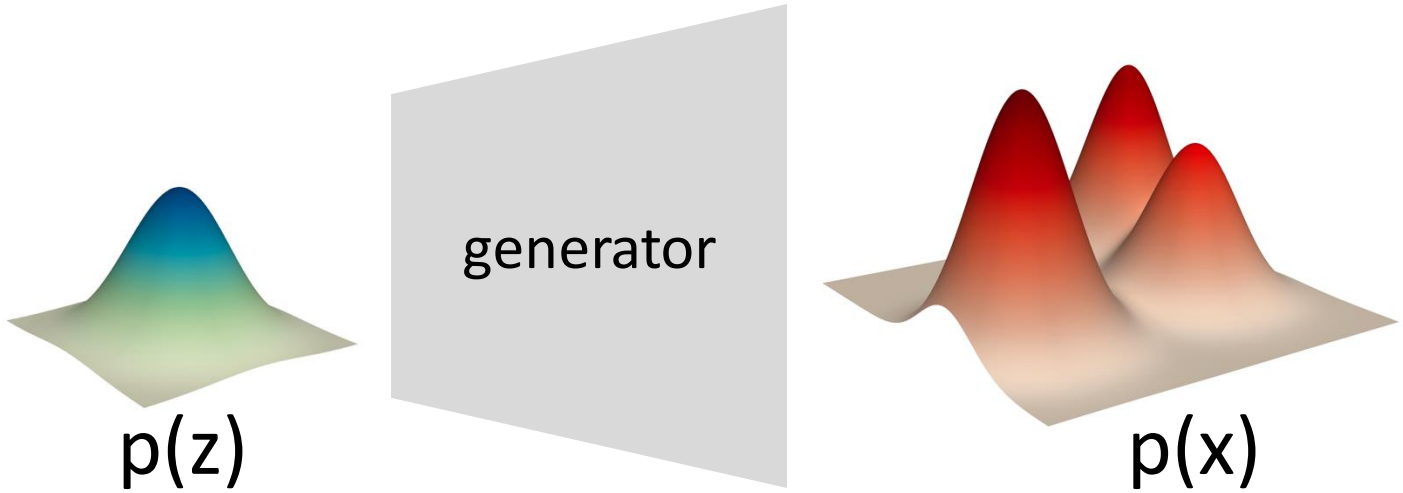
- z - latent variables
- x - observed variables



Latent Variable Models

Assuming a data generation process:

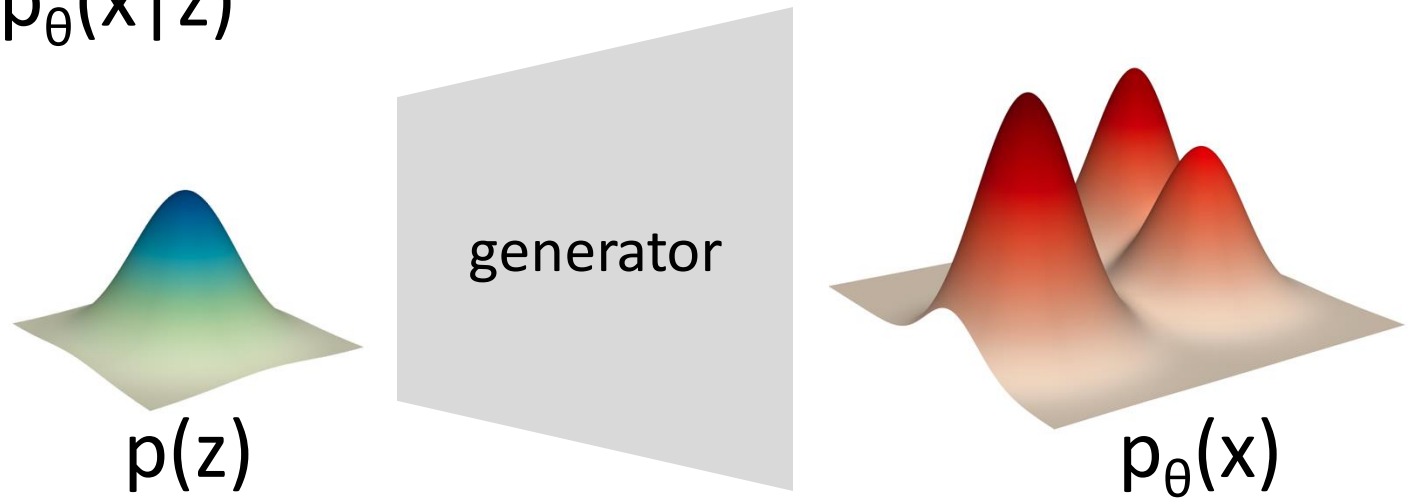
- z - latent variables
- x - observed variables



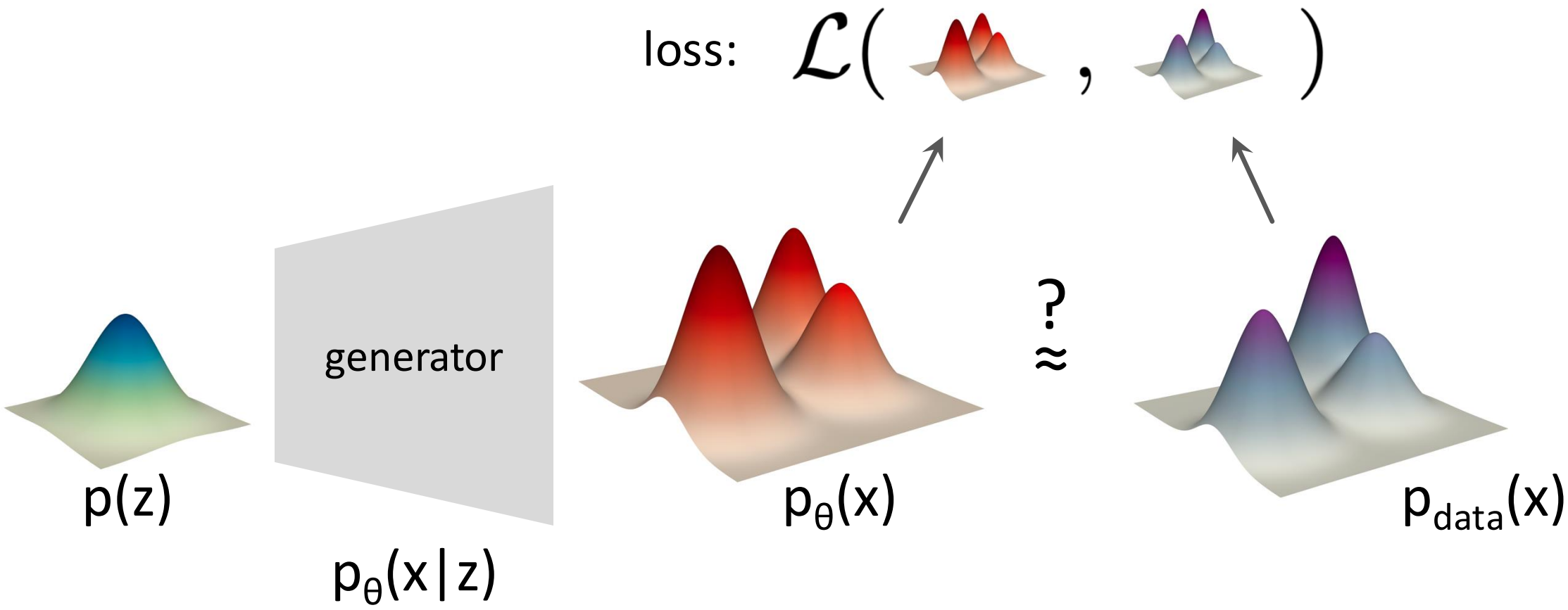
Latent Variable Models

Represent a distribution by a neural network

- θ - learnable parameters
- represent a conditional: $p_\theta(x|z)$



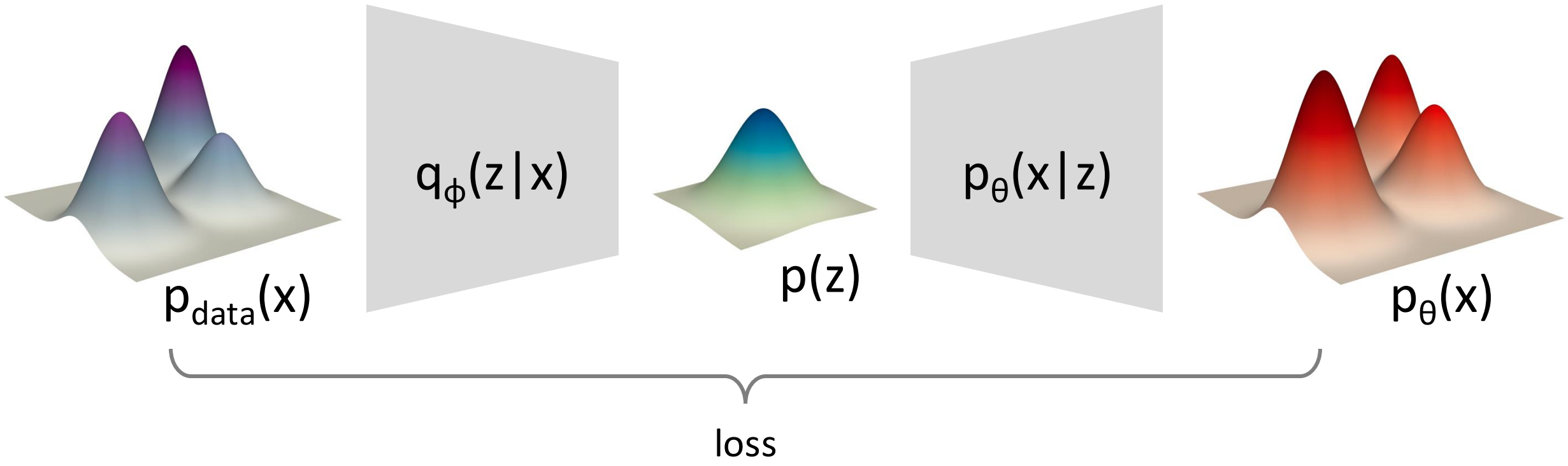
Measuring How Good A Distribution Is



- A core topic in generative modeling: make it **differentiable**, **computable**, and **tractable**

Variational Autoencoder: Overview

- **Encoder**: data to latent, parameterized by ϕ
- **Decoder**: latent to data, parameterized by θ



Maximum Likelihood Estimation (MLE)

Minimize KL-Divergence:

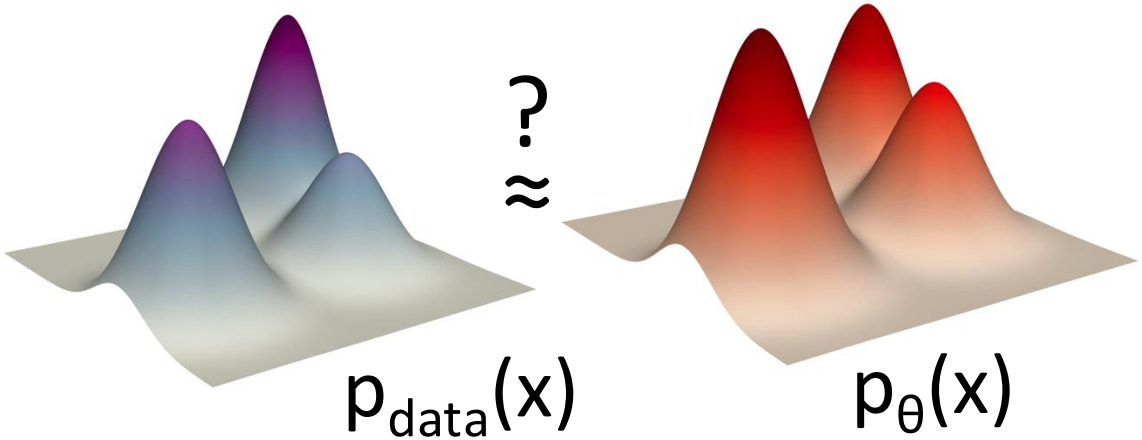
$$\min_\theta \; \mathcal{D}_{\mathrm{KL}}(\, p_{data} \, || \, p_\theta \,)$$

⇒ **Maximize** likelihood:

$$\max_\theta \; \mathbb{E}_{x \sim p_{data}} \log p_\theta(x)$$

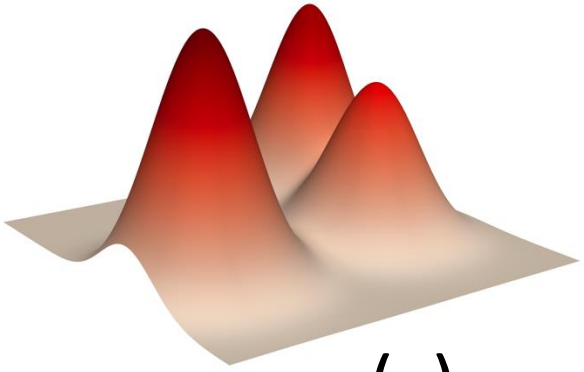
tl; dr

$$\begin{aligned}
& \arg\min_\theta \; \mathcal{D}_{\mathrm{KL}}(\, p_{data} \, || \, p_\theta \,) \\
= \; & \arg\min_\theta \; \sum_x p_{data}(x) \log \frac{p_{data}(x)}{p_\theta(x)} \\
= \; & \arg\min_\theta \; \sum_x -p_{data}(x) \log p_\theta(x) + const \\
= \; & \arg\max_\theta \; \sum_x p_{data}(x) \log p_\theta(x) \\
= \; & \arg\max_\theta \; \mathbb{E}_{x \sim p_{data}} \log p_\theta(x)
\end{aligned}$$



Variational Autoencoder (VAE)

We want to maximize $\mathbb{E}_{x \sim p_{data}} \log p_\theta(x)$



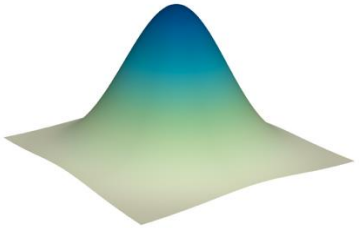
$p_\theta(x)$

Variational Autoencoder (VAE)

We want to maximize $\mathbb{E}_{x \sim p_{data}} \log p_\theta(x)$

with $p_\theta(x)$ represented as:

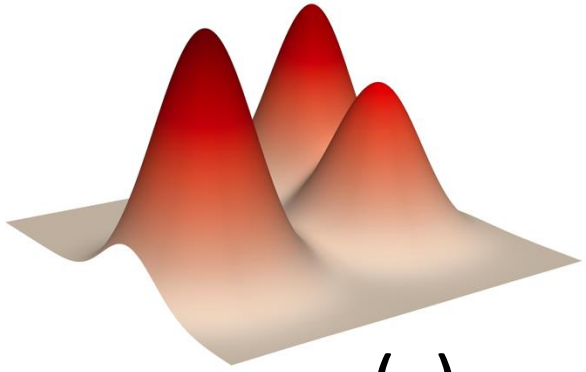
$$p_\theta(x) = \int_z p_\theta(x|z)p(z)dz$$



p(z)

generator

$p_\theta(x|z)$



$p_\theta(x)$

Variational Autoencoder (VAE)

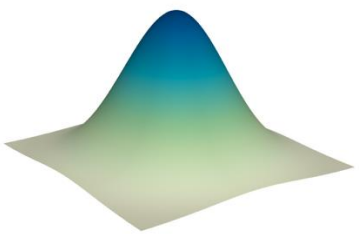
We want to maximize $\mathbb{E}_{x \sim p_{data}} \log p_\theta(x)$

with $p_\theta(x)$ represented as:

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

Two sets of unknowns:

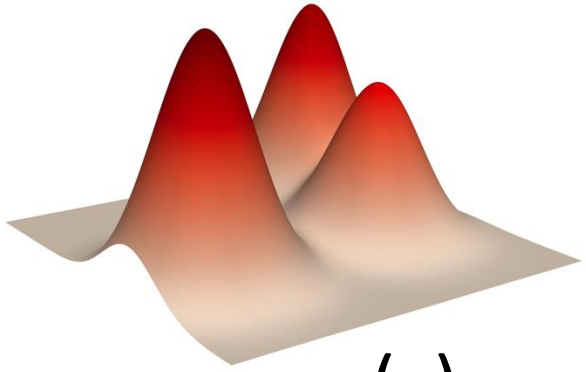
- We need to optimize for θ
- We can't control "true" p(z)



p(z)

generator

$p_\theta(x|z)$



$p_\theta(x)$

Idea: introduce a "controllable" distribution $q_\phi(z)$

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

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

$$= \int_z q(z) \log p_\theta(x) dz$$

- valid for any distribution q(z)

$$= \int_z q(z) \log \Big(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \Big) dz$$

- Bayes' rule

$$= \int_z q(z) \log \Big(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \Big) dz$$

$$= \int_z q(z) \Big(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \Big) dz$$

$$= \mathbb{E}_{z \sim q(z)} \Big[\log p_\theta(x|z) \Big] - \mathcal{D}_{\mathrm{KL}} \Big(q(z) || p_\theta(z) \Big) + \mathcal{D}_{\mathrm{KL}} \Big(q(z) || p_\theta(z|x) \Big)$$

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

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

$$= \int_z q(z) \log p_\theta(x) dz$$

- valid for any distribution q(z)

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \right) dz$$

- Bayes' rule

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \right) dz$$

- just algebra

$$= \int_z q(z) \left(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \right) dz$$

- just algebra

$$= \mathbb{E}_{z \sim q(z)} \left[\log p_\theta(x|z) \right] - \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z) \right) + \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z|x) \right)$$

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

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

$$= \int_z q(z) \log p_\theta(x) dz$$

- valid for any distribution q(z)

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \right) dz$$

- Bayes' rule

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \right) dz$$

- just algebra

$$= \int_z q(z) \left(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \right) dz$$

- just algebra

$$= \mathbb{E}_{z \sim q(z)} \left[\log p_\theta(x|z) \right] - \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z) \right) + \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z|x) \right)$$

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

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

$$= \int_z q(z) \log p_\theta(x) dz$$

- valid for any distribution q(z)

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \right) dz$$

- Bayes' rule

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \right) dz$$

- just algebra

$$= \int_z q(z) \left(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \right) dz$$

- just algebra

$$= \mathbb{E}_{z \sim q(z)} \left[\log p_\theta(x|z) \right] - \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z) \right) + \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z|x) \right)$$

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

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

$$= \int_z q(z) \log p_\theta(x) dz$$

- valid for any distribution q(z)

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \right) dz$$

- Bayes' rule

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \right) dz$$

- just algebra

$$= \int_z q(z) \left(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \right) dz$$

- just algebra

$$= \mathbb{E}_{z \sim q(z)} \left[\log p_\theta(x|z) \right] - \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z) \right) + \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z|x) \right)$$

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

$$
\begin{aligned}
& \log p_\theta(x) \\
= & \int_z q(z) \log p_\theta(x) dz \\
= & \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \right) dz \\
= & \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \right) dz \\
= & \int_z q(z) \left(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \right) dz \\
= & \mathbb{E}_{z \sim q(z)} \left[\log p_\theta(x|z) \right] - \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z) \right) + \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z|x) \right)
\end{aligned}
$$

- valid for any distribution q(z)
- Bayes' rule
- just algebra
- just algebra

Evidence Lower Bound (ELBO)

- Rewrite log-likelihood using latent z

intractable $\log p_\theta(x)$

$$= \int_z q(z) \log p_\theta(x) dz$$

- valid for any distribution q(z)

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \right) dz$$

- Bayes' rule

$$= \int_z q(z) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} \frac{q(z)}{q(z)} \right) dz$$

$$= \int_z q(z) \left(\log p_\theta(x|z) + \log \frac{p_\theta(z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)} \right) dz$$

$$= \mathbb{E}_{z \sim q(z)} \left[\log p_\theta(x|z) \right] - \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z) \right) + \mathcal{D}_{\mathrm{KL}} \left(q(z) || p_\theta(z|x) \right)$$

tractable tractable intractable

Evidence Lower Bound (ELBO)

intractable $\log p_\theta(x) - \mathcal{D}_{\mathrm{KL}}\left(q(z)||p_\theta(z|x)\right)$ intractable

$$= \mathbb{E}_{z\sim q(z)}\left[\log p_\theta(x|z)\right] - \mathcal{D}_{\mathrm{KL}}\left(q(z)||p_\theta(z)\right)$$

tractable tractable

Evidence Lower Bound (ELBO)

intractable $\log p_\theta(x) - \mathcal{D}_{\mathrm{KL}}\left(q(z)||p_\theta(z|x)\right)$ intractable

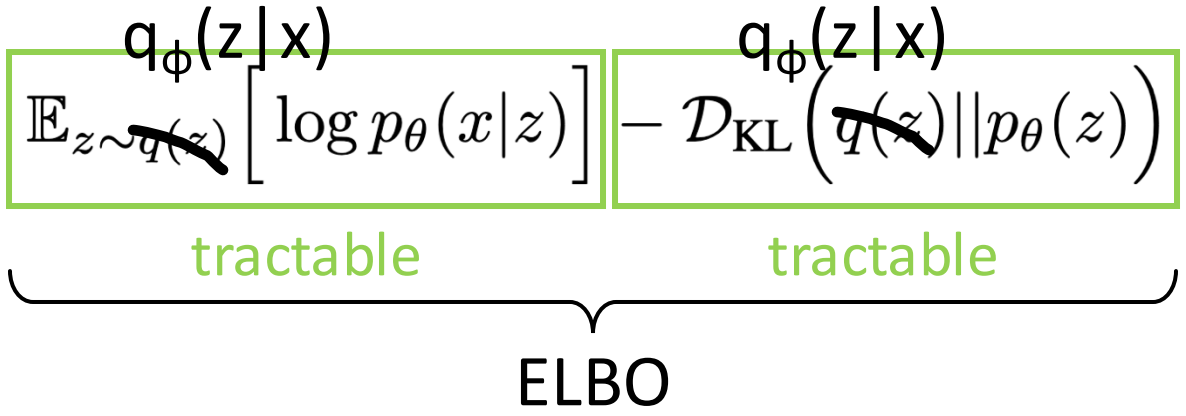
$$= \underbrace{\mathbb{E}_{z\sim q(z)}\left[\log p_\theta(x|z)\right] - \mathcal{D}_{\mathrm{KL}}\left(q(z)||p_\theta(z)\right)}_{\text{ELBO}}$$

tractable tractable

ELBO:

- It's lower bound of $\log p_\theta(x)$
- This formulation holds for <u>any</u> distribution q(z)

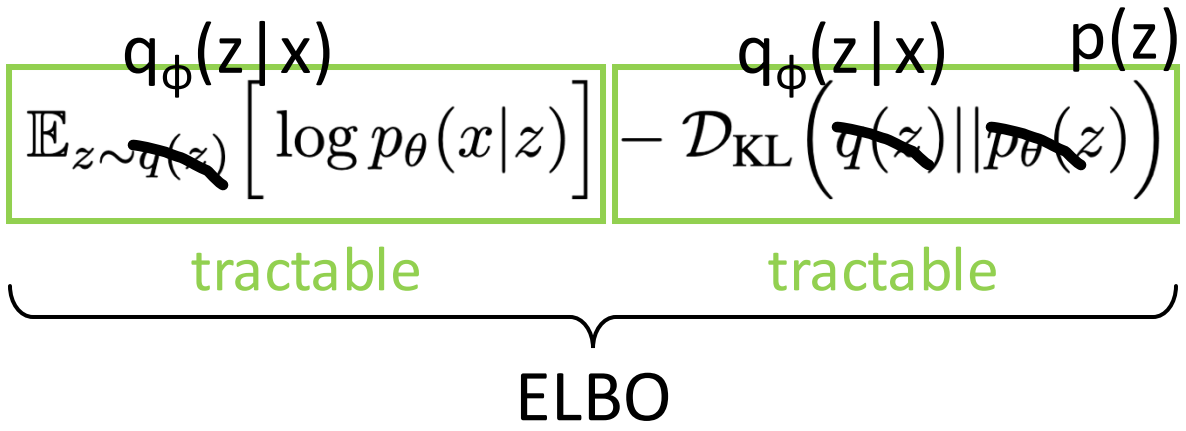
Evidence Lower Bound (ELBO)



ELBO:

- It's lower bound of $\log p_\theta(x)$
- This formulation holds for <u>any</u> distribution q(z) , so ...
- we can parameterize q(z) by $q_\phi(z|x)$

Evidence Lower Bound (ELBO)



*__Note__: ELBO is not specific to VAE.

ELBO:

- It's lower bound of $\log p_\theta(x)$
- This formulation holds for <u>any</u> distribution q(z) , so ...
- we can parameterize q(z) by $q_\phi(z|x)$
- and let $p_\theta(z)$ be a fixed, known prior p(z) (e.g., Gaussian)

Variational Autoencoder

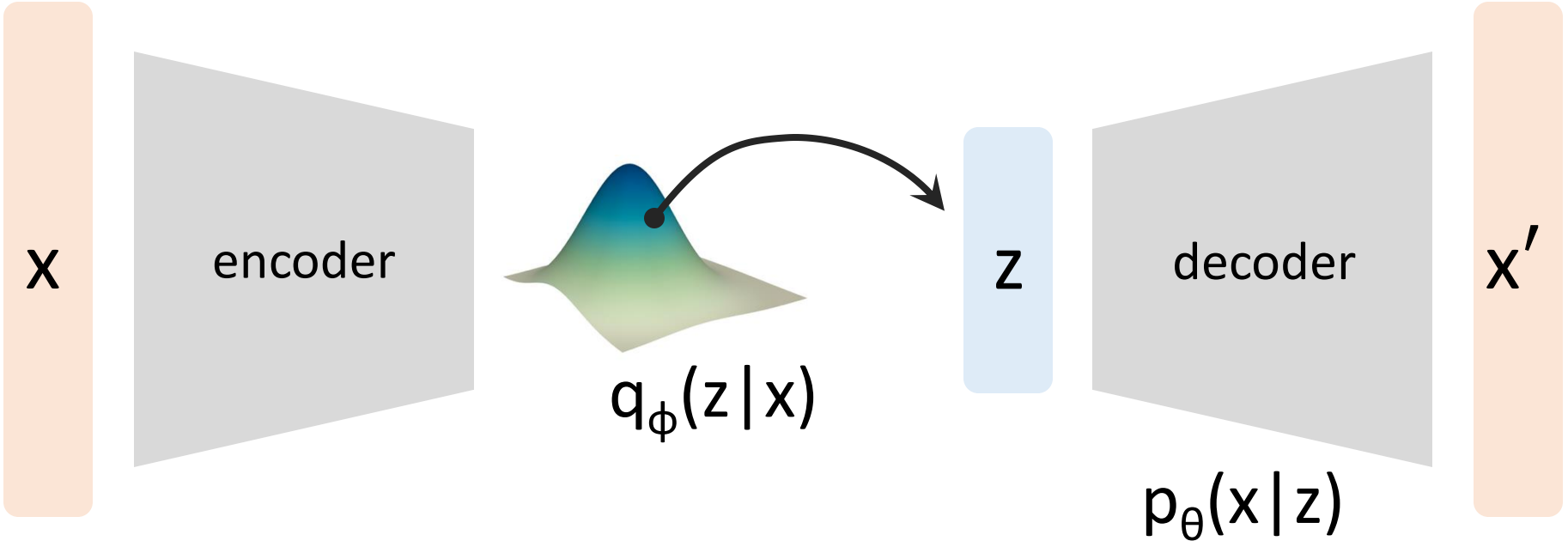
Maximize ELBO ⇒ minimize:

$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$

Variational Autoencoder

Maximize ELBO ⇒ minimize:

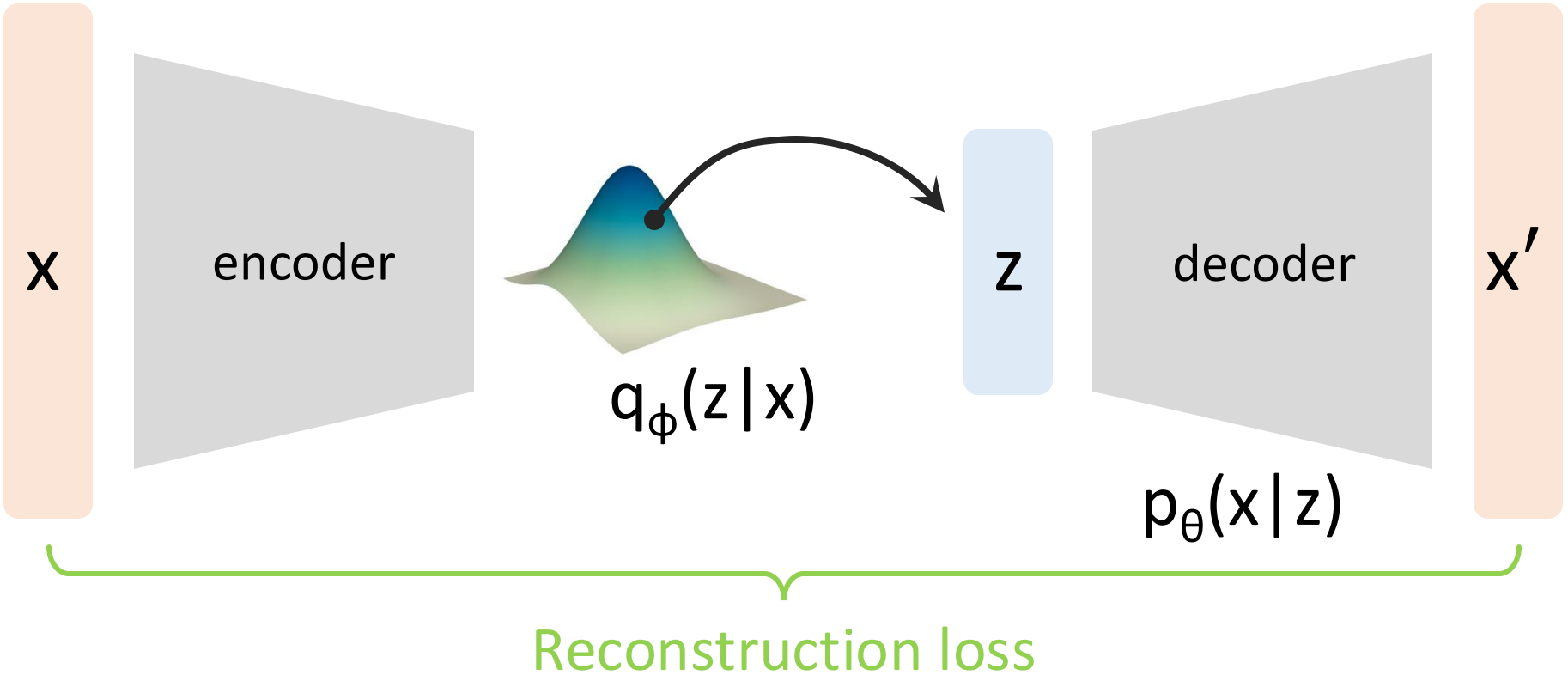
$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$



Variational Autoencoder

Maximize ELBO ⇒ minimize:

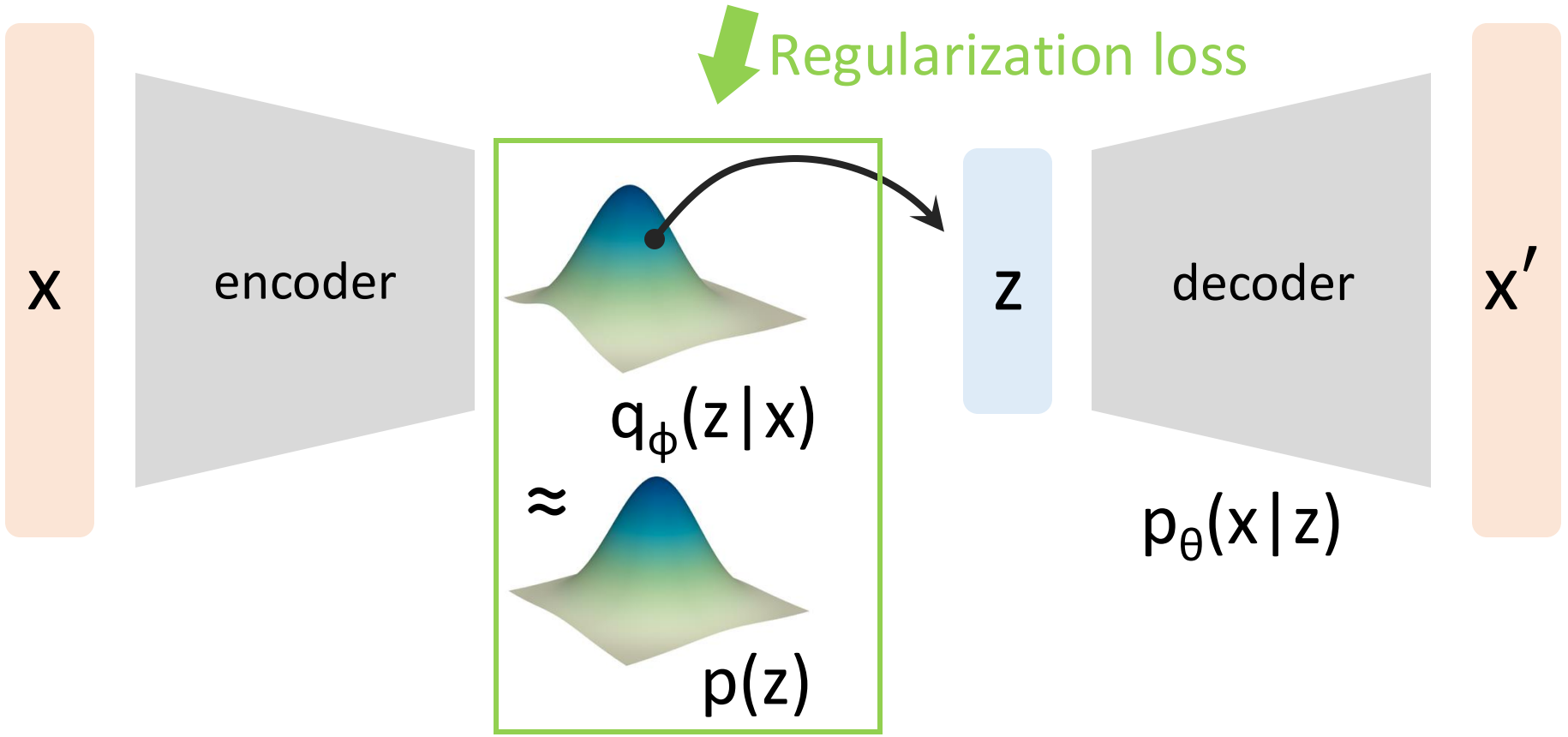
$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$



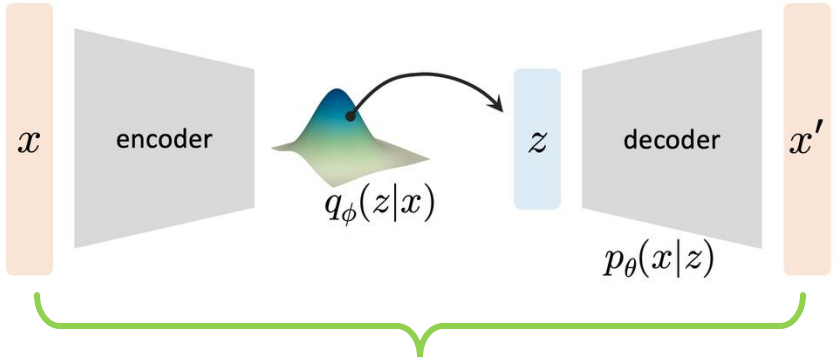
Variational Autoencoder

Maximize ELBO ⇒ minimize:

$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$



Variational Autoencoder



Reconstruction loss

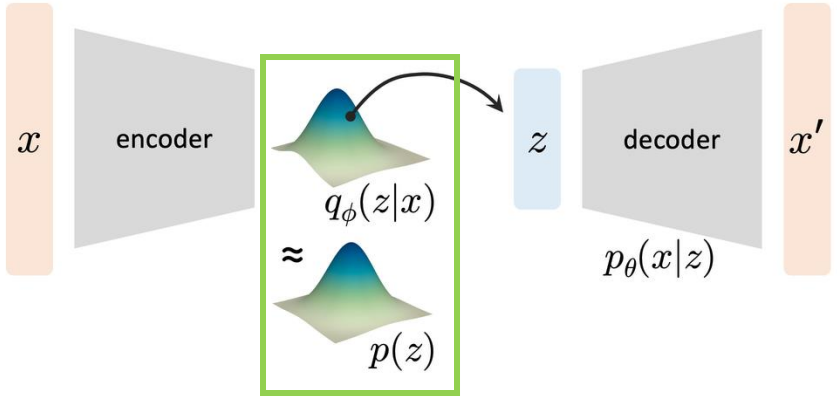
$$-\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right]$$

Example: L2 loss

- one-step Monte Carlo: $z \sim q_\phi(z|x)$
- map z by decoder net: $g_\theta(z) \to x'$ network estimates distribution's parameters (Gaussian's mean)
- model $p_\theta(x|z)$ by Gaussian: $p_\theta(x|z) = \mathcal{N}(x \mid x', \sigma_0^2)$ (assume fixed std)
- negative log likelihood: $\frac{1}{2\sigma_0^2}||x - x'||^2 + const$

- L2 loss ⇒ a Gaussian neighborhood around data point x

Variational Autoencoder



Regularization loss

$$\mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$

Example: Gaussian prior

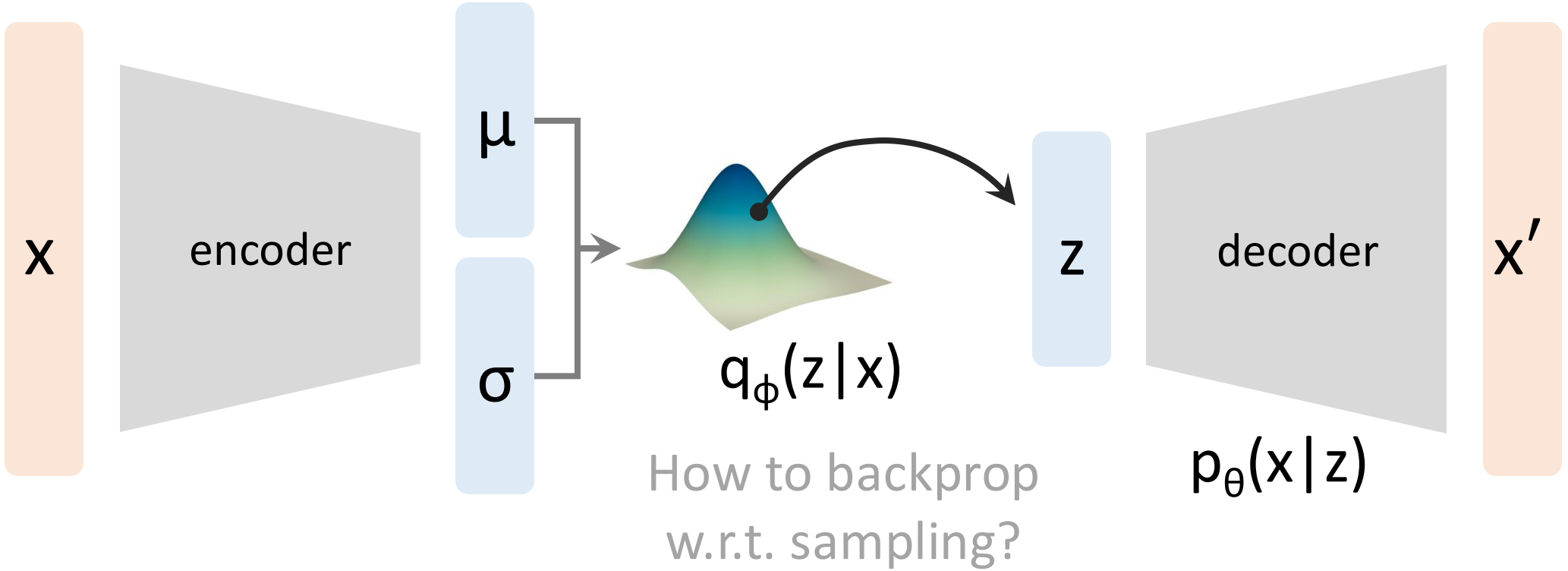
- let $p(z) = \mathcal{N}(z \mid 0, \mathbf{I})$
- model $q_\phi(z|x)$ by Gaussian: $\mathcal{N}(z \mid \mu, \sigma)$
- map x by encoder net: $f_\phi(x) \rightarrow \mu, \sigma$ again, network estimates distribution's parameters
- compute loss analytically: $\mathcal{D}_{\mathrm{KL}}\Big(\mathcal{N}(z \mid \mu, \sigma) \,||\, \mathcal{N}(z \mid 0, \mathbf{I})\Big)$

- fixed covariance ⇒ L2 loss on μ

Variational Autoencoder

Maximize ELBO ⇒ minimize:

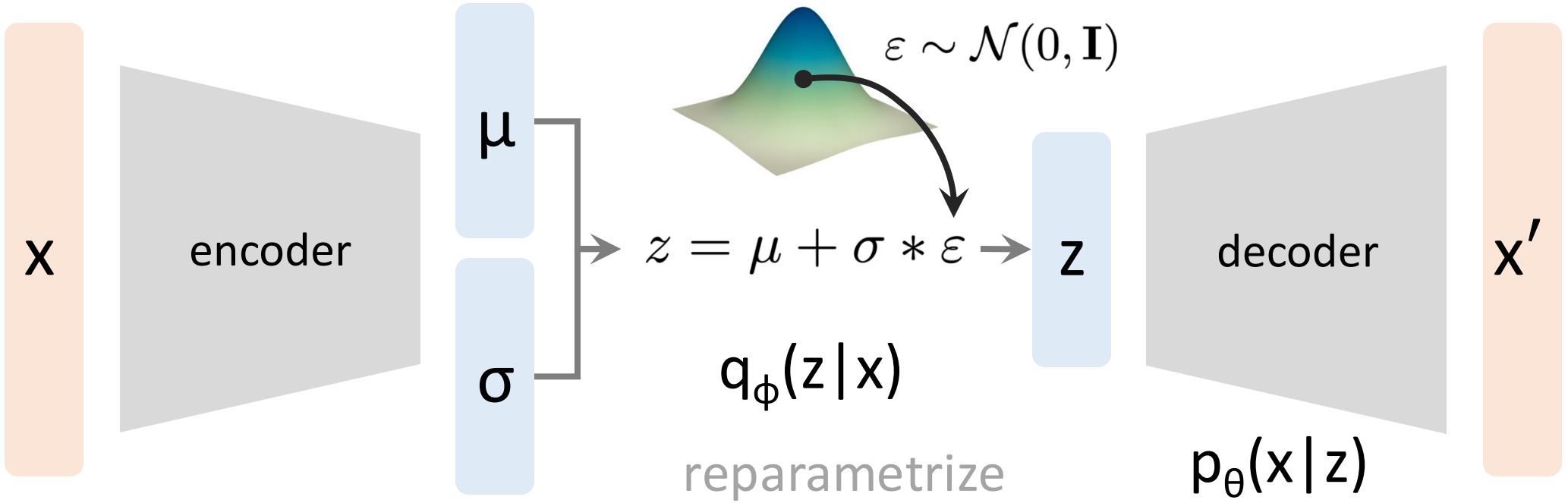
$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$



Variational Autoencoder

Maximize ELBO ⇒ minimize:

$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$



Variational Autoencoder

... so far, we have discussed an objective on **one** x:

$$\mathcal{L}_{\theta,\phi}(x) = -\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)$$

Overall loss is expectation over data:

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}_{x\sim p_{data}(x)}\Big[-\mathbb{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] + \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(z|x)||p(z)\Big)\Big]$$

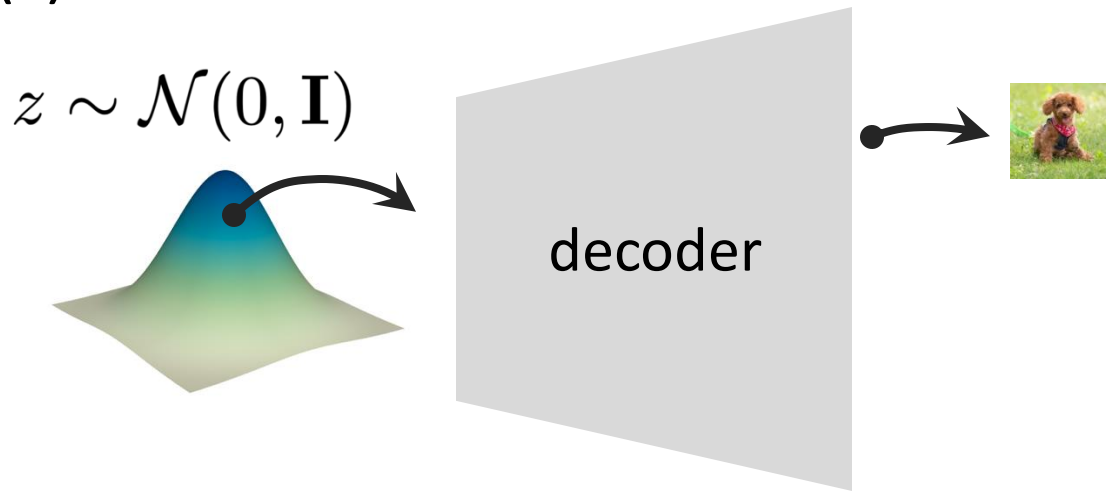
The formulation of "x-**conditional**" then "**marginalization**" is a common strategy

- Diffusion, Flow Matching, ...

Variational Autoencoder

Inference (generation):

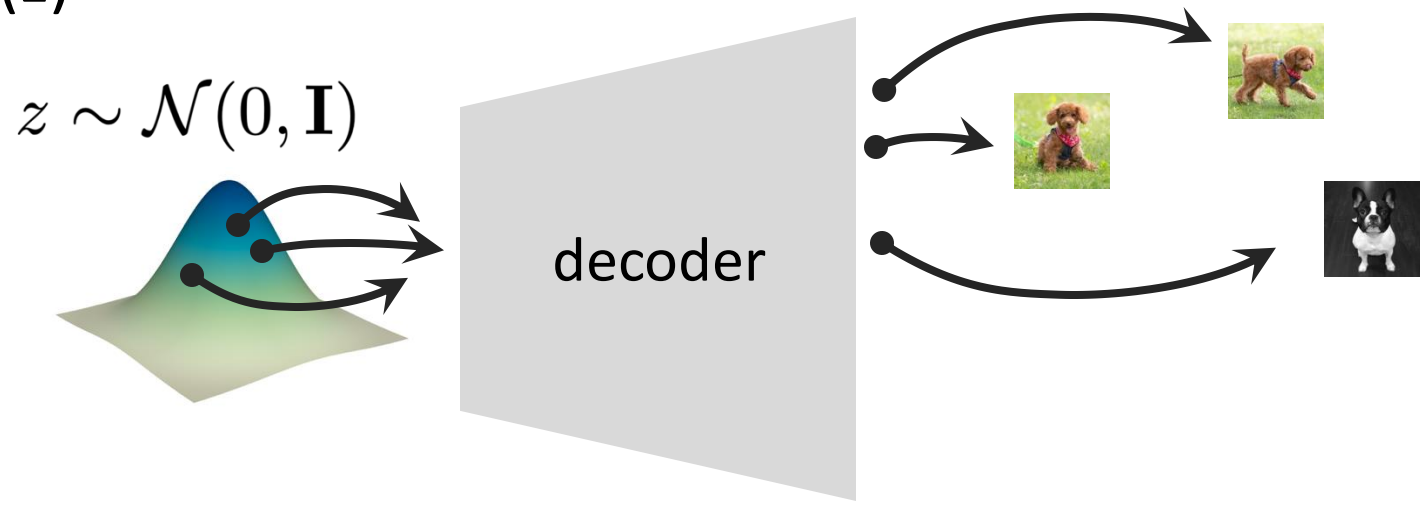
- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net: $g_\theta(z)$



Variational Autoencoder

Inference (generation):

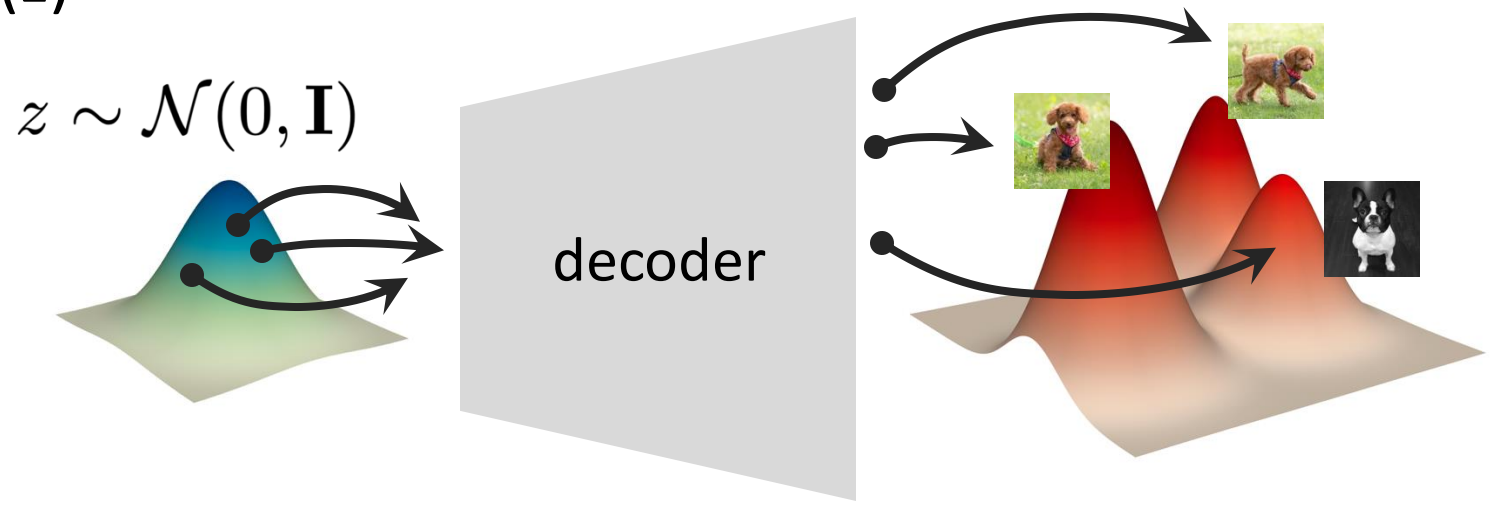
- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net: $g_\theta(z)$



Variational Autoencoder

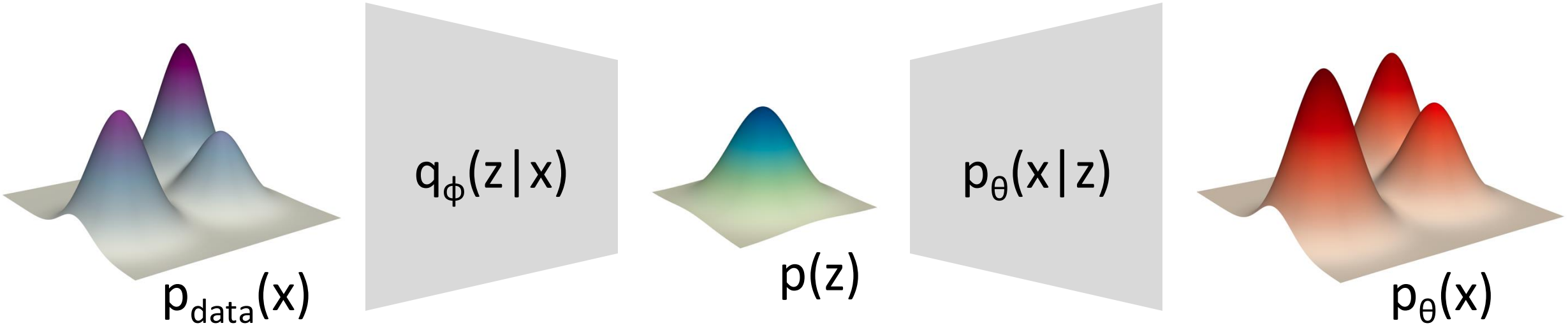
Inference (generation):

- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net: $g_\theta(z)$



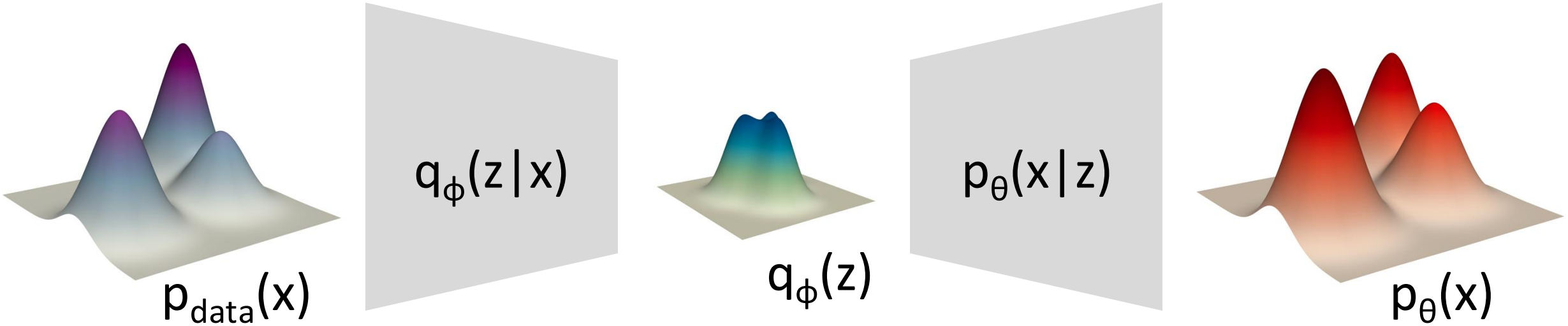
A view of “Autoencoding Distributions”

- **Encoder**: data to latent
- **Decoder**: latent to data



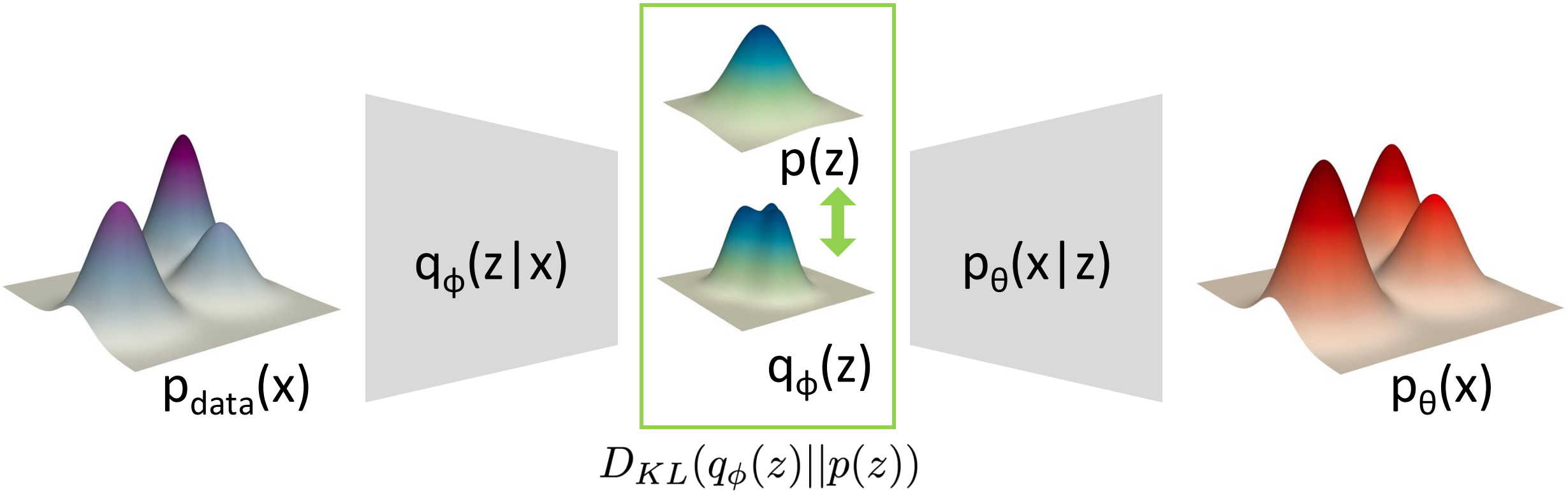
A view of “Autoencoding Distributions”

- encoded latent distribution: $q_\phi(z) = \int_x q_\phi(z|x) p_{data}(x) dx$



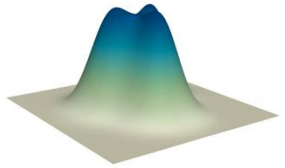
A view of “Autoencoding Distributions”

- encoded latent distribution: $q_\phi(z) = \int_x q_\phi(z|x)p_{data}(x)dx$

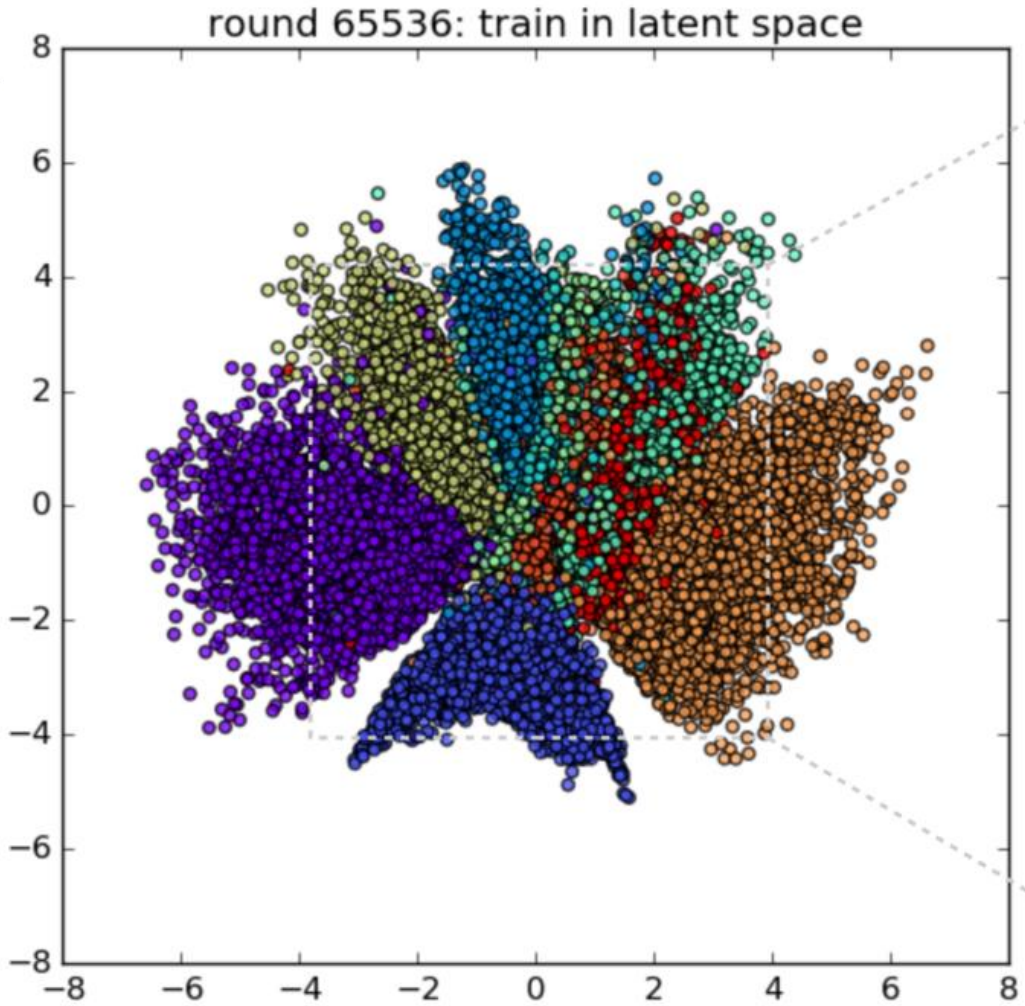


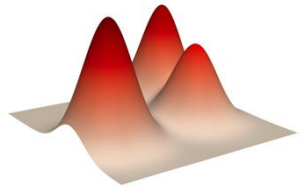
E.g., see “*InfoVAE: Information Maximizing Variational Autoencoders*”, 2017

VAE: 2D latent space on MNIST

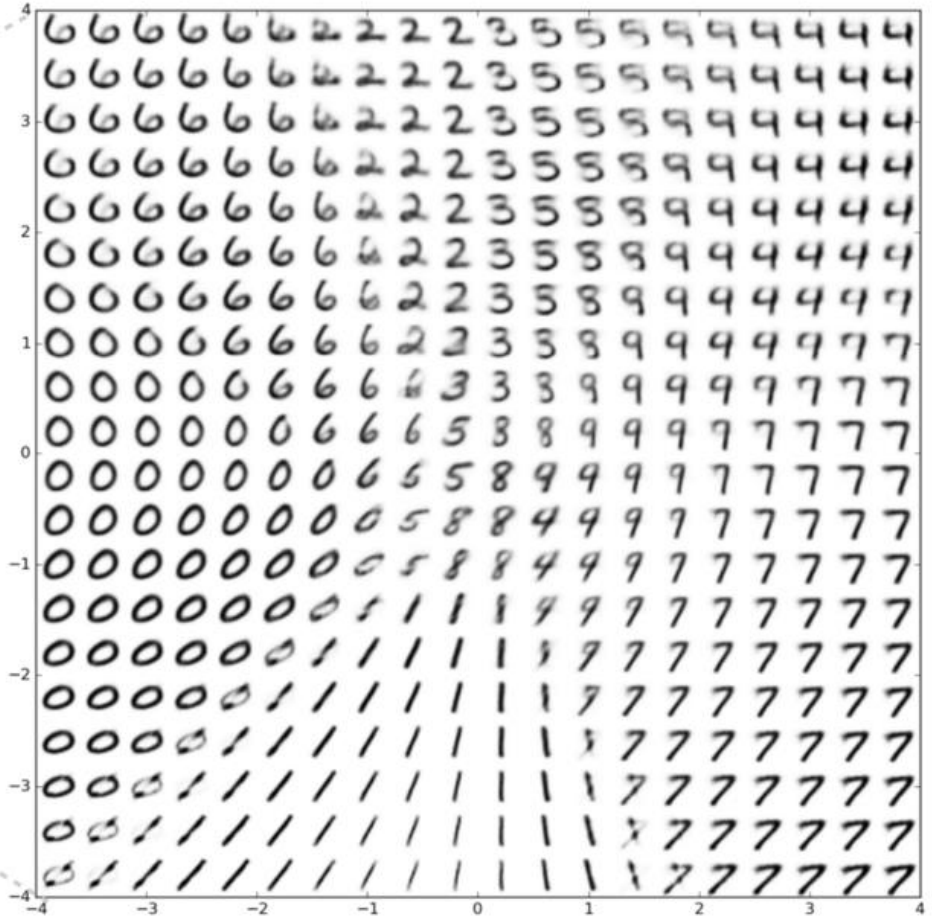


$q_\phi(z)$





$p_\theta(x)$



"Introducing Variational Autoencoders (in Prose and Code)"
https://blog.fastforwardlabs.com/2016/08/12/introducing-variational-autoencoders-in-prose-and-code.html

VAE results on 784-d MNIST data

(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Too strict to model the 784-d (28x28) joint distribution by independent distributions

Figure from: Kingma and Welling. “Auto-Encoding Variational Bayes”, ICLR 2014

VAE: 2D latent space on “Frey Face” dataset

Kingma and Welling. Auto-Encoding Variational Bayes, ICLR 2014

Issues of VAE

- Gaussian prior is too **strict** an assumption
- Reconstruction loss assumes **independence** across output elements
- Difficult to model **high-dimensional**, **complex** distributions

In practice - VAE is often used as a module (tokenizer), and the prior is learned by other models (AR or Diffusion).