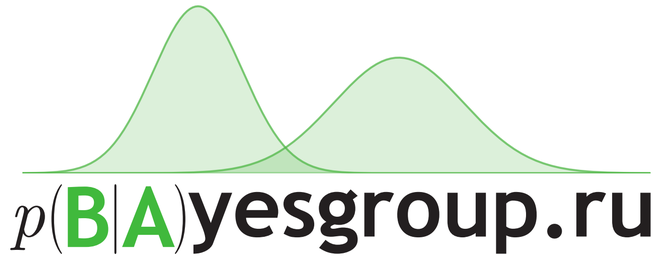
Variational autoencoders

Nadia Chirkova

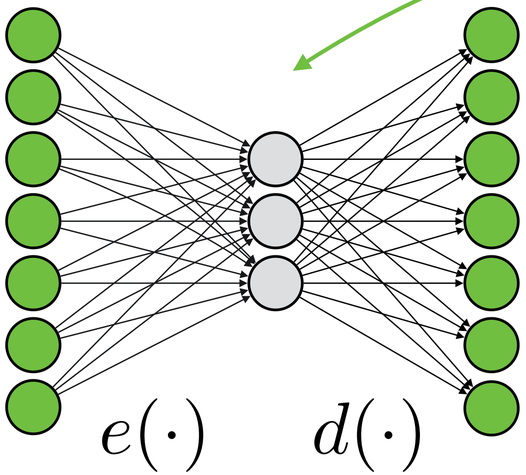
Higher School of Economics, Samsung-HSE Laboratory
Moscow, Russia

SAMSUNG
Research



Deep autoencoders: reminder

x



$e(\cdot)$ $d(\cdot)$

$d(e(x))$

low-dim representation

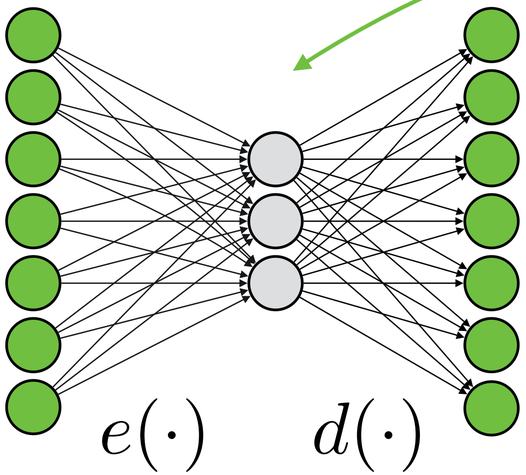
$z = e(x)$

Optimization problem:

$$\sum_{i=1}^{N} \|x_i - d(e(x_i))\|^2 \to \min_{d,e}$$

Deep autoencoders: reminder

x



$e(\cdot)$ $d(\cdot)$

$d(e(x))$

low-dim representation

$z = e(x)$

Go Bayesian?
p(xlz), p(zlx) ?

Optimization problem:

$$\sum_{i=1}^{N} \|x_i - d(e(x_i))\|^2 \to \min_{d,e}$$

Variational autoencoders: motivation

face 1
from
data

decoded faces
at linear segment
[e(face 1), e(face2)]

face 2
from
data

Xianxu Hou, Ke Sun, Linlin Shen, Guoping Qiu. mproving Variational Autoencoder with Deep Feature Consistent and Generative Adversarial Training, Neurocomputing, 2019

Latent variable models: reminder

$$X = \{x_1, \dots, x_N\}, \; x_i \in \mathbb{R}^d, \; i = 1, \dots, N$$

$$p(x|\theta) - ? \qquad p(X|\theta) \to \max_\theta$$

images

x_1

x_2

x_3

tabular data

	Age	Attrition	BusinessTravel	DailyRate	Department
x_1	41	Yes	Travel_Rarely	1102	Sales
x_2	49	No	Travel_Frequently	279	Research & Development
x_3	37	Yes	Travel_Rarely	1373	Research & Development
x_4	33	No	Travel_Frequently	1392	Research & Development

Latent variable models: reminder

$$X = \{x_1, \dots, x_N\}, \; x_i \in \mathbb{R}^d \underset{i = 1, \dots, N}{}$$

$$Z = \{z_1, \dots, z_N\}$$ — latent variables (one per object)

$$p(x, z|\theta) = p(x|z, \theta_1)p(z|\theta_2)$$

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

(or sum in case of discrete z)

$$p(x|\theta) - ? \qquad p(X|\theta) \to \max_\theta$$

images

x_1

x_2

x_3

tabular data

	Age	Attrition	BusinessTravel	DailyRate	Department
x_1	41	Yes	Travel_Rarely	1102	Sales
x_2	49	No	Travel_Frequently	279	Research & Development
x_3	37	Yes	Travel_Rarely	1373	Research & Development
x_4	33	No	Travel_Frequently	1392	Research & Development

Latent variable models: reminder

$X = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$

$Z = \{z_1, \dots, z_N\}$ — latent variables

$$p(x, z|\theta) = p(x|z, \theta_1)p(z|\theta_2)$$

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

(or sum in case of discrete z)

How can l. v. models be useful?

- Understanding data structure: $p(z|x, \theta) = \frac{p(x|z, \theta_1)p(z|\theta_2)}{\int (x|\tilde{z}, \theta_1)p(\tilde{z}|\theta_2)d\tilde{z}}$

e. g. what is an object's embedding?

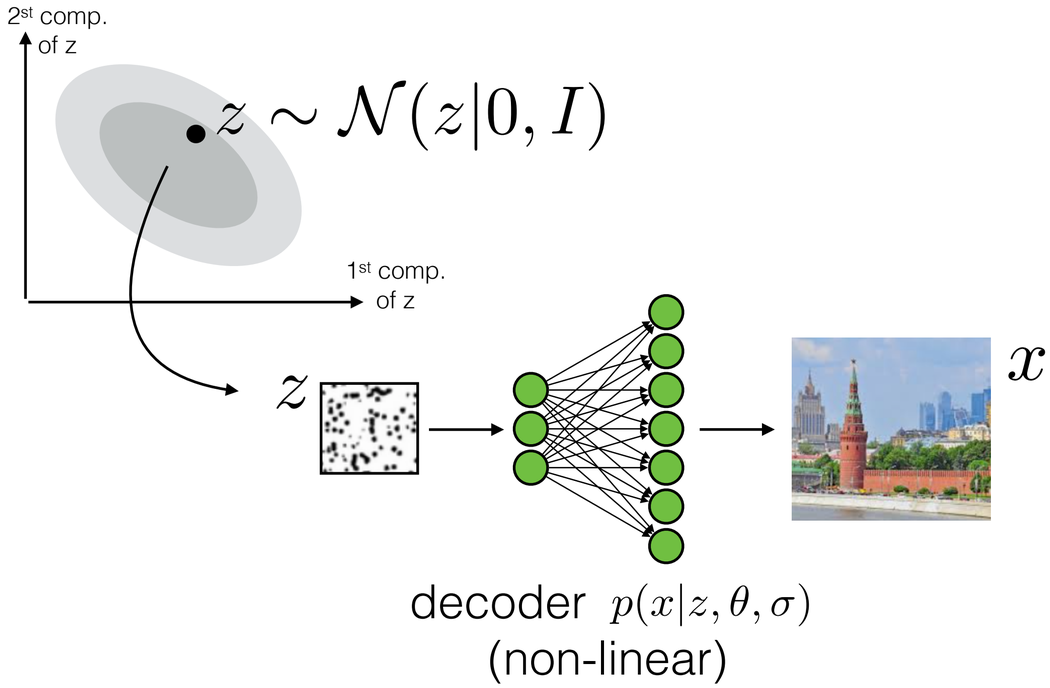
- Generating new objects: $z \sim p(z|\theta_2), \quad x \sim p(x|z, \theta_1)$

e. g. generating new images using decoder

Variational autoencoder as a generative model

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d$$
$$i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}}$$
$$i = 1, \ldots, N$$



Assume each object
is generated by the decoder
from an embedding

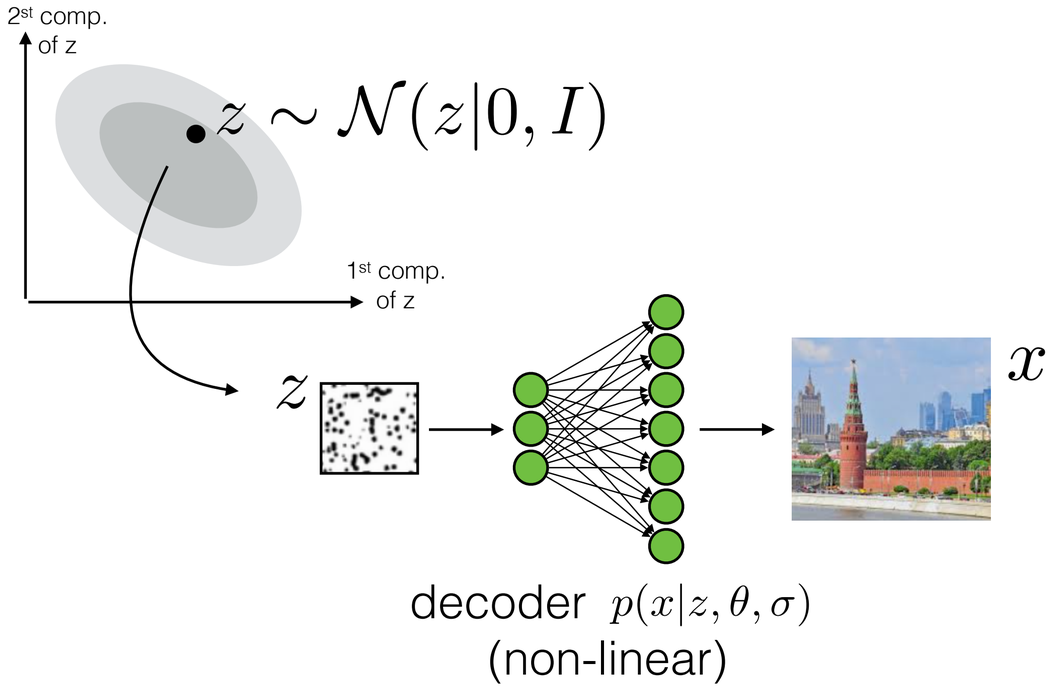
Latent variables z:

what is an object's embedding?

Variational autoencoder as a generative model

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d \quad i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}} \quad i = 1, \ldots, N$$

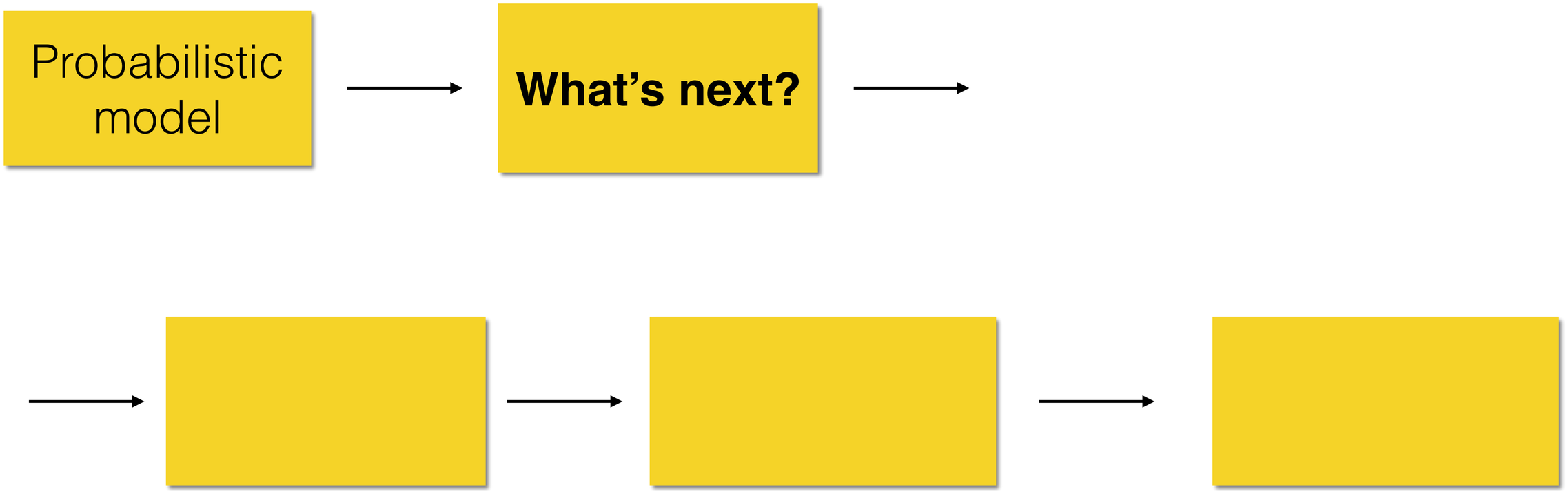


$$p(x, z|\theta, \sigma) = p(x|z, \theta, \sigma)p(z)$$

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

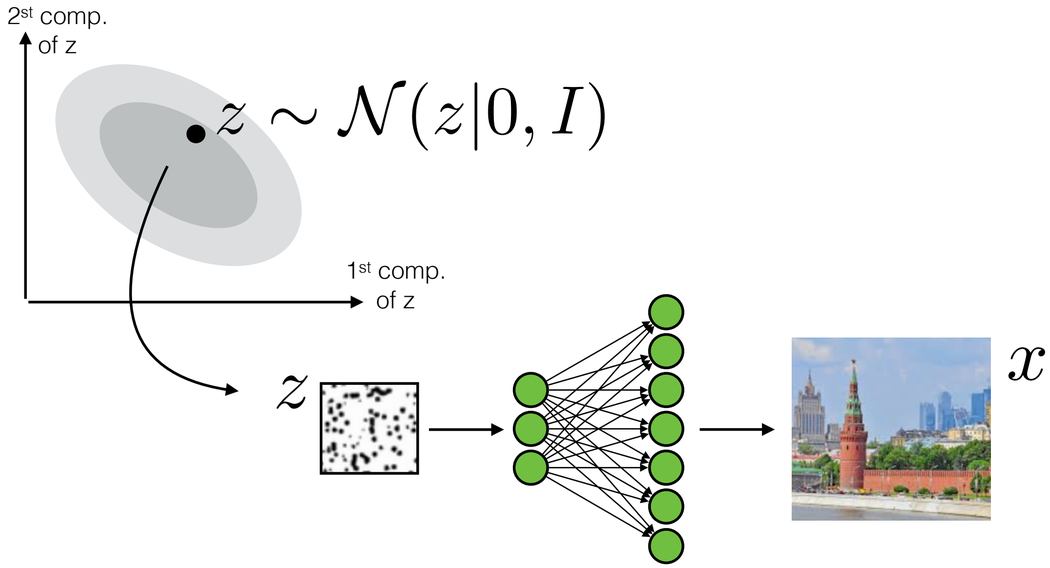
VAE: outline



Variational autoencoder (VAE)

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}^d \quad i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}} \quad i = 1, \dots, N$$



decoder $p(x|z, \theta, \sigma)$
(non-linear)

$$p(x, z|\theta, \sigma) = p(x|z, \theta, \sigma)p(z)$$

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

$$p(z|x, \theta, \sigma) - ?$$

$$\theta, \sigma - ?$$

understanding data structure

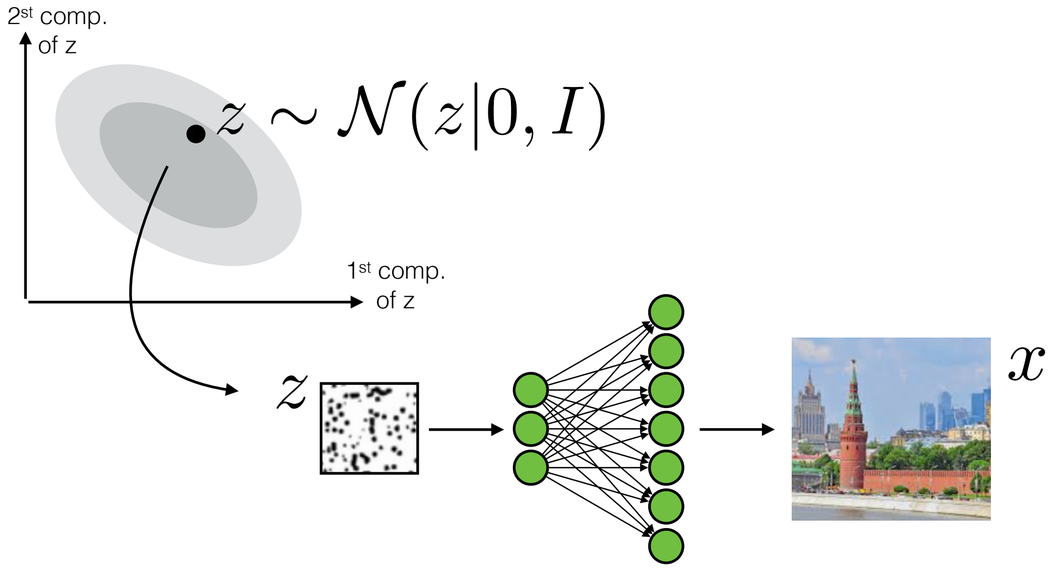
generating new objects

*** Let's omit σ for brevity**

Variational autoencoder (VAE)

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d \quad i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}} \quad i = 1, \ldots, N$$



decoder $p(x|z, \theta, \sigma)$
(non-linear)

$$p(x, z|\theta, \sigma) = p(x|z, \theta, \sigma)p(z)$$

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

$$p(X|\theta) \to \max_\theta$$

$$p(X|\theta) = \int p(X|Z, \theta)p(Z)dZ$$

*** Let's omit σ for brevity**

Variational lower bound: reminder

$$\log p(X|\theta) = \mathcal{L}(q(Z), \theta) + KL(q(Z)||p(Z|X, \theta))$$

(holds for any $q(Z)$)

data likelihood

variational lower bound

Variational lower bound: reminder

$$\log p(X|\theta) = \mathcal{L}(q(Z), \theta) + KL(q(Z)||p(Z|X, \theta))$$

(holds for any $q(Z)$)

data likelihood — **variational lower bound** — $\geqslant 0$

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \rightarrow \max_{q,\theta}$$

data likelihood — **variational lower bound**

Variational lower bound: reminder

$$\log p(X|\theta) = \mathcal{L}(q(Z), \theta) + KL(q(Z)||p(Z|X, \theta))$$

(holds for any $q(Z)$)

data likelihood

variational lower bound

$$\geqslant 0$$

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \rightarrow \max_{q,\theta}$$

data likelihood we wish to optimize

variational lower bound we will optimize

Variational lower bound (ELBO) for VAE

$$\mathcal{L}(q(Z), \theta) \rightarrow \max_{q, \theta}$$

variational lower bound

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i)} \log p(x_i | z_i, \theta) - KL(q(z_i) || p(z_i)) \right)$$

variational lower bound

reconstructon term

regularizer

Variational lower bound (ELBO) for VAE

$$\mathcal{L}(q(Z), \theta) \to \max_{q, \theta}$$

variational lower bound

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i)} \log p(x_i|z_i, \theta) - KL(q(z_i)||p(z_i)) \Big)$$

variational lower bound

reconstruction term

regularizer

Task: $\log p(x_i|z_i, \theta) - ?$

Variational lower bound (ELBO) for VAE

$$\mathcal{L}(q(Z), \theta) \to \max_{q, \theta}$$

variational lower bound

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i)} \log p(x_i|z_i, \theta) - KL(q(z_i)||p(z_i)) \right)$$

variational lower bound

reconstructon term

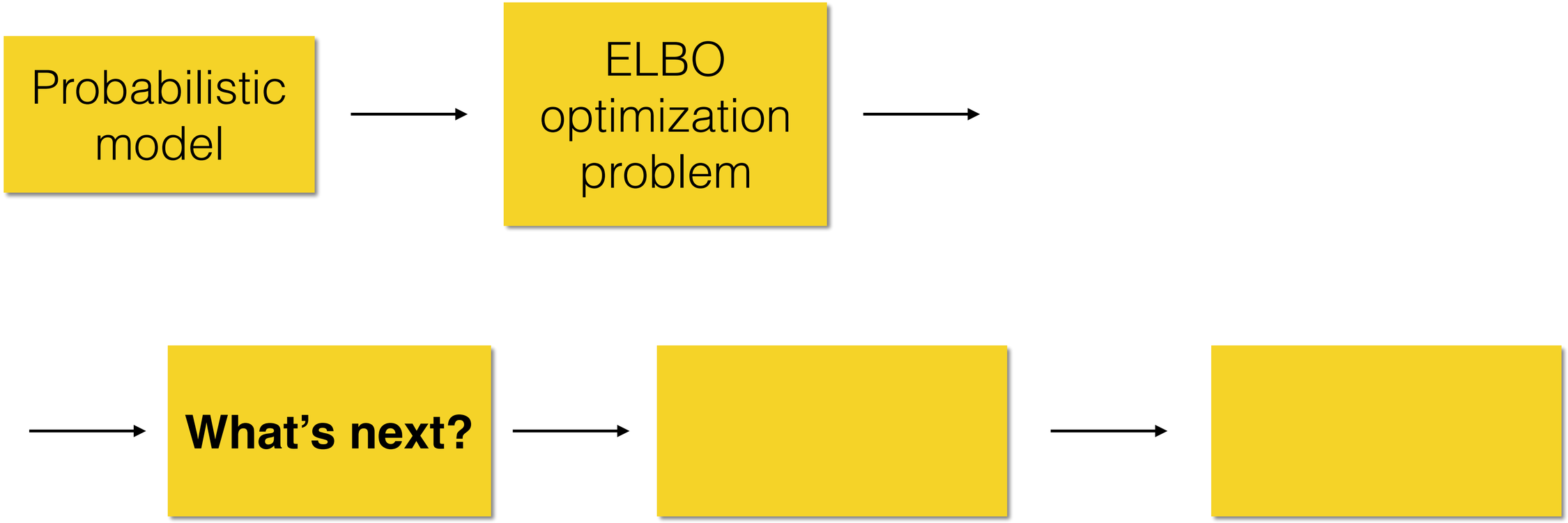
regularizer

$$\log p(x_i|z_i, \theta, \sigma) = \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2\sigma^2} \|x_i - d_\theta(z_i)\|^2 \right) \right) =$$

reconstructon term

$$\mathrm{C} - \log \sigma - \frac{1}{2\sigma^2} \|x_i - d_\theta(z_i)\|^2$$

VAE: outline


Probabilistic model
ELBO optimization problem
What's next?

Optimizing variational lower bound

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i)} \log p(x_i | z_i, \theta) - KL(q(z_i) || p(z_i)) \right) \rightarrow \max_{q, \theta}$$

?

EM-algorithm?

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i)} \log p(x_i|z_i, \theta) - KL(q(z_i)||p(z_i)) \Big) \to \max_{q,\theta}$$

E-step:
(expectation)

$$\mathcal{L}(q(Z), \theta) \to \max_{q} \quad \Leftrightarrow \quad KL(q(Z)||p(Z|X, \theta)) \to \min_{q}$$

$$q(Z) = p(Z|X, \theta)$$

M-step:
(maximization)

$$\mathcal{L}(q(Z), \theta) \to \max_{\theta}$$

EM-algorithm?

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i)} \log p(x_i|z_i, \theta) - KL(q(z_i)||p(z_i)) \Big) \to \max_{q,\theta}$$

E-step:
(expectation)

$$\mathcal{L}(q(Z), \theta) \to \max_{q} \quad \Leftrightarrow \quad KL(q(Z)||p(Z|X, \theta)) \to \min_{q}$$

$$q(Z) = p(Z|X, \theta)$$ — **intractable!**

M-step:
(maximization)

$$\mathcal{L}(q(Z), \theta) \to \max_{\theta}$$

Optimizing variational lower bound

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i)} \log p(x_i | z_i, \theta) - KL(q(z_i) || p(z_i)) \right) \rightarrow \max_{q, \theta}$$

?

Optimizing variational lower bound

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i)} \log p(x_i|z_i, \theta) - KL(q(z_i)||p(z_i)) \Big) \rightarrow \max_{q,\theta}$$

?

Optimizing w.r.t. a distribution?

Let's choose q in some parametric family: $q(z|x, \phi)$

Example: $q(z|x, \phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$

Optimizing variational lower bound

$$\mathcal{L}(q(Z), \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i)} \log p(x_i|z_i, \theta) - KL(q(z_i)||p(z_i)) \Big) \to \max_{q,\theta}$$

?

Optimizing w.r.t. a distribution?

Let's choose q in some parametric family: $q(z|x, \phi)$

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \Big) \to \max_{\theta,\phi}$$

Optimizing w.r.t. parameters!

VAE: model and optimization problem

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \Big) \to \max_{\theta,\phi}$$

Model:

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

Approximate posterior:

$$q(z|x, \phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$

Approximate posterior in VAE

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \right) \to \max_{\theta,\phi}$$

Model:

$$p(x|z, \theta, \sigma) = \mathcal{N}(x|d_\theta(z), \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

Approximate posterior:

$$q(z|x, \phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$

$$q(z_i|x_i, \phi) \approx p(z_i|x_i, \theta)$$

because with fixed θ

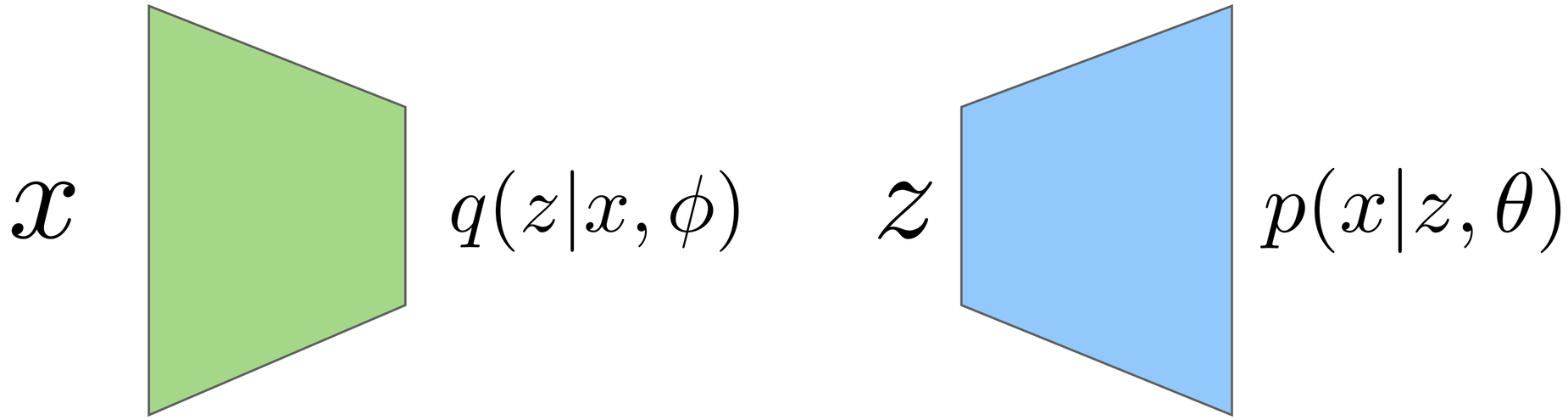
$$\mathcal{L}(q(Z), \theta) \to \max_q$$

$$\Updownarrow$$

$$KL(q(Z)||p(Z|X, \theta)) \to \min_q$$

VAE: architecture

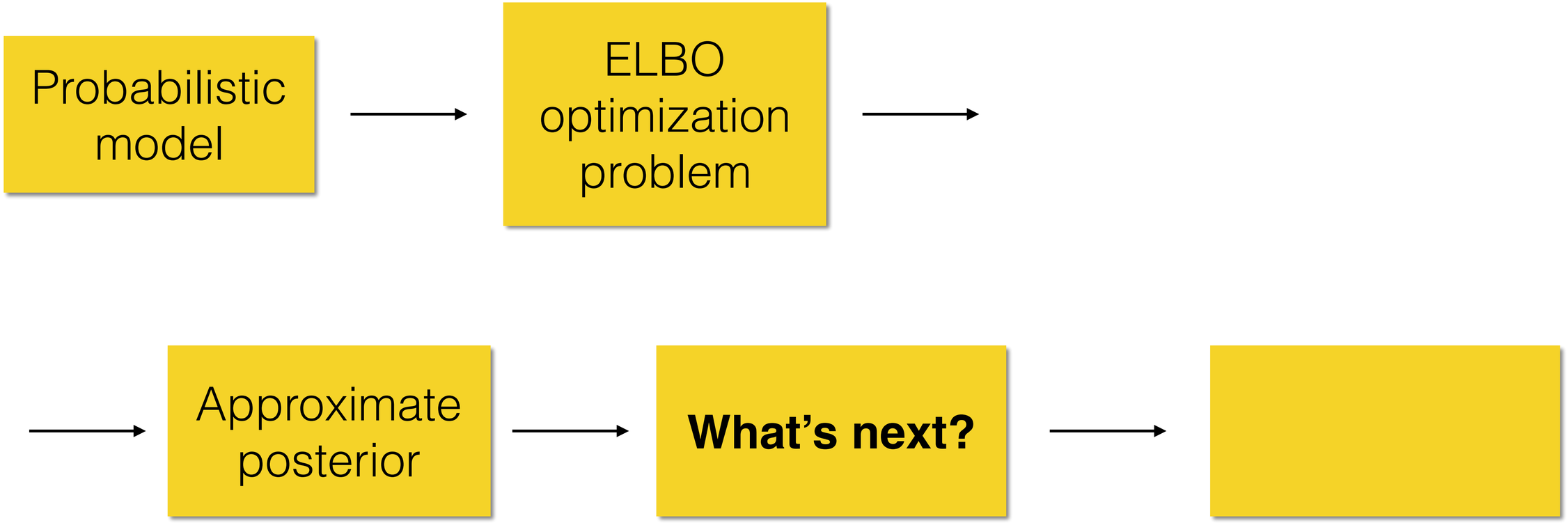
$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \right) \to \max_{\theta,\phi}$$



Encoder
$e_\phi(x)$

Decoder
$d_\theta(z)$

VAE: outline


Probabilistic model
ELBO optimization problem
Approximate posterior
What's next?

Optimizing variational lower bound (continued)

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \right) \to \max_{\theta,\phi}$$

?

?

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$

$p(z)) = \mathcal{N}(z|0, I)$

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$
$$p(z)) = \mathcal{N}(z|0, I) \qquad z \in \mathbb{R}^d$$

$$KL(q(z|x,\phi)||p(z)) = \underbrace{\int q(z|x,\phi) \log \frac{q(z|x,\phi)}{p(z)} dz}_{\text{expectation}}$$

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$
$$p(z)) = \mathcal{N}(z|0, I) \qquad z \in \mathbb{R}^d$$

$$KL(q(z|x,\phi)||p(z)) = \int q(z|x,\phi) \log \frac{q(z|x,\phi)}{p(z)} dz \quad =$$

$$= \mathbb{E}_{q(z|x,\phi)} \big(\log q(z|x,\phi) - \log p(z)\big) = \quad ?$$

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$
$$p(z)) = \mathcal{N}(z|0, I) \qquad z \in \mathbb{R}^d$$

$$KL(q(z|x,\phi)||p(z)) = \int q(z|x,\phi) \log \frac{q(z|x,\phi)}{p(z)} dz \quad =$$

$$= \mathbb{E}_{q(z|x,\phi)} \big(\log q(z|x,\phi) - \log p(z)\big) =$$

$$= -\cancel{\log \sqrt{2\pi}} - \log \sigma - \frac{1}{2\sigma^2} \mathbb{E}_q \|z - e_\phi(x)\|^2 + \cancel{\log \sqrt{2\pi}} + \frac{1}{2} \mathbb{E}_q \|z\|^2$$

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$
$$p(z)) = \mathcal{N}(z|0, I) \qquad z \in \mathbb{R}^d$$

$$KL(q(z|x,\phi)||p(z)) = \int q(z|x,\phi) \log \frac{q(z|x,\phi)}{p(z)} dz \quad =$$

$$= \mathbb{E}_{q(z|x,\phi)} \big(\log q(z|x,\phi) - \log p(z)\big) =$$

$$= -\log \sqrt{2\pi} - \log \sigma - \frac{1}{2\sigma^2} \mathbb{E}_q \|z - e_\phi(x)\|^2 + \log \sqrt{2\pi} + \frac{1}{2} \mathbb{E}_q \|z\|^2$$

$\|0\|^2 + \sigma^2 d$ $\qquad$ $\|e_\phi(x)\|^2 + \sigma^2 d$

Moments of normal distribution

$$p(z) = \mathcal{N}(z|\mu, \sigma^2 I), \quad z \in \mathbb{R}^d \qquad \Leftrightarrow \quad p(z_j) = \mathcal{N}(z_j|\mu_j, \sigma) \underset{j=1,\dots,d}{}$$

$$\mathbb{E}\|z\|^2 = \sum_{j=1}^{d} \mathbb{E} z_j^2 = \sum_{j=1}^{d} (\mu_j^2 + \sigma^2) = \|\mu\|^2 + d\sigma^2$$

$$p(z-\mu) = \mathcal{N}(z-\mu|0, \sigma^2 I) \quad \Rightarrow \quad \mathbb{E}\|z-\mu\|^2 = d\sigma^2$$

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$
$$p(z)) = \mathcal{N}(z|0, I) \qquad z \in \mathbb{R}^d$$

$$KL(q(z|x,\phi)||p(z)) = \int q(z|x,\phi) \log \frac{q(z|x,\phi)}{p(z)} dz \quad =$$

$$= \mathbb{E}_{q(z|x,\phi)} \big(\log q(z|x,\phi) - \log p(z)\big) =$$

$$= -\log \sqrt{2\pi} - \log \sigma - \frac{1}{2\sigma^2} \underbrace{\mathbb{E}_q \|z - e_\phi(x)\|^2}_{\|0\|^2 + \sigma^2 d} + \log \sqrt{2\pi} + \frac{1}{2} \underbrace{\mathbb{E}_q \|z\|^2}_{\|e_\phi(x)\|^2 + \sigma^2 d}$$

KL-divergence between normal distributions

$KL(q(z|x,\phi)||p(z))-?$

$$q(z|x,\phi) = \mathcal{N}(z|e_\phi(x), \sigma^2 I)$$
$$p(z)) = \mathcal{N}(z|0, I) \qquad z \in \mathbb{R}^d$$

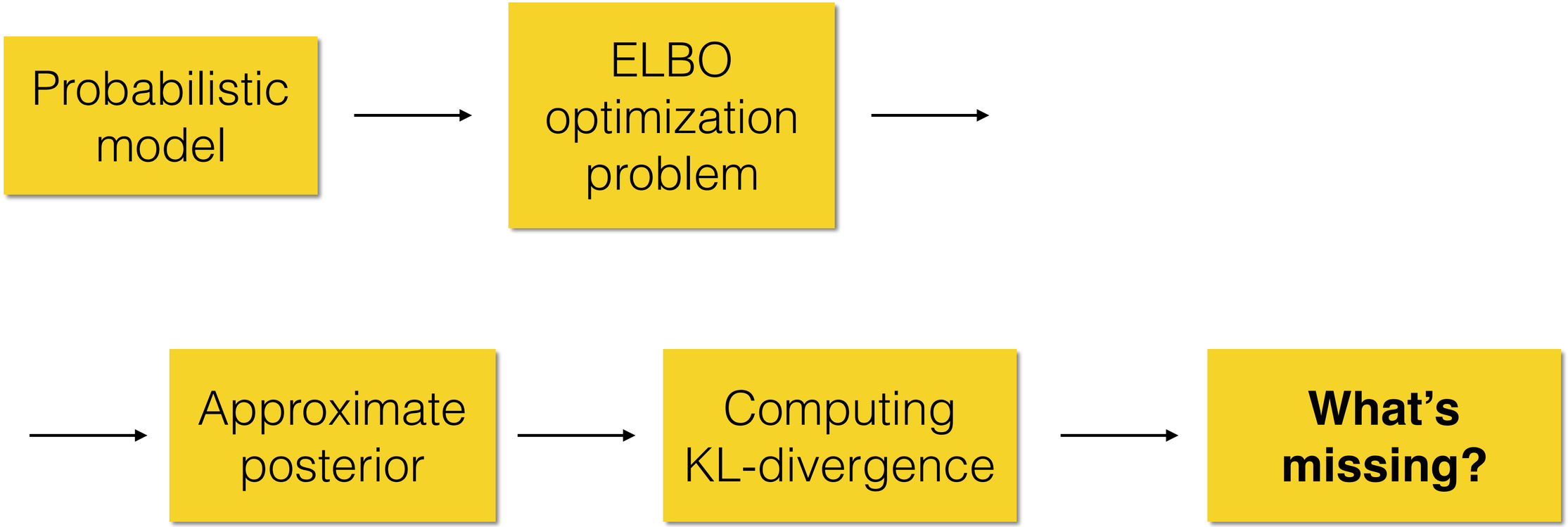
$$KL(q(z|x,\phi)||p(z)) = \int q(z|x,\phi) \log \frac{q(z|x,\phi)}{p(z)} dz \quad =$$

$$= \mathbb{E}_{q(z|x,\phi)} \big(\log q(z|x,\phi) - \log p(z)\big) =$$

$$= -\cancel{\log \sqrt{2\pi}} - \log \sigma - \frac{1}{2\sigma^2} \mathbb{E}_q \|z - e_\phi(x)\|^2 + \cancel{\log \sqrt{2\pi}} + \frac{1}{2} \mathbb{E}_q \|z\|^2$$

$$= -\log \sigma - \frac{1}{2} d + \frac{1}{2} \|e_\phi(x)\|^2 + \frac{1}{2} \sigma^2 d$$

VAE: outline


Probabilistic model
ELBO optimization problem
Approximate posterior
Computing KL-divergence
What's missing?

Optimizing variational lower bound (continued)

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \right) \to \max_{\theta,\phi}$$

?

Doubly stochastic variational inference for VAE

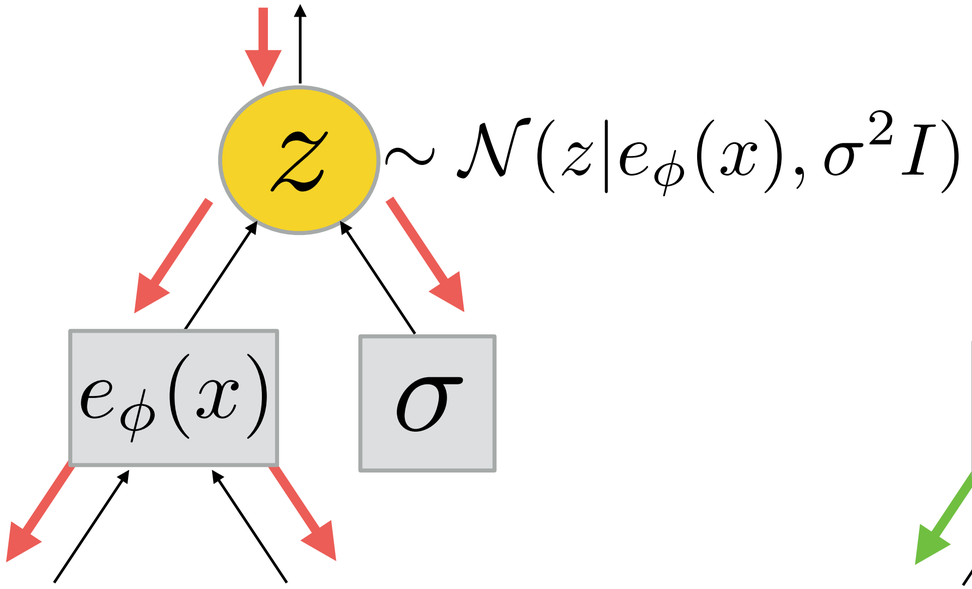
$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i,\theta) - KL(q(z_i|x_i,\phi)||p(z_i)) \Big) \to \max_{\theta,\phi}$$

sample mini-batch
of the training objects

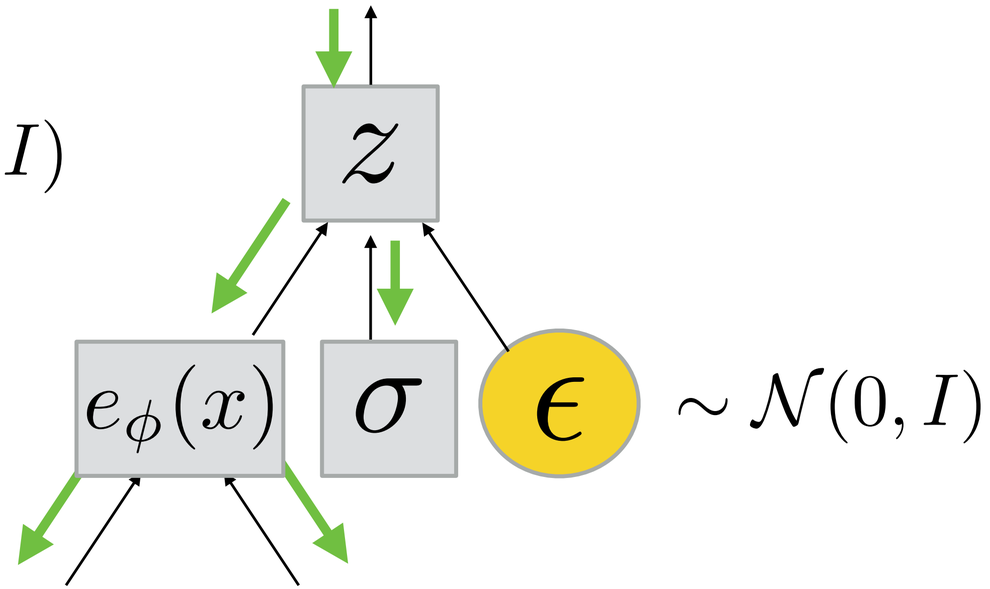
sample z
from approx. posterior
using
reparametrization trick

Reparametrization trick: reminder

$$z \sim \mathcal{N}(z|e_\phi(x), \sigma^2 I) \quad \Leftrightarrow \quad z = e_\phi(x) + \sigma\epsilon, \;\; \epsilon \sim \mathcal{N}(\epsilon|0, I)$$



Gradients propagate through randomness

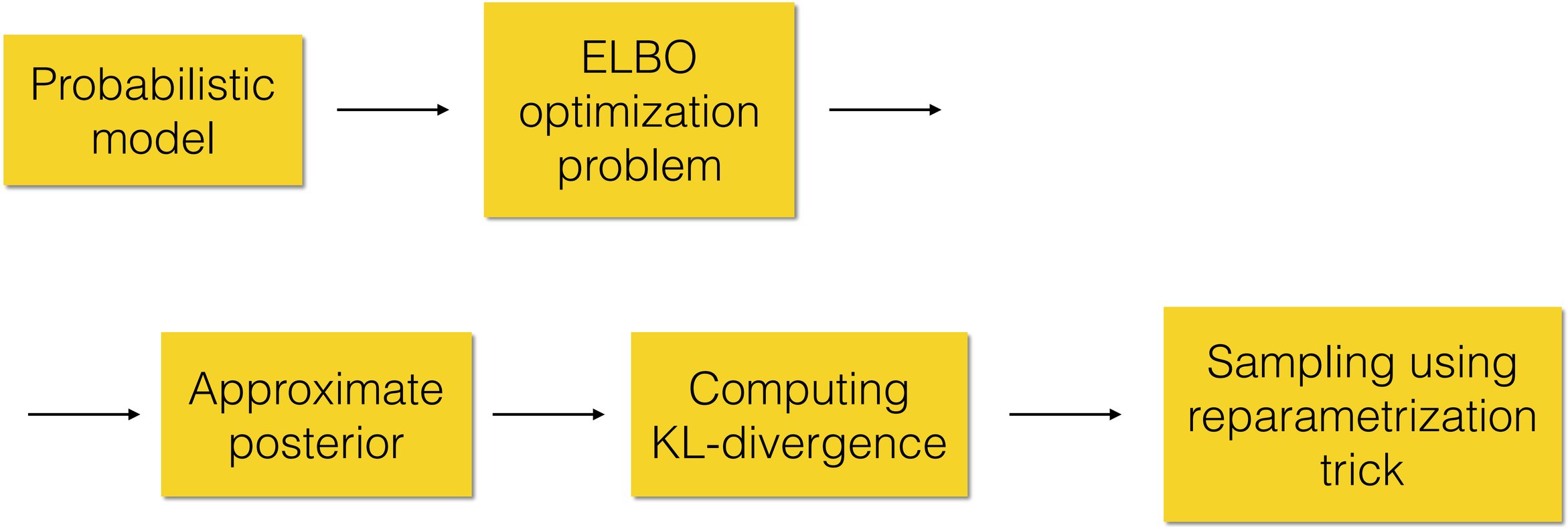


Gradients propagate only through deterministic nodes

Diederik P Kingma, Max Welling . Auto-Encoding Variational Bayes. ICLR 2013
Image inspired by https://jaan.io/what-is-variational-autoencoder-vae-tutorial/

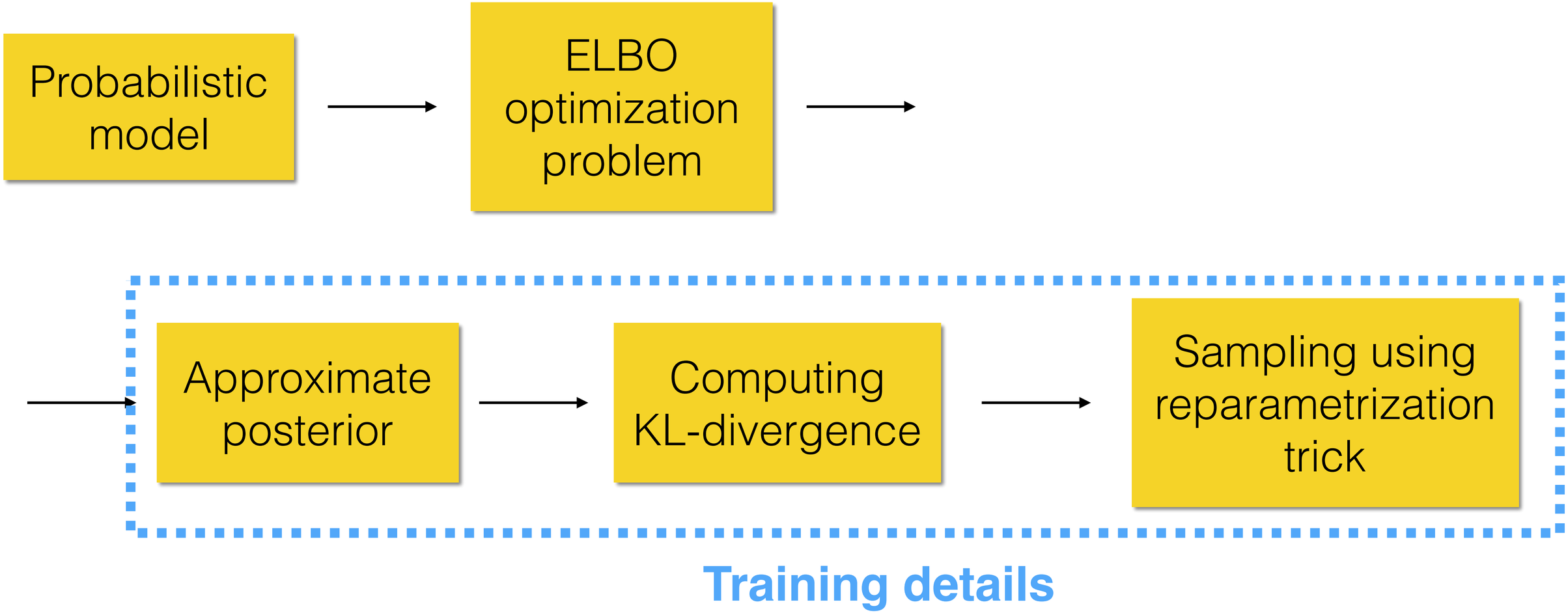
VAE: outline

Gathered them all!



VAE: outline

Gathered them all!

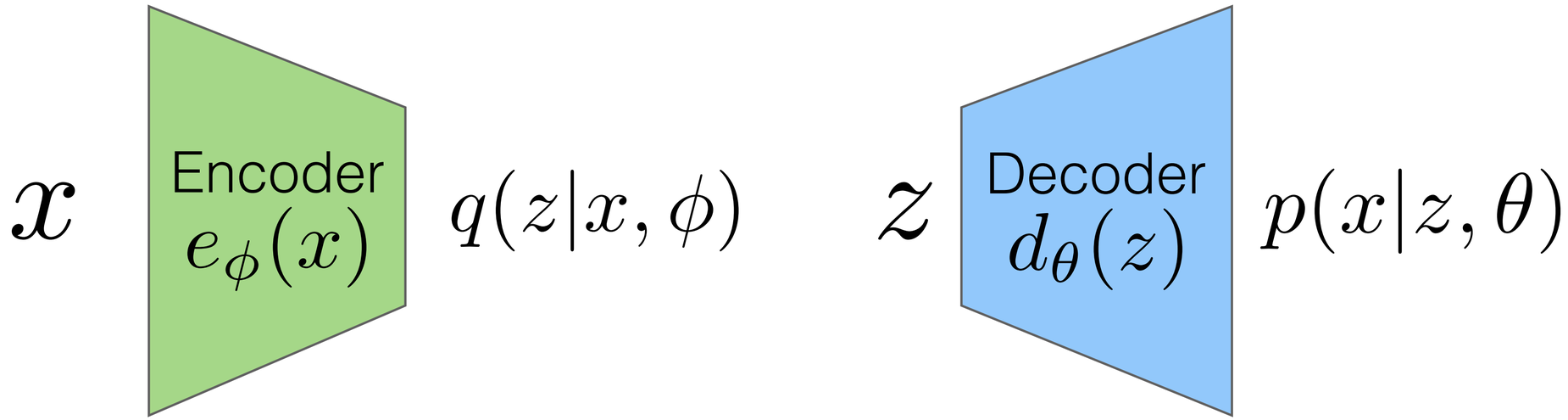


VAE: loss and architecture

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \Big) \to \max_{\theta,\phi}$$

reconstructon term

regularizer



VAE: loss and architecture (with normal distributions)

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \frac{1}{2\sigma^2} \left\| d_\theta\big(e_\phi(x_i) + \sigma\epsilon\big) - x_i \right\|^2 + \frac{1}{2} \left\| e_\phi(x_i) \right\|^2 + \frac{1}{2}\sigma^2 d \right) \to \min_{\theta, \phi, \sigma}$$

noisy embedding

reconstructon term

embedding regularizer

x

Encoder $e_\phi(x)$

$+\sigma\epsilon$

$\epsilon \sim \mathcal{N}(0, I)$

z

Decoder $d_\theta(z)$

similar to x ?

VAE: training algorithm

Input: mini-batch $\tilde{X} \in \mathbb{R}^{m \times D}$

Result: updating weights σ, θ and ϕ

1. Encode mini-batch: $e_\phi(\tilde{X}) \in \mathbb{R}^{m \times d}$
2. Sample noise: $\epsilon_i \sim \mathcal{N}(\epsilon|0, I_d)\,, \quad E \in \mathbb{R}^{m \times d}$
3. Compute (noisy) embeddings: $\tilde{Z} = e_\phi(\tilde{X}) + \sigma E$
4. Compute KL-divergence regularizer: $\mathcal{L}_{\mathrm{KL}} = -m \log \sigma + \frac{1}{2}\|e_\phi(\tilde{X})\|^2 + \frac{1}{2}\sigma^2 md$
5. Compute reconstruction term $\mathcal{L}_{\mathrm{data}} = -m \log \sigma - \frac{m}{2\sigma^2}\|\tilde{X} - d_\theta(\tilde{Z})\|^2$
6. Compute gradients of $\mathcal{L}_{\mathrm{data}} - \mathcal{L}_{\mathrm{KL}}$ w. r. t. σ, θ and ϕ and perform gradient *ascent* step

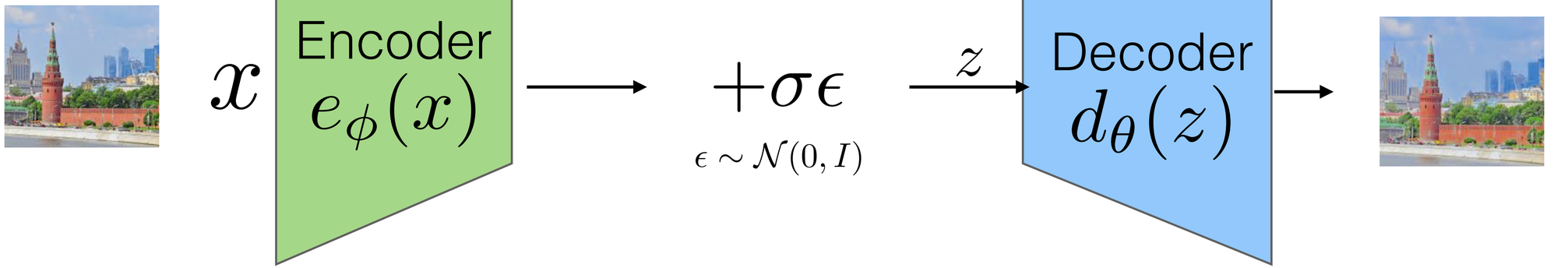
VAE: loss and architecture (with normal distributions)

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \frac{1}{2\sigma^2} \left\| d_\theta \left(e_\phi(x_i) + \sigma\epsilon \right) \right\|^2 + \frac{1}{2} \left\| e_\phi(x_i) \right\|^2 + \frac{1}{2}\sigma^2 d \right) \to \max_{\theta, \phi, \sigma}$$

noisy embedding

reconstructon term

embedding regularizer



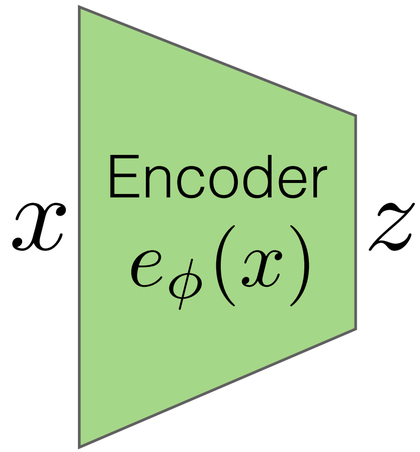
VAE: testing stage

Sampling / computing pdf during training,
Using mean values during testing

Embedding objects:

$$z = e_\phi(x)$$

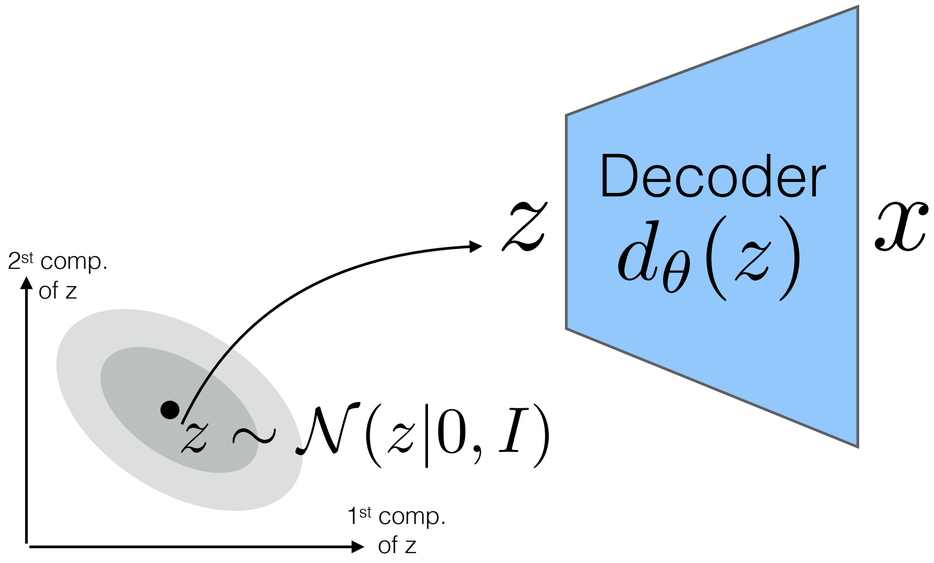
Only encoder needed!



Generating new objects:

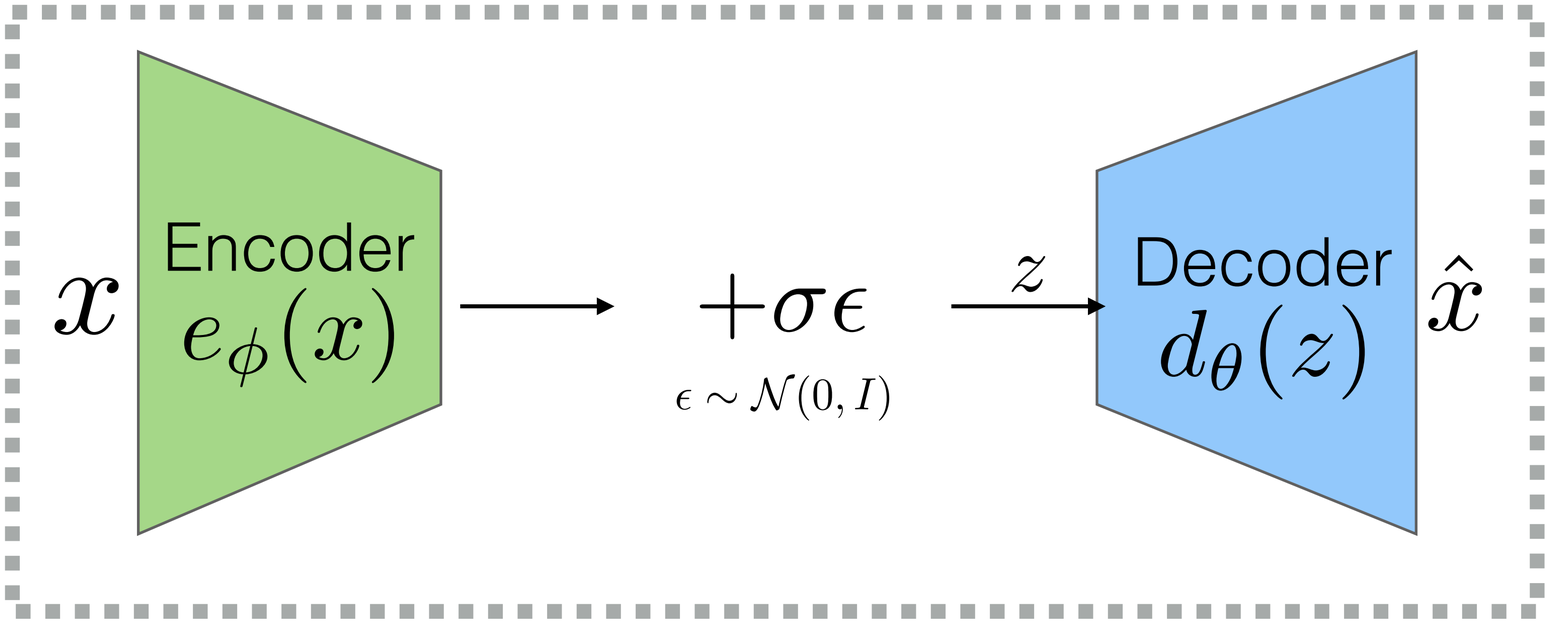
1. $z \sim \mathcal{N}(z|0, I)$
2. $x = d_\theta(z)$

Only decoder needed!

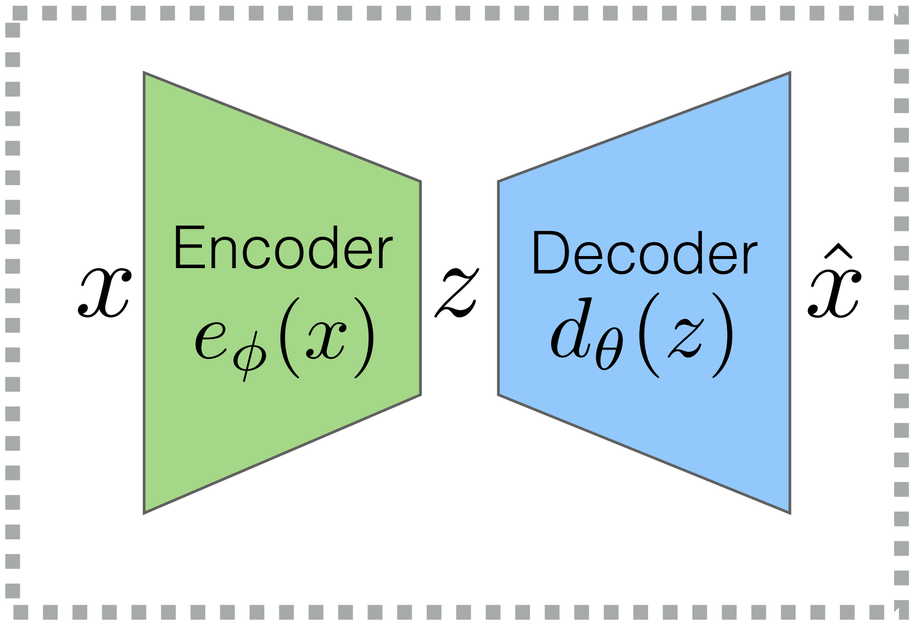


VAE vs AE

Variational autoencoder

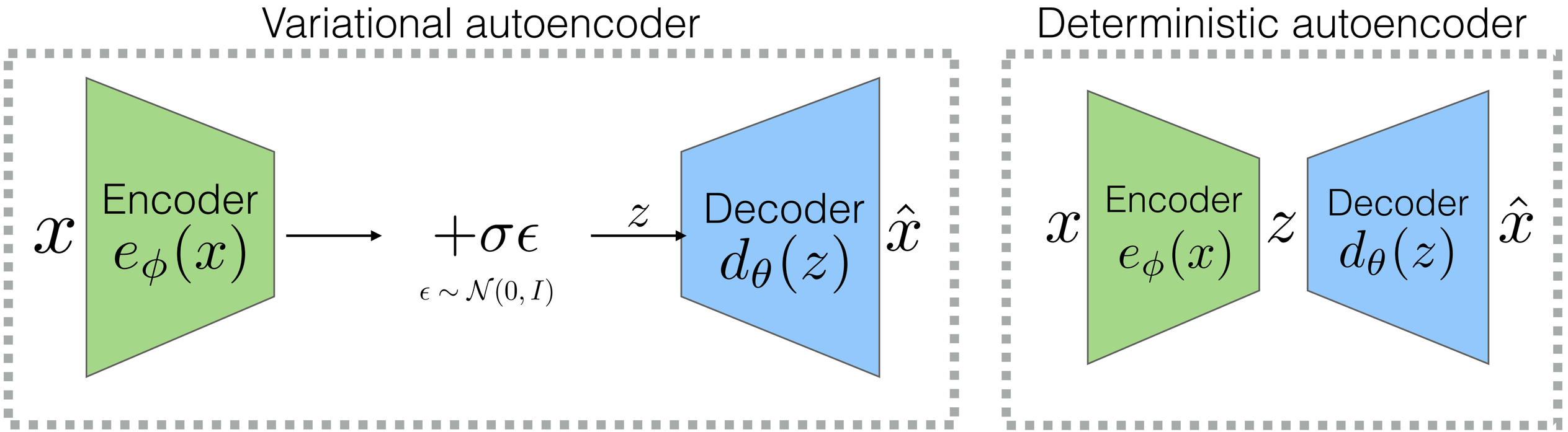


Deterministic autoencoder



- KL-term regularizes embeddings z
- Decoder trains to reconstruct x from noisy z
- Smooth embedding space

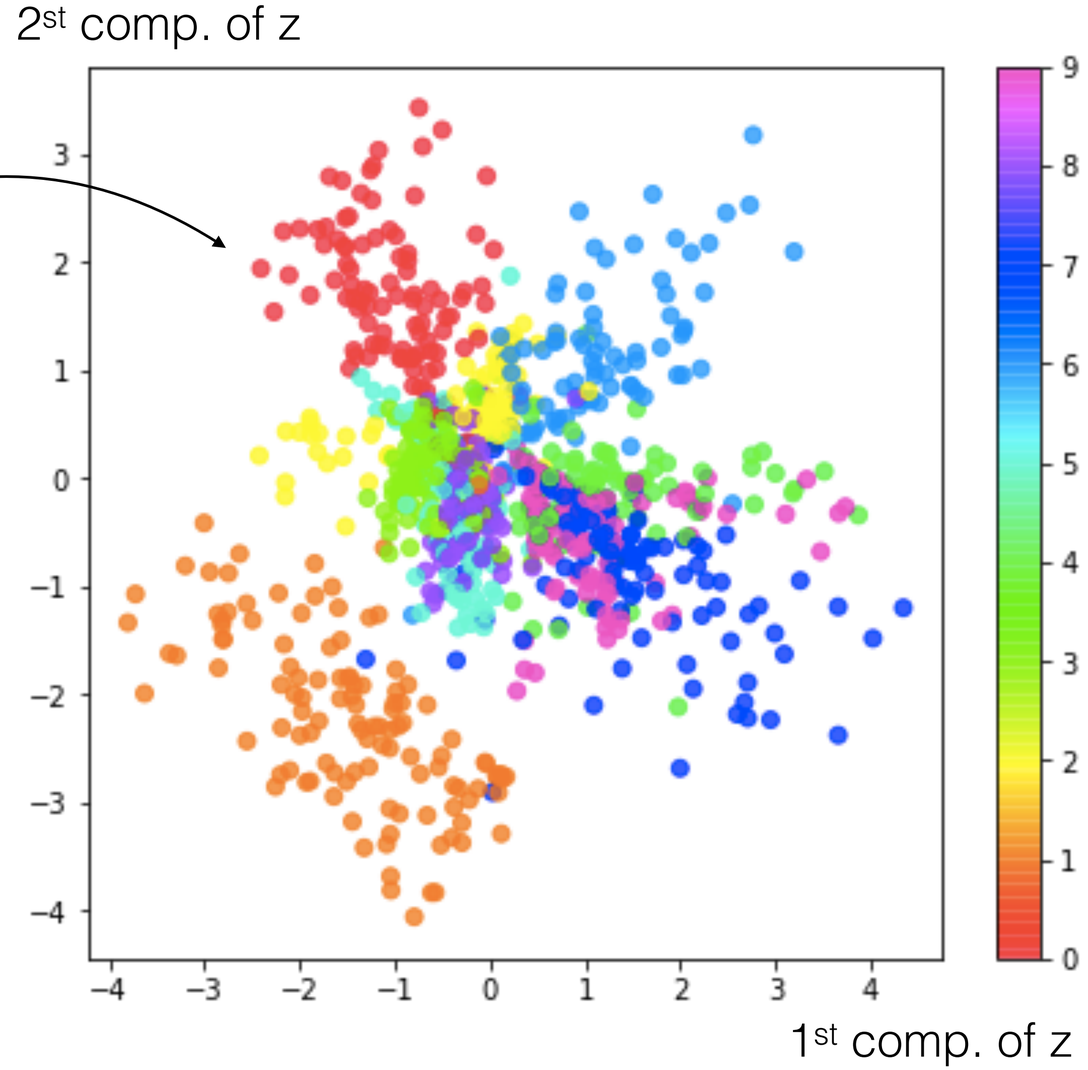
VAE vs AE



What can be VAE used for?

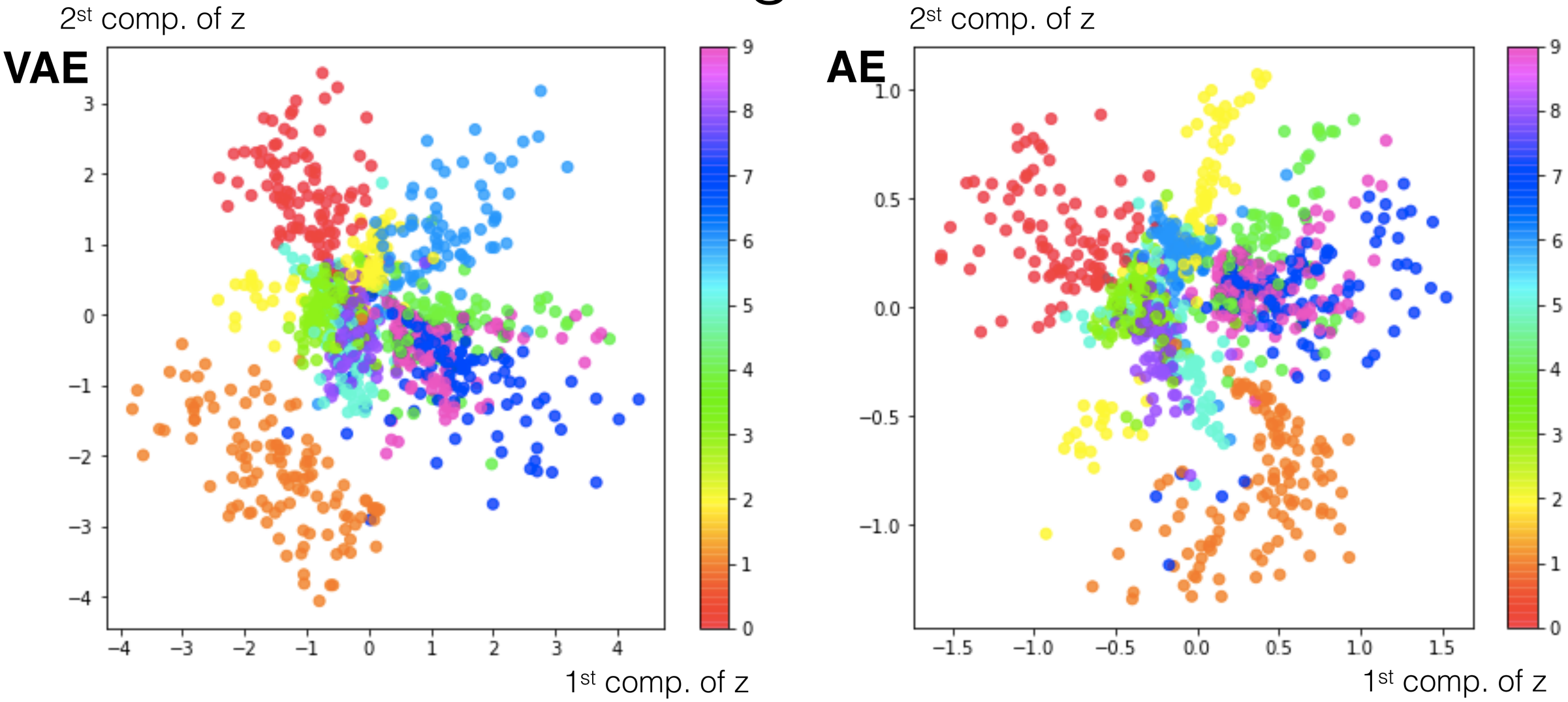
VAE: embedding visualization

$e_\phi(x_i)$



Images from Oleg Ivanov's assignment, github.com/nadiinchi/dl_labs

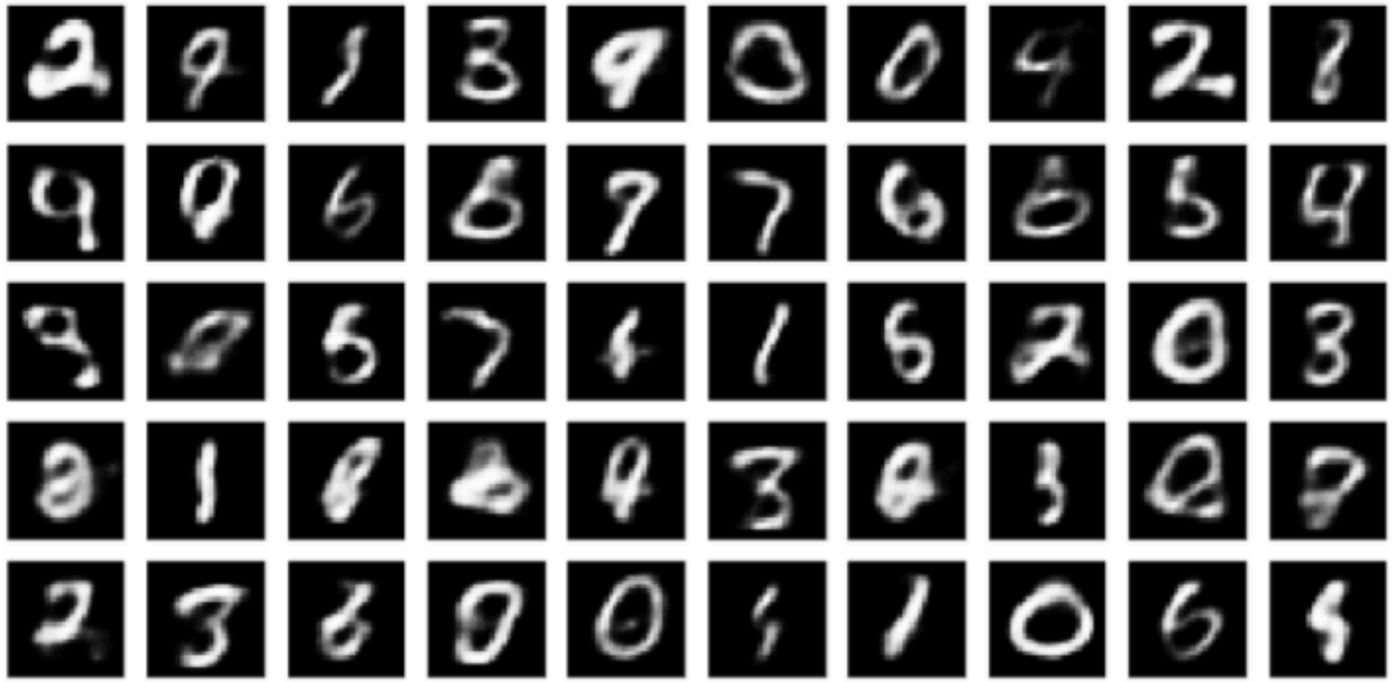
VAE vs AE: embedding visualization



Images from Oleg Ivanov's assignment, github.com/nadiinchi/dl_labs

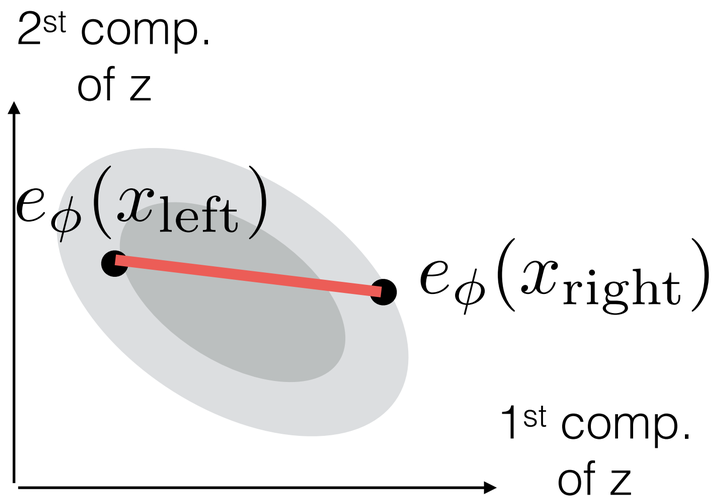
VAE: samples visualization

1. $z \sim \mathcal{N}(z|0, I)$
2. $x = d_\theta(z)$



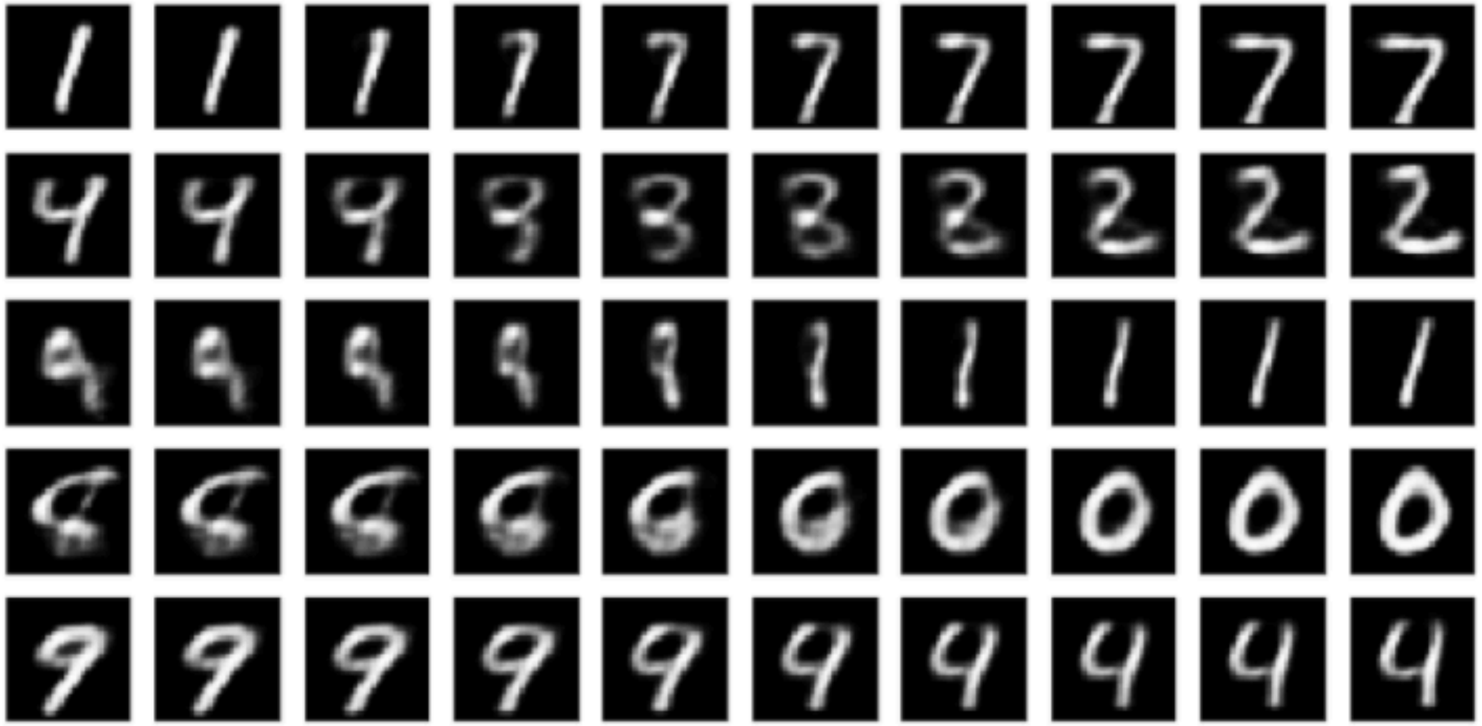
Images from Kirill Struminsky's assignment, deepbayes.ru/2019

VAE: embedding interpolation



Images:

$$x = d_\theta(z)$$



segment connecting $e_\phi(x_{\text{left}})$ and $e_\phi(x_{\text{right}})$

Images from Kirill Struminsky's assignment, deepbayes.ru/2019

VAE vs AE: embedding interpolation

VAE

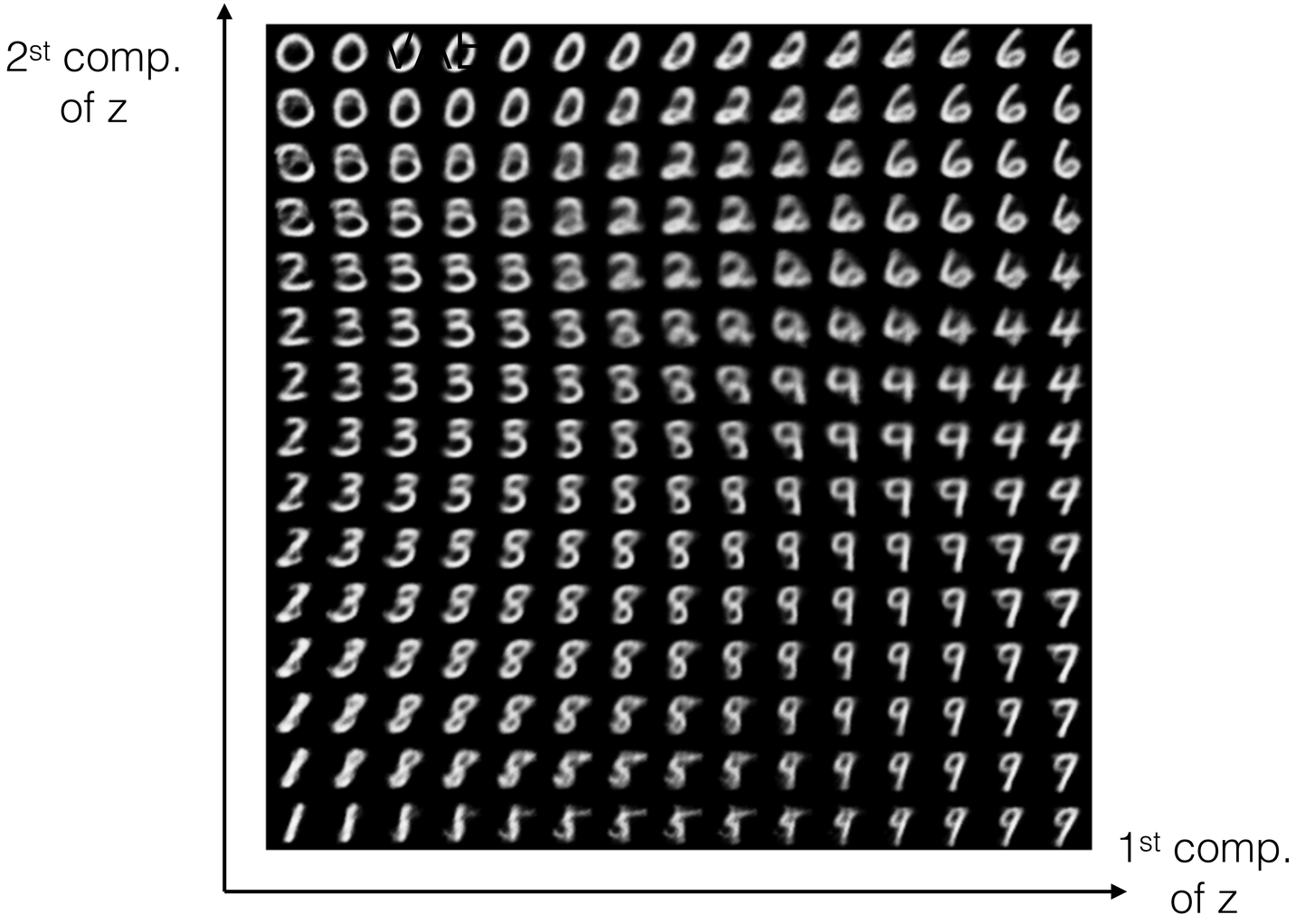
segment connecting $e_\phi(5)$ and $e_\phi(2)$

AE

Images from Kirill Struminsky's assignment, deepbayes.ru/2019

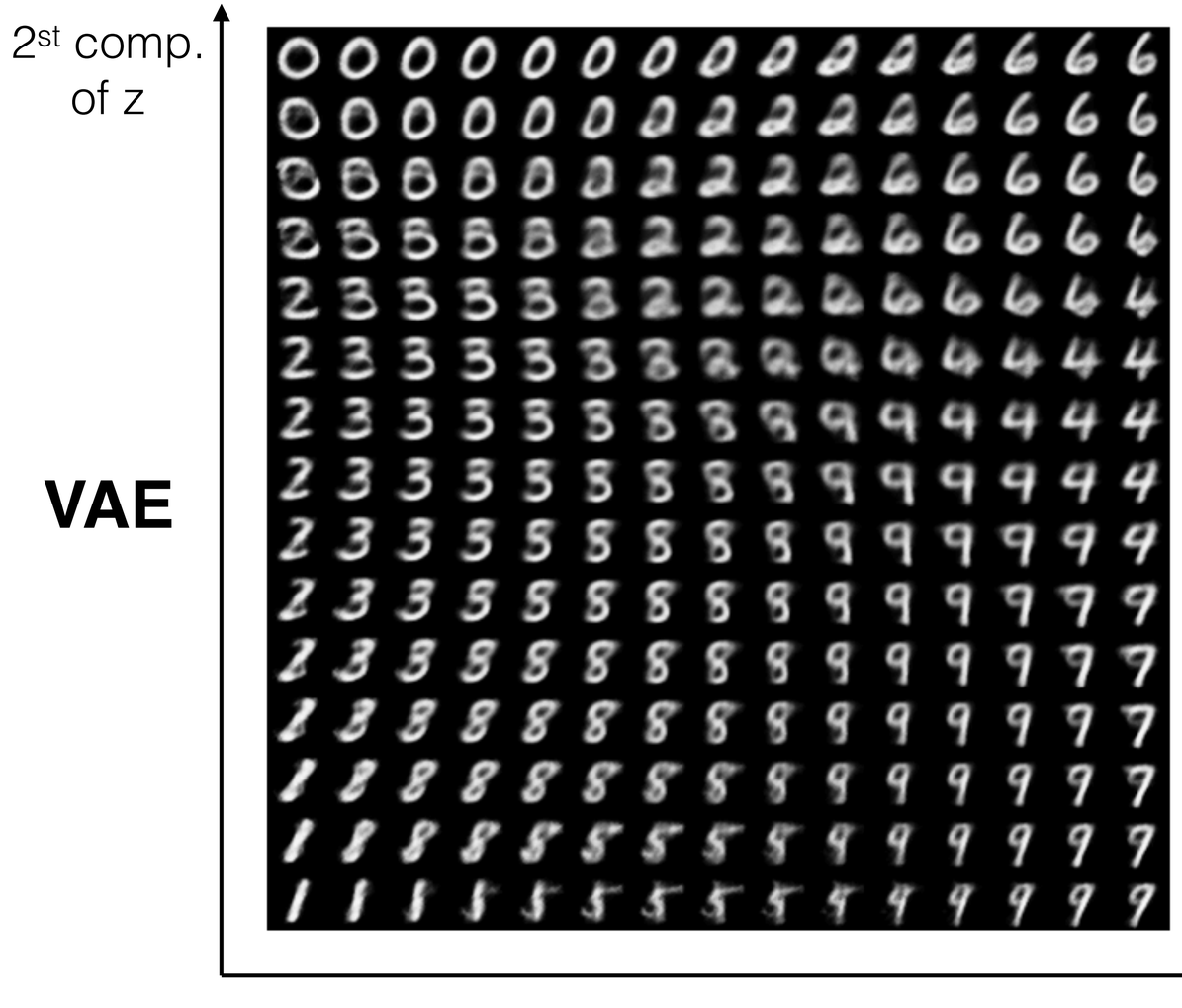
VAE: embedding “carpet”

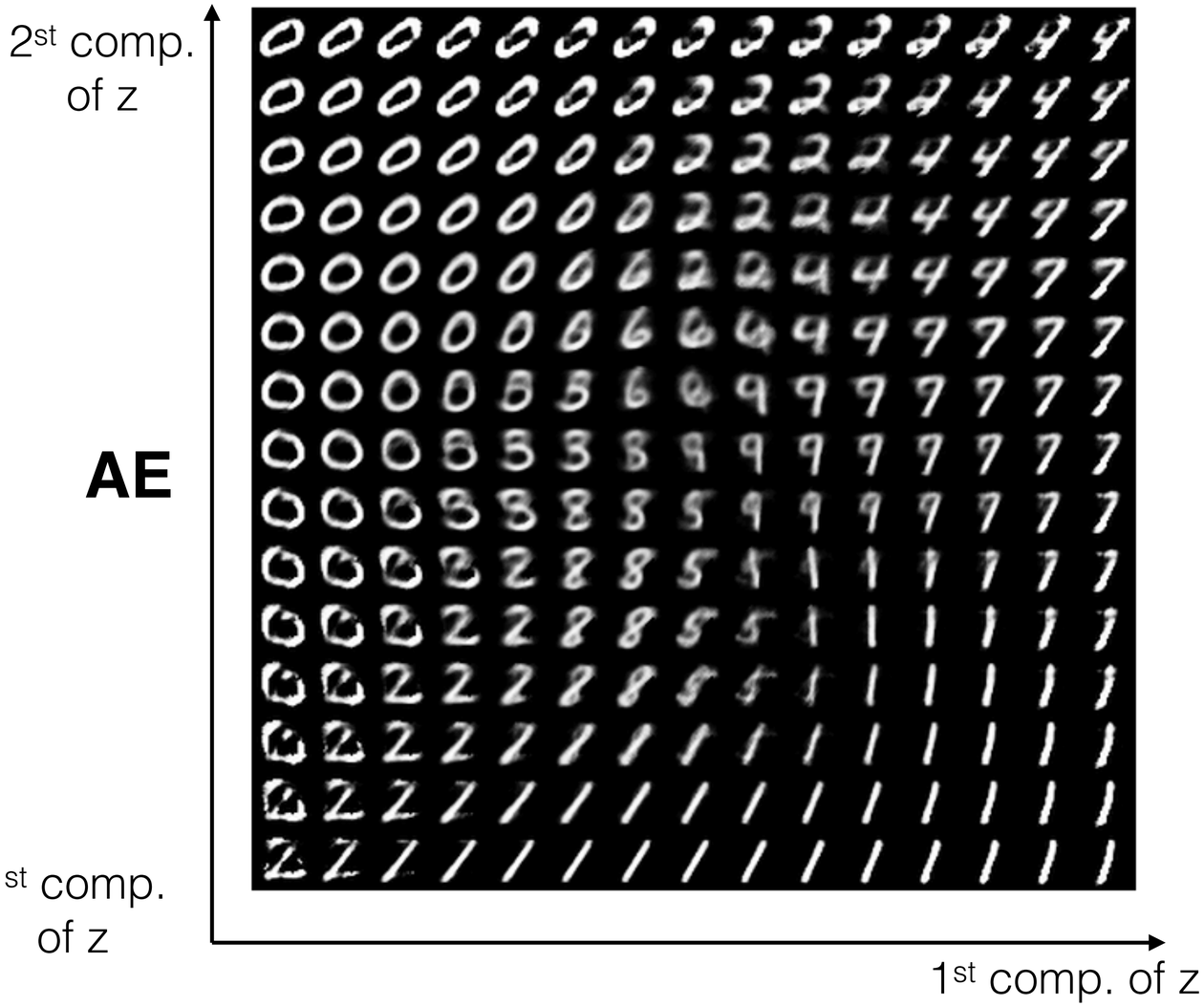
$$x = d_\theta(z)$$



Images from Oleg Ivanov’s assignment, github.com/nadiinchi/dl_labs

VAE vs AE: embedding “carpet”

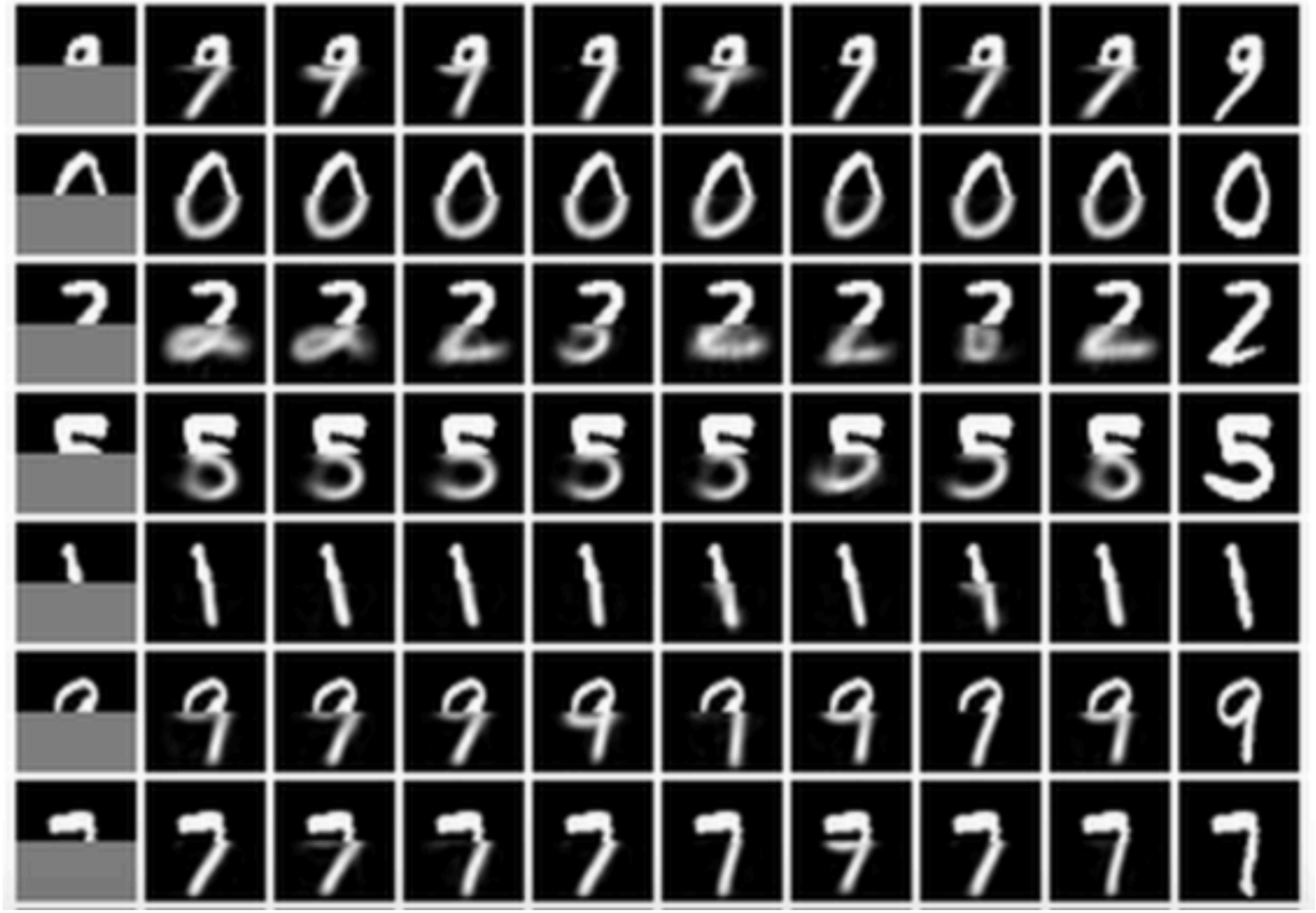




Images from Oleg Ivanov’s assignment, github.com/nadiinchi/dl_labs

VAE for missing data imputation

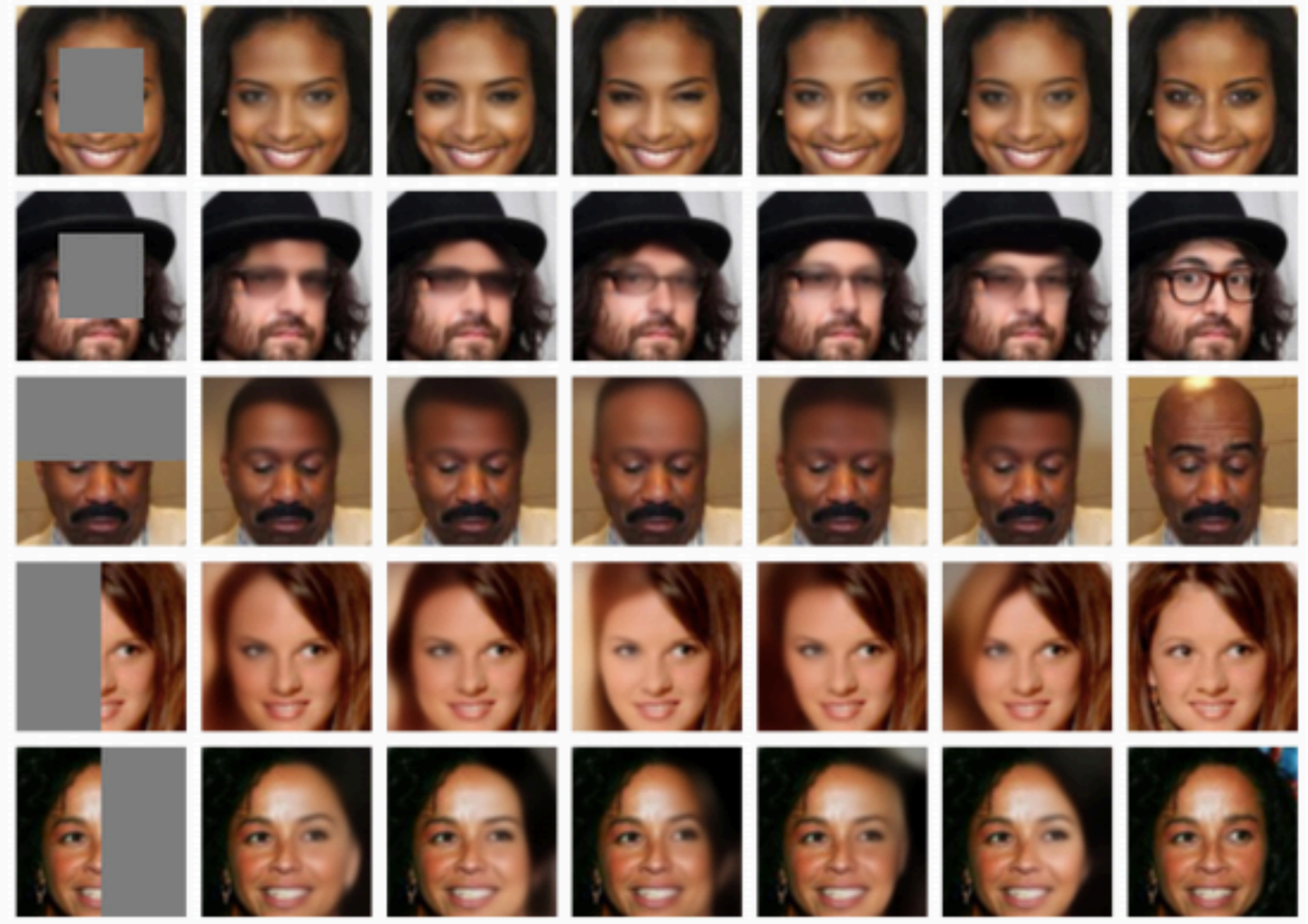
Closing one part
of image
and generating
various inpaintings



Oleg Ivanov, Michael Figurnov, Dmitry Vetrov. Variational Autoencoder with Arbitrary Conditioning, ICLR 2019

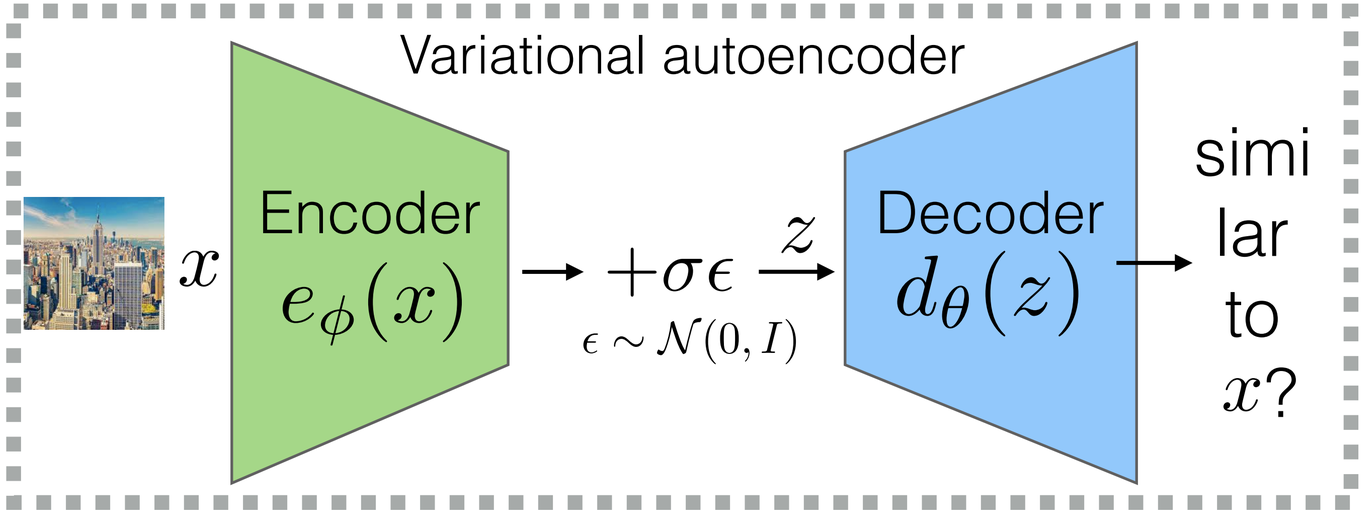
VAE for missing data imputation

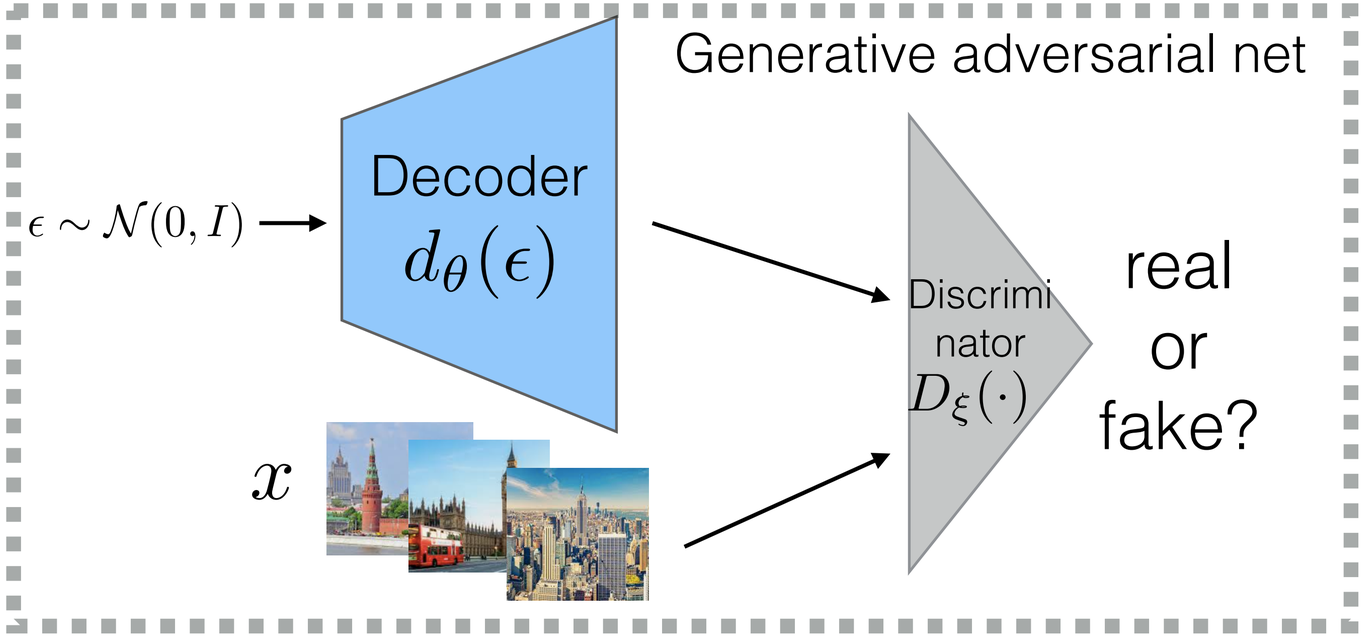
Closing one part
of image
and generating
various inpaintings



Oleg Ivanov, Michael Figurnov, Dmitry Vetrov. Variational Autoencoder with Arbitrary Conditioning, ICLR 2019

VAE vs GAN





$$\sum_{i=1}^{N}\left(\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\frac{1}{2\sigma^2}\overbrace{\left\|d_\theta\left(e_\phi(x_i)+\sigma\epsilon\right)-x_i\right\|^2}^{\text{reconstruction term}}+\right.$$

$$\left.\text{+ some term with } \sigma + \mathcal{L}_{\text{KL}}(\theta,\sigma)\right) \to \min_{\theta,\phi}$$

$$\frac{1}{N}\sum_{i=1}^{N}\log D_\xi(x_i)+$$

$$+\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\log\left(1-D_\xi(d_\theta(\epsilon))\right) \to \max_\xi \min_\theta$$

VAE vs GAN

Variational autoencoder:

+ cover the whole data space
- blurry images

$$\sum_{i=1}^{N}\left(\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\frac{1}{2\sigma^2}\overbrace{\left\|d_\theta\left(e_\phi(x_i)+\sigma\epsilon\right)-x_i\right\|^2}^{\text{reconstruction term}}+\right.$$

$$\left.\underset{\text{with }\sigma}{\text{+ some term}}+\mathcal{L}_{\mathrm{KL}}(\theta,\sigma)\right)\to\min_{\theta,\phi}$$

Generative adversarial net:

+ sharp images
- may lose some data modes

$$\frac{1}{N}\sum_{i=1}^{N}\log D_\xi(x_i)+$$

$$+\,\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\log\left(1-D_\xi(d_\theta(\epsilon))\right)\to\max_{\xi}\min_{\theta}$$

Combine in one model !?

VAE + GAN: interpolation visualization

face 1
from
data

decoded faces
at linear segment
[e(face 1), e(face2)]

face 2
from
data

Xianxu Hou, Ke Sun, Linlin Shen, Guoping Qiu. mproving Variational Autoencoder with Deep Feature Consistent and Generative Adversarial Training, Neurocomputing, 2019

Variational autoencoders: summary

- *Autoencoder point of view*:
  VAE allows reconstructing objects from noisy embeddings
  and learns more smooth latent space than AE

- *Generative model point of view*:
  VAE is a deep generative model capable of reconstructing embedded objects
  and, as a result, covering the whole data space

- *Bayesian methods point of view*:
  VAE is a non-linear latent variable model
  trained using doubly stochastic variational inference

A lot of model variants exist!

Discrete latent variables

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i,\theta) - KL(q(z_i|x_i,\phi)||p(z_i)) \Big) \to \max_{\theta,\phi}$$

General task:
(with “abstract” notation)

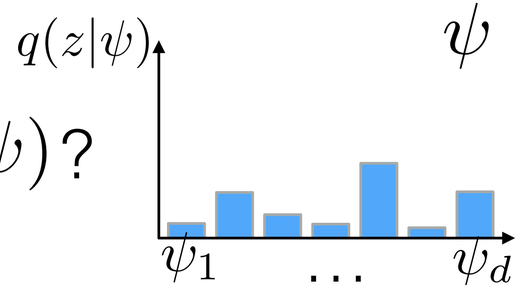
$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

reparametrization trick

$$= \frac{\partial}{\partial \psi} \mathbb{E}_{p(\epsilon)} f(z = g(\psi, \epsilon)) \approx \frac{\partial}{\partial \psi} f(z = g(\psi, \hat{\epsilon})), \hat{\epsilon} \sim p(\epsilon)$$

How to reparametrize categorical $q(z|\psi)$?

$$z = 1, \dots, d$$



ψ — probabilities of categories

Discrete latent variables

$$\mathcal{L}(\phi, \theta) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z_i|x_i,\phi)} \log p(x_i|z_i, \theta) - KL(q(z_i|x_i, \phi)||p(z_i)) \right) \to \max_{\theta,\phi}$$

General task:
(with "abstract" notation)

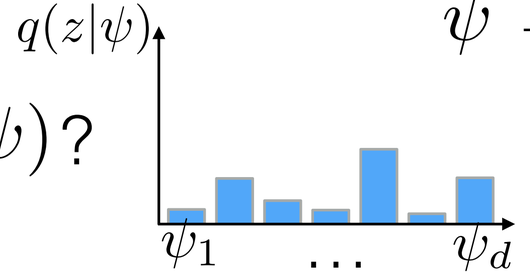
$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

non-differentiable in case of discrete z!
(piece-wise function)

$$= \frac{\partial}{\partial \psi} \mathbb{E}_{p(\epsilon)} f(z = g(\psi, \epsilon)) \approx \frac{\partial}{\partial \psi} f(z = g(\psi, \hat{\epsilon})), \hat{\epsilon} \sim p(\epsilon)$$

How to reparametrize categorical $q(z|\psi)$?

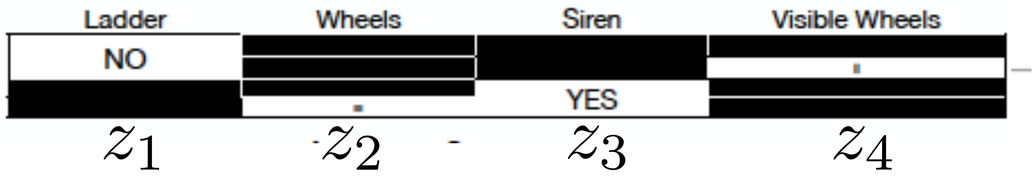
$$z = 1, \ldots, d$$

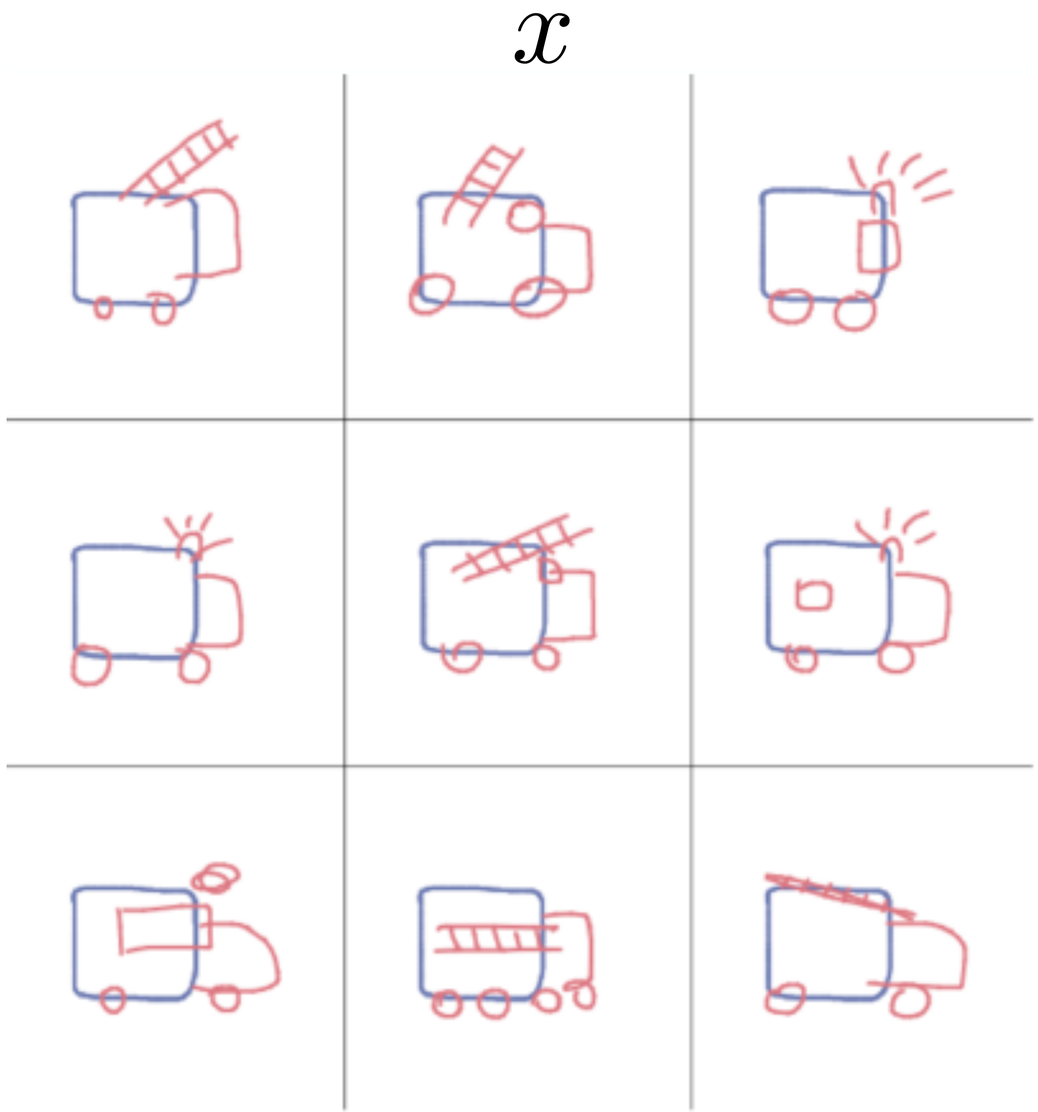


Discrete latent variables: motivation

Real latent variables:

Categorical latent variables:





A slide from Kirill Struminsky's presentation

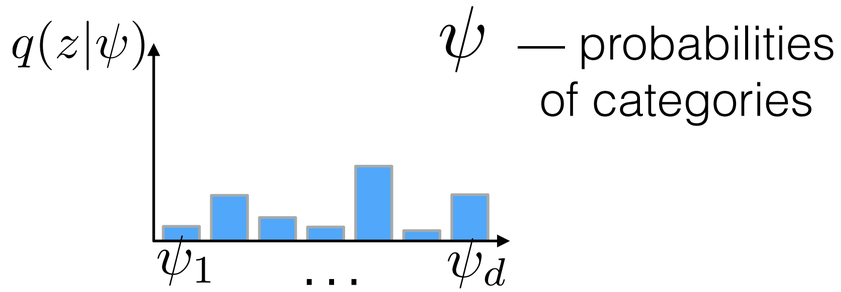
Discrete latent variables: examples

- Categorical latent variables
- Text as latent variable
- Hard attention in images or NLP (binary attention masks)
- Discrete actions in RL
- Binary weight in neural network (quantization)
- Automatically tuning dropout rates

A slide from Kirill Struminsky's presentation

Gumbel argmax trick

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$



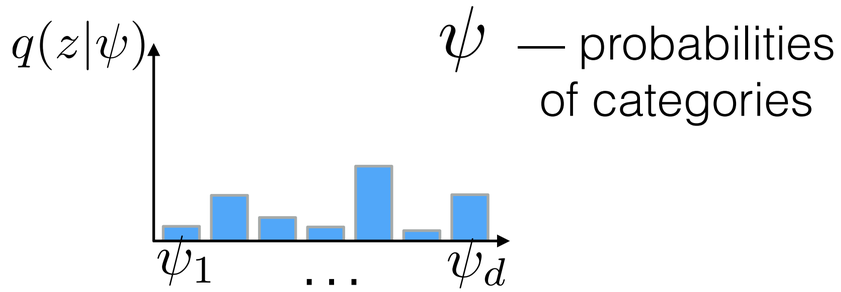
$$z \sim q(z|\phi) \quad \Leftrightarrow$$

1. $\epsilon_j = -\log(-\log u_j),\ u_j \sim [0,1], j = 1, \dots, d$
2. $\hat{z} = \text{argmax}_{j=1,\dots,d}(\psi_j + \epsilon_j)$

non-differentiable

Gumbel *softmax* trick

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$



$$z \sim q(z|\psi) \quad \Leftrightarrow$$

1. $\epsilon_j = -\log(-\log u_j),\ u_j \sim [0,1], j = 1, \ldots, d$
2. $\hat{z} = \text{softmax}_{j=1,\ldots,d}((\psi_j + \epsilon_j)/\tau)$

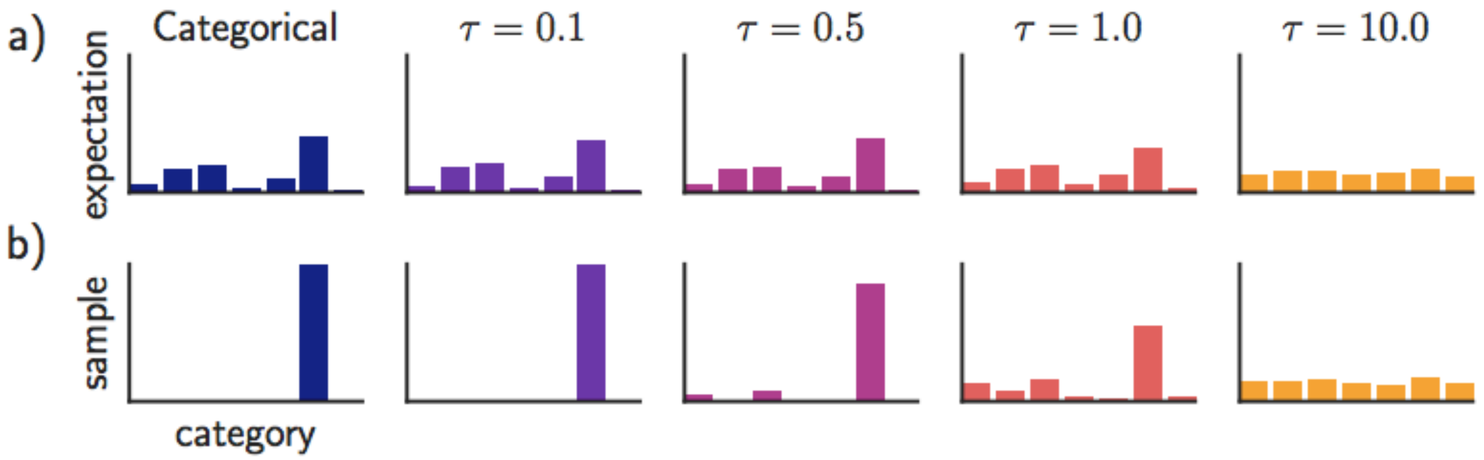
differentiable!

$$\Rightarrow \frac{\partial}{\partial \phi} \mathbb{E}_{q(z|\psi)} f(z) \approx \frac{\partial}{\partial \psi} f(\hat{z})$$

differentiable!

Eric Jang, Shixiang Gu, Ben Poole. Categorical Reparameterization with Gumbel-Softmax. ICLR 2017

Gumbel *softmax* trick



Temperature τ may be annealed during training

1. $\epsilon_j = -\log(-\log u_j),\ u_j \sim [0,1], j = 1, \ldots, d$
2. $\hat{z} = \text{softmax}_{j=1,\ldots,d}((\psi_j + \epsilon_j)/\tau)$

differentiable!

$$\Rightarrow \frac{\partial}{\partial \phi} \mathbb{E}_{q(z|\phi)} f(z) \approx \frac{\partial}{\partial \phi} f(\hat{z})$$

differentiable!

Eric Jang, Shixiang Gu, Ben Poole. Categorical Reparameterization with Gumbel-Softmax. ICLR 2017

VAE with discrete latent variables and Gumbel softmax trick

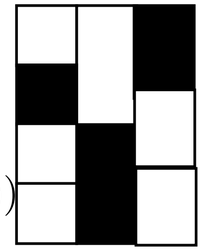
x

Encoder
$e_\phi(x)$

sampling
z
$z \sim q(z|\psi = e_\phi(x))$

During testing:

z_1, z_2, z_3

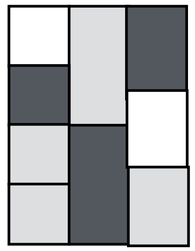


Decoder
$d_\theta(z)$

similar to x ?

During training:

z_1, z_2, z_3



REINFORCE estimator

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

REINFORCE estimator

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) = \frac{\partial}{\partial \psi} \sum_{z=1}^{d} q(z|\psi) f(z)$$

$$= \sum_{z=1}^{d} \frac{\partial}{\partial \psi} q(z|\psi) f(z)$$

Cannot approximate with sampling!

REINFORCE estimator

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

log-derivative trick

$$\frac{\partial}{\partial \psi} \log q(z|\psi) = \frac{1}{q(z|\psi)} \frac{\partial}{\partial \psi} q(z|\psi)$$

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) = \frac{\partial}{\partial \psi} \sum_{z=1}^{d} q(z|\psi) f(z)$$

$$= \sum_{z=1}^{d} \frac{\partial}{\partial \psi} q(z|\psi) f(z)$$

Cannot approximate with sampling!

REINFORCE estimator

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

log-derivative trick

$$\frac{\partial}{\partial \psi} \log q(z|\psi) = \frac{1}{q(z|\psi)} \frac{\partial}{\partial \psi} q(z|\psi)$$

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) = \frac{\partial}{\partial \psi} \sum_{z=1}^{d} q(z|\psi) f(z)$$

$$= \sum_{z=1}^{d} \frac{\partial}{\partial \psi} q(z|\psi) f(z) =$$

$$= \sum_{z=1}^{d} q(z|\psi) \frac{\partial}{\partial \psi} \log q(z|\psi) f(z)$$

Can approximate with sampling!

REINFORCE estimator

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

log-derivative trick

$$\frac{\partial}{\partial \psi} \log q(z|\psi) = \frac{1}{q(z|\psi)} \frac{\partial}{\partial \psi} q(z|\psi)$$

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) = \frac{\partial}{\partial \psi} \sum_{z=1}^{d} q(z|\psi) f(z)$$

$$= \sum_{z=1}^{d} \frac{\partial}{\partial \psi} q(z|\psi) f(z) =$$

$$= \sum_{z=1}^{d} q(z|\psi) \frac{\partial}{\partial \psi} \log q(z|\psi) f(z) \approx \frac{\partial}{\partial \psi} \log q(\hat{z}|\psi) f(\hat{z}), \hat{z} \sim q(z|\psi)$$

Sampling discrete latent variables

$$\frac{\partial}{\partial \psi} \mathbb{E}_{q(z|\psi)} f(z) - ?$$

REINFORCE

$$\frac{\partial}{\partial \psi} \log q(\hat{z}|\psi) f(\hat{z})$$

$$\hat{z} \sim q(z|\psi)$$

+ unbiased

- very high variance

Gumbel-softmax

$$\frac{\partial}{\partial \psi} f(\hat{z})$$

1. $\epsilon_j = -\log(-\log u_j),\ u_j \sim [0, 1]$
2. $\hat{z} = \text{softmax}_{j=1,\ldots,d}((\psi_j + \epsilon_j)/\tau)$

+ low variance

- biased

Usually used in practice!