Variational Autoencoders

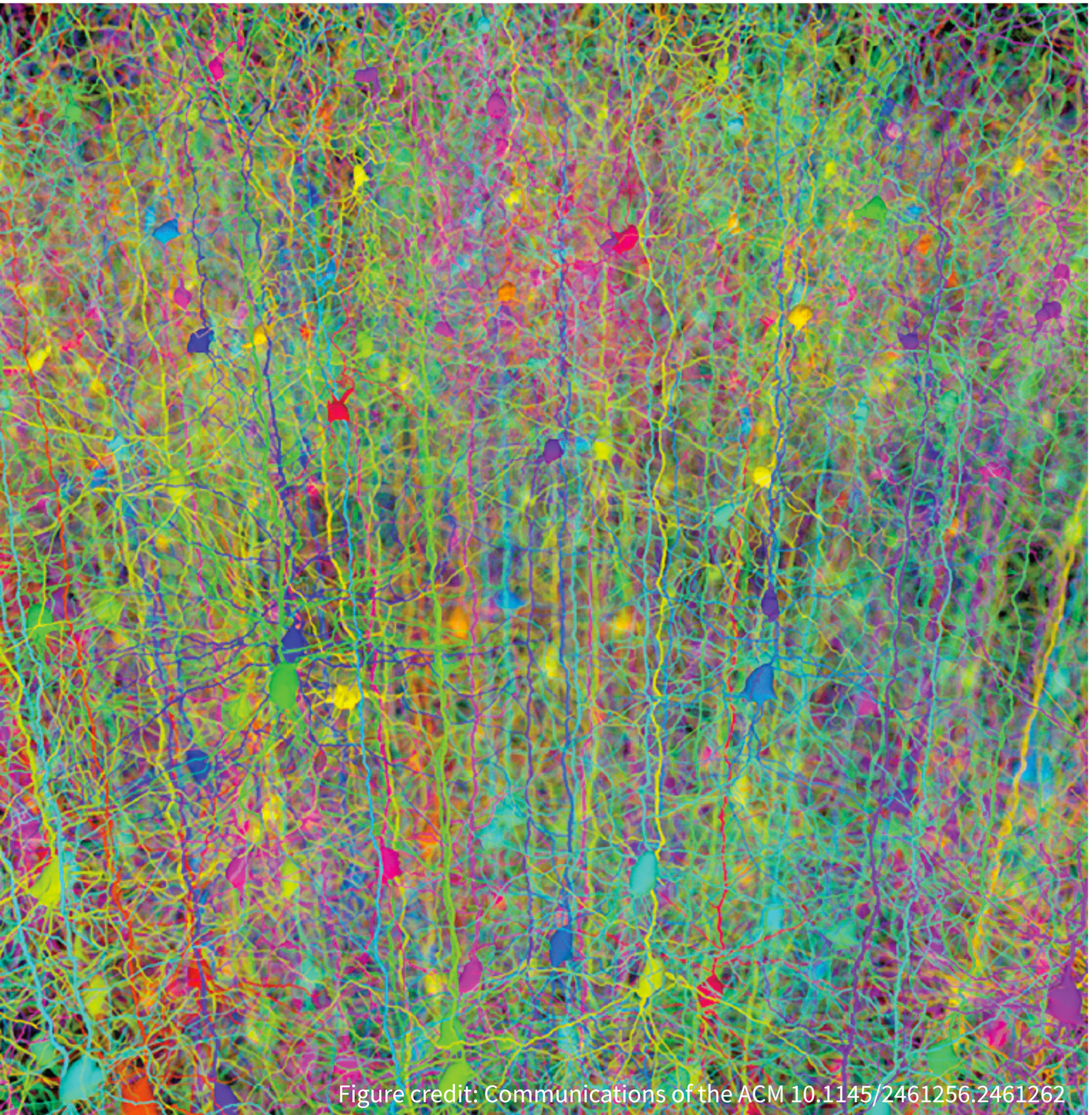


Figure credit: Communications of the ACM 10.1145/2461256.2461262

GRAHAM TAYLOR

VECTOR INSTITUTE

SCHOOL OF ENGINEERING
UNIVERSITY OF GUELPH

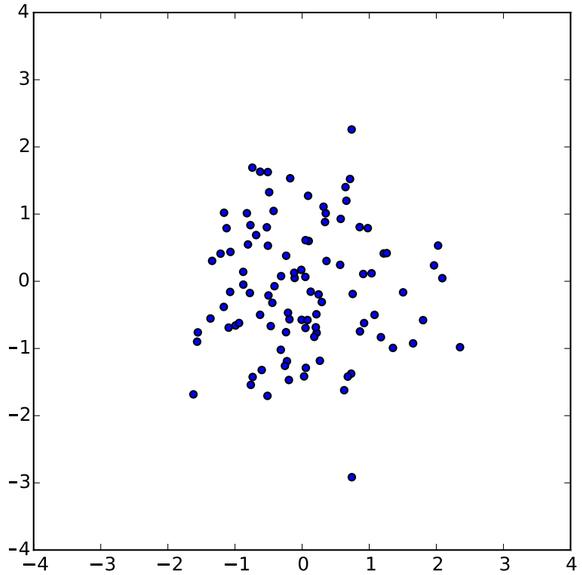
CANADIAN INSTITUTE
FOR ADVANCED RESEARCH

UNIVERSITY of GUELPH
CHANGING LIVES
IMPROVING LIFE

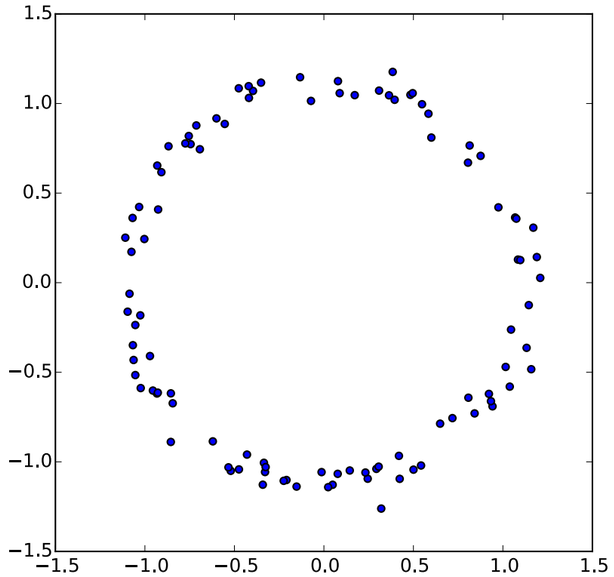
CIFAR
CANADIAN
INSTITUTE
FOR
ADVANCED
RESEARCH

Intuition

Given a random variable with one (simple) distribution, we can create another random variable with a completely different (complex) distribution



$$g(\boldsymbol{z}) = \boldsymbol{z}/10 + \boldsymbol{z}/||\boldsymbol{z}||$$



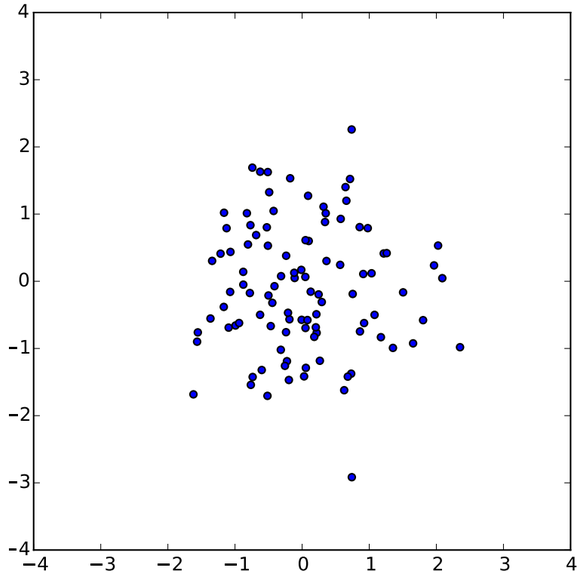
$\boldsymbol{z}$

$\boldsymbol{x} = g(\boldsymbol{z})$

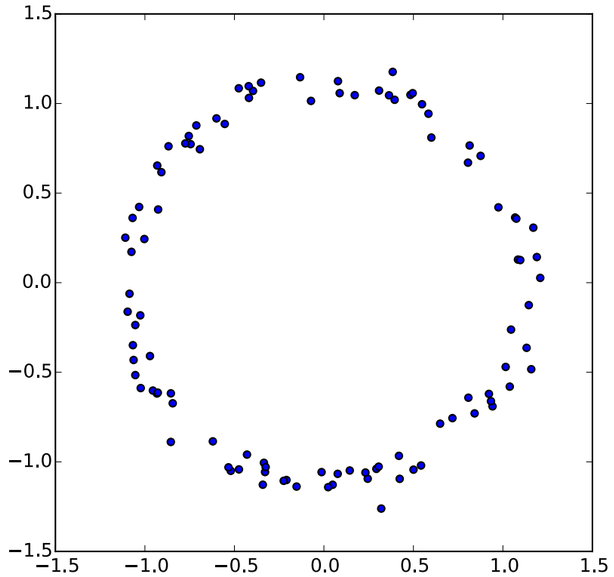
via Doersch 2016

Intuition

Given a random variable with one (simple) distribution, we can create another random variable with a completely different (complex) distribution



$$f(\mathbf{z};\boldsymbol{\theta})$$



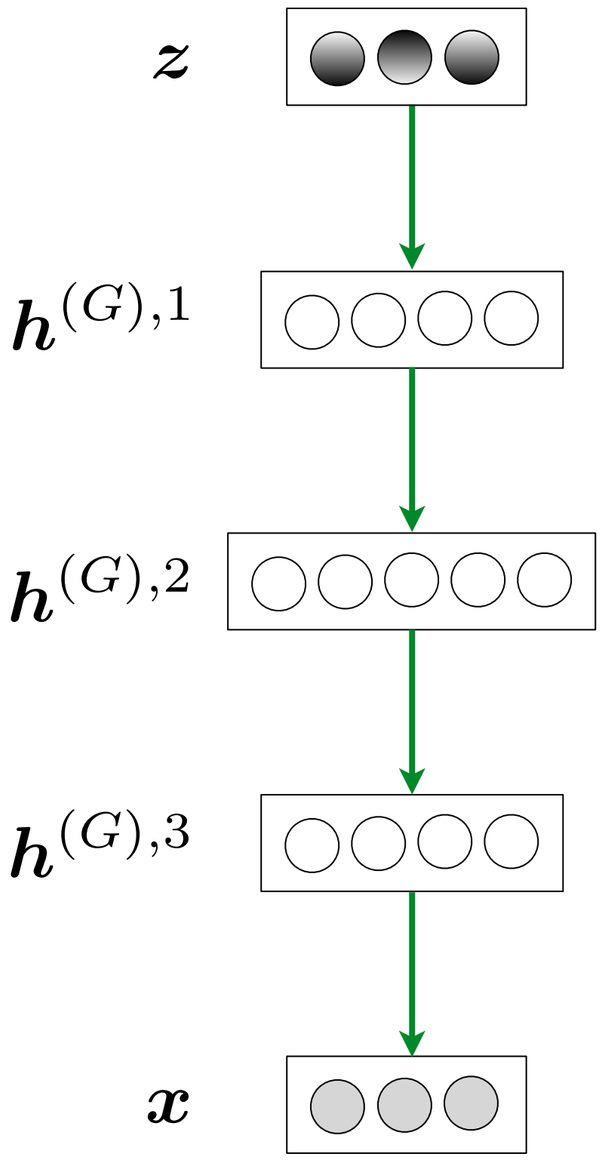
$$\boldsymbol{z} \sim \mathcal{N}(0, 1)$$

$$\boldsymbol{x} \sim \mathcal{N}\left(f\left(\boldsymbol{z};\boldsymbol{\theta}\right), \sigma^2 I\right)$$

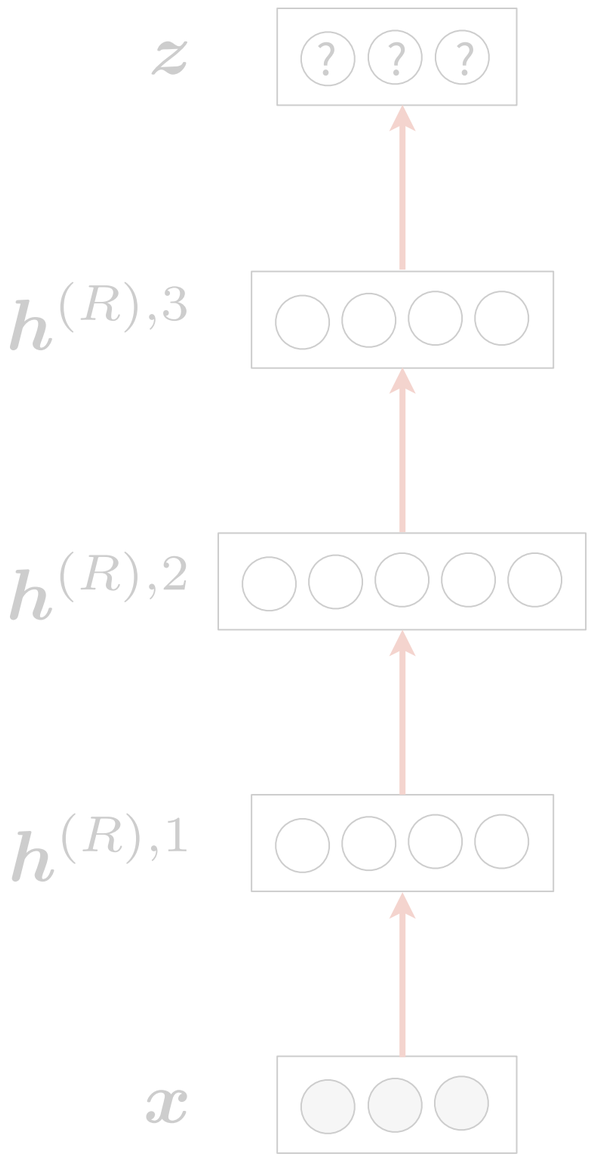
via Doersch 2016

(Kingma and Welling, 2014) (Rezende et al., 2014) (Mnih and Gregor, 2014)

Variational Autoencoder

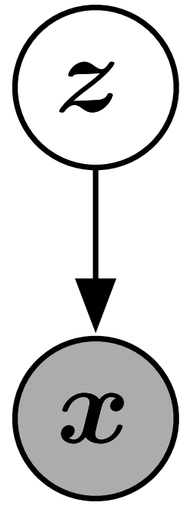


Generative Network (Decoder)



Recognition Network (Encoder)

ML Objective

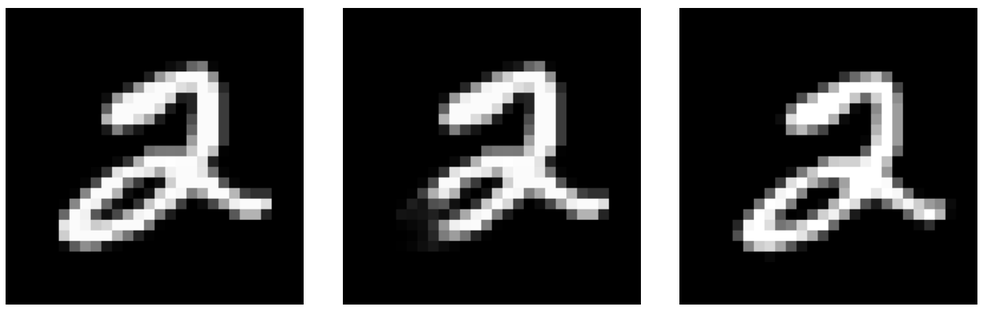


- Want to train to maximize $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})p(\boldsymbol{z})d\boldsymbol{z}$

- How to define the latent variables z?
- How to deal with the integral over z?

Estimating Likelihood?

- What about estimating $p(\boldsymbol{x}) \approx \frac{1}{N}\sum_n p(\boldsymbol{x}|\boldsymbol{z}_n)$?



Target | Bad example | Good example

via Doersch 2016

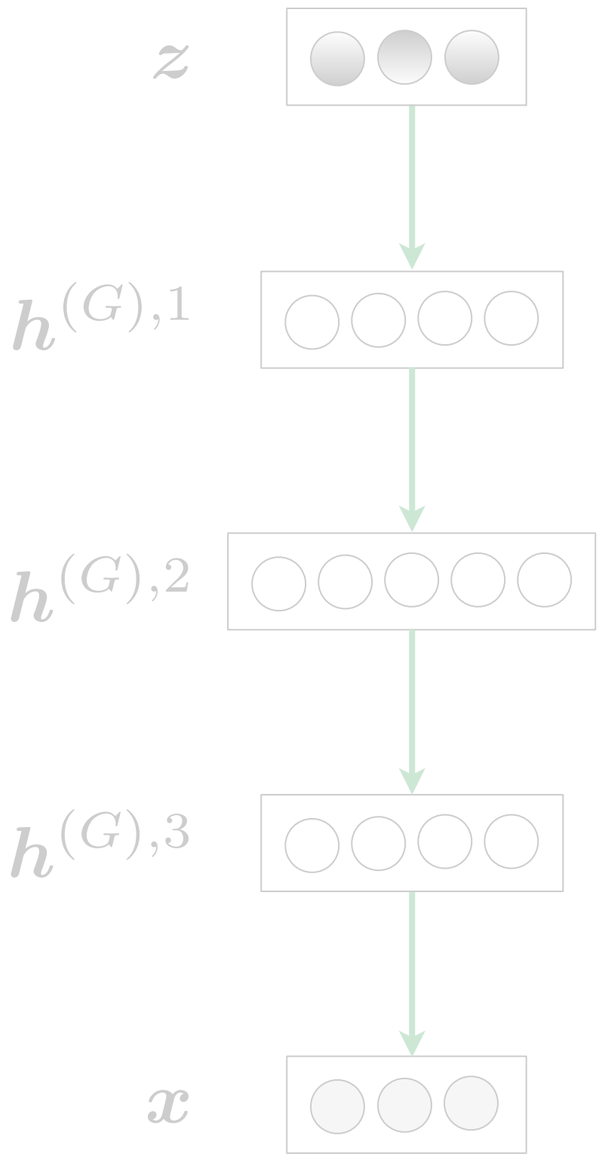
Approximate Posterior

- In practice, for most z, $p(x|z)$ will be nearly zero and contribute almost nothing to the estimate for $p(x)$
- The key idea behind the VAE is to attempt to sample values of z that are likely to have produced x and compute $p(x)$ just from those z
- This means we need a new function $q(z|x)$ which can take a value of x and give a distribution over z values that are likely to produce x

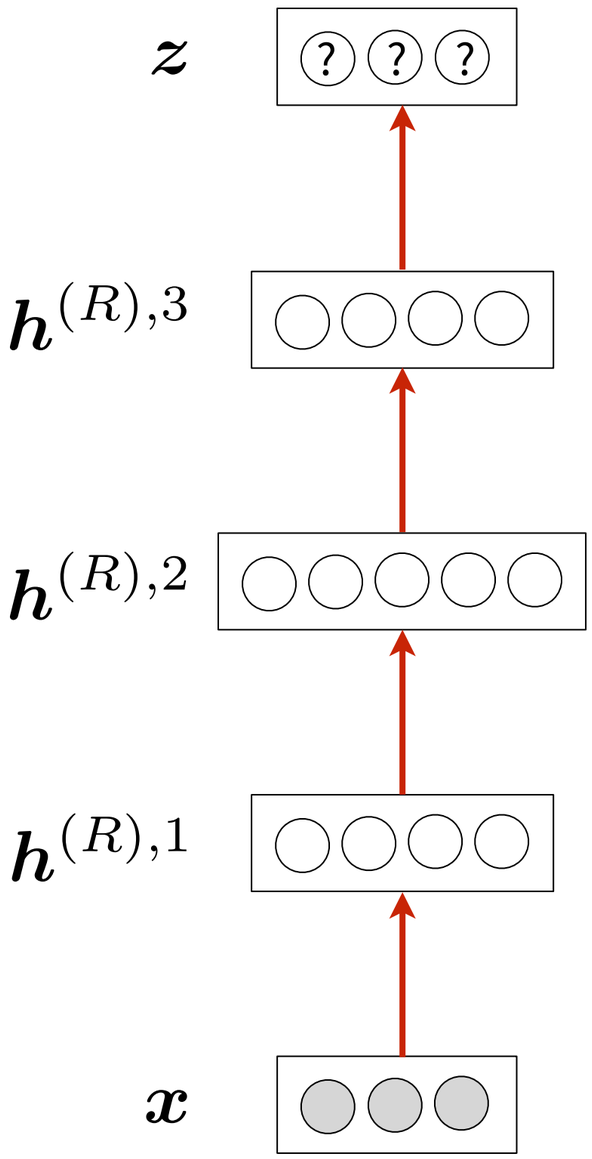
via Doersch 2016

(Kingma and Welling, 2014) (Rezende et al., 2014) (Mnih and Gregor, 2014)

Variational Autoencoder



Generative Network (Decoder)



Recognition Network (Encoder)

The Math (1)

via Doersch 2016

The Math (1)

- Start with the Kullback-Leibler (KL) divergence between $p(\boldsymbol{z}|\boldsymbol{x})$ and $q(\boldsymbol{z})$ for some arbitrary $q(\cdot)$

$$\mathcal{D}\left[q\left(\boldsymbol{z}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log q\left(\boldsymbol{z}\right) - \log p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right]$$

via Doersch 2016

The Math (1)

- Start with the Kullback-Leibler (KL) divergence between $p(\boldsymbol{z}|\boldsymbol{x})$ and $q(\boldsymbol{z})$ for some arbitrary $q(\cdot)$

$$\mathcal{D}\left[q\left(\boldsymbol{z}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log q\left(\boldsymbol{z}\right) - \log p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right]$$

- We can get both $p(\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{z})$ into this equation by applying Bayes' rule to $p(\boldsymbol{z}|\boldsymbol{x})$

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

Comes out of the expectation, does not depend on $\boldsymbol{z}$

via Doersch 2016

The Math (2)

via Doersch 2016

The Math (2)

- Starting point:

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

via Doersch 2016

The Math (2)

- Starting point:

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

- Negating both sides, rearranging, and spotting another KL-divergence yields

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}\right)||p\left(\boldsymbol{z}\right)\right]$$

via Doersch 2016

The Math (2)

- Starting point:

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

- Negating both sides, rearranging, and spotting another KL-divergence yields

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}\right)||p\left(\boldsymbol{z}\right)\right]$$

- Note that x is fixed, and $q(\cdot)$ can be any distribution. It makes sense to construct a $q(\cdot)$ that depends on x

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

via Doersch 2016

The VAE Objective

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

via Doersch 2016

The VAE Objective

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

Why is this a sensible objective?

via Doersch 2016

The VAE Objective

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

Why is this a sensible objective?

- The LHS has the thing we want to maximize plus an error term
  - The error term will be small for high-capacity $q\left(\cdot\right)$

via Doersch 2016

The VAE Objective

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

Why is this a sensible objective?

- The LHS has the thing we want to maximize plus an error term
  - The error term will be small for high-capacity $q\left(\cdot\right)$
- The RHS is something we can optimize using SGD

via Doersch 2016

The VAE Objective

Reconstruction term

Regularizer term

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

Decode $\boldsymbol{z}$ into $\boldsymbol{x}$

Encode $\boldsymbol{x}$ into $\boldsymbol{z}$

Why is this a sensible objective?

- The LHS has the thing we want to maximize plus an error term
  - The error term will be small for high-capacity $q\left(\cdot\right)$
- The RHS is something we can optimize using SGD
- Note that the RHS looks something like an autoencoder

via Doersch 2016

The VAE Objective

Reconstruction term Regularizer term

$$\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

Decode $\boldsymbol{z}$ into $\boldsymbol{x}$ Encode $\boldsymbol{x}$ into $\boldsymbol{z}$

Why is this a sensible objective?

- The LHS has the thing we want to maximize plus an error term
  - The error term will be small for high-capacity $q\left(\cdot\right)$
- The RHS is something we can optimize using SGD
- Note that the RHS looks something like an autoencoder

via Doersch 2016

Optimizing VAE Objective

via Doersch 2016

Optimizing VAE Objective

- How can we perform SGD on the RHS of our objective?

$$\mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right]-\mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

via Doersch 2016

Optimizing VAE Objective

- How can we perform SGD on the RHS of our objective?

$$\mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right]-\mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

- We have to get a bit more formal for $q\left(\boldsymbol{z}|\boldsymbol{x}\right)$

via Doersch 2016

Optimizing VAE Objective

- How can we perform SGD on the RHS of our objective?

$$\mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right]-\mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

- We have to get a bit more formal for $q\left(\boldsymbol{z}|\boldsymbol{x}\right)$
- The usual choice is to let: $q(\boldsymbol{z}|\boldsymbol{x})=\mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{\Sigma}(\boldsymbol{x};\boldsymbol{\theta})\right)$ so

$$\mathcal{D}\left[\mathcal{N}\left(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right)||\mathcal{N}\left(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)\right]=$$
$$\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right)+\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)-k+\log\left(\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_0}\right)\right)$$

via Doersch 2016

Optimizing VAE Objective

- How can we perform SGD on the RHS of our objective?

$$\mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right]-\mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$$

- We have to get a bit more formal for $q\left(\boldsymbol{z}|\boldsymbol{x}\right)$
- The usual choice is to let: $q(\boldsymbol{z}|\boldsymbol{x})=\mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{\Sigma}(\boldsymbol{x};\boldsymbol{\theta})\right)$ so

$$\mathcal{D}\left[\mathcal{N}\left(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right)||\mathcal{N}\left(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)\right]=$$
$$\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right)+\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)-k+\log\left(\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_0}\right)\right)$$

- This simplifies to:

$$\mathcal{D}\left[\mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{x}),\boldsymbol{\Sigma}(\boldsymbol{x})\right)||\mathcal{N}\left(\mathbf{0},I\right)\right]=\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x})\right)+\boldsymbol{\mu}(\boldsymbol{x})^T\boldsymbol{\mu}(\boldsymbol{x})-k+\log\det\left(\boldsymbol{\Sigma}(\boldsymbol{x})\right)\right)$$

via Doersch 2016

Stochastic Gradient Descent

- Think about what we really want to optimize:

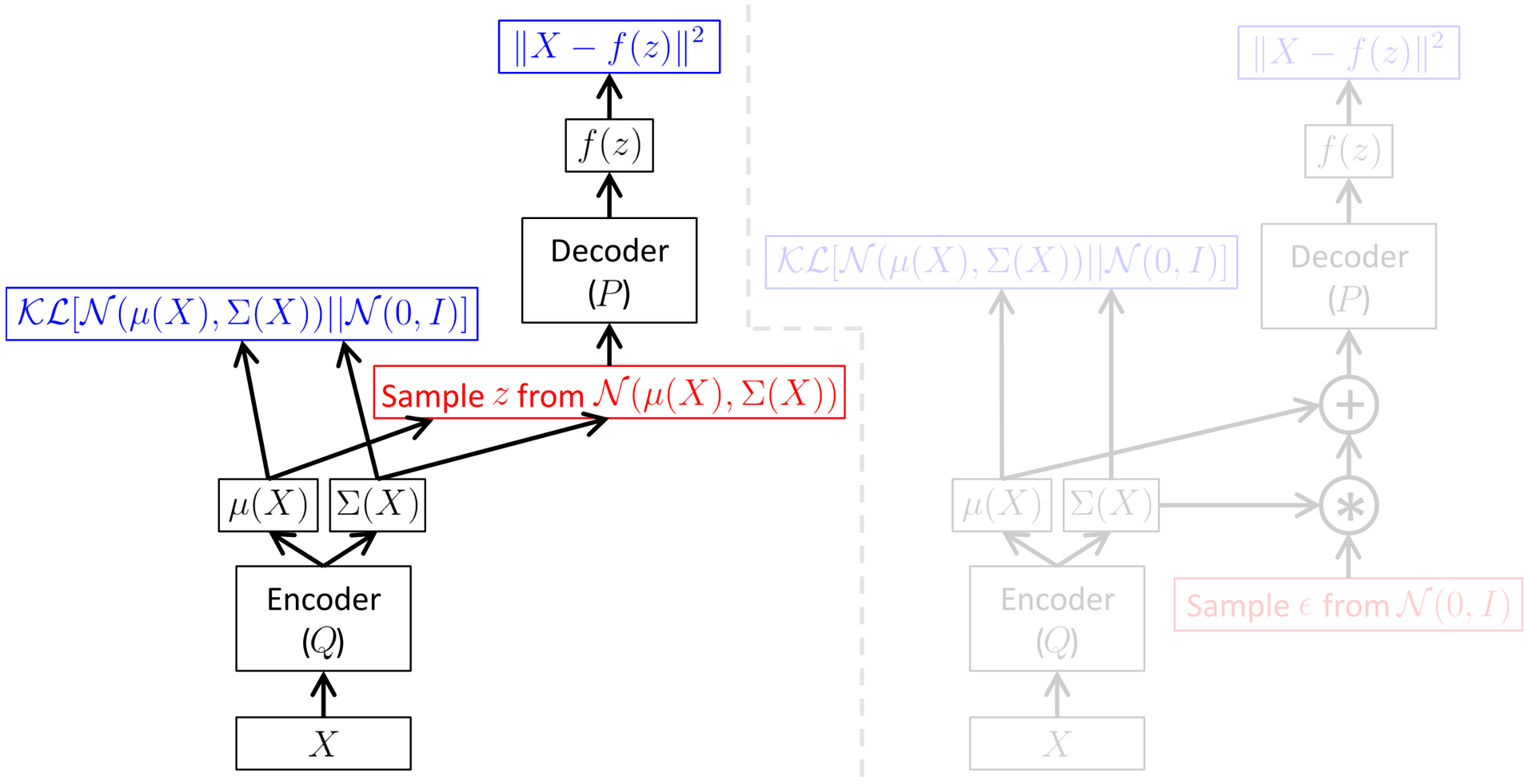
$$\mathbb{E}_{\boldsymbol{x}\sim D}\left[\log p\left(\boldsymbol{x}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}|\boldsymbol{x}\right)\right]\right] =$$
$$\mathbb{E}_{\boldsymbol{x}\sim D}\left[\mathbb{E}_{\boldsymbol{z}\sim q}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]\right]$$

- If we take the gradient of this equation, the gradient can be moved inside the expectations

- Therefore, as usual in SGD, we can sample a single value of x and a single value of $\boldsymbol{z}\sim q\left(\boldsymbol{z}|\boldsymbol{x}\right)$ and compute the gradient of $\log p\left(\boldsymbol{x}|\boldsymbol{z}\right) - \mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]$

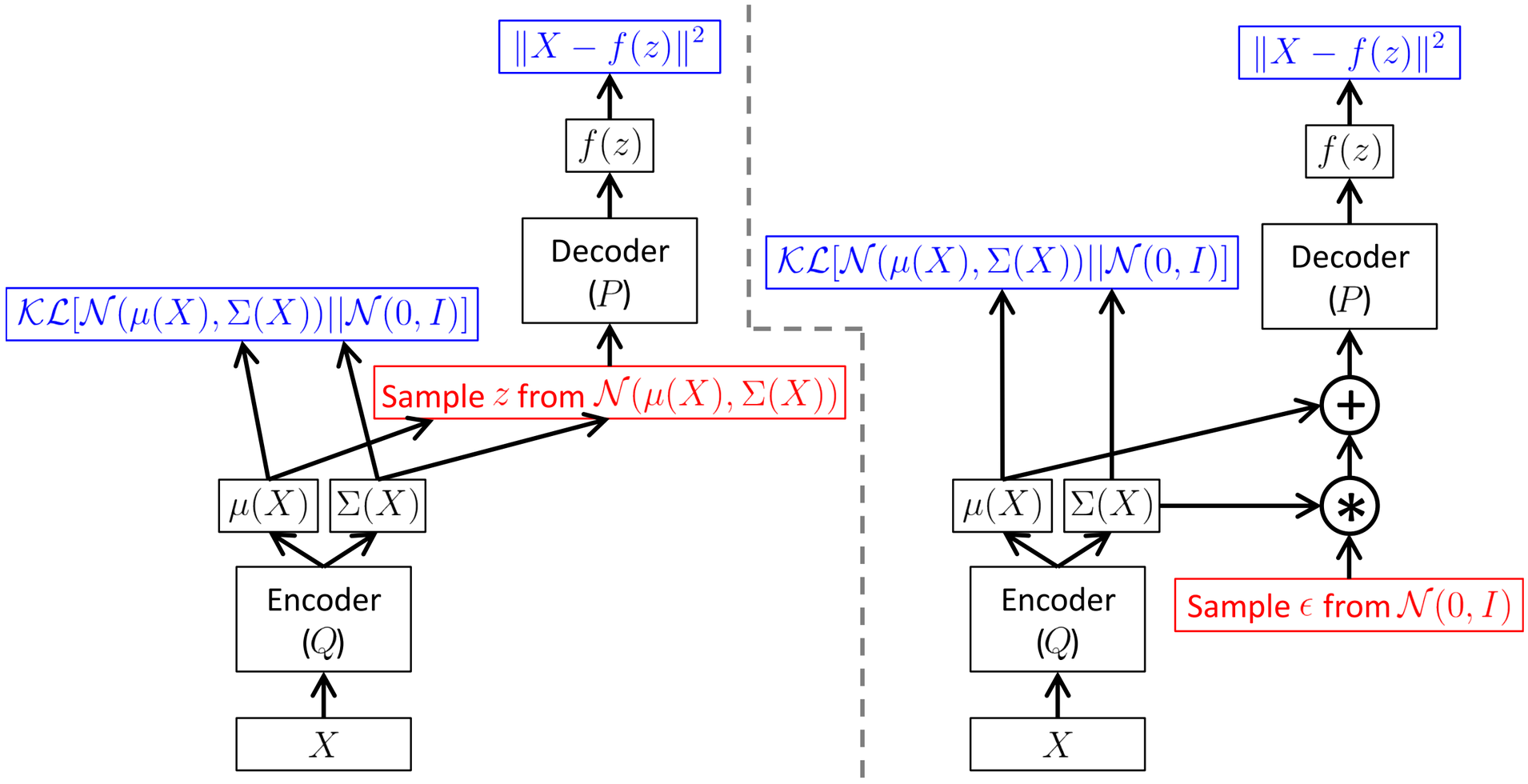
via Doersch 2016

The Repameterization Trick



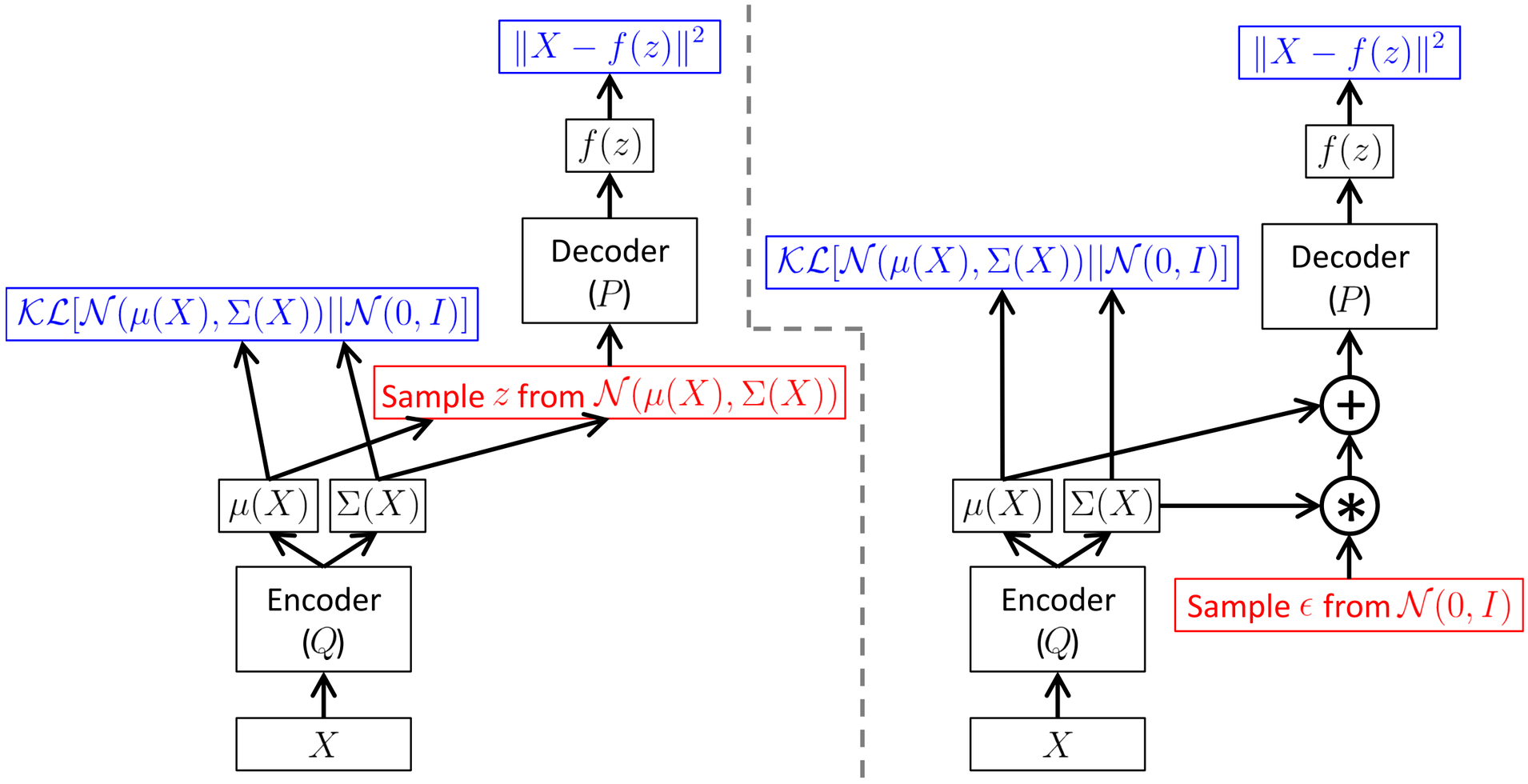
via Doersch 2016

The Repameterization Trick



via Doersch 2016

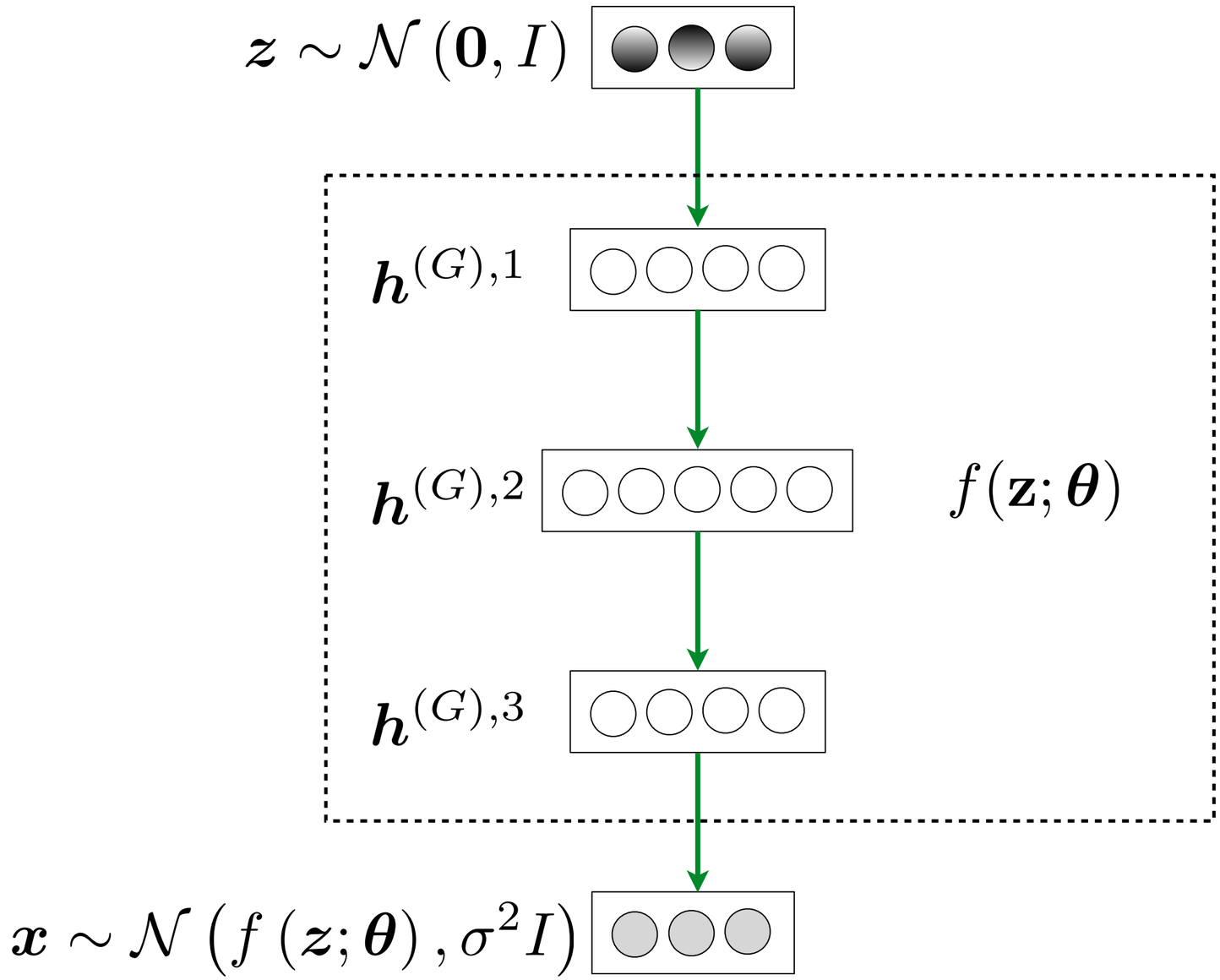
The Repameterization Trick



$$\mathbb{E}_{\boldsymbol{x}\sim D}\left[\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},I)}\left[\log p\left(\boldsymbol{x}|\boldsymbol{z}=\boldsymbol{\mu}(\boldsymbol{x})+\boldsymbol{\Sigma}^{1/2}(\boldsymbol{x})\cdot\epsilon\right)\right]-\mathcal{D}\left[q\left(\boldsymbol{z}|\boldsymbol{x}\right)||p\left(\boldsymbol{z}\right)\right]\right]$$

via Doersch 2016

Test Time / Sampling

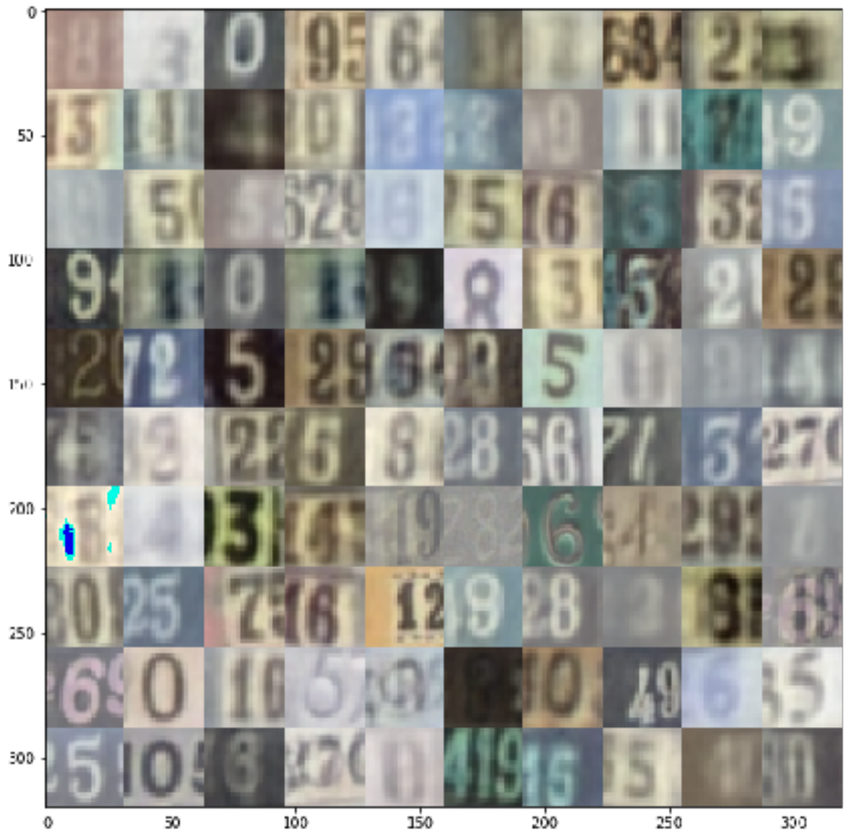


Decoder

References and Demos

- Carl Doersch's tutorial (these slides mainly follow his presentation): https://arxiv.org/abs/1606.05908

- Practical tips on implementation (and way more links): http://bjlkeng.github.io/posts/a-variational-autoencoder-on-the-svnh-dataset/



Generation from VAE
trained on SVHN
Courtesy of Brian Keng