Generative models

Variational AE and flows

Formation SCAI

Philippe Esling (esling@ircam.fr)

Maître de conférences – Sorbonne Université

Equipe représentations musicales (IRCAM, Paris)

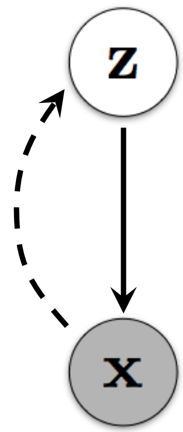
Variational inference

We take back our inference problem

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

$p(\mathbf{z})$ prior

posterior $p(\mathbf{z}|\mathbf{x})$



$p(\mathbf{x}|\mathbf{z})$ likelihood

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

Could we transform this integral

Using a simpler and tractable distribution $q(\mathbf{z}|\mathbf{x})$?

Variational inference

Starting from our intractable integral

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

The goal of Variational Inference is to *approximate* this by using *optimization* instead of derivation.

- Select a family of *parametric* distributions $q_\phi \in \mathcal{Q}$

Optimize the parameters ϕ in order to solve

$$\underset{\phi}{\operatorname{argmin}}\ \mathcal{D}_{KL}\left[q_\phi || p\right]$$

Modeling the joint probability

$$\log\, p(\mathbf{x}) = \log \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

$$\log p(\mathbf{x}) = \log \int p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}q(\mathbf{z}|\mathbf{x})d\mathbf{z}$$

Introducing $q(\mathbf{z}|\mathbf{x})$

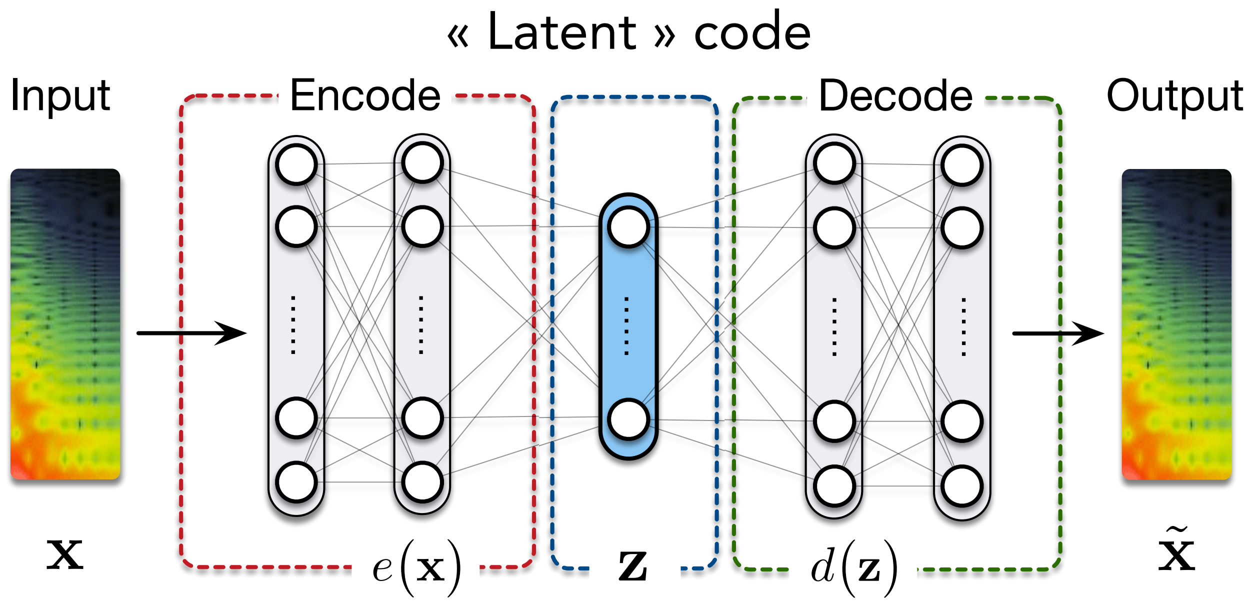
$$\geq \int q(\mathbf{z}|\mathbf{x}) \log \left(p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right)d\mathbf{z}$$

Jensen's inequality
$f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$

$$\geq \underbrace{\int q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}|\mathbf{z})d\mathbf{z}}_{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]} - \underbrace{\int q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}d\mathbf{z}}_{D_{KL}\left[q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right]}$$

$$\log p(\mathbf{x}) \geq \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction}} - \underbrace{D_{KL}\left[q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right]}_{\text{regularization}}$$

Auto-encoding



$$\mathcal{J}_{AE}(\theta) = \sum_{\mathbf{x}\in\mathcal{D}_n} \mathcal{L}\left(\mathbf{x}, d\left(e\left(\mathbf{x}\right)\right)\right)$$

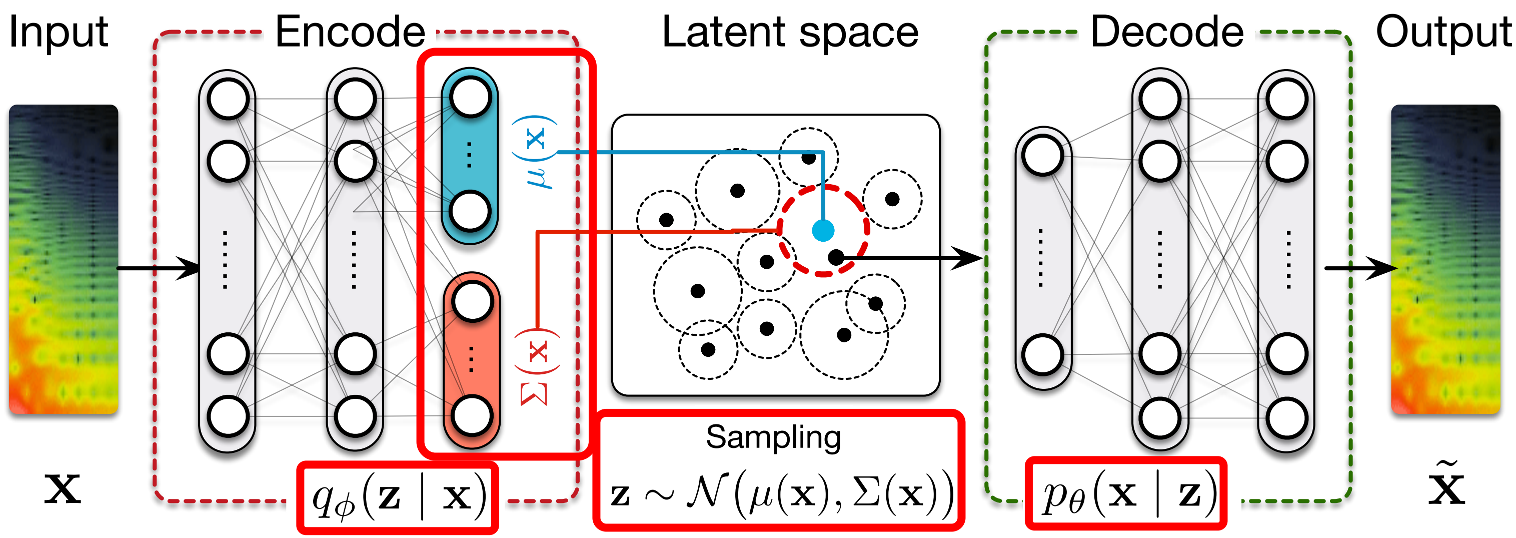
Very limited model because

- Defines a **deterministic mapping** $\tilde{\mathbf{x}} = d(e(\mathbf{x}))$
- No robust **theoretical properties on the latent** code ($\mathbf{z}$ almost decorative)
- Very **limited generalization** abilities (does not model $p(\mathbf{x}, \mathbf{z})$)

How can we model the intractable $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$

Introducing variational inference ☺

Variational auto-encoders



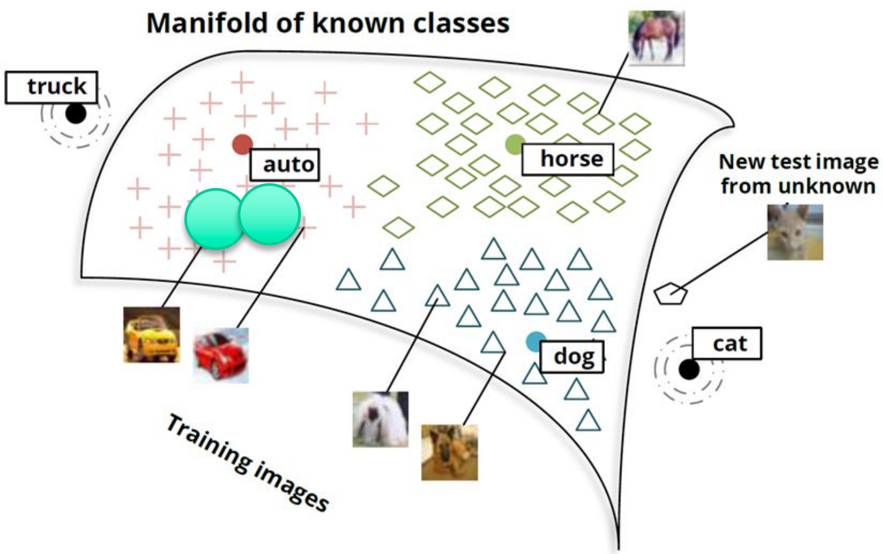
We want to minimize approximation error $q^*(\mathbf{z}|\mathbf{x}) = \underset{q(\mathbf{z}|\mathbf{x})\in\mathcal{Q}}{arg\,min} \mathcal{D}_{KL}\left[q\left(\mathbf{z}|\mathbf{x}\right) \parallel p\left(\mathbf{z}|\mathbf{x}\right)\right]$

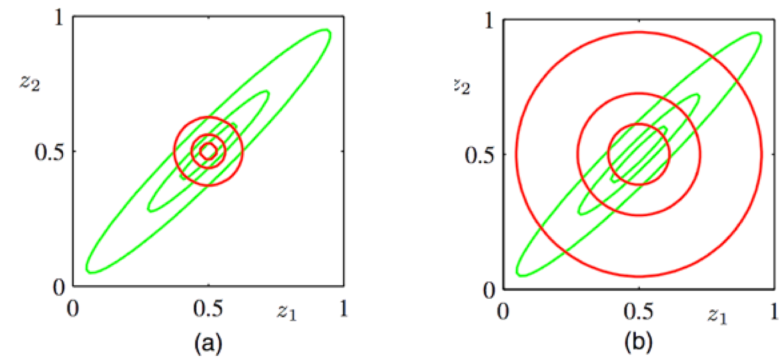
Encode a sound using $q_\phi(\mathbf{z}|\mathbf{x})$ to find the mean and variance in the space
Sample from this gaussian distribution to obtain $\mathbf{z} \sim \mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$
Decode the corresponding position using $p_\theta(\mathbf{x}|\mathbf{z})$

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{reconstruction}} - \beta \cdot \underbrace{D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z})\right]}_{\text{regularization}}$$

[Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ICLR 2015*, *1050*, 1.]

Variational auto-encoders





optimizing

$$D_{KL}\left[q(\mathbf{z}) \parallel p(\mathbf{z})\right]$$

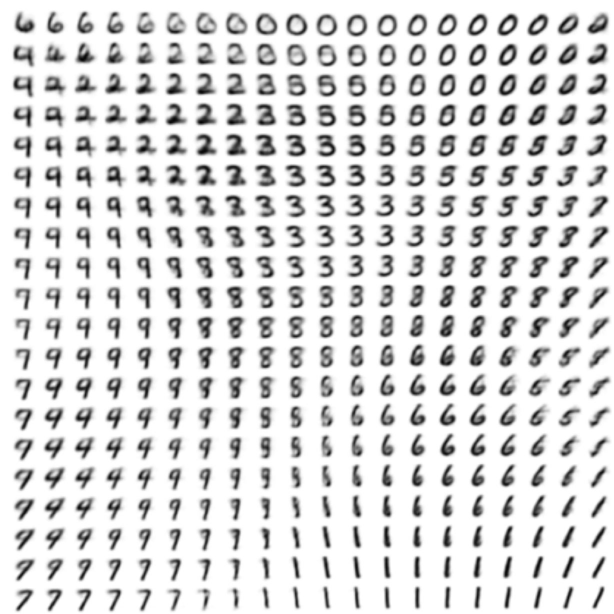
optimizing

$$D_{KL}\left[p(\mathbf{z}) \parallel q(\mathbf{z})\right]$$

Mean-field family $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(\mathbf{z}_j)$

[Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ICLR 2015*, *1050*, 1.]

Variational auto-encoders

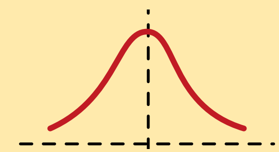


[Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ICLR 2015*, *1050*, 1.]

Normalizing flows

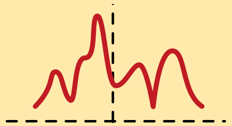
Probabilistic inference mostly deal with simple distributions

- Provide easier analytical solutions
- Implicit assumption of a simple explanation (capacity)



$\mathbf{z}_0 \sim \mathcal{N}(\mu, \sigma^2)$

The almighty Gaussian



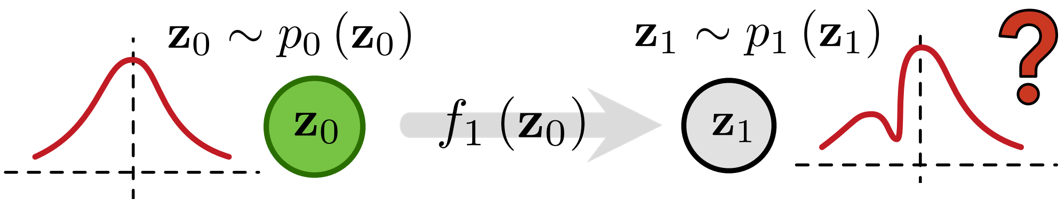
$\mathbf{z}_k \sim p_k(\mathbf{z}_k)$

Real data ?

- Too simplistic assumption in most cases
- Real data follows largely more complex distributions
- These distributions are usually intractable

How to use complex distributions but keep analytical simplicity ?

Given a random variable $\mathbf{z}_0 \sim p_0(\mathbf{z}_0)$ we can transform it to obtain



We want this to be a distribution

Only need it to sum (integrate) to 1 !

Change of volume is given by the determinant of the Jacobian $p_1(\mathbf{z}_1) = p_0(\mathbf{z}_0)\left|\det \frac{\partial f_1}{\partial \mathbf{z}_0}\right|^{-1}$

So in fact we can use any transform, but with a simple determinant of Jacobian

Planar flow proposed in the original paper

$$f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^\top \mathbf{z} + b)$$

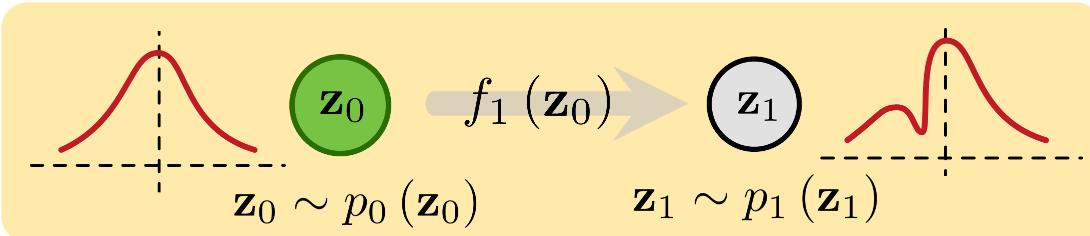
Given $\psi(\mathbf{z}) = h'(\mathbf{w}^\top \mathbf{z} + b)\mathbf{w}$

We have $\left|\det \frac{\partial f}{\partial \mathbf{z}}\right| = |1 + \mathbf{u}^\top \psi(\mathbf{z})|$

[Danilo Rezende and Shakir Mohamed, "Variational inference with normalizing flows," ICML Conference, 2015]

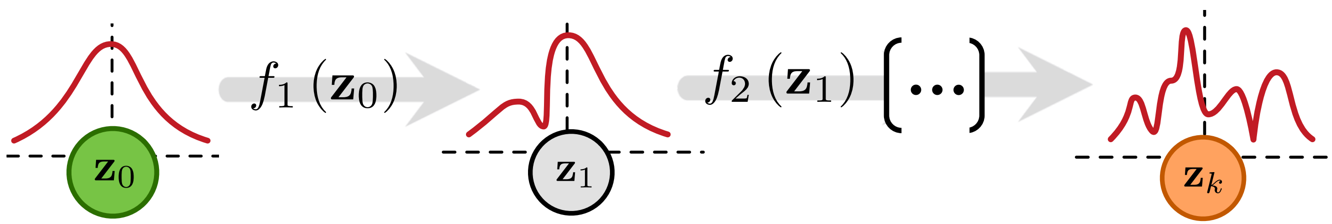
Normalizing flows

Now we know how to transform $\mathbf{z}_0 \sim p_0(\mathbf{z}_0)$



$$p_1(\mathbf{z}_1) = p_0(\mathbf{z}_0)\left|\det\frac{\partial f_1}{\partial \mathbf{z}_0}\right|^{-1}$$

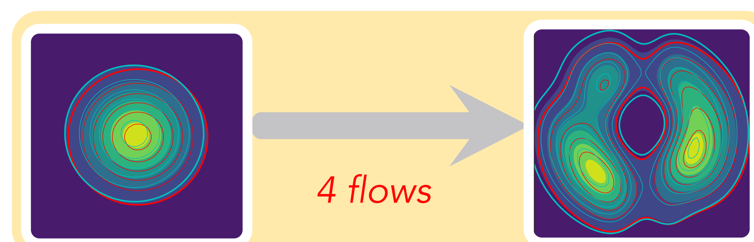
If we chain these transforms, so that $\mathbf{z}_k = f_k \circ \ldots \circ f_1(\mathbf{z}_0)$



We obtain the final distribution simply by reapplying the same reasoning

$$p_k(\mathbf{z}_k) = p_0(f_1^{-1} \circ \ldots \circ f_k^{-1}(\mathbf{z}_k)) \prod_{i=1}^{k}\left|\det\frac{\partial f_i^{-1}}{\partial \mathbf{z}_i}\right| = p_0(\mathbf{z_0}) \prod_{i=1}^{k}\left|\det\frac{\partial f_i}{\partial \mathbf{z}_{i-1}}\right|^{-1}$$

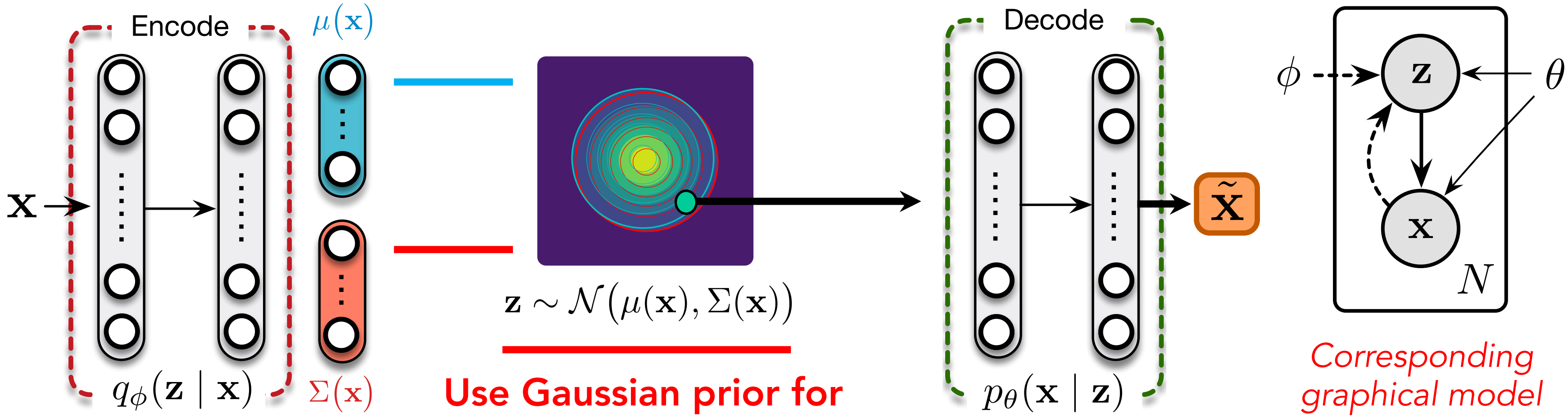
- Learn increasingly complex distributions
- Just applying simple invertible transforms
- Of course it holds for multiple dimensions



[Danilo Rezende and Shakir Mohamed, "Variational inference with normalizing flows," ICML Conference, 2015]

Variational Auto-Encoders (VAE)

A (*very brief*) recap on Variational Auto-Encoders (VAE)



Use Gaussian prior for analytical simplicity

Too simplistic for real data distributions

Postulate variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$

We want to minimize approximation error $q^*(\mathbf{z}|\mathbf{x}) = \underset{q(\mathbf{z}|\mathbf{x})\in\mathcal{Q}}{arg\,min} \mathcal{D}_{KL}\left[q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}|\mathbf{x})\right]$

$$\mathcal{L}_{\theta,\phi} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{reconstruction}} - \beta \cdot \underbrace{D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z})\right]}_{\text{regularization}}$$

Encode the input using $q_\phi(\mathbf{z}|\mathbf{x})$ to find the mean and variance

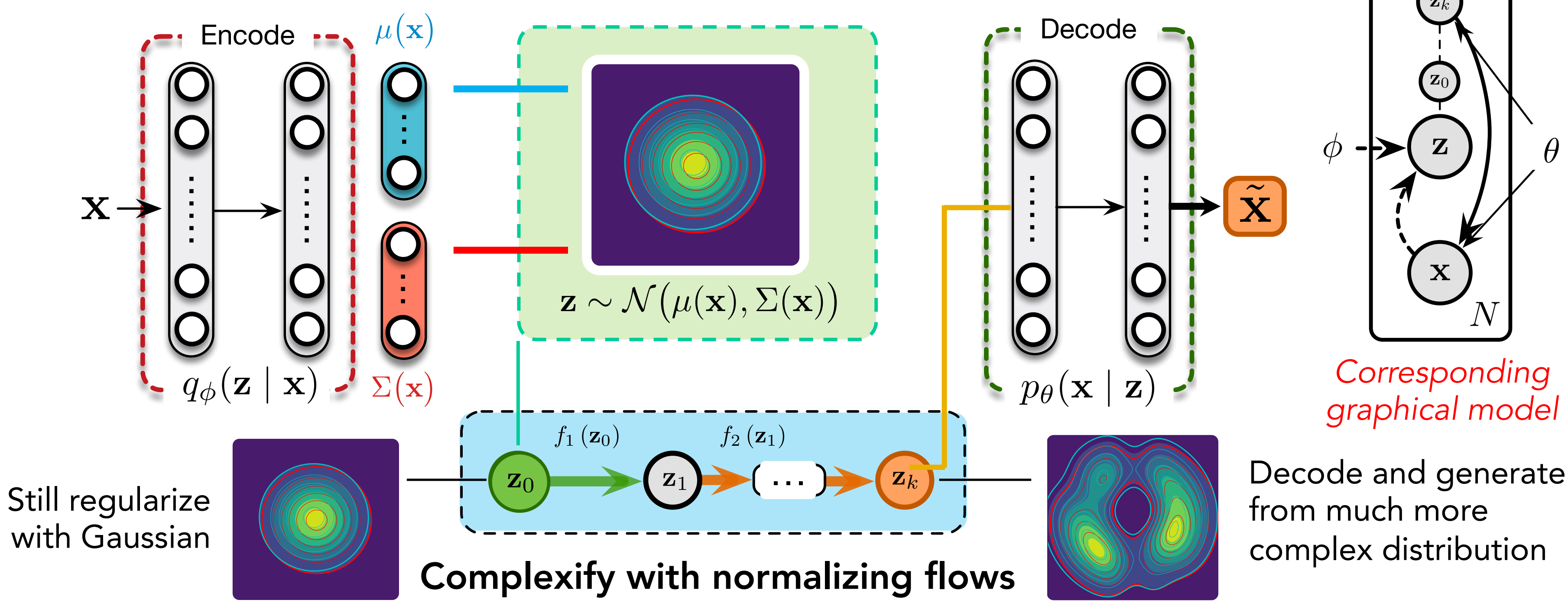
Sample from this gaussian distribution to obtain $\mathbf{z} \sim \mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$

Decode the corresponding position using $p_\theta(\mathbf{x}|\mathbf{z})$

[Derdrik P. Kingma and Max Welling "*Auto-encoding variational Bayes*" ICML Conference, 2014]

Normalizing flows in variational inference

Alleviate this limitation and model more complex distributions



Introduce normalizing flow in the optimization objective

$$\mathcal{L} = \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\ln q_0(\mathbf{z}_0)\right] - \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\log p(\mathbf{x}, \mathbf{z}_K)\right] - \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\sum_{i=1}^{k} \log \left|\det \frac{\partial f_i}{\partial \mathbf{z}_{i-1}}\right|\right]$$

Still regularize on a simple prior (Analytical Gaussian simplicity)

Decode from the complex distribution

Ensure that all latent are distributions (normalizing flow)