Introduction to Deep Generative Modeling

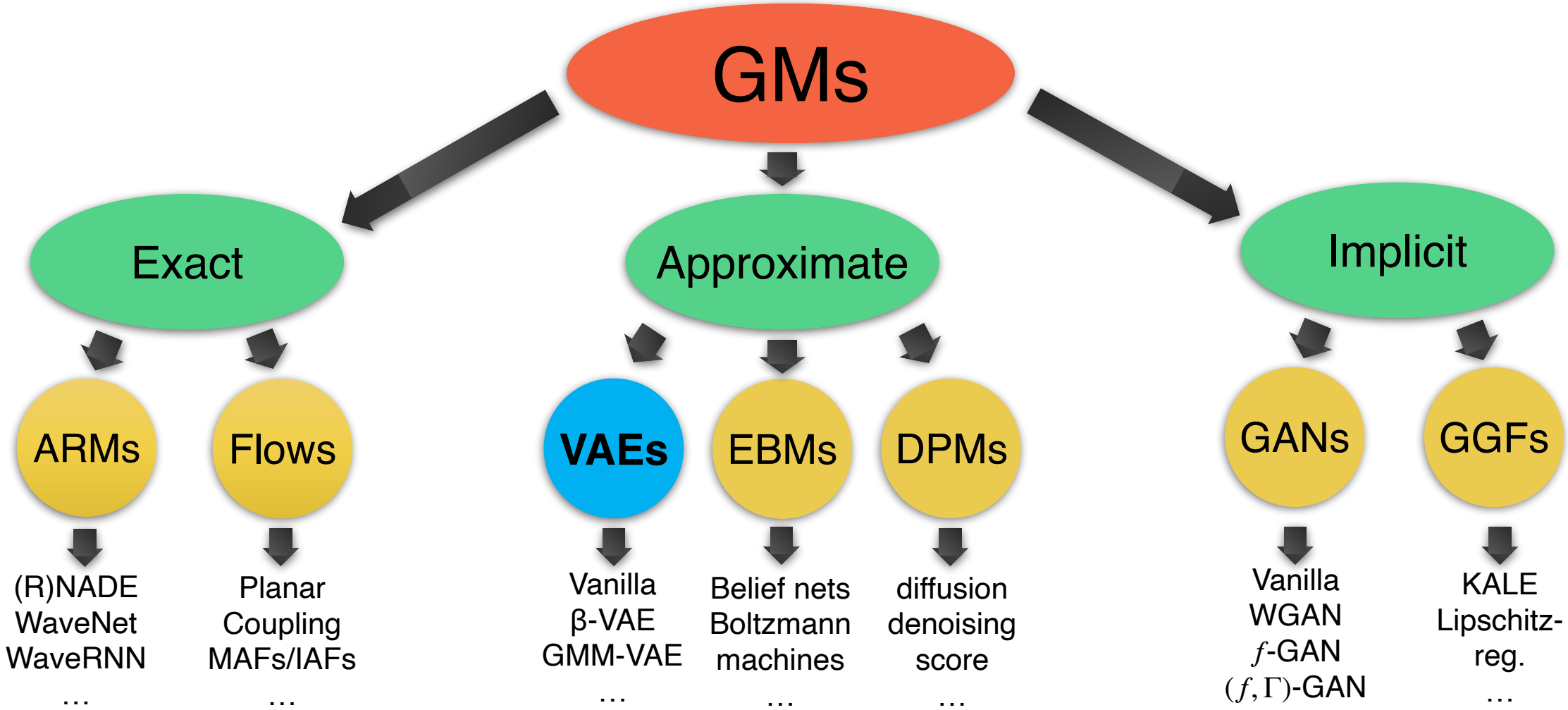
Lecture #11

HY-673 – Computer Science Dep., University of Crete

Professors: Yannis Pantazis, Yannis Stylianou

Teaching Assistant: Michail Raptakis

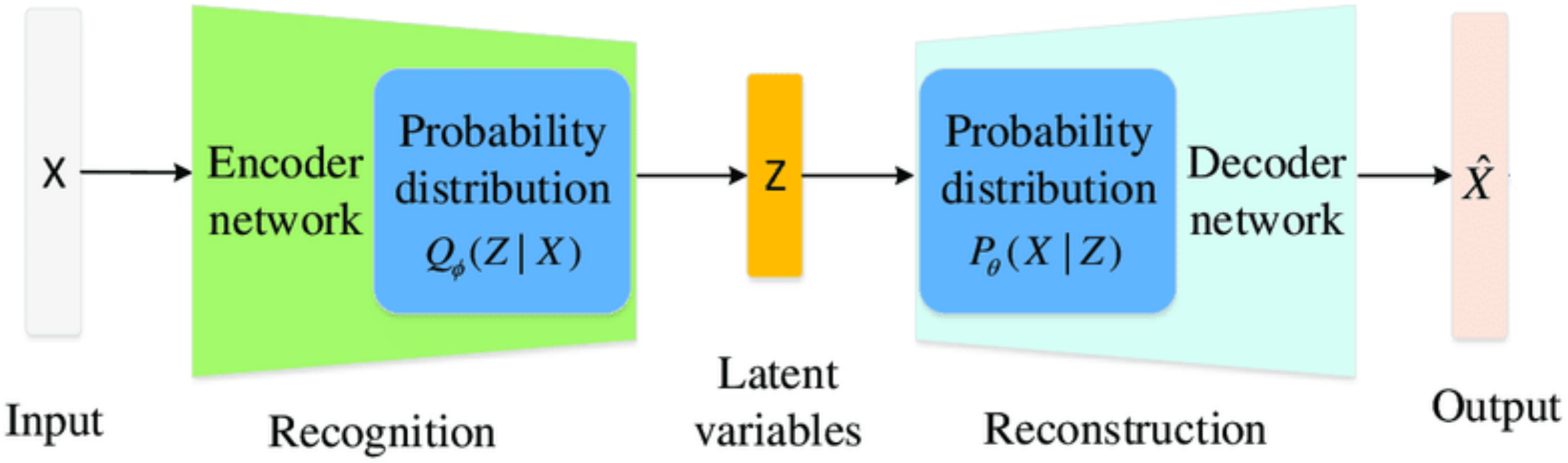
Taxonomy of GMs



Recap: Latent Variable Models

- Latent Variable Models:
  - Allow us to define complex models $p_\theta(x)$ in terms of simple building blocks $p_\theta(x|z)$.
  - Natural for unsupervised learning tasks (clustering, unsupervised representation learning, feature disentanglement, etc.).
  - No free lunch: Much more difficult to learn compared to fully observed autoregressive models.

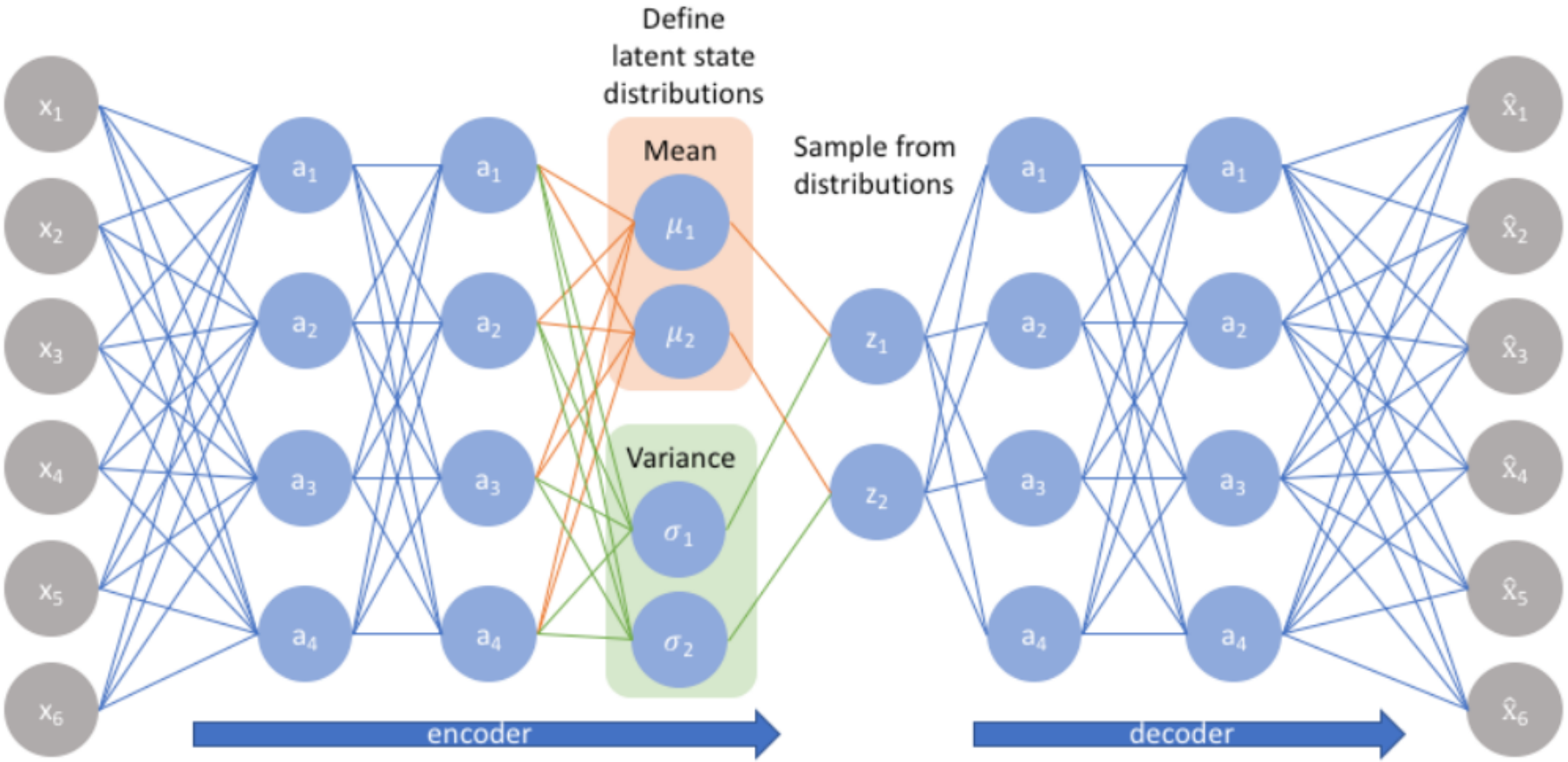
Recap: Variational Inference



- *Key idea in variatioanl inference.* approximate the intractable posterior with a (parametric) inference model:

$$q_\phi(z|x) \approx p_\theta(z|x)$$

Recap: A Concrete VAE



Prior, likelihood, approximate posterior: all **Gaussians**.

Recap: ELBO

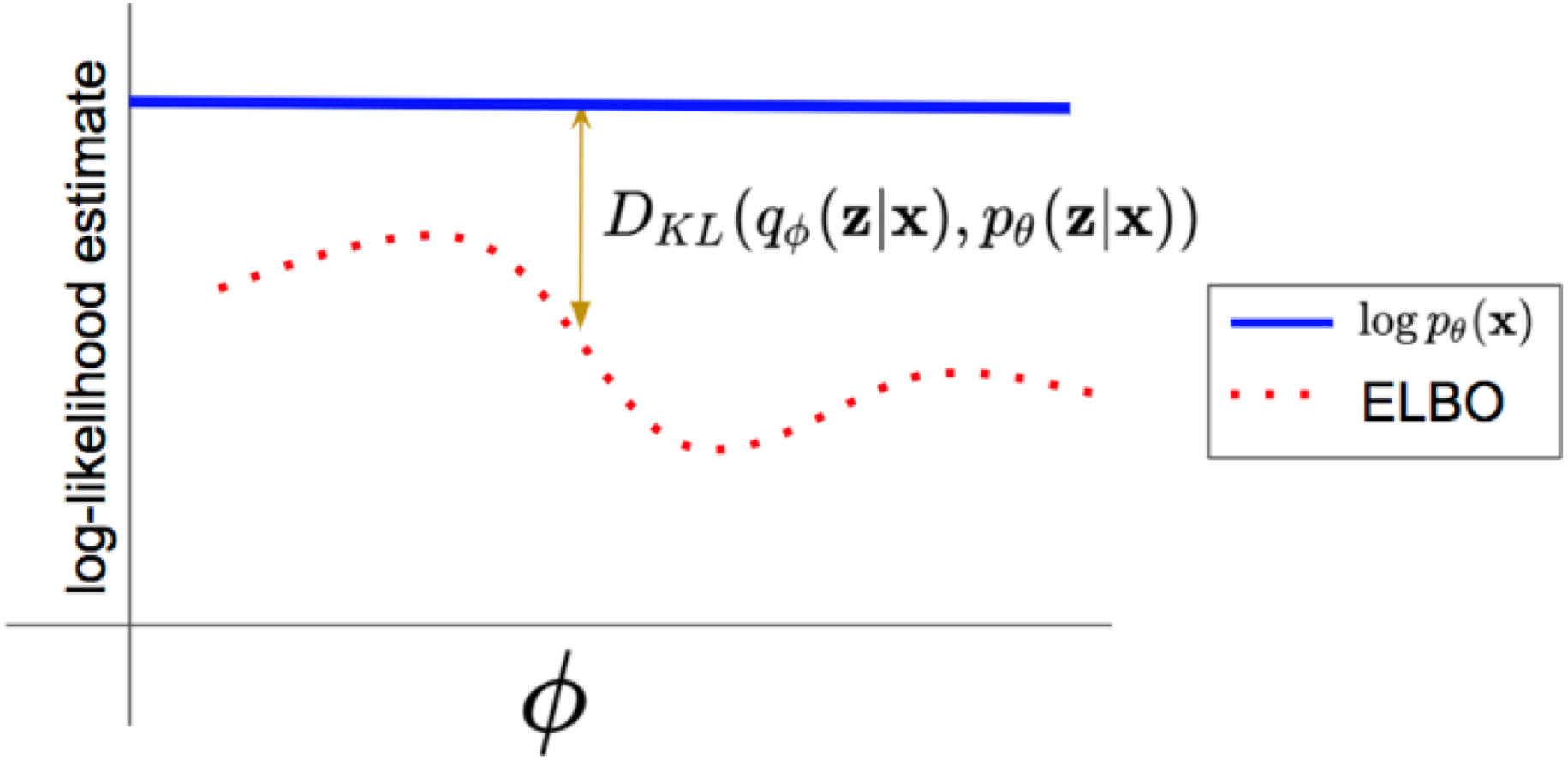
- Evidence or the (marginal) likelihood for a single data x equals to

$$\log p_\theta(x) = \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]}_{= \mathcal{L}_{\theta,\phi}(x)\ (\text{ELBO})} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]}_{= D_{KL}(q_\phi(z|x)||p_\theta(z|x))}$$

- Since $D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \geq 0$, it holds that

$$\boxed{\log p_\theta(x) \geq \mathcal{L}_{\theta,\phi}(x)}$$

Recap: ELBO



The better $q_\phi(z|x)$ can approximate the posterior $p_\theta(z|x)$, the smaller $D_{KL}(q_\phi(z|x)||p_\theta(z|x))$ we can achieve, thus, the closer ELBO will be to $\log p_\theta(x)$.

Next: Jointly optimizer over θ and ϕ to maximize the ELBO over a dataset.

Amortized Variational Inference

Maximize ELBO w.r.t. both θ and ϕ:

$$\max_{\theta,\phi} \sum_{x_i \in \mathcal{D}} \mathcal{L}_{\theta,\phi}\left(x_i\right).$$

Amortization: When we learn a **single** parametric function f_ϕ that maps each x to a set of (good) parameters for the approximate posterior distribution.

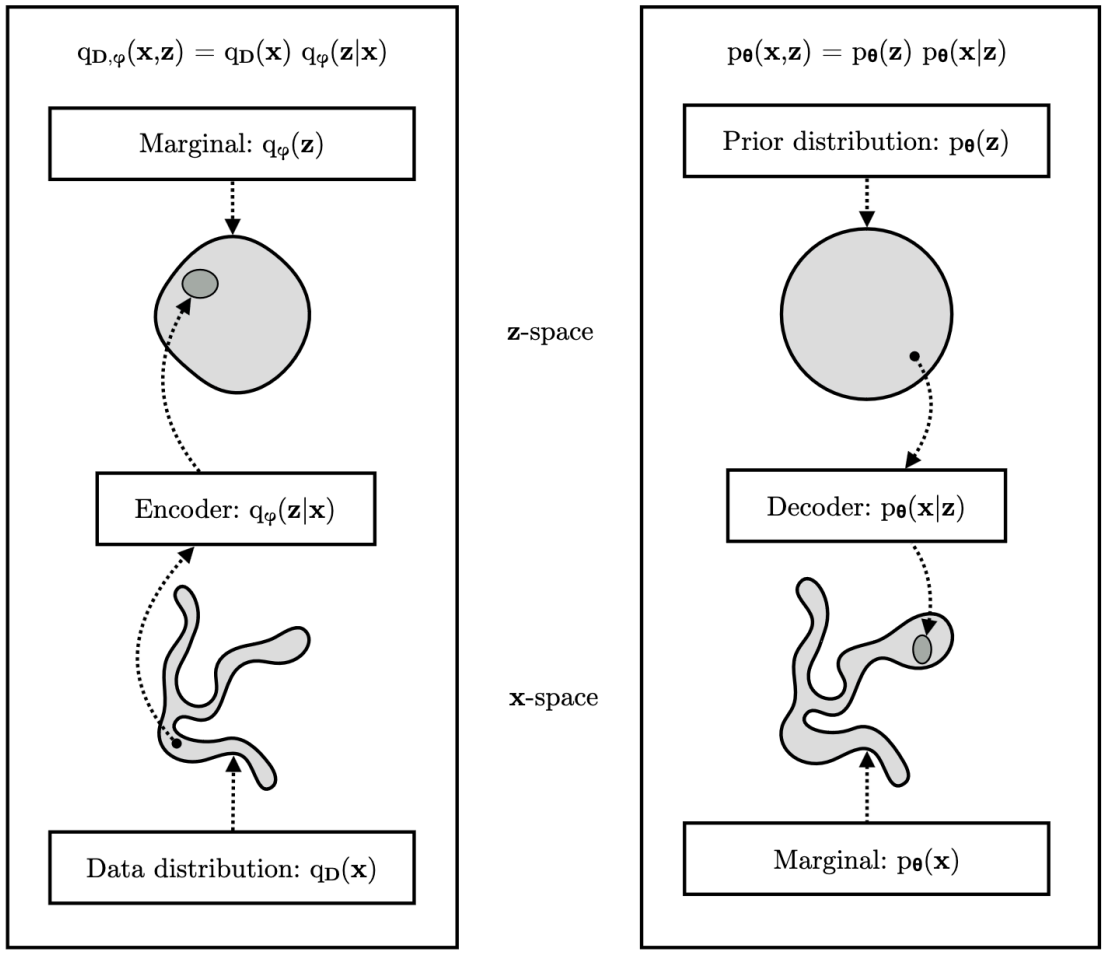
Rewriting ELBO

- ELBO is rewritten as

$$\mathcal{L}_{\theta,\phi}(x) := \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{\text{negative likelihood (reconstruction error)}} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(z)}{q_\phi(z|x)}\right]}_{= -D_{KL}(q_\phi(z|x)||p_\theta(z)) \text{ (regularization term)}}$$

- There are cases where $D_{KL}(q_\phi(z|x)||p_\theta(z))$ can be computed analytically (i.e., Gaussians).

ELBO as Kullback-Leibler Divergence (KLD)



$$\text{ML objective} = -\ D_{KL}(\ q_D(\mathbf{x})\ ||\ p_\theta(\mathbf{x})\ )$$
$$\text{ELBO objective} = -\ D_{KL}(\ q_{D,\varphi}(\mathbf{x},\mathbf{z})\ ||\ p_\theta(\mathbf{x},\mathbf{z})\ )$$

- ELBO can be written as the KL divergence between the joint distributions plus a cosntant.
- Racall likelihood maximization is equivalent to KL divergence minimization.
- ELBO can be viewed as a maximum likelihood objective in an *augmented space*.

Recap: Reparametrization Trick

- Want to compute a gradient with respect to ϕ of:

$$\mathbb{E}_{q_\phi(z|x)}[f(z)] = \int f(z) q_\phi(z|x) dz,$$

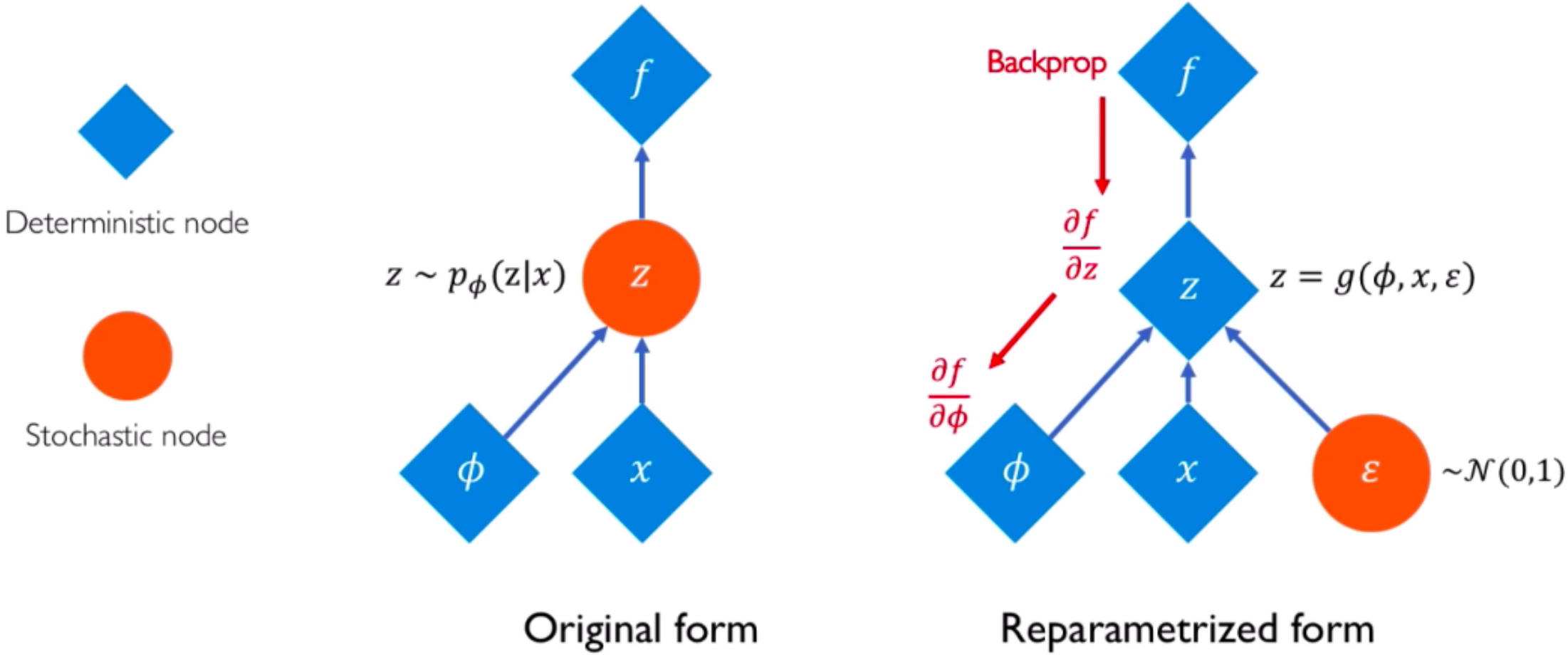
- Suppose $q_\phi(z|x) = \mathcal{N}\left(\mu_\phi(x), \text{diag}(\sigma^2_\phi(x))\right)$ is a Gaussian with $\mu_\phi(x), \sigma_\phi(x)$ be neural nets. It holds:

$$\mathbb{E}_{q_\phi(z|x)}[f(z)] = \mathbb{E}_{\epsilon \sim p(\epsilon)}\left[f(g(\epsilon, \phi, x))\right] := \int f(\mu_\phi(x) + \sigma_\phi(x)\epsilon) p(\epsilon) d\epsilon$$

- Thus, the gradient w.r.t. ϕ becomes

$$\nabla_\phi \mathbb{E}_{q_\phi(z|x)}[f(z)] = \nabla_\phi \mathbb{E}_{p(\epsilon)}\left[f(g(\epsilon, \phi, x))\right] = \mathbb{E}_{p(\epsilon)}\left[\nabla_\phi f(g(\epsilon, \phi, x))\right] \approx \nabla_\phi f(g(\epsilon_i, \phi, x_i))$$

Recap: Reparametrization Trick



A Concrete VAE

1. Prior distribution is isotropic/spherical Gaussian (independent of θ): $p(z) = \mathcal{N}(0, I)$.

2. Stochastic decoder is Gaussian: $p_\theta(x|z) = \mathcal{N}\left(\mu_\theta(z), \text{diag}(\sigma^2_\theta(z))\right)$ where $\mu_\theta(z), \sigma_\theta(z)$ are the outputs of a neural network.

3. Stochastic encoder (i.e., inference or recognition model) is Gaussian: $q_\phi(z|x) = \mathcal{N}\left(\mu_\phi(x), \text{diag}(\sigma^2_\phi(x))\right)$ where $\mu_\phi(x), \sigma_\phi(x)$ are also the outputs of a neural network.

A Concrete VAE – ELBO statistical (i.e., Monte Carlo) approximation

Computation of unbiased estimate of single-datapoint (x) ELBO

1. $(\mu, \log(\sigma)) \leftarrow \text{EncoderNN}_\phi(x)$
2. $\epsilon \sim \mathcal{N}(0, I_{d'})$
3. $z = \epsilon \odot \sigma + \mu$
4. $\mathcal{L}_{q_\phi(z|x)} \leftarrow -\frac{1}{2}\sum_{i=1}^{d'}\left(z_i^2 + \log(2\pi) + \log(\sigma_i)\right)$
5. $\mathcal{L}_{p_\theta(z)} \leftarrow -\frac{1}{2}\sum_{i=1}^{d'}\left(z_i^2 + \log(2\pi)\right)$
6. $(\bar{\mu}, \log(\bar{\sigma})) \leftarrow \text{DecoderNN}_\theta(z)$
7. $\mathcal{L}_{p_\theta(x|z)} \leftarrow -\frac{1}{2}\sum_{i=1}^{d}\left((x_i - \bar{\mu}_i)^2/\bar{\sigma}_i^2 + \log(2\pi) + \log(\bar{\sigma}_i)\right)$
8. $\mathcal{L} \leftarrow \mathcal{L}_{p_\theta(z)} + \mathcal{L}_{p_\theta(x|z)} - \mathcal{L}_{q_\phi(z|x)}$

Another Concrete VAE

1. Prior distribution is isotropic/spherical Gaussian (independent of θ): $p(z) = \mathcal{N}(0, I)$.

2. Stochastic decoder is factorized Bernoulli: $p_\theta(x|z) = \mathcal{B}\left(r_\theta(z)\right)$ where $r_\theta(z) \in [0,1]^d$ is the output of a neural network.

3. Stochastic encoder (i.e., inference or recognition model) is Gaussian: $q_\phi(z|x) = \mathcal{N}\left(\mu_\phi(x), \text{diag}(\sigma^2_\phi(x))\right)$ where $\mu_\phi(x), \sigma_\phi(x)$ are also the outputs of a neural network.

A Concrete VAE – ELBO statistical (i.e., Monte Carlo) approximation

Computation of unbiased estimate of single-datapoint ELBO

1. $(\mu, \log(\sigma)) \leftarrow \text{EncoderNN}_\phi(x)$
2. $\epsilon \sim \mathcal{N}(0, I_{d'})$
3. $z = \epsilon \odot \sigma + \mu$
4. $\mathcal{L}_{q_\phi(z|x)} \leftarrow -\frac{1}{2} \sum_{i=1}^{d'} \left(z_i^2 + \log(2\pi) + \log(\sigma_i)\right)$
5. $\mathcal{L}_{p_\theta(z)} \leftarrow -\frac{1}{2} \sum_{i=1}^{d'} \left(z_i^2 + \log(2\pi)\right)$
6. $r \leftarrow \text{DecoderNN}_\theta(z)$
7. $\mathcal{L}_{p_\theta(x|z)} \leftarrow \sum_{i=1}^{d} \left(x_i \log(r_i) + (1 - x_i) \log(1 - r_i)\right)$
8. $\mathcal{L} \leftarrow \mathcal{L}_{p_\theta(z)} + \mathcal{L}_{p_\theta(x|z)} - \mathcal{L}_{q_\phi(z|x)}$

Four Points for Intervention

We can design:

1. The prior distribution, $p_\theta(z)$,
2. The conditional (likelihood), $p_\theta(x|z)$,
3. The approximate posterior, $q_\phi(z|x)$ and
4. The "loss" function, $-\mathcal{L}$.

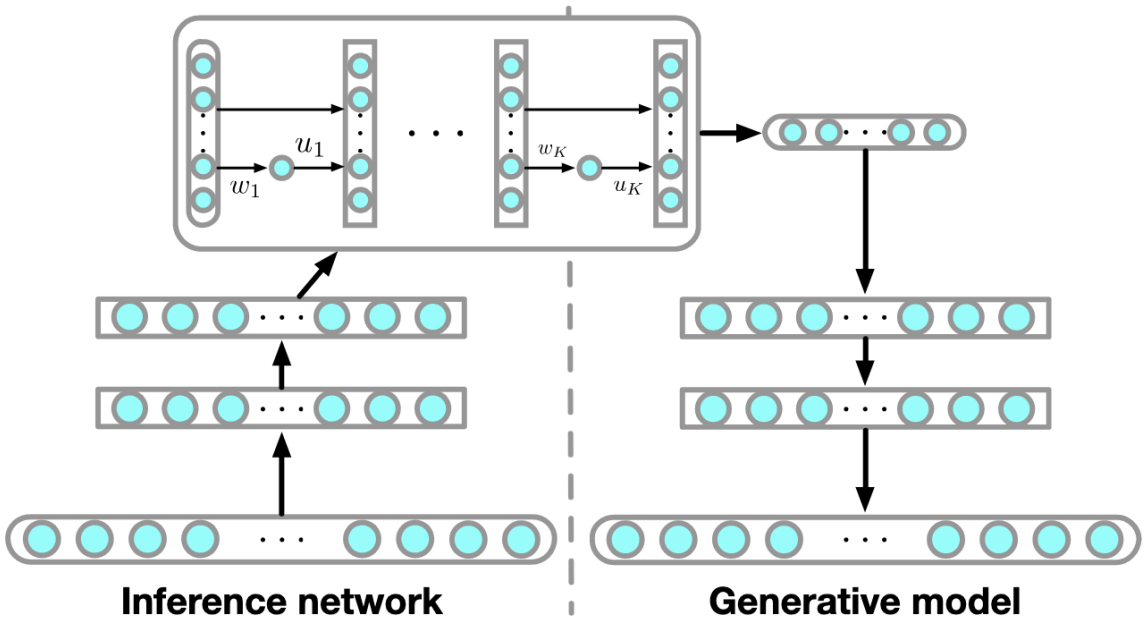
Changing the prior distribution

We can choose $p_\theta(z)$ between

- Fully Gaussian

$$z = (L \odot L_{\text{mask}} + \text{diag}(\sigma))\,\epsilon + \mu$$

- Various normalizing flows
  - Planar flows
  - Inverse autoregressive flows (IAFs)



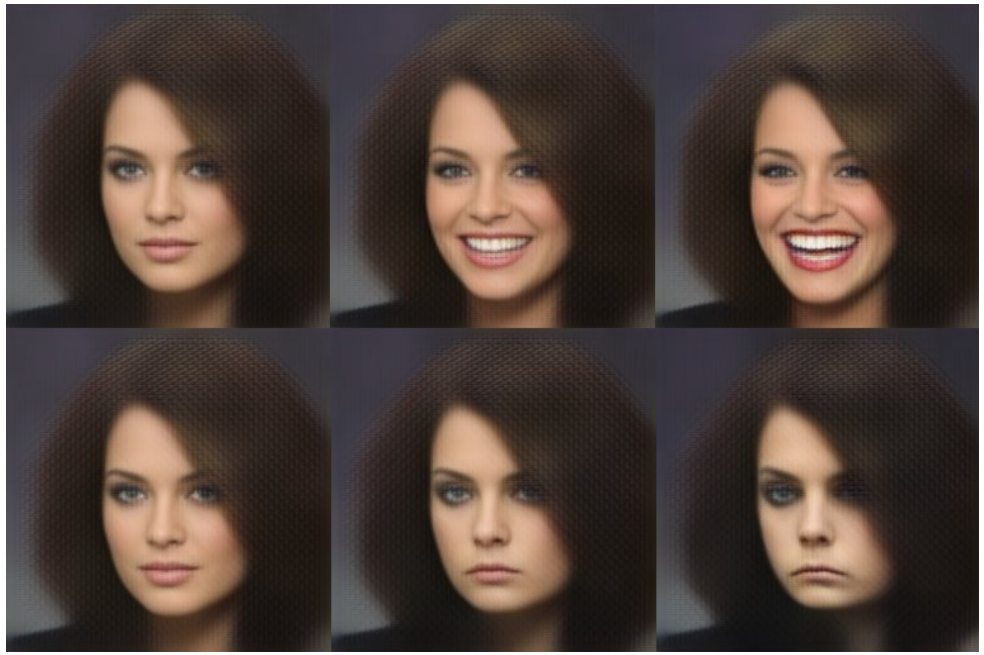
Common property: They have an analytic (i.e., tractable) probability density function.

Changing the Prior Distribution

- Till now, we have discussed only for continuous latent variables.
- However, discrete latent variables are also important (e.g., GMMs).
- Three popular solutions are:
  - VQ-VAE: Neural Discrete Representation Learning, Oord et al.
  - The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables, Maddison et al.
  - Categorical Reparameterization with Gumbel-Softmax, Jang et al.

Changing the "Loss"

- An attractive property of VAEs is that they are capable of constructing independent features.
  - Also known as **disentagled representations.**



Smile feature:
from neutral to happy to sad

- Are there systematic ways to guide the training towards constructing indepedent features?

Changing the "Loss"

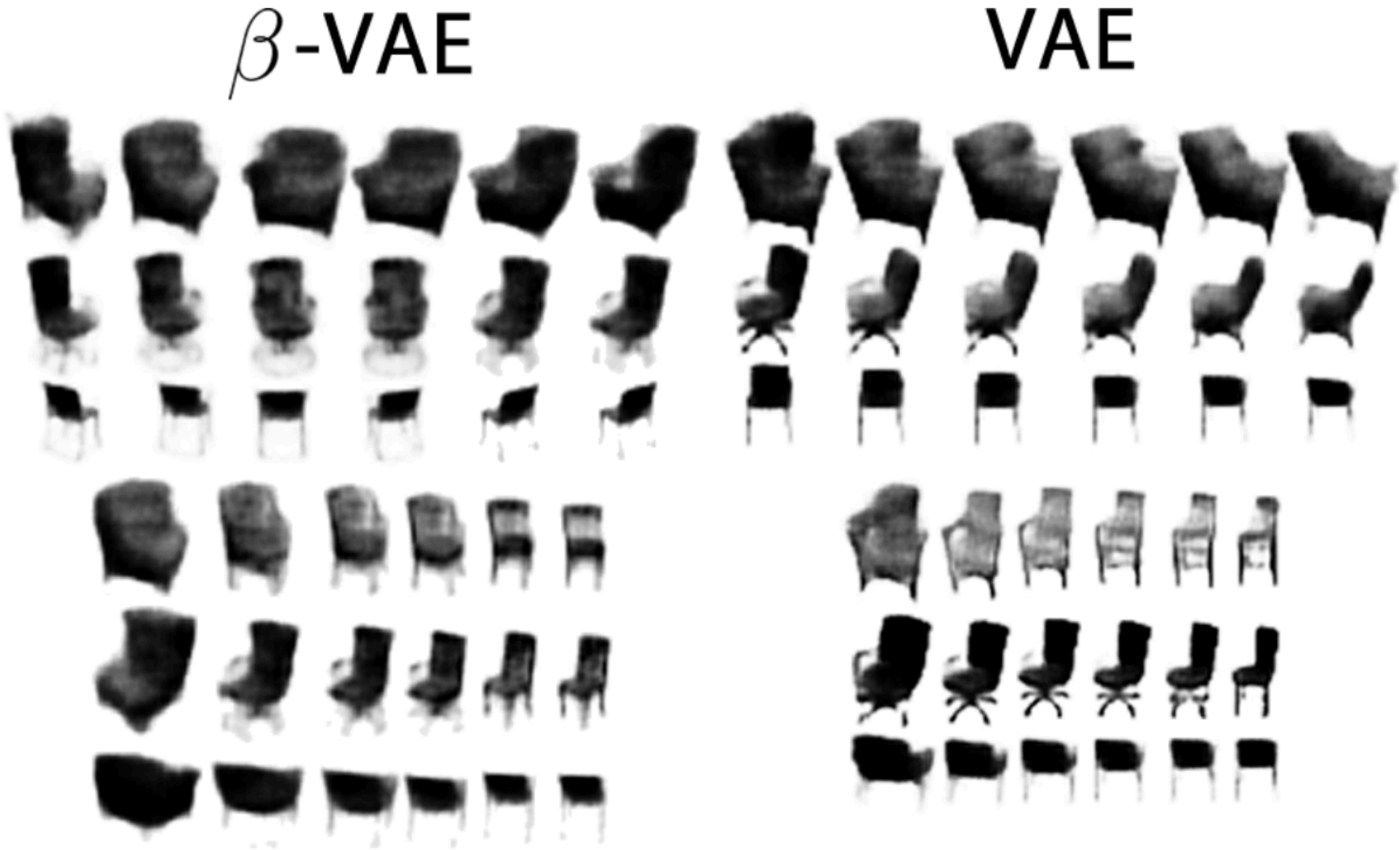
- β-VAE puts more weight to the regularization term of ELBO

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

- InfoVAE additionally tries to maximize the *Mutual Information* between the data x and the latent variable z.

$$\bar{\mathcal{L}}_{\theta,\phi} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta D_{KL}(q_\phi(z|x)||p_\theta(z)) + \alpha\mathcal{I}(x,z)$$

Changing the “Loss”



VAE-Generated Images are Blurry

- One possible explanation is that we perform maximization of the ELBO and not of the log-likelihood.
  - ELBO's gap is large.
- Blurriness can be alleviated to some degree by choosing a sufficiently flexible inference model (i.e., encoder) and/or a sufficiently flexible generative model (i.e., decoder).

Posterior Collapse

- *Posterior collapse* or variational overpruning may occur due to: Sufficiently powerful decoders which ignore latent codes due to the tradeoff between recostruction error and KL prior penalty minimization (Bowman et al., 2015, Chen et al., 2016, Zhao et al., 2017, Alemi et al., 2018).
- Sufficiently powerful decoders means that

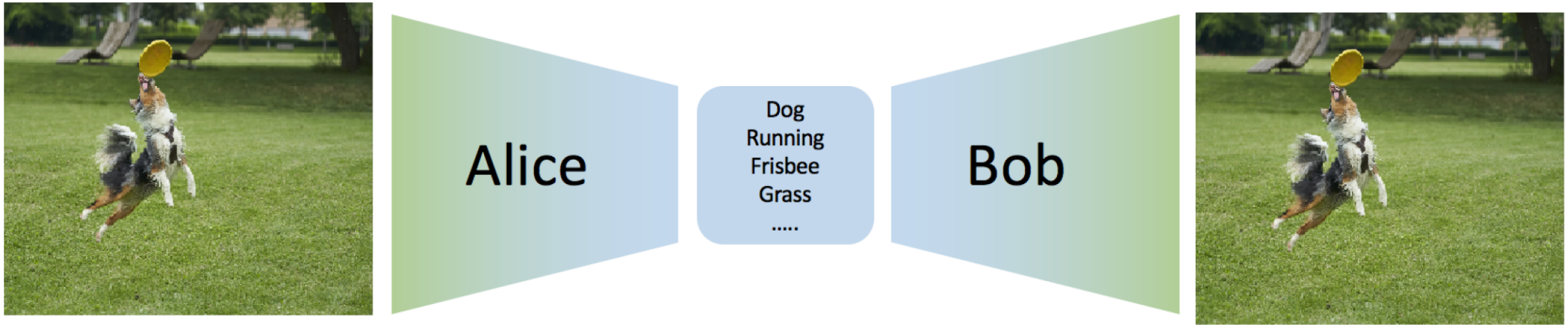
$$\exists \theta^* \text{ such that } \; p_{\theta^*}(x|z) = p_d(x)$$

- Some proposed solutions:
  - δ-VAE (δ adjusts the prior's parameters).
  - Free bits (guarantees a minimum KL prior penalty).
  - Adding skip connections.

Selected Applications of VAEs

1. Multi-modal VAE (both x_1 and x_2 are generated from z).
2. VAE with missing data.
3. Hierarchical VAE (z has an hierarchical structure).
4. VQ-VAE-2: Generates large-scale, diverse and high-fidelity images.
5. VAEs with sequential encoders/decoders for text generation, sketch drawing, molecular design, etc.
6. β-VAE for disentangled latent feature construction in underwater source localization.

Learning Deep Generative Models

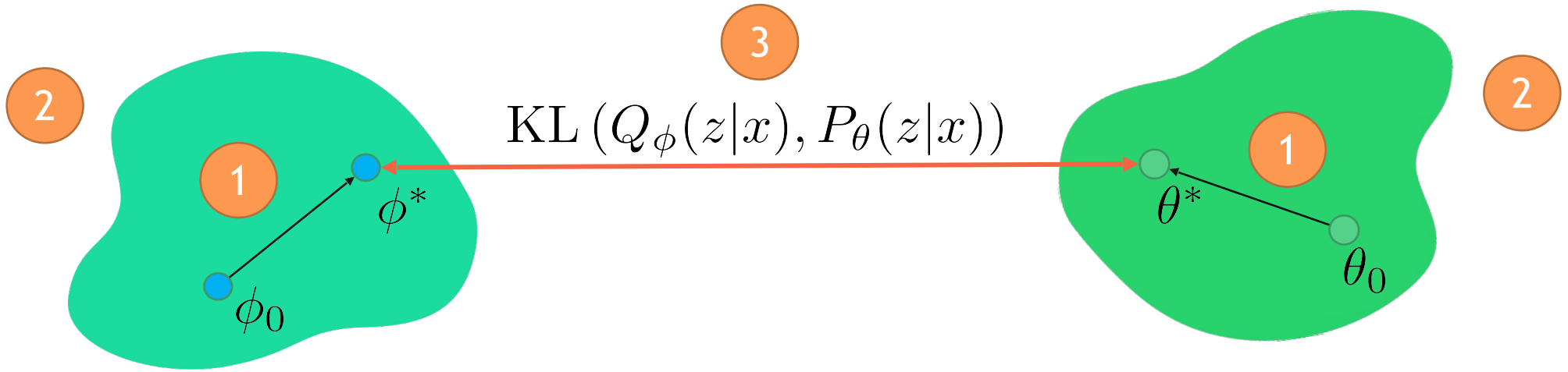


- This scheme works well if $\mathbb{E}_{q_\phi(z|x)}\left[\log p(x|z;\theta)\right]$ is large.
- The term $D_{\mathrm{KL}}\left(q_\phi(z|x)||p(z)\right)$ forces the distribution over messages to have a specific shape $p(z)$. If Bob knows $p(z)$, then he can generate realistic messages $\hat{z} \sim p(z)$ and the corresponding image, as if he had received them from Alice!

Summary of VAEs

1. Combine two relatively simple models to get a more flexible one: $p_\theta(x) = \int p_\theta(x, z)dz$.
2. Directed model permits ancestral sampling (efficient generation): $z \sim p_\theta(z),\ x \sim p_\theta(x|z)$.
3. However, log-likelihood is generally intractable, hence learning is difficult.
4. Joint learning of a model θ and an amortized inference component ϕ to achive tractability via ELBO optimization.
5. Latent representations for any x can be inferred via $q_\phi(z|x)$.
6. Latent representations may reveal high-level controllable features of the data.
7. Some issues with the quality of the generated samples as well as with posterior collapse exist.

Research Directions



Improving variational learning via:

1. Better optimization techniques.
2. More expressive techniques.
3. Alternate loss functions.

Model Families - Encoder

Amortization (Gershman & Goodman, 2015; Kingma; Rezende; ...)

- Scalability: Efficient learning and inference on massive datasets.
- Regularization effect: Because of joint training, it also implicitly regularizes the model θ (Shu et al., 2018)..

Augmenting variational posteriors:

- Monte Carlo methods: Importance Sampling (Burda et al., 2015), MCMC (Salimans et al., 2015, Hoffman, 2017, Levy at al., 2018), Sequential Monte Carlo (Maddison et al., 2017, Le et al., 2018, Naesseth et al., 2018), Rejection Sampling (Grover et al., 2018).
- Normalizing flows (Rezende & Mohammed, 2015, Kingma et al., 2016).

Model Families - Decoder

- Powerful Decoders $p(x|z;\theta)$ such as DRAW (Gregor et al., 2015), PixelCNN (Gulrajani et al., 2016).
- Parametrized, learned priors $p(z;\theta)$ (Nalusnich et al., 2016, Tomczak & Welling, 2018, Graves et al., 2018).

Variational Objectives & References

Tighter ELBO does **not** imply:

- Better samples: Sample quality and likelihoods are uncorrelated (Theis et al., 2016).
- Informative latent codes: Powerful decoders can ignore latent codes due to tradeoff in minimizing recostruction error vs. KL prior penaly (Bowman et al., 2015, Chen et al., 2016, Zhao et al., 2017, Alemi et al., 2018).

Alternatives to KL divergence:

- Renyi's alpha-divergences (Li & Turner, 2016).
- Integral probability metrics, such as, maximum mean discrepancy, Wasserstein distance (Dziugaite et al., 2015; Zhao et. al 2017; Tolstikhin et al., 2018).

References: https://deepgenerativemodels.github.io

References

1. Probabilistic Machine Learning: Advanced Topics (*Chapter 20*)
   Kevin P Murphy, The MIT Press (2023)

2. An Introduction to Variational Autoencoders, D. Kingma & M. Welling, Foundations and Trends in ML, 2019. (A coherent and accessible introduction to variational autoencoders - Highly recommended read!)
   `https://arxiv.org/abs/1906.02691`

3. Auto-Encoding Variational Bayes, D. Kingma & M. Welling, ICLL, 2014.

4. `https://github.com/matthewvowels1/Awesome-VAEs`

5. `https://deepgenerativemodels.github.io`

Introduction to Deep Generative Modeling

Lecture #11

HY-673 – Computer Science Dep., University of Crete

Professors: Yannis Pantazis, Yannis Stylianou

Teaching Assistant: Michail Raptakis