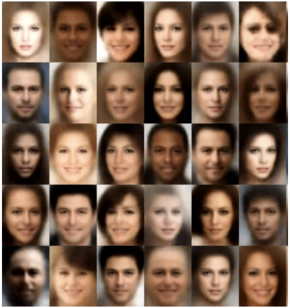
Why is this a classic?

- Merges deep learning and probabilistic models
- Introduces new techniques: **amortized inference**, **reparametrisation trick** and the use of lower bound **(ELBO)** to jointly optimise encoder and decoder
- Competing work by Razende 2014.

Generative modelling



- Probabilistic models
- We assume that data points are i.i.d. samples from a probability density function $x \sim p(x)$ over some space X.
- Generative modelling involves sampling new data from $p(x)$.

Key idea

Approximate $p(x)$ with a parametrised density $q_\theta(x)$, then optimise θ by minimising a distributional loss θ.

Typically $\mathcal{L}(\theta) = D(p||q_\theta)$ is defined by a divergence (such as KL), rather than a distance, as the metric structure of X is typically unknown.

Divergences are functions that satisfy

$$\begin{aligned} D(p||q) &\geq 0 \quad \text{(non-negativity)} \\ D(p||q) &= 0 \iff p = q \quad \text{(positivity)} \end{aligned}$$

Note that symmetry, i.e., $D(p||q) \neq D(q||p)$ and triangle inequality are not satisfied in general. So divergences are not a distances.

Examples

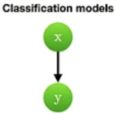
Generating . . .

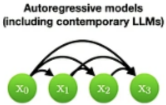
- the next word in a sentence involves sampling from $p(x_n|x_{n-1}, \ldots, x_{n-k})$
- a 3D molecule from an amino-acid sequence
- an image from a noise input and text prompt
- trajectory of prosthetic arm $y(t)$ from neural signals $x(t)$

Life before VAEs: large-scale supervised learning

The classic paradigm before VAE was to perform supervised learning in signal space X.

We can understand this as a method for learning graphical models





Maximum likelihood estimation

Assume $q_\theta(x)$ is a deep NN. For i.i.d samples $x_i \sim q_\theta(x)$, the MLE becomes

$$
\begin{aligned}
\theta^* &= \arg\max_\theta \prod_{i=1}^{N} q(x_i) \quad \text{(i.i.d.)} \\
&= \arg\max_\theta \sum_{i=1}^{N} \log q(x_i) \\
&= \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} -\log q(x_i) \\
&\approx \arg\min_\theta \mathbb{E}_{N\to\infty, x\sim p}[-\log q(x)] \quad \text{(l.l.n.)}
\end{aligned}
$$

Taking the KL divergence between q and the 'true' model p

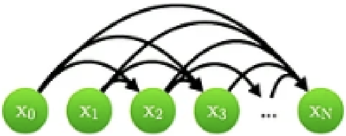
$$
\begin{aligned}
D_{KL}(p||q) &= \int p(x)\log p(x)dx - \int p(x)\log q(x)dx \\
&= \mathbb{E}_{x\sim p}[\log p(x) - \log q(x)] \geq 0
\end{aligned}
$$

Thus:

- the MLE approximates the 'true' model
- minimising KL is equivalent to maximising the log-likelihood

The issue with MLE

- Generation directly in high-D **signal** space X is intractable - most points do not yield valid samples
  - But... can use autoregression, predicting one dimension at a time, e.g., LLMs, but this does not scale to high-D

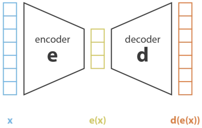


Introducing a latent variable and optimise in latent space

- Find a low-D **code** space Z, which ‘parametrises’ the signal
- One way to achieve this is using an autoencoder that produces latent embeddings $z = e(x)$, such that $||x - d(e(z))||$ is minimised

Questions/issues:

- Overfitting - What should the geometry and dimension of latent space be?
- Lossy compression - What signal features should we care about?



Hence autoencoders (on their own) do not suffice to generate new examples.

Deep latent variable models

- observed data x, latent (unseen) variables $z \implies$ probabilistic model $p(x, z) = p(x|z)p(z)$, where $p(x|z)$ is the likelihood, often called the generative model
- Advantages: fast sampling through factorisation, potentially interpretable z and controlled generation

MLE approach: The marginal log likelihood becomes

$$p(x) = \prod_{x_i} \sum_{z} p(x_i, z)$$

$$\log(p(x)) = \sum_{x_i} \log(\sum_{z} p(x_i, z))$$

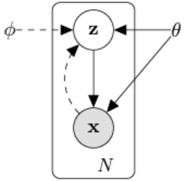
This is only tractable if z is discrete and can take a few values, but not in general.

MAP approach: via a variational Bayesian approach. Approximate the posterior by a free energy-based models $q_\theta(z|x) = e^{f_\theta(z)}/Z_\theta$, where f is an energy function (NN). Thus, Z_θ is intractable in general.

VAEs

Assume, the prior and likelihood (aka. generative model) have parametric forms $p_\theta(z)$ and $p_\theta(x|z)$.

Amortization: introduce **inference** (aka. posterior/encoder/recognition) model (NN) $q_\phi(z|x)$ to approximate $p_\theta(z|x)$. This posterior links data and model.



Decomposing the marginal likelihood

$$
\begin{aligned}
D_{KL}(q_\phi(z|x) \| p_\theta(z|x)) &= \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} dz \\
&= - \int_z q_\phi(z|x) \log \frac{p_\theta(z|x)}{q_\phi(z|x)} dz \\
&= - \int_z q_\phi(z|x) \log \frac{p_\theta(z,x)}{q_\phi(z|x)p(x)} dz \\
&= - \left(\int_z q_\phi(z|x) \log \frac{p_\theta(z,x)}{q_\phi(z|x)} dz - \int_z q_\phi(z|x) \log p_\theta(x) dz \right) \\
&= - \int_z q_\phi(z|x) \log \frac{p_\theta(z,x)}{q_\phi(z|x)} dz + \log p_\theta(x)
\end{aligned}
$$

Hence the marginal likelihood under the generative model be decomposed as a sum of a variational free energy and the KL divergence between the approximate and true posteriors,

ELBO

Then we can define.

$$\mathcal{L}(\theta, \phi, z) := \int_z q_\phi(z|x) \log \frac{p_\theta(z, x)}{q_\phi(z|x)} dz$$
$$= \mathbf{E}_{q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z))$$

This is called the **ELBO** (Evidence Lower BOund). It is a lower bound on the marginal likelihood of the data under the generative model

$$\log p_\theta(x) = \log \int_z p_\theta(z, x) dz = \log \int_z p_\theta(x, z) \frac{q_\phi(z|x)}{q_\phi(z|x)} dz = \log \mathbb{E}_{q_\phi}(z|x) \frac{p_\theta(x, z)}{q_\phi(z|x)}$$
$$\geq \mathbf{E}_{q_\phi(z|x)} \log \frac{p_\theta(z, x)}{q_\phi(z|x)} \quad \text{(Jensen's inequality)}$$
$$= \mathcal{L}(\theta, \phi, z)$$

Note:

- ELBO is a trade-off between maximising the likelihood of observations and staying close to the prior
- measure of additional information required to express the posterior relative to the prior
- replaces the intractable MLE with a lower bound to jointly

In practice

Assume $p_\theta(z) = \mathcal{N}(0, I)$ and $q_\theta(z|x) = \mathcal{N}(\mu(x), \sigma(x)I)$, where $\mu(x), \sigma(x)$ are outputs of an encoder (shallow MLP).

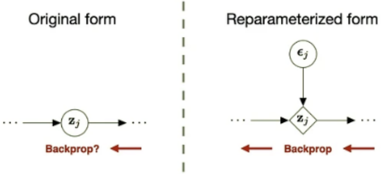
In this case,

$$\begin{aligned}\log p_\theta(x) &= \mathbf{E}_{q_\theta(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z)) \\ &= \mathbf{E}_{q_\theta(z|x)} \left(-\frac{||x-\mu(x)||^2}{2\sigma(x)}\right) - D_{KL}(\mathcal{N}(\mu(x), \sigma(x)I)||\mathcal{N}(0, I))\end{aligned}$$

The reparametrization trick

To train a VAE, one needs to sample from the inference model $q_\phi(z|x)$, evaluate a generative model $p_\theta(x|z)$ to get x and backpropagate the error through the inference model. This involves computing gradients with respect to θ and ϕ of $\mathcal{L}$.

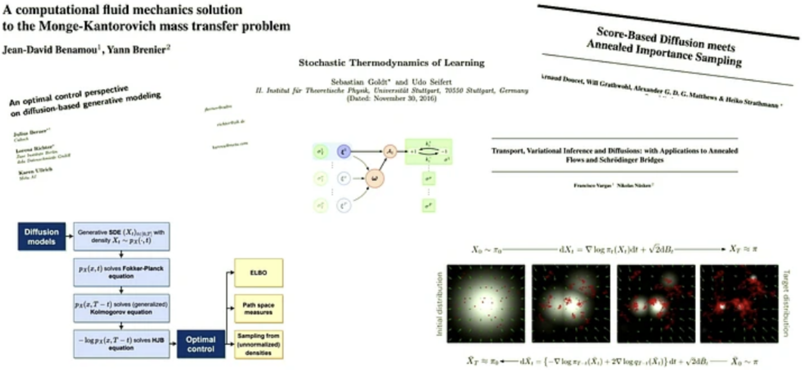
Reparametrise the random variable $z \sim q_\phi(z|x)$ as $z = g_\phi(\epsilon, x)$, where g is a differentiable transformation, with $\epsilon \sim p(\epsilon)$. Now the noise is a 'parameter' of a deterministic function, which can be differentiated.



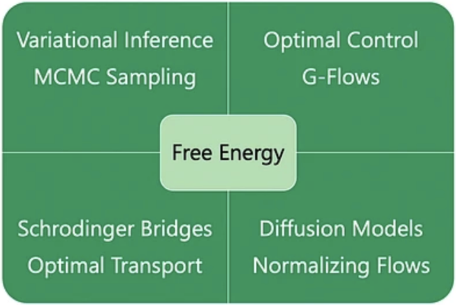
Applications

- lossy compression (Habibian 2019)
- cellular responses to drug perturbations (Rampasek 2017)
- latent neural encodings (Pandarinath 2018)
- genetics (Fraiser 2021)

Connections



Connections



- Objective is to minimize KL(Q||P) (a.k.a. Free Energy)

$$\mathcal{F}(Q;x) \equiv E_Q[-\log(P(x|z)P(z))] - S(Q)$$

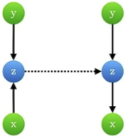
- Q & P are Markov Chains:

$$Q(Z) = Q_0(z_0)\prod_{t=1}^{T} F_t(z_t|z_{t-1})$$
$$P(Z) = P_T(z_T)\prod_{t=1}^{T} B_t(z_{t-1}|z_t)$$

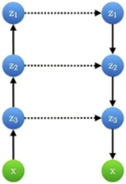
Limitations/extensions

Extensions - different graphical models - not just Gaussian distributions

Limitations - hard to optimise - multi-level VAEs can get stuck in local optima



Class-conditional VAEs



Hierarchical VAEs

Sources

- https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73
- https://iclr.cc/virtual/2024/test-of-time/21444
- https://yunfanj.com/blog/2021/01/11/ELBO.html
- https://arxiv.org/pdf/1906.02691