Variational Autoencoders

by Paul Hand
Northeastern University

Outline:

- Generative Models and Autoencoders
- Variational Lower Bound (VLB)
- Optimizing VLB + Variational Autoencoders

Resource: Kingma and Welling 2019, Chapter 2,
"Introduction to Variational Autoencoders"

Generative Models

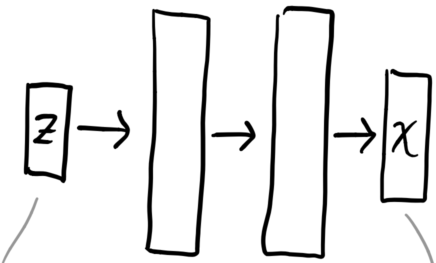
A model that can sample from a
learned distribution

(a) 2-D latent space

(Kingma and Welling, 2014)

(Razavi et al. 2019)

In some generative models, samples are generated by a net applied to a random latent code



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

random

deterministic

induced random variable

latent code

typically $z \sim N(0, I)$

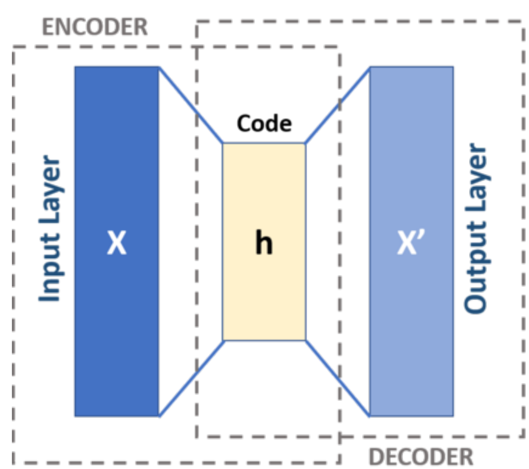
typically low-dimensional

sample/image

high dimensional

Autoencoders

Autoencoders attempt to reconstruct input signals/images by learning mappings to and from a code



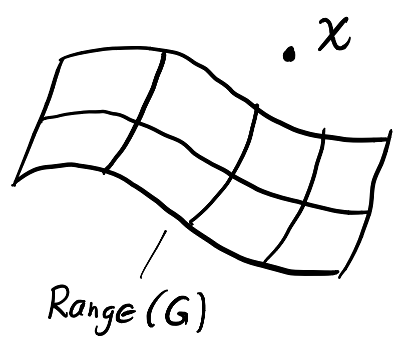
Want: $x' = D_\theta(E_\phi(x)) \approx x$

$$\min_{\theta,\phi} \sum_{i=1}^{n} \| D_\theta(E_\phi(x_i)) - x_i \|^2 \quad \text{w/ } \{X_i\}_{i=1\cdots n} \text{ is dataset}$$

A (plain) autoencoder is not a generative model as it does not define a distribution

Training a low latent-dimensional generative model by likelihood

Given data $\{X_i\}_{i=1\cdots n}$, train a gen. model to maximize the likelihood of the observed data



If gen. model

$G_\theta : \mathbb{R}^k \to \mathbb{R}^d$ w/ $k<d$,

$z \mapsto x$

then $p(x)=0$ almost everywhere

So, can't directly optimize likelihood

To have nonzero likelihood everywhere, define noisy observation model

$$P_\theta(x \mid z) = \mathcal{N}(x ; G_\theta(z), \eta I)$$

Under a simple prior $p(z)$, this induces a joint distribution $p_\theta(x, z)$

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

intractable to evaluate at each iteration

optimize a lower bound instead

Variational Lower Bound

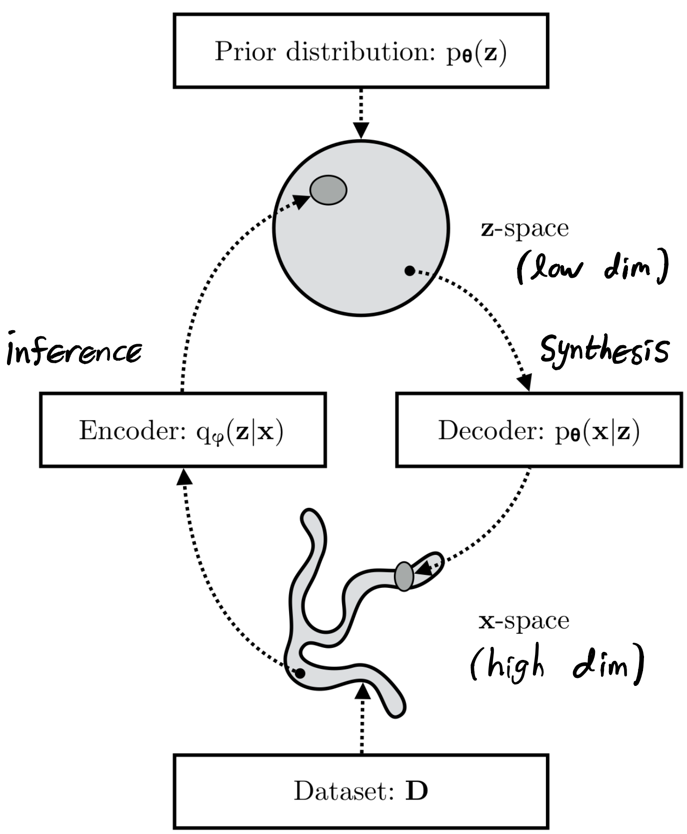
Setup:

Data generated by

$z \sim p(z)$ prior

$x \sim p_\theta(x|z)$

Use $q_\varphi(z|x)$ as proxy for $p_\theta(z|x)$



(Kingma and Welling 2019)

Find: lower bound to $p_\theta(x)$

$$\log p_\theta(x) = \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \log p_\theta(x) = \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \log \frac{p_\theta(x,z)}{p_\theta(z|x)}$$

$$= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \log\left(\frac{p_\theta(x,z)}{q_\varphi(z|x)} \cdot \frac{q_\varphi(z|x)}{p_\theta(z|x)}\right)$$

$$= \underbrace{\mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \log \frac{p_\theta(x,z)}{q_\varphi(z|x)}}_{\mathcal{L}_{\theta,\varphi}(x)} + \underbrace{\mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \log \frac{q_\varphi(z|x)}{p_\theta(z|x)}}_{D_{KL}(q_\varphi(z|x)\,\|\,p_\theta(z|x))}$$

Variational Lower Bound (VLB)
Evidence lower bound (ELBO)

Define: $D_{KL}(q\|p) = \mathop{\mathbb{E}}_{z\sim q} \log \frac{q(z)}{p(z)}$

Note:
- $D_{KL}(q\|p) \neq D_{KL}(p\|q)$
- Measure of how far p is from q
- $D_{KL}(q\|p) \geq 0$ (and is 0 iff $p=q$)

So, $\log p_\theta(x) \geq \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \log \frac{p_\theta(x,z)}{q_\varphi(z|x)} = \mathcal{L}_{\theta,\varphi}(x)$

Interpretation of VLB

$$\mathcal{L}_{\theta,\varphi}(x) = \mathbb{E}_{z\sim q_\varphi(z|x)} \log \frac{p_\theta(x,z)}{q_\varphi(z|x)} = \mathbb{E}_{z\sim q_\varphi(z|x)} \log p_\theta(x|z) \frac{p(z)}{q_\varphi(z|x)}$$

$$= \underbrace{\mathbb{E}_{z\sim q_\varphi(z|x)} \log p_\theta(x|z)}_{\text{reconstruction error}} + \underbrace{\mathbb{E}_{z\sim q_\varphi(z|x)} \log \frac{p(z)}{q_\varphi(z|x)}}_{\text{regularization}}$$

First term: $p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), \eta I)$

$$\Rightarrow \log p_\theta(x|z) = -\frac{1}{2\eta}\|x - G_\theta(z)\|^2 + \text{constant}$$

So $\mathbb{E}_{z\sim q_\varphi} \log p_\theta(x|z)$ is expected ℓ_2 reconstruction error under the encoder model

Maximizing VLB encourages q_φ to be point mass

Second term: $\mathbb{E}_{z\sim q_\varphi(z|x)} \log \frac{p(z)}{q_\varphi(z|x)} = -D_{KL}(q_\varphi(z|x) \,\|\, p(z))$

So maximizing VLB $\mathcal{L}_{\theta,\varphi}$ pushes $q_\varphi(z|x)$ toward $p(z)$. Prevents q_φ from being a point mass.

Makes $q_\varphi(z|x)$ more like standard normal

Maximizing VLB $\mathcal{L}_{\theta,\varphi}$:

- Roughly maximizes $p(x)$
- Minimizes KL divergence of $q_\varphi(z|x)$ and $p_\theta(z|x)$, making q_φ better

Main Idea: Instead of optimizing $\sum_{i=1}^{n} \log p_\theta(x_i)$,

optimize $\sum_{i=1}^{n} \mathcal{L}_{\theta,\varphi}(x_i)$

w/ $\mathcal{L}_{\theta,\varphi}(x) = \mathbb{E}_{z \sim q_\varphi(z|x)} \log \frac{p_\theta(x,z)}{q_\varphi(z|x)}$

Optimizing Variational Lower Bound

$$\max_{\theta,\varphi} \sum_{i=1}^{n} \mathcal{L}_{\theta,\varphi}(x_i)$$

One possibility:

for each x_i, find best $q_\varphi(z|x_i)$ by multiple gradient steps in φ. Then gradient ascend in θ.

Expensive inference updates

Instead:

amortize the inference costs

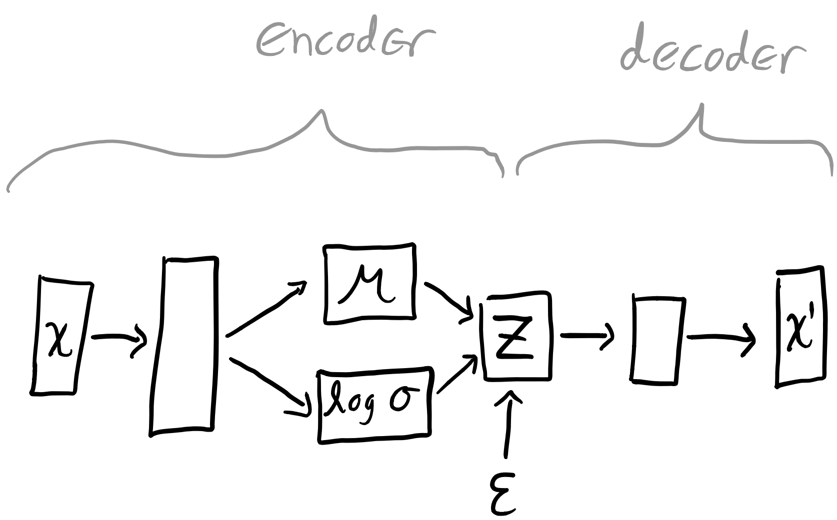
by learning an inference net

$x \mapsto (\mu, \Sigma)$ w/ $q_\varphi(z|x) = \mathcal{N}(z; \mu(x), \Sigma(x))$

or $\sigma(x) I$

Parameters of inference model are shared between data points

Variational Autoencoder architecture



random noise — reparameterization trick

- separate random source from diff'able quantities

Stochastic Gradient Optimization of VLB

Dataset $\mathcal{D} = \{x_i\}_{i=1 \dots n}$

Solve $\max_{\theta, \varphi} \sum_{x_i \in \mathcal{D}} \mathcal{L}_{\theta,\varphi}(x_i)$

w/ $\mathcal{L}_{\theta,\varphi}(x) = \mathbb{E}_{z \sim q_\varphi(z|x)} \log \frac{p_\theta(x,z)}{q_\varphi(z|x)}$

Computing $\nabla_{\theta,\varphi} \mathcal{L}_{\theta,\varphi}(x_i)$ is intractable, but there are unbiased estimators

Easy to get unbiased $\nabla_\theta \mathcal{L}_{\theta,\varphi}$:

$$\nabla_\theta \mathcal{L}_{\theta,\varphi}(x) = \nabla_\theta \mathbb{E}_{z \sim q_\varphi(z|x)} \left[\log p_\theta(x,z) - \log q_\varphi(z|x)\right]$$

$$= \mathbb{E}_{z \sim q_\varphi} \nabla_\theta \log p_\theta(x,z)$$

$$\simeq \nabla_\theta \log p_\theta(x,z) \quad \text{w/ } z \sim q_\varphi(z|x)$$

unbiased estimate

Not as easy to get unbiased $\nabla_\varphi \mathcal{L}_{\theta,\varphi}$:

$$\nabla_\varphi \mathcal{L}_{\theta,\varphi}(x) = \nabla_\varphi \mathbb{E}_{z \sim q_\varphi(z|x)}\left[\log p_\theta(x,z) - \log q_\varphi(z|x)\right]$$

$$\neq \mathbb{E}_{z \sim q_\varphi} \nabla_\varphi \left(\log p_\theta(x,z) - \log q_\varphi(z|x)\right)$$

Recall, $q_\varphi(z|x) = \mathcal{N}(z; \mu(x), \sigma(x) I)$

$$= \mu(x) + \sigma(x) \cdot \varepsilon, \quad \text{w/ } \varepsilon \sim \mathcal{N}(0, I)$$

$$\mathcal{L}_{\theta,\varphi}(x) = \mathbb{E}_{z \sim q_\varphi(z|x)}\left(\log p_\theta(x,z) - \log q_\varphi(z|x)\right)$$

$$= \mathbb{E}_{\varepsilon \sim p(\varepsilon)}\left(\log p_\theta(x,z) - \log q_\varphi(z|x)\right)$$

Form estimator of $\mathcal{L}_{\theta,\varphi}(x)$ as $\tilde{\mathcal{L}}_{\theta,\varphi}(x)$ by:

$$\varepsilon \sim p(\varepsilon)$$
$$z = \mu_\varphi(x) + \sigma_\varphi(x)\varepsilon = g(\varphi, x, \varepsilon)$$
$$\tilde{\mathcal{L}}_{\theta,\varphi}(x) = \log p_\theta(x,z) - \log q_\varphi(z|x)$$

Unbiased estimate of $\nabla_\varphi \mathcal{L}_{\theta,\varphi}(x)$:

$$\nabla_\varphi \tilde{\mathcal{L}}_{\theta,\varphi}(x)$$

Note: $\mathbb{E}_{\varepsilon\sim p(\varepsilon)} \tilde{\mathcal{L}}_{\theta,\varphi}(x) = \mathcal{L}_{\theta,\varphi}(x)$

So $\mathbb{E}_{\varepsilon\sim p(\varepsilon)} \nabla_\varphi \tilde{\mathcal{L}}_{\theta,\varphi}(x) = \nabla_\varphi \mathbb{E}_{\varepsilon\sim p(\varepsilon)} \tilde{\mathcal{L}}_{\theta,\varphi} = \nabla_\varphi \mathcal{L}_{\theta,\varphi}$

Optimize VAE parameters w/ stochastic gradients.

Can extend models to be more sophisticated, eg $\Sigma(x)$ vs $\sigma I(x)$ as inference model.

Key points:

- optimize a lower bound to likelihood
- lower bound has terms for reconstruction and for regularization
- maintain an inference model for $z|x$ in place of intractable true distribution

- reparameterization trick allows backpropagating on mean and variance of inference model
- VAEs have been trained with photorealistic outputs