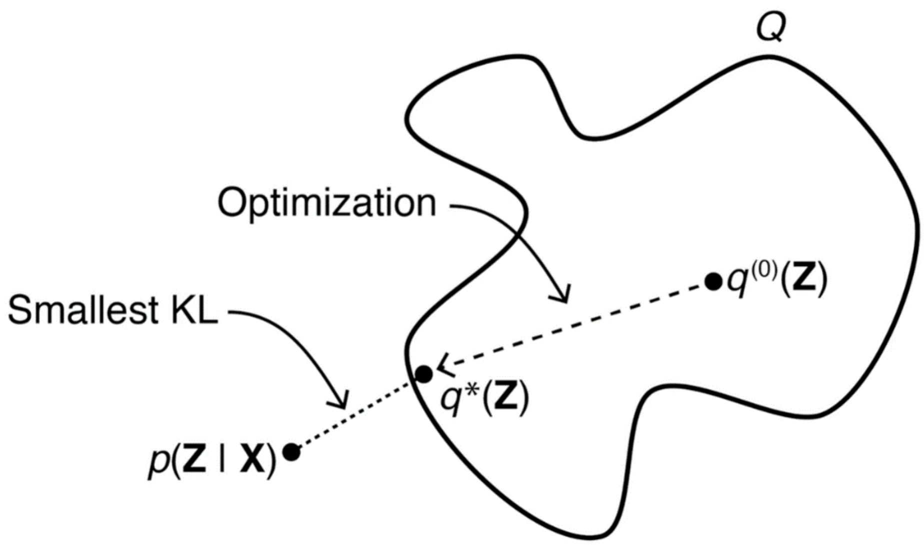
Variational Inference

Variational Inference



Variational Inference: Idea

we are generally interested in getting insights from observed data

inferring latent variables from observed data

a conditional probabilistic distribution

$$x \sim p(x) \quad z \sim p(z|x)$$ (posterior) (inference)

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int_z p(x,z)\,dz}$$

$p(x)$ is generally intractable

idea: use a family of easy distribution to approximate the hard distribution

(variational)

$z \sim p(z)$

$x \sim p(x|z)$

(generation)

Latent Graphical Model | Variational Inference: Idea | Variational Inference: Example Setup | Variational Inference: KL, ELBO & Example | Take Away

Variational Inference: Example (1)

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

bayes' rule

$$z \backsim p(z) = \begin{cases} e^{-z}, & z \geq 0 \\ 0 & , z < 0 \end{cases} = e^{-z} I(z \geq 0)$$

$$x \backsim p(x|z) = N(x, \mu = z, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(x-z)^2)}$$

$$p(x, z) = p(x|z)p(z) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(x-z)^2)} e^{-z} I(z \geq 0)$$ (joint distribution)

$$p(x) = \int_0^\infty p(x, z)\, dz = \underbrace{\int_0^\infty e^{-z} \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(x-z)^2)}\, dz}_{\text{the integral has no closed-form solution}}$$

thus the posterior has no closed-form solution

Variational Inference: Example (2)

$$p(z|x) \sim p(x, z) = p(x|z)p(z) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(x-z)^2)} e^{-z} I(z \geq 0)$$

$$= \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}z^2 + (x-1)z - \frac{1}{2}x^2)} I(z \geq 0)$$

$$= \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(z-(x-1))^2 + \frac{1}{2}(x-1)^2 - \frac{1}{2}x^2)} I(z \geq 0)$$

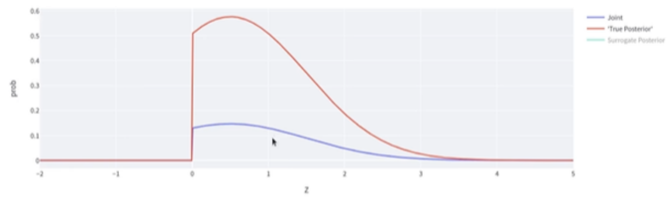
$$= \frac{1}{\sqrt{2\pi}} e^{(\frac{1}{2}(x-1)^2 - \frac{1}{2}x^2)} e^{(-\frac{1}{2}(z-(x-1))^2)} I(z \geq 0)$$

$$\sim \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(z-(x-1))^2)} I(z \geq 0)$$

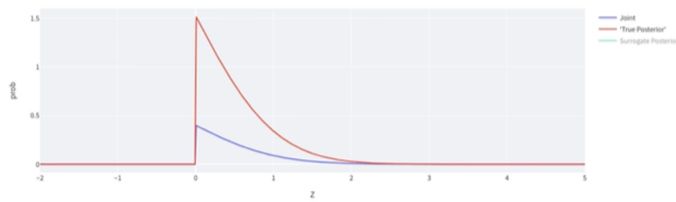
posterior proportional to a gaussian curve for *z* larger or equal to zero

Variational Inference: Example (3)

$$p(z|x) \sim \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}(z-(x-1))^2)} I(z \geq 0)$$



$x = 1.5$



$x = 0$

let's now use an "easier" distribution to approximate the intractable posterior

Latent Graphical Model | Variational Inference: Idea | Variational Inference: Example Setup | Variational Inference: KL, ELBO & Example | Take Away

Approximate Distribution: ELBO

$$q_\theta(z) \approx p(z|x)$$

$$D(q_\theta(z) \| p(z|x)) = E_{z\sim q}[\log \frac{q_\theta(z)}{p(z|x)}] = E_{z\sim q}[\log q_\theta(z) - \log p(z|x)]$$

$$= E_{z\sim q}[\log q_\theta(z) - \log \frac{p(z,x)}{p(x)}]$$

$$= E_{z\sim q}[\log q_\theta(z) - \log p(z,x)] + \log p(x)$$

evidence lower bound (elbo)

$$\log p(x) = E_{z\sim q}[\log p(z,x) - \log q_\theta(z)] + D(q_\theta(z) \| p(z|x))$$

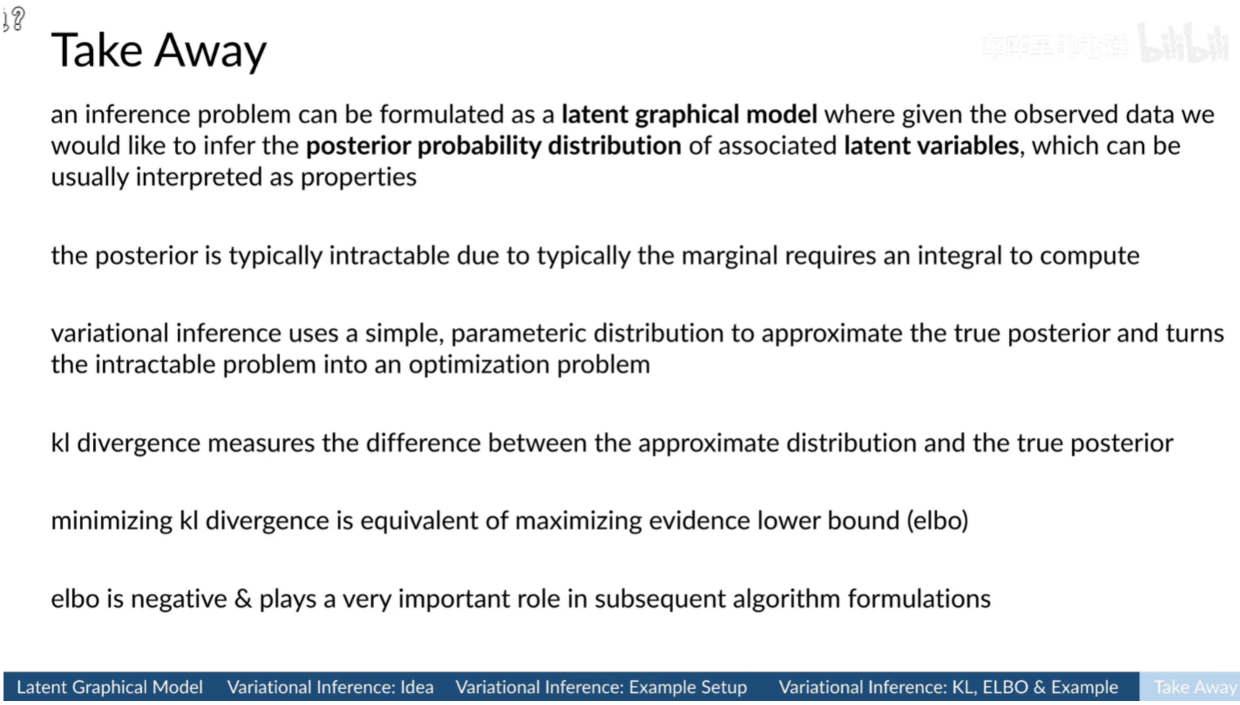
$$\log p(x) \geq E_{z\sim q}[\log p(z,x) - \log q_\theta(z)] \equiv \mathcal{L}_q$$

objective: minimizing kl divergence ⟶ maximizing elbo (variational optimization: optimizing over functions)

Latent Graphical Model | Variational Inference: Idea | Variational Inference: Example Setup | Variational Inference: KL, ELBO & Example | Take Away

Note: here (in standard variational inference), we know p(z|x), which is the real distribution, because we know p(x) and p(z,x). But in VAE, which is introduced later, we don't know p(z|x), so we use q(z|x) to approximate it, which is called amortized variational inference. q(z|x) is the conditional probability the encoder (the neural network) gives, which is based on a signal data piece. We have many pieces of data to derive different q(z|x) and optimize it (called "amortized").

q(z|x) is a variational distribution, and is commonly a Gaussian parametrized by the encoder network outputs.



Variational AutoEncoder

ELBO (1)

original form

$$\mathcal{L}_q = E_{z\sim q}[\log p(z, x) - \log q(z)]$$

$q(z) \approx p(z|x)$ approximate distribution

z latent variable x observed variable

variational autoencoder changes

approximation distribution $q_\varphi(z|x)$ encoder (amortized variational inference)

likelihood distribution $p_\theta(x|z)$ decoder

$$\mathcal{L}_q = E_{z\sim q}[\log p_\theta(x|z)p(z) - \log q_\varphi(z|x)]$$
$$= E_{z\sim q}[\log p_\theta(x|z) + \log p(z) - \log q_\varphi(z|x)]$$

ELBO (2)

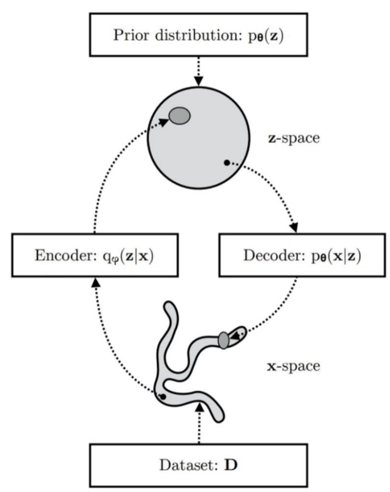
$$\mathcal{L}_q = E_{z\sim q}[\log p_\theta(x|z) + \log p(z) - \log q_\varphi(z|x)]$$
$$= E_{z\sim q}[\log p_\theta(x|z)] - E_{z\sim q}[\log \frac{q_\varphi(z|x)}{p(z)}]$$
$$= E_{z\sim q}[\log p_\theta(x|z)] - D(q_\varphi(z|x)||p(z))$$

$E_{z\sim q}[\log p_\theta(x|z)]$ reconstruction likelihood

$D(q_\varphi(z|x)||p(z))$ latent prior similarity
$p(z) = N(0, I)$

maximizing reconstruction likelihood
while miminizing kl divergence

(approximating $\log p(x)$)



ref. 2

Objective (1)

$q_\varphi(z|x) = N(z; \mu_\varphi(x), \Sigma_\varphi(x))$

$p(z) = N(z; 0, I)$

kl divergence of gaussian

$$D(N(\mu_0, \Sigma_0)||N(\mu_1, \Sigma_1)) = \frac{1}{2}(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^{\mathrm{T}}\Sigma_1^{-1}(\mu_1 - \mu_0) - k + \log(\frac{\det\Sigma_1}{\det\Sigma_0}))$$

$D(q_\varphi(z|x)||p(z)) = \frac{1}{2}(\mathrm{tr}(\Sigma_\varphi(x)) + \mu_\varphi(x)^{\mathrm{T}}\mu_\varphi(x) - k - \log(\det\Sigma_\varphi(x)))$ k is the dimension

encoder objective

Objective (2)

$p_\theta(x|z) = N(x; \mu_\theta(z), \sigma^2 I)$

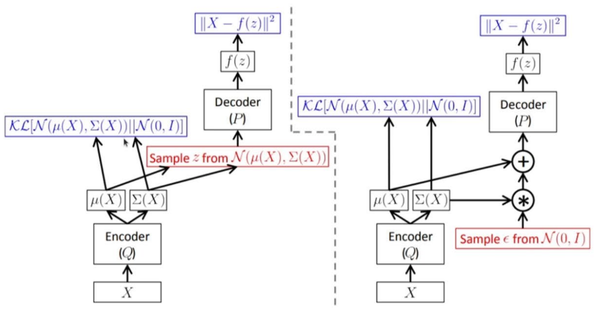
$\log p_\theta(x|z) = -c||x - \mu_\theta(z)||^2 + d$ $\nabla_\theta ||x - \mu_\theta(z)||^2$ decoder objective

but

$E_{z\sim q_\varphi}[\log p_\theta(x|z)]$ $q_\varphi(z|x) = N(z; \mu_\varphi(x), \Sigma_\varphi(x))$

cannot back propagate

Reparameterization



ref. 3

$$
\begin{aligned}
q_\varphi(z|x) &= N(z; \mu_\varphi(x), \Sigma_\varphi(x)) \\
&= \mu_\varphi(x) + \Sigma_\varphi^{1/2}(x) N(0, I) \\
&= \mu_\varphi(x) + \Sigma_\varphi^{1/2}(x)\varepsilon
\end{aligned}
$$

$$E_{\varepsilon\sim N(0,I)}[\log p_\theta(x|z = \mu_\varphi(x) + \Sigma_\varphi^{1/2}(x)\varepsilon)]$$

Loss Fuction

reconstruction loss

$$\mathcal{L}_1 = -\log p_\theta(x|z) \sim \|x - \mu_\theta(z)\|^2$$

latent prior similarity loss

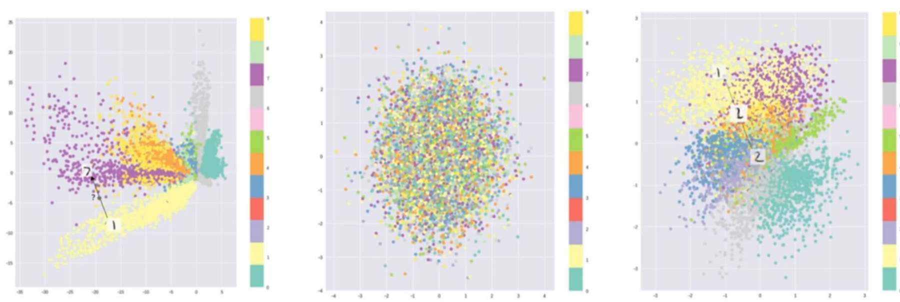
$$\mathcal{L}_2 = D(q_\varphi(z|x)\|p(z)) = \frac{1}{2}(\mathrm{tr}(\Sigma_\varphi(x)) + \mu_\varphi(x)^{\mathrm{T}}\mu_\varphi(x) - k - \log(\det\Sigma_\varphi(x)))$$

total loss

$$\mathcal{L} = \mathcal{L}_1 + \alpha\mathcal{L}_2$$

Latent Space Visualizations

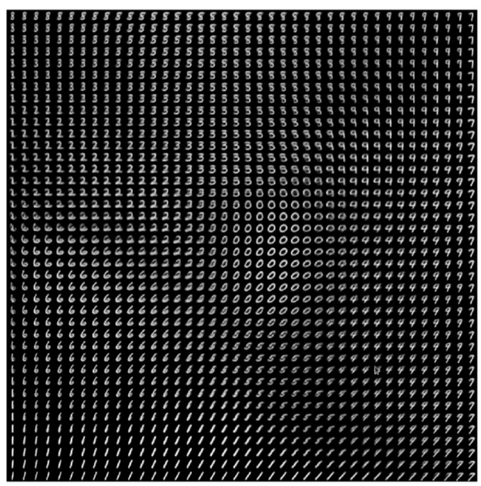
Only reconstruction loss | Only KL divergence | Combination



ref. 1

VAE on MNIST

Image Generation Visualizations



ref. 1

VAE on MNIST

Take Away

variational autoencoder can be understood as an instansiation of variational inference with encoder-decoder based framework

latent space prior is chosen to be unit gaussian

variational autoencoder can be used to generate data by sampling latent variables from unit gaussian

loss function can be derived from elbo which constitutes reconstruction loss and latent prior similarity loss, which can be treated as a regularization term

reparameterization trick is used to enable proper backprobagation

latent space visualization shows the latent prior constraint makes the learned latent variable distribution relatively continuous and has somewhat of a topology in it