To turn DLVM's intractable learning problem into something tractable, introduce a parametric inference model $q_\phi(z|x)$ which is encoder.

ϕ parameters of inference model; variational parameters.

optimize variational parameters ϕ: $q_\phi(z|x) \approx p_\theta(z|x)$.

Encoder has a directed graphical model, can be factorized:

$$q_\phi(z|x) = q_\phi(z_1, \dots, z_M | x) = \prod_{j=1}^{M} q_\phi(z_j | Pa(z_j), x)$$

where $Pa(z_j)$ is set of parents of z_j in directed graph.

Once we have factorization $q_\phi(z|x)$ can be specifically parametrized using (deep) neural networks, where variational parameters ϕ are weights, biases of neural network.

The optimization objective of VAE is evidence lower bound (ELBO), also called variational lower bound.

For any choice of inference model $q_\phi(z|x)$ including choice of ϕ,

$$\log p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x)\right]$$

$$= \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{p_\theta(z|x)}\right)\right]$$

$$= \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z) \cdot q_\phi(z|x)}{q_\phi(z|x) \cdot p_\theta(z|x)}\right)\right]$$

Maximizing ELBO w.r.t. θ, ϕ will concurrently optimize:

① approximately maximize marginal likelihood $p_\theta(x)$, so generative model will be better.

② minimize KL divergence of approximation $q_\phi(z|x)$ from true posterior $p_\theta(z|x)$ so $q_\phi(z|x)$ becomes better.

Jointly optimize ϕ and θ using SGD.

Reparametrization trick.

for continuous latent variables and a differentiable encoder/decoder, the ELBO can be differentiated w.r.t. both ϕ and θ if we reparametrize.

First express r.v. $z \sim q_\phi(z|x)$ as a differentiable/invertible transformation of another r.v. ε : Given ϕ, x:

$$z = g(\varepsilon, \phi, x)$$ where ε is independent of x and ϕ.

Now expectations in terms of ε :

$$\mathbb{E}_{q_\phi(z|x)}[f(z)] = \mathbb{E}_{p(\varepsilon)}[f(z)]$$

and since expectation and gradient commute by linearity:

$$\nabla_\phi \mathbb{E}_{q_\phi(z|x)}[f(z)] = \nabla_\phi \mathbb{E}_{p(\varepsilon)}[f(z)]$$
$$= \mathbb{E}_{p(\varepsilon)}[\nabla_\phi f(z)]$$
$$\simeq \nabla_\phi f(z)$$ where we can estimate by Monte Carlo draws of $\varepsilon \sim p(\varepsilon)$ and doing $z = g(\varepsilon, \phi, x)$

under reparameterization, we can replace expectation w.r.t. $q_\phi(z|x)$ with one w.r.t. $p(\varepsilon)$, so ELBO becomes:

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

$$= \mathbb{E}_{p(\varepsilon)}\left[\log p_\varepsilon(x,z) - \log q_\phi(z|x)\right]$$

where $z = g(\varepsilon, \phi, x)$

so we can get simple Monte Carlo estimator $\hat{\mathcal{L}}_{\theta,\phi}(x)$ of individual datapoint ELBO through a single sample $\varepsilon \sim p(\varepsilon)$

$$\varepsilon \sim p(\varepsilon)$$
$$z = g(\varepsilon, \phi, x)$$
$$\hat{\mathcal{L}}_{\theta,\phi}(x) = \log p_\theta(x,z) - \log q_\phi(z|x)$$

this can be optimized using SGD.

VAEs have some problems:

→ approximate inference distribution is often different from true posterior (from ELBO objective).

can modify the ELBO objective itself to balance correct inference and fitting training data?

InfoVAE

ELBO loss is

$$\mathcal{L}_{ELBO} = -D_{KL}\left(q_\phi(z) \| p(z)\right) - \mathbb{E}_{q(z)}\left[D_{KL}\left(q_\phi(x|z) \| p_\theta(x|z)\right)\right]$$

- allow some weighting parameter λ to balance these two terms.
- introduce a mutual information term.

$$\mathcal{L}_{infoVAE} : -\lambda D_{KL}\left(q_\phi(z) \| p(z)\right) - \mathbb{E}_{q(z)}\left(D_{KL}\left(q_\phi(x|z) \| p_\theta(x|z)\right)\right] + \alpha I_q(x;z)$$

where $I_q(x;z)$ is mutual information between x and z under $q_\phi(x,z)$

the mutual information maximization encourages the model to use the latent code and avoids "information preference problem" when latent variables z ignored in favor of the data variables x.

$$\mathcal{L}_{\text{infoVAE}} = -\lambda D_{KL}\left(q_\phi(z) \| p(z)\right) - \mathbb{E}_{q(z)}\left[D_{KL}\left(q_\phi(x|z) \| p_\theta(x|z)\right)\right] + \alpha I(x;z)$$

$$= \mathbb{E}_{q_\phi(x,z)}\left[-\lambda \log \frac{q_\phi(z)}{p(z)} - \log \frac{q_\phi(x|z)}{p_\theta(x|z)} - \alpha \log \frac{q_\phi(z)}{q_\phi(z|x)}\right]$$

$$= \mathbb{E}_{q_\phi(x,z)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z)^{\lambda+\alpha-1} \, p_D(x)}{p(z)^\lambda \, q_\phi(z|x)^{\alpha-1}}\right]$$

$$= \mathbb{E}_{q_\phi(x,z)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z)^{\lambda+\alpha-1} \, q_\phi(z|x)^{1-\alpha} \, p_D(x)}{p(z)^{\lambda+\alpha-1} \, p(z)^{\alpha-1}}\right]$$

$$= \boxed{\mathbb{E}_{p_D(x)} \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - (1-\alpha)\, \mathbb{E}_{p_D(x)} D_{KL}\left(q_\phi(z|x) \| p(z)\right) - (\alpha+\lambda-1)\, D_{KL}\left(q_\phi(z) \| p(z)\right)} - \mathbb{E}_{p_D}\left[\log p_D(x)\right]$$

constant.