Discrete Variational Autoencoders

K. Struminsky[1]

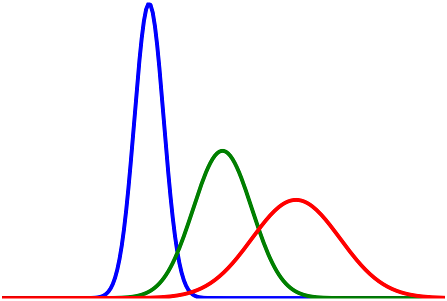
[1]Faculty of Computer Science
HSE

September 2016

Unsupervised learning of probabilistic models

Model distribution over datapoints x with parametric family:

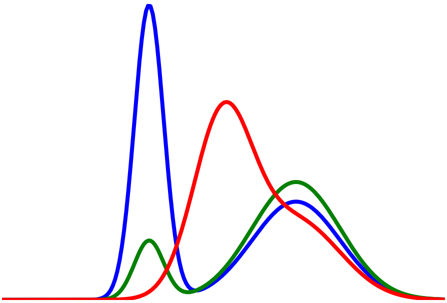
$$p(x|\theta)$$



Latent variable models

A common trick to augment distribution family

$$p(x|\theta) = \int p(x|z, \theta) p(z|\theta) \mathrm{d}z$$



Fitting latent variable models

Maximizing likelihood

$$\sum_{i=1}^{N} \log p(x_i|\theta) \to \max_{\theta}$$

What if we try gradient descent?

$$\frac{\partial}{\partial\theta} \log \int p(x, z|\theta)\mathrm{d}z = \mathbb{E}_{p(z|x,\theta)} \left[\frac{\partial}{\partial\theta} \log p(x, z|\theta) \right]$$

Need posterior distribution $p(z|x, \theta)$

Variational inference

Whenever $p(z|x, \theta)$ is intractable we can use VI

- Set distribution family $q(z|x, \phi), \phi \in \Phi$
- Set objective
  $\mathcal{L}(x, \theta, \phi) = \mathbb{E}_{q(z|x,\phi)} \log p(x|z, \theta) - KL(q(z|x, \phi)||p(z|\theta))$
- Solve $\mathcal{L}(x, \theta, \phi) \to \max_{\phi,\theta}$
- Motivation: $\mathcal{L}(x, \theta, \phi) \leq p(x|\theta)$, tight iff $q(z|x, \phi) = p(z|x, \theta)$

ELBO gradients

$$\nabla_\theta \mathcal{L}(x, \theta, \phi) = \mathbb{E}_q \nabla_\theta \log p(x, z|\theta)$$
$$\nabla_\phi \mathcal{L}(x, \theta, \phi) = \mathbb{E}_q \nabla_\phi \log q(z|x, \phi) \cdot (logp(x, z|\phi) - \log q(z|x, \phi) - 1)$$

Here we use $\nabla_\phi q(z|x, \phi) = q(z|x, \phi) \nabla_\phi \log q(z|x, \phi)$.

Expectations are intractable

Can't compute expectations analytically, adopt stochastic approximations:

$$\nabla_\theta \mathcal{L}(x, \theta, \phi) \approx \frac{1}{N} \sum_{z \sim q} \nabla_\theta \log p(x, z|\theta)$$

$$\nabla_\phi \mathcal{L}(x, \theta, \phi) \approx \frac{1}{N} \sum_{z \sim q} [\nabla_\phi \log q(z|x, \phi)(logp(x, z|\phi) - \log q(z|x, \phi) - 1)]$$

- Pros: weak assumptions on $q(z|x, \phi)$
- Cons: variance is too high [1]

---

[1] More on this problem https://arxiv.org/abs/1602.06725

Variational Autoencoders

(Kingma, Welling 2014)[2] introduced the following probabilistic model:

- $p(z) = \mathcal{N}(z|0, I)$
- $p(x|z, \theta) = \mathcal{N}(x|\mu_\theta(z), \sigma_\theta(z))$ (decoder)
- $q(z|x, \phi) = \mathcal{N}(z|\mu_\phi(x), \sigma_\phi(x))$ (encoder)

Where $\mu_\theta(z), \sigma_\theta(z), \mu_\phi(x), \sigma_\phi(x)$ are defined by neural networks.

[2] https://arxiv.org/abs/1312.6114

Training variational autoencoders

$$\mathcal{L}(x,\theta,\phi) = \underbrace{\mathbb{E}_q \log p(x|z,\theta)}_{\text{autoencoding term}} - \underbrace{KL(q(z|x,\phi)||p(z|\theta))}_{\text{KL term}}$$

Reparametrization for autoencoding term:

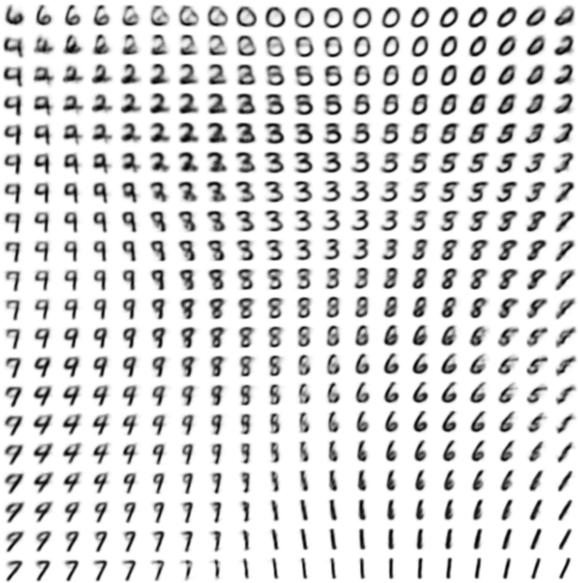
$$z \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi(x)) \leftrightarrow z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon = g(\epsilon,\phi) \text{for} \epsilon \sim \mathcal{N}(0,I)$$

Integration by substitution:

$$\mathbb{E}_q \log p(x|z,\theta) = \mathbb{E}_\epsilon \log p(x|\mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \theta)$$

$$\nabla_\phi \mathbb{E}_q \log p(x|z,\theta) \approx \frac{1}{N} \sum_{\epsilon \sim \mathcal{N}(0,I)} \nabla_\phi \log p(x|\mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \theta)$$

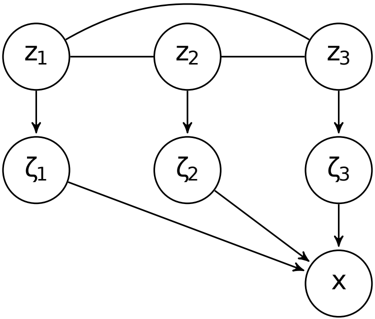
KL-term is computed analytically



Discrete VAE: model [3]

For $z \in \{0,1\}^n$, $\zeta \in [0,1]^n$

- $p(\zeta, z|\theta) = r(\zeta|z)p(z|\theta)$
- $p(z|\theta) = \frac{exp\{z^T Wz + b^T z\}}{Z_p}$
- $r(\zeta|z) = \prod r(\zeta_i|z_i)$
- $r(\zeta_i|z_i)$ is fixed by design
- $p(x|\zeta, z, \theta) = p(x|\zeta, \theta)$



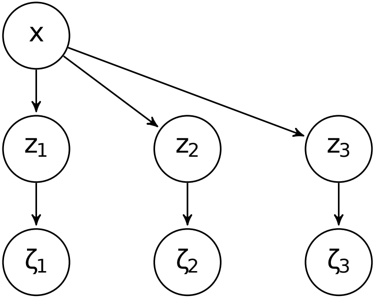
[3] As presented in https://arxiv.org/abs/1609.02200

Discrete VAE: approximating posterior

$$q(\zeta, z|\theta) = r(\zeta|z)q(z|x, \phi)$$

$$q(z|x, \phi) = \prod q(z_i|x, \phi)$$

$r(\zeta|z)$ is the same as in previous slide



Choice of $r(\zeta|z)$

$$r(\zeta_i|z_i = 0) = \delta(\zeta_i)$$
$$r(\zeta_i|z_i = 1) = \frac{\beta e^{\beta\zeta_i}}{e^{\beta} - 1}$$

CDF of ζ_i is a smooth function of $q := q(z_i = 1|x, \phi)$:

$$F_{q(\zeta_i|x,\phi)}(\zeta_i) = \underbrace{(1 - q) \cdot 1 + q\frac{e^{\beta\zeta_i} - 1}{e^{\beta} - 1}}_{\text{average of CDFs over } z_i}$$

CDF is invertible, thus for $\rho \sim U[0, 1]$ we can write:

$$\mathbb{E}_{q(\zeta_i|x,\phi)} f(\zeta_i) = \mathbb{E}_{\rho} f(F^{-1}_{q(\zeta_i|x,\phi)}(\rho))$$

Autoencoding term gradient: derivation

Sum out z:

$$\mathbb{E}_{q(\zeta,z|x,\phi)}\left[\log p(x|\zeta,z,\theta)\right] = \mathbb{E}_{q(\zeta|x,\phi)}\left[\log p(x|\zeta,\theta)\right]$$

Use the trick from the previous slide:

$$\mathbb{E}_{q(\zeta|x,\phi)}\left[\log p(x|\zeta,\theta)\right] = \int_0^1 \log p(x|F^{-1}_{q(\zeta|x,\phi)}(\rho),\theta)\mathrm{d}\rho$$

Get gradient stochastic estimates:

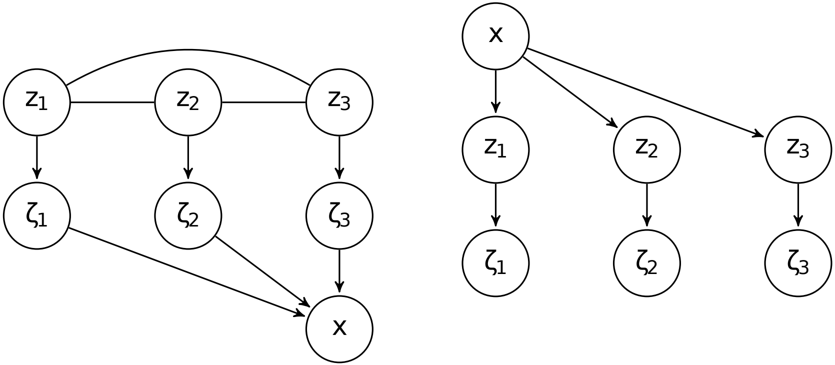
$$\frac{\partial}{\partial\phi}\mathbb{E}_{q(\zeta|x,\phi)}\left[\log p(x|\zeta,\theta)\right] \approx \mathbb{E}_{\rho\sim U(0,1)^n}\frac{\partial}{\partial\phi}\log p(x|F^{-1}_{q(\zeta|x,\phi)}(\rho),\theta)$$

Autoencoding term gradient: stochastic estimate

Finally we get

$$\frac{\partial}{\partial \phi}\mathbb{E}_{q(\zeta,z|x,\phi)}\left[\log p(x|\zeta,z,\theta)\right] \approx \frac{1}{N}\sum_{\rho\sim U(0,1)^n}\frac{\partial}{\partial \phi}\log p(x|F^{-1}_{q(\zeta|x,\phi)}(\rho),\theta)$$

Recall the model and the approximating posterior



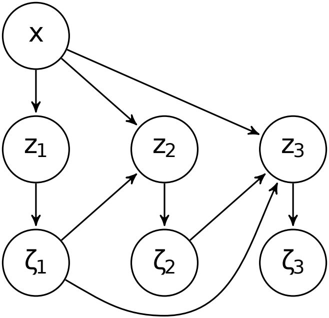
Original posterior q is factorial in z, thus it is too weak.

Hierarchical distribution for approximating posterior [4]

Original posterior q is factorial in z, thus it is too weak.

$$q(z_1, \zeta_1, ..., z_k, \zeta_k | x, \phi) = \prod_{1 \leq j \leq k} r(\zeta_j | z_j) q(z_j | \zeta_{i<j}, x, \phi)$$

$$q(z_j = 1 | \zeta_{i<j}, x, \phi) = \frac{\exp\{g_j(\zeta_{i<j}, x, \phi)\}}{1 + exp\{g_j(\zeta_{i<j}, x, \phi)\}}$$



[4] See appendix A for AE term gradint

KL-term gradients: θ

Straightforward computation gives

$$\frac{\partial}{\partial\theta} KL[q||p] = \mathbb{E}_{q(z,\zeta|x,\phi)}\left[\frac{\partial E_p(z,\theta)}{\partial\theta}\right] - \mathbb{E}_{p(z|\theta)}\left[\frac{\partial E_p(z,\theta)}{\partial\theta}\right]$$

for energy $E_p = -(z^T Wz + b^T z)$.

KL-term gradients: ϕ

$$\frac{\partial}{\partial \phi} KL[q||p] = \mathbb{E}_\rho \left[(g(x,\zeta) - b)^T \frac{\partial q}{\partial \phi} - z^T \cdot W \cdot \left(\frac{1-z}{1-q} \odot \frac{\partial q}{\partial \phi} \right) \right]$$

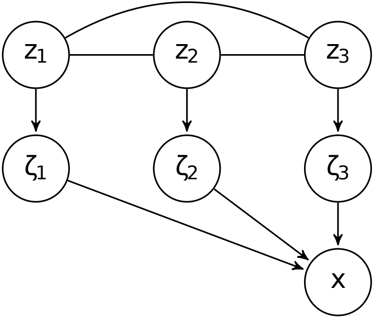
where g is taken from the definition of $q(z_i = 1|x, \phi)$
Major computation steps: $KL[q||p] = H(q) + \mathbb{E}_q(\log p)$
For $KL[q||p]$:

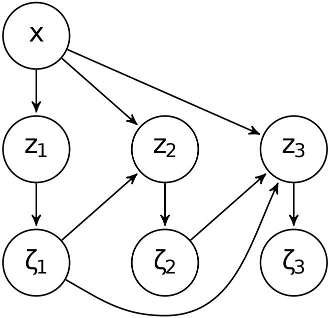
- Reparametrize expectation over ζ
- Explicitly compute expectations over z

For $\mathbb{E}_q \log p$:

- Explicitly use $r(\zeta|z=0) = \delta(\zeta)$ to compute gradients

Recall the model and the approximating posterior

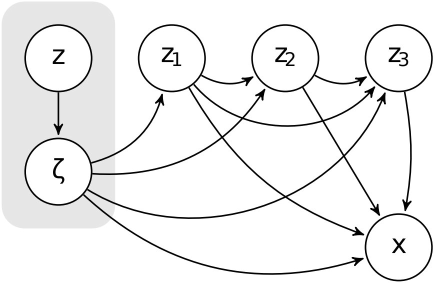




Even more layers are coming!

Continuous latent variables [5]

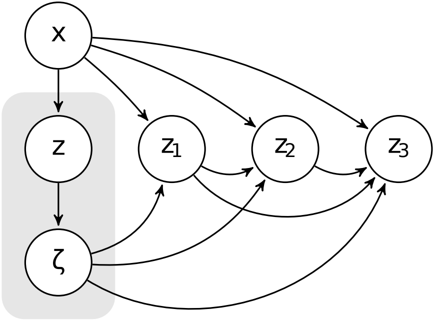
- $\mathfrak{z}_0 := \zeta$
- $p(\mathfrak{z}_0, ..., \mathfrak{z}_n|\theta) = \prod_{0 \leq m \leq n} p(\mathfrak{z}_m|\mathfrak{z}_{l<m}, \theta)$



[5] z_i on the scheme was supposed to be $\mathfrak{z}_i$

Continuous latent variables [6]

- $\mathfrak{z}_0 := \zeta$
- $q(\mathfrak{z}_0, ..., \mathfrak{z}_n | x, \phi) = \prod_{0 \leq m \leq n} q(\mathfrak{z}_m | \mathfrak{z}_{l<m}, x, \phi)$



[6] z_i on the scheme was supposed to be $\mathfrak{z}_i$

ELBO for final model

Continuous variables add new terms to ELBO:

$$\mathcal{L}(x, \theta, \phi) = \mathbb{E}_{q(\mathfrak{z}|x,\phi)} \log p(x|\mathfrak{z}, \theta) - \sum_m \mathbb{E}_{q(\mathfrak{z}_{l<m}|x,\phi)} KL\left(q(\mathfrak{z}_m|\mathfrak{z}_{l<m}, x, \phi) || p(\mathfrak{z}_m|\mathfrak{z}_{l<m}, \theta)\right)$$

New terms can be differentiated as in recurrent VAEs (Chung, 2015) [7]

[7] https://arxiv.org/abs/1506.02216

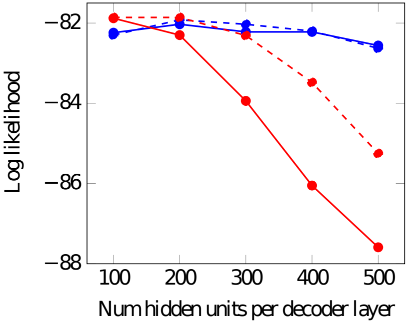
Model performance evaluation

MNIST	LL
DBN	-84.55
IWAE	-82.90
Ladder VAE	-81.74
Discrete VAE	**-80.04**

OMNIGLOT	LL
DBN	-100.45
IWAE	-103.38
RBM	-100.46
Ladder VAE	-102.11
Discrete VAE	**-97.43**

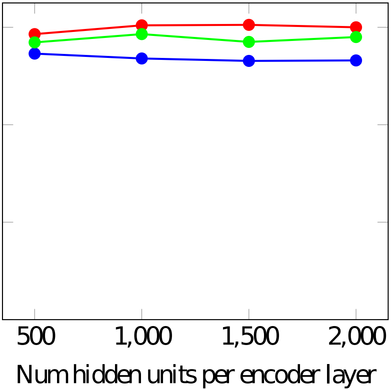
Caltech-101	LL
IWAE	-117.2
RBM	-107.8
Discrete VAE	**-97.6**

A tendency to overfit

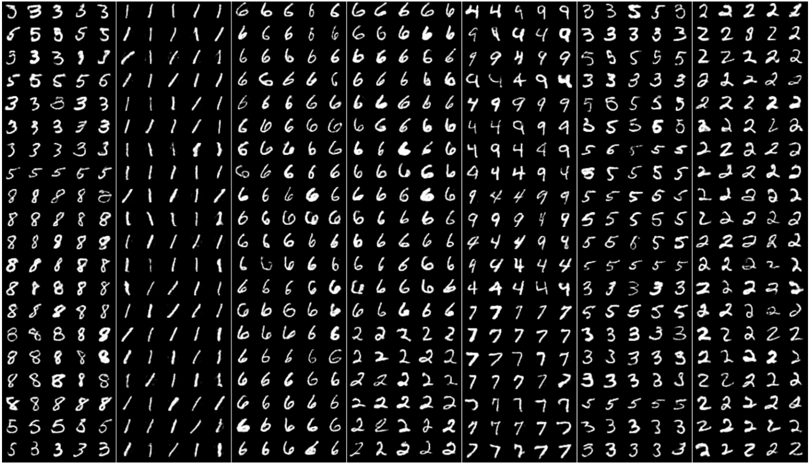


(a) Prior

Fitting approximating posterior



(b) Approximating posterior



For each column fix z, sample $\mathfrak{z}$