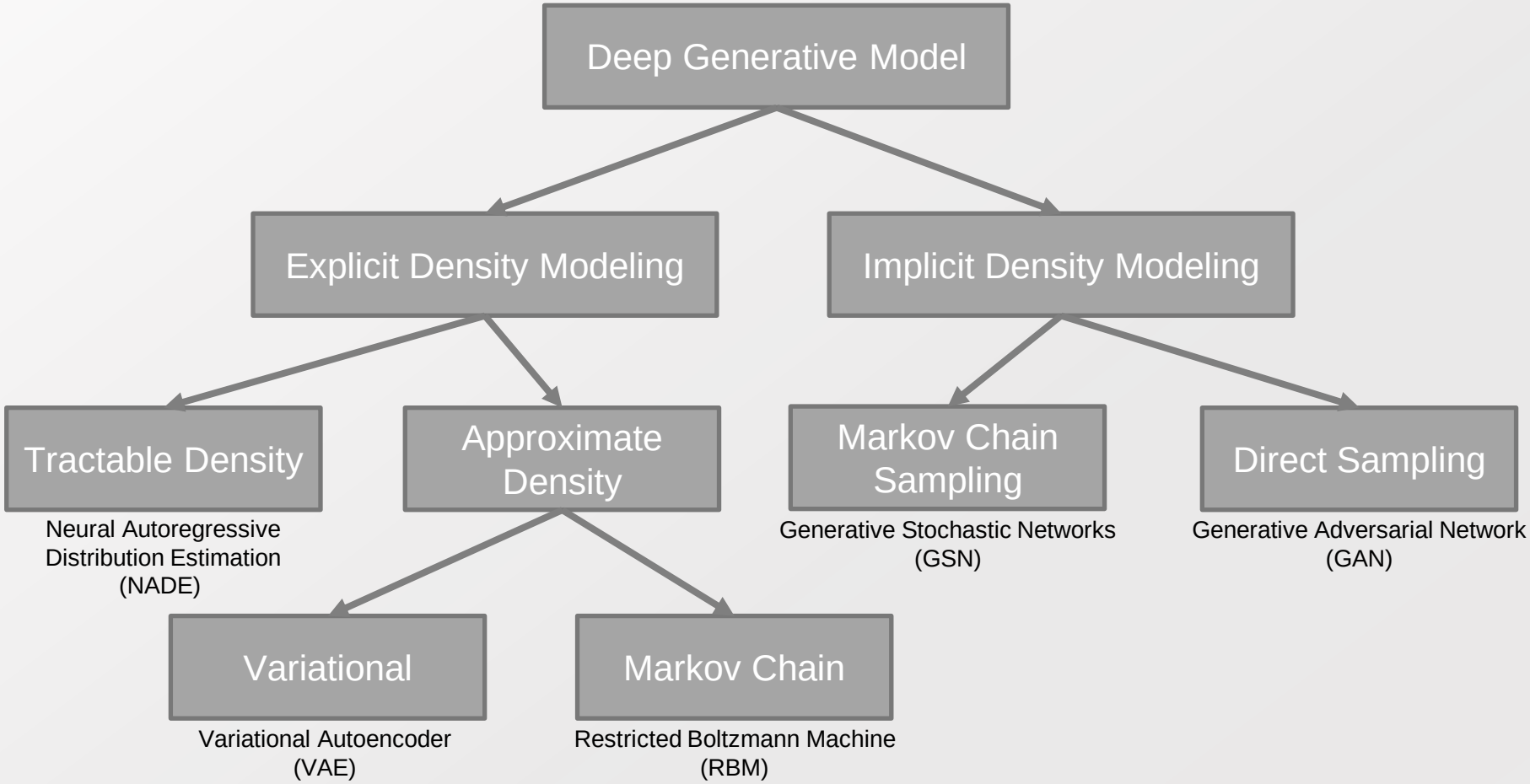
Deep Generative Models
Variational Autoencoder

Il-Chul Moon
Department of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

EXPLICIT DISTRIBUTION MODELS

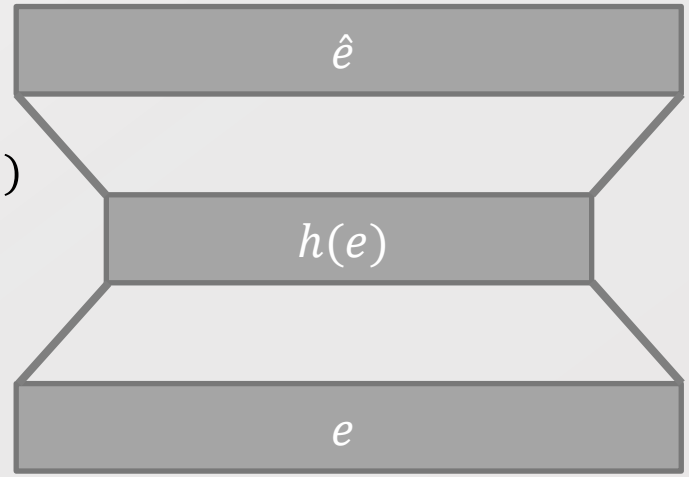
Taxonomy of Deep Generative Model

- Deep Learning + Generative Modeling
  - Why model a problem in a generative approach?
  - Good learning requires a generation of previous and new examples



Autoencoders

- Feed forward neural network trained to generate its input at the output layer
  - Simplest autoencoder : identity matrix if the hidden layer has an identical dimension to the input and the output
  - What-if we limit the dimension of the hidden layer?
    - Compression on the evidence → Latent modeling
- Typical structure
  - Decoder
    - $\hat{e} = o\big(\hat{a}(e)\big) = \sigma(c + W^{*}h(e))$
  - Encoder
    - $h(e) = g\big(a(e)\big) = \sigma(b + We)$
    - $h(e)$ becomes the feature representation
- Loss function
  - $l\big(f(e)\big) = -\sum_{d=1}^{D}(e_d \log(\widehat{e_d}) + (1 - e_d)\log(1 - \widehat{e_d}))$
    - Cross Entropy for binary cases
  - $l\big(f(e)\big) = \frac{1}{2}\sum_{d=1}^{D}(\widehat{e_d} - e_d)^2$
    - Sum of squared error
    - Linear activation at the output of the decoder
- Decoder and encoder can be multi-layered perceptrons, MLP



Detour: Setting the Minimum Criteria

- $\ln P(E) = \ln \sum_H P(H,E) = \ln \sum_H Q(H|E) \frac{P(H,E)}{Q(H|E)}$
- Since, log is a concave function
- $\ln \sum_H Q(H|E) \frac{P(H,E)}{Q(H|E)}$

$$\geq \sum_H Q(H|E) \ln \left[\frac{P(H,E)}{Q(H|E)} \right]$$

$$= \sum_H Q(H|E) \ln P(H,E) - Q(H|E) \ln Q(H|E)$$

$$= \sum_H Q(H|E) \{\ln P(E|H) + \ln P(H)\} - Q(H|E) \ln Q(H|E)$$

$$= \sum_H Q(H|E) \ln P(E|H) - Q(H|E) \frac{\ln Q(H|E)}{\ln P(H)}$$

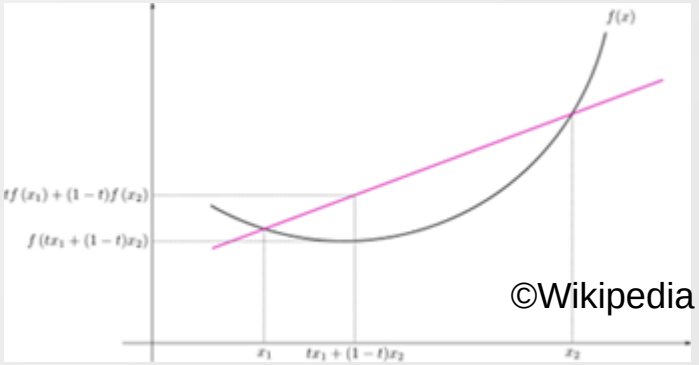
$$= E_{Q(H|E)} \ln P(E|H) - KL(Q(H|E) \parallel P(H))$$

- Using the Jensen's Inequality
- The right hand side is well known function in the statistics community
  - KL divergence

$$KL(Q\|P) = -\sum_i Q(i) \ln \left[\frac{P(i)}{Q(i)} \right]$$

Minimizing KL Divergence → Finding the true $\ln P(E)$

Jensen's Inequality



When $\varphi(x)$ is concave

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_j}\right) \geq \frac{\sum a_i \varphi(x_i)}{\sum a_j}$$

When $\varphi(x)$ is convex

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_j}\right) \leq \frac{\sum a_i \varphi(x_i)}{\sum a_j}$$

Detour: Optimizing the Lower Bound

- $\ln P(E|\theta) \geq \sum_H Q(H|E,\lambda)\ln P(H,E|\theta) - Q(H|E,\lambda)\ln Q(H|E,\lambda)$
  $= \sum_H Q(H|E)\ln P(E|H,\theta) - Q(H|E)\ln \frac{Q(H|E)}{P(H|\theta)}$
  $= E_{Q(H|E)}\ln P(E|H) - KL(Q(H|E) \parallel P(H|\theta))$
  - The lower bound of this equation is
  - $L(\lambda,\theta) = \sum_H Q(H|E,\lambda)\ln P(H,E|\theta) - Q(H|E,\lambda)\ln Q(H|E,\lambda)$
  - How to optimize the above?
    - Selecting a good λ
      - Suppose that we setup λ to make $Q(H|E,\lambda) = P(H|E,\theta)$
      - $\sum_H P(H|E,\theta)\ln P(H,E|\theta) - P(H|E,\theta)\ln P(H|E,\theta)$
        $= \sum_H P(H|E,\theta)\ln P(H|E,\theta)P(E|\theta) - P(H|E,\theta)\ln P(H|E,\theta)$
        $= \sum_H P(H|E,\theta)\ln P(E|\theta) = \ln P(E|\theta)\sum_H P(H|E,\theta) = \ln P(E|\theta)$
      - Proven lower bound
    - Readjusting θ is needed
- This results in two sets of parameters to optimize
  - Good match for the EM approach
  - (E Step): $\lambda^{t+1} = argmax_\lambda L(\lambda^t,\theta^t)$
  - (M Step): $\theta^{t+1} = argmax_\theta L(\lambda^{t+1},\theta^t)$
- However, still updating λ^t is a conceptual idea…

Detour: Amortized Analysis and Inference

- Amortized analysis
  - Complexity analysis. Let's assume that we have a stack with an array A. If we have a full array, and if we have to push an item?
    - We may create a new larger array, and we pop and push, repeatedly. This is an expensive operation compared to pop and push.
    - We need to "amortize" the payment of the expensive operation over the payment of the cheap operation
  - Need an assumption on the data size and sequence
- Amortized Inference
  - Previous inferences
    - Expensive inference
      - MCMC, Variational Inference on PGM : approximately solve an intractable integral with samplings or optimizations
  - Amortized inference
    - The inference on $P(H|E_1)$ must provide information on $P(H|E_2)$
    - Basic principle of quick human perception based upon experience
  - Need to learn the weights on calculating the posterior with prior experience
    - Then, use the weights for the inference on the later experience

Detour: Amortized Inference

- Probabilistic graphical model
  - $P(H,E) = P(E|H)\prod_i P(H_i|H_{Parent(H_i)})$
- Variational inference
  - Instead of approximating the posterior $P(H|E,\lambda)$, we make a variational distribution of $q_E(H;\phi)$ that has a simpler form with additional variational parameters, ϕ.
  - Then, optimize the KL divergence between $P(H|E,\lambda)$ and $q_E(H;\phi)$
  - As a common simpler form of $q_E(H;\phi)$, we choose the mean field assumption
    - $q_E^{MF}(H;\phi) = \prod_i q(H_i;\phi_i)$
- Amortized variational inference
  - Instead of defining a parametric $q_E(H;\phi)$ that describes the nature of E
    - Which could be a probability distribution of the model design
  - We define a general $q(H|E;\phi)$ that takes the conditional information of E
    - Which could be a general function approximation
  - $q^{MF}(H|E;\phi)$=$\prod_i q(H_i;NN_i(E;\phi_i))$
    - Training NN_i with E to reduce the difference between $q(H|E;\phi)$ and $P(H|E,\lambda)$

Variational Autoencoder

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

- Variational autoencoder (VAE) is a probabilistic autoencoder.
  - Obj. = expected negative reconstruction error + regularizer

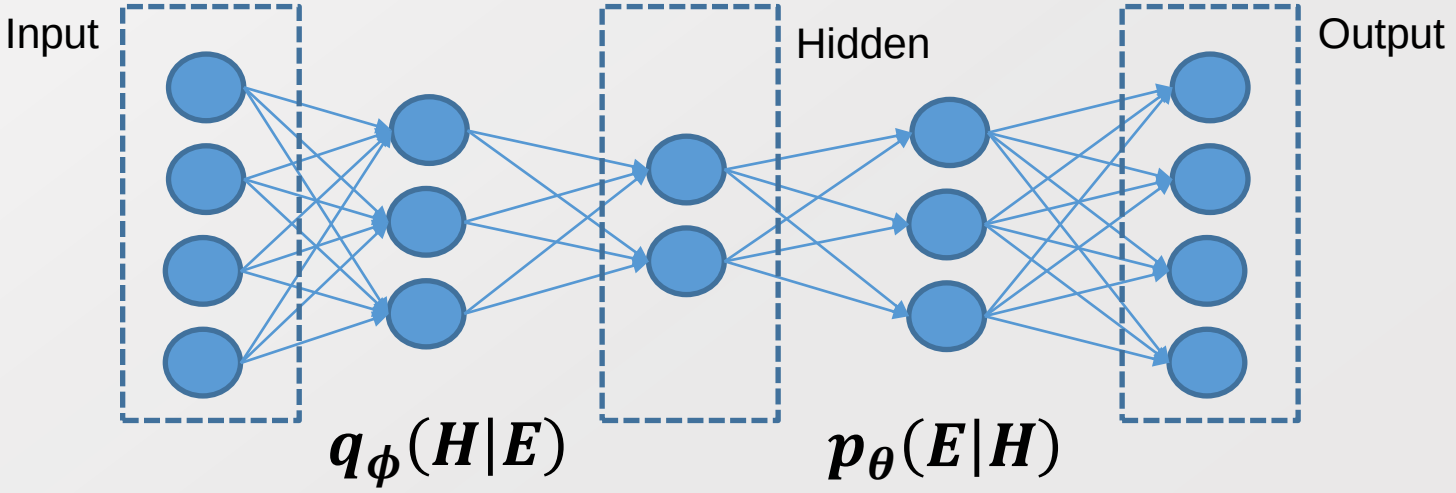
$$\mathcal{L} = \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)] - D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right)$$

Original distribution p: $H \Rightarrow E$ **generates data E from latent (hidden) variable H**

Variational distribution q: $E \Rightarrow H$ **generates hidden representation H from data E**

Probabilistic (stochastic) decoder

Probabilistic (stochastic) encoder



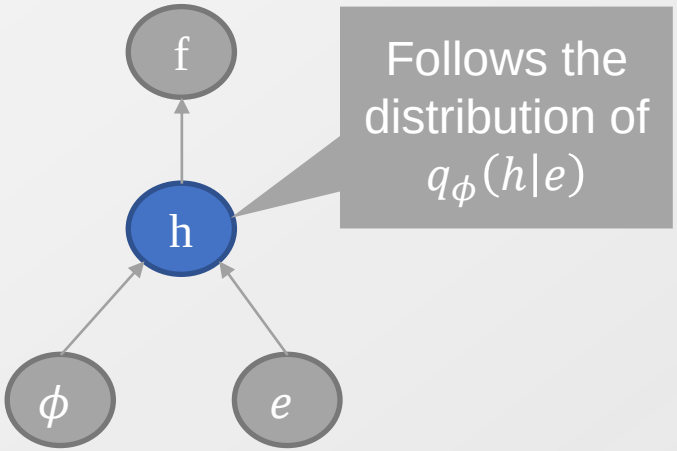
Objective Function in VAE

- Evidence lower bound
  - $\mathbb{E}_{q_\phi(h|e)}[\log p_\theta(e|h)] - D_{KL}\left(q_\phi(h|e)||p_\theta(h)\right)$
- Reparameterization Trick or Stochastic Gradient Variational Bayes
  - $q_\phi(h|e)$ can be reparameterized to differentiable and deterministic variable.
  - For $\tilde{z}\sim q_\phi(h|e)$: $\tilde{z} = g_\phi(\epsilon, e)$ with $\epsilon\sim p(\epsilon)$
    - A differentiable transformation g_ϕ
    - Applying amortized inference $q(H|E;\phi)=\prod_i q(H_i; NN_i(E;\phi_i))$ instead of $q_E(H;\phi)$
    - Consider $q_\phi(h|e)\sim N(\mu(e;\phi), \Sigma(e;\phi))$
      - $g_\phi(\epsilon, e) = \mu(e;\phi) + \epsilon\sqrt{\Sigma(e;\phi)}, \epsilon\sim N(0,1)$
      - $\mu_i(e;\phi) = NN_i^\mu(e|E;\phi_i^\mu)$, $\Sigma_i(e;\phi) = NN_i^\Sigma(e|E;\phi_i^\Sigma)$
  - Monte-Carlo estimates for some function $f(h)$ w.r.t. $q_\phi(h|e)$:
    - $\mathbb{E}_{q_\phi(H|E)}[f(h)] = \mathbb{E}_{p(\epsilon)}\left[f\left(g_\phi(\epsilon, e)\right)\right] \approx \frac{1}{D}\sum_{d=1}^{D} f\left(g_\phi\left(\epsilon^{(d)}, e\right)\right)$
    - where $\epsilon^{(d)}\sim p(\epsilon)$

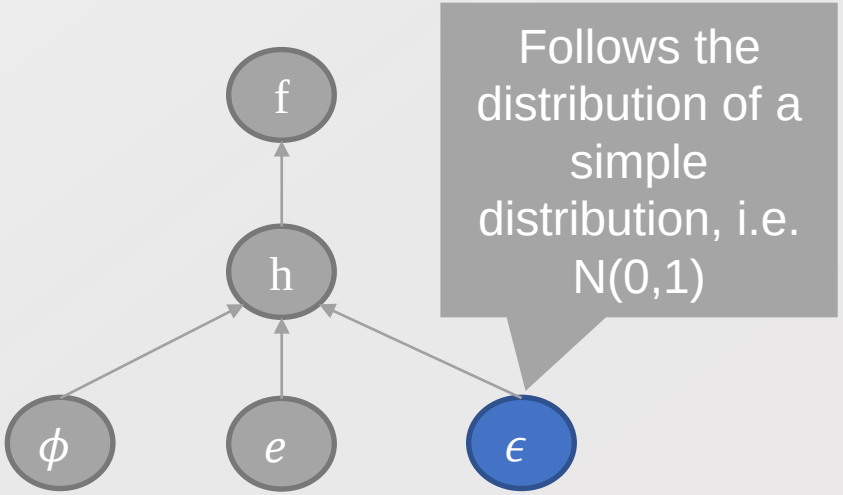
Reparameterization as Structure

- Reparameterization Trick
  - $\mathbb{E}_{q_\phi(H|E)}[f(h)] = \mathbb{E}_{p(\epsilon)}\left[f\left(g_\phi(\epsilon, e)\right)\right] \approx \frac{1}{L}\sum_{l=1}^{L} f\left(g_\phi\left(\epsilon^{(l)}, e\right)\right)$
  - $\tilde{h} = g_\phi(\epsilon, e)$ with $\epsilon \sim p(\epsilon)$
- Enabling the back-propagation by removing the random nodes from the back-propagation learned variable

$\mathbb{E}_{q_\phi(H|E)}[f(h)]$



$\mathbb{E}_{p(\epsilon)}\left[f\left(g_\phi(\epsilon, e)\right)\right]$



Derivation of Evidence Lower Bound for Gaussian VAE (1)

- $\mathcal{L} = -D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) + \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)]$
  - ELBO requires the instantiation of $q_\phi(H|E)$, $p_\theta(H)$ and $p_\theta(E|H)$.
    - From variational inference, we know that it would be better to match the variational distribution and the model distribution on the latent variable
    - Gaussian VAE assumes $q_\phi(H|E)$ and $p_\theta(H)$ to follow the Gaussian distribution
      - $q_\phi(H|E) \sim N(\mu, \sigma^2 I)$, $p_\theta(H) \sim N(0, I)$
  - After deciding the PDF of $q_\phi(H|E)$ and $p_\theta(H)$, we can calculate the KL divergence
    - $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) = -\int q_\phi(H|E)\log p_\theta(H)\,dH + \int q_\phi(H|E)\log q_\phi(H|E)\,dH$
      - $-\int q_\phi(H|E)\log p_\theta(H)\,dH = -\int q_\phi(H|E)\log \frac{1}{(2\pi(1)^2)^{\frac{1}{2}}}\exp\left(-\frac{(H-0)^2}{2(1)^2}\right)dH$

        $= -\int q_\phi(H|E)\log\frac{1}{(2\pi(1)^2)^{\frac{1}{2}}}\,dH - \int q_\phi(H|E)\left\{-\frac{(H-0)^2}{2(1)^2}\right\}dH$

        $= -\log\frac{1}{(2\pi(1)^2)^{\frac{1}{2}}}\int q_\phi(H|E)dH + \int\frac{1}{2}H^2 q_\phi(H|E)dH = \frac{1}{2}\log 2\pi + \int\frac{1}{2}H^2 q_\phi(H|E)dH$

        $= \frac{1}{2}\log 2\pi + \frac{1}{2}(\sigma^2+\mu^2)$
        - $\sigma^2 = \int H^2 q_\phi(H|E)dH - \left\{\int H^2 q_\phi(H|E)dH\right\}^2 = \int H^2 q_\phi(H|E)dH - \mu^2 \rightarrow \int H^2 q_\phi(H|E)dH = \sigma^2 + \mu^2$

$$KL(Q\|P) = -\sum_i Q(i)\ln\left[\frac{P(i)}{Q(i)}\right]$$

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

$$\boldsymbol{N(x|\mu,\Sigma)} = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

Derivation of Evidence Lower Bound for Gaussian VAE (2)

$$KL(Q\|P) = -\sum_i Q(i)\ln\left[\frac{P(i)}{Q(i)}\right]$$

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

$$\boldsymbol{N(x|\mu,\Sigma)} = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

- $\mathcal{L} = -D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) + \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)]$
  - $q_\phi(H|E) \sim N(\mu, \sigma^2 I)$, $p_\theta(H) \sim N(0, I)$
  - $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) = -\int q_\phi(H|E)\log p_\theta(H)\,dH + \int q_\phi(H|E)\log q_\phi(H|E)\,dH$
    - $\int q_\phi(H|E)\log q_\phi(H|E)\,dH = \int q_\phi(H|E)\log\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left(-\frac{(H-\mu)^2}{2\sigma^2}\right)dH$

      $= \int q_\phi(H|E)\log\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}dH + \int q_\phi(H|E)\left\{-\frac{H^2-2\mu H+\mu^2}{2\sigma^2}\right\}dH$

      $= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\{\int H^2 q_\phi(H|E)dH - 2\mu\int H q_\phi(H|E)dH + \mu^2\int q_\phi(H|E)dH\right\}$

      $= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\{\sigma^2+\mu^2-2\mu\times\mu+\mu^2\} = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2} = -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}$
      - $\sigma^2 = \int H^2 q_\phi(H|E)dH - \left\{\int H^2 q_\phi(H|E)dH\right\}^2 = \int H^2 q_\phi(H|E)dH - \mu^2 \rightarrow \int H^2 q_\phi(H|E)dH = \sigma^2 + \mu^2$
- $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) = \frac{1}{2}\log 2\pi + \frac{1}{2}(\sigma^2+\mu^2) - \frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}$

  $= \frac{1}{2}(\sigma^2+\mu^2-1) - \log\sigma$

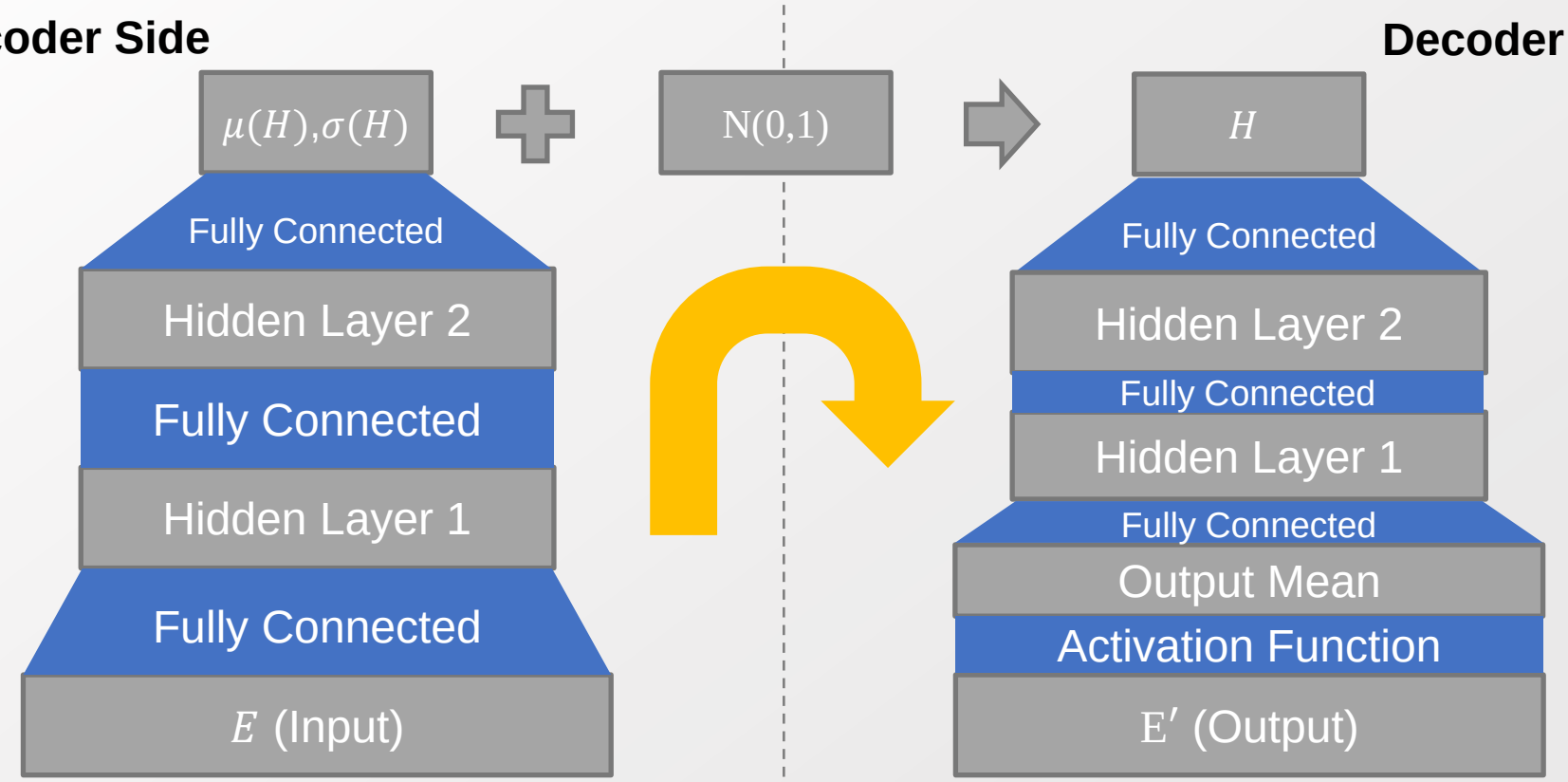
Derivation of Evidence Lower Bound for Gaussian VAE (3)

- $\mathcal{L} = -D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) + \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)]$
  - $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) = \frac{1}{2}(\sigma^2 + \mu^2 - 1) - \log\sigma$
  - $\mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)] = \frac{1}{D}\sum_{d=1}^{D}\left(\log p_\theta\left(E\middle|H^{(d)}\right)\right)$
    - Is a negative cross entropy and can be computed empirically through Monte-carlo
    - $= \frac{1}{D}\sum_{d=1}^{D}\left(\log p_\theta\left(E\middle|E^{(d)}\right)\right)$
      - VAE has a deterministic decoder to project $H^{(d)}$ to $E^{(d)}$
      - Depending upon the assumption of E_i and $p_\theta\left(E\middle|E^{(d)}\right)$
        - If E_i is continuous and $p_\theta\left(E_i\middle|E_i^{(d)}\right) \sim \mathrm{N}\left(E_i^{(d)}, 1^2\right)$
          - $\log p_\theta\left(E\middle|E^{(d)}\right) = C \times \log\exp(-\frac{(E_i - E_i^{(d)})^2}{2}) = -C \times (E_i - E_i^{(d)})^2$
        - If E_i is discrete and $p_\theta\left(E_i\middle|E_i^{(d)}\right) \sim \mathrm{Bern}\left(E_i^{(d)}\right)$
          - $\log p_\theta\left(E\middle|E^{(d)}\right) = \log\{E^{(d)^{E_i}}(1 - E^{(d)})^{1-E_i}\} = E_i \log E^{(d)} + (1 - E_i)\log(1 - E^{(d)})$

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

$\boldsymbol{N(x|\mu,\Sigma)}$
$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$

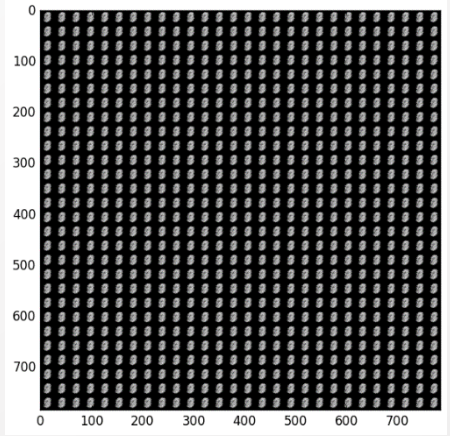
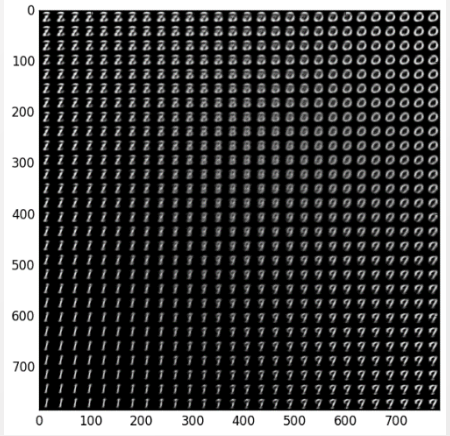
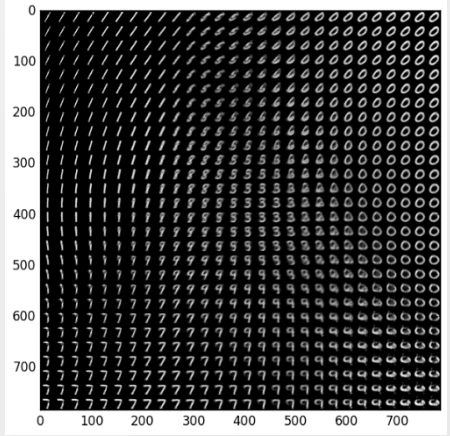
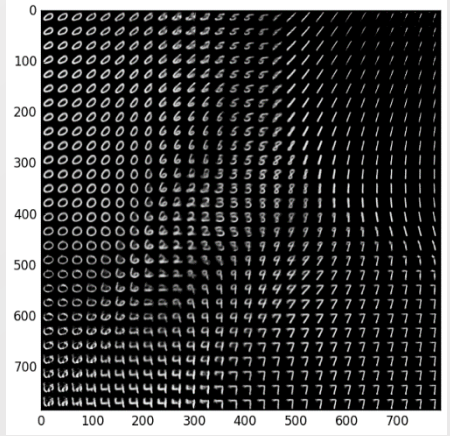
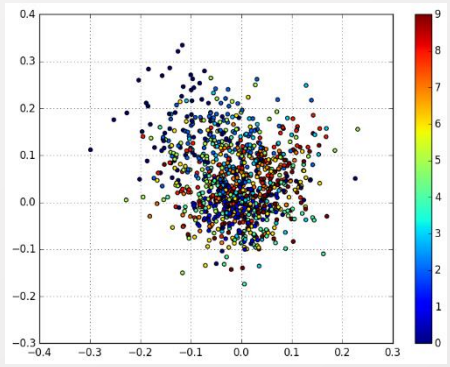
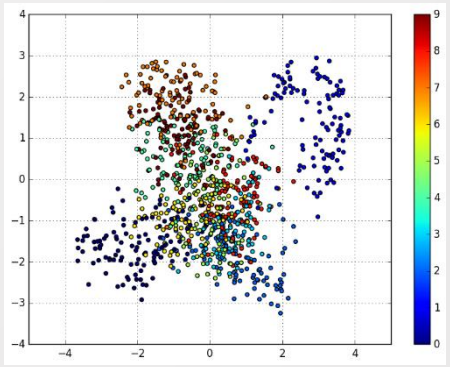
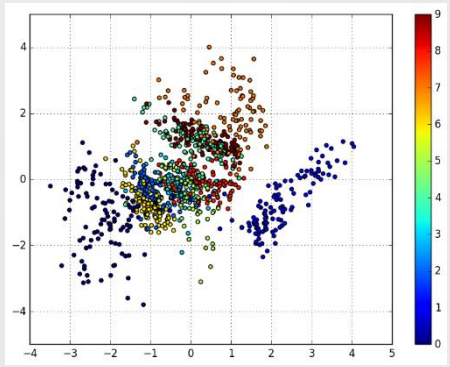
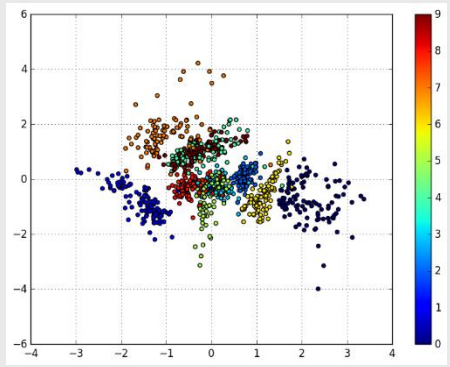
Feed-Forward Neural Net. Of VAE



- Learning the variational q distribution with a neural network
  - For training with the back-propagation and the gradient method
    - Re-parameterization trick is used
  - The latent needs to be continuous

Manifold Learning Result with VAE

- 2 Dimension of Z variable
  - Reconstruction from the manifold
  - Training instances in the 2 dimension

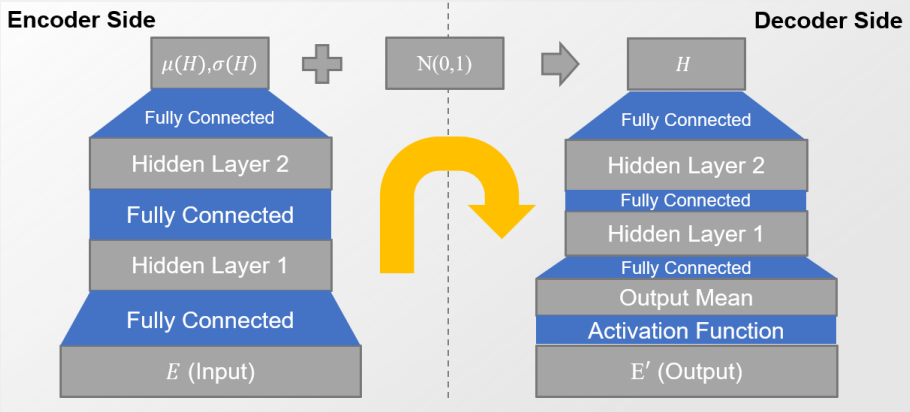
Iteration 1	Iteration 25	Iteration 125	Iteration 625
			
			

Variational Autoencoder for Document Modeling

- Neural Variational Document Model
  - Input as the bag of words
  - Change the decoder side

$$P_\theta(e_i|h) = \frac{\exp\{-E(e_i; h, \theta)\}}{\sum_{j=1}^{|V|} \exp\{-E(e_j; h, \theta)\}}$$

$$E(e_i; h, \theta) = -h^T R e_i - b_{e_i}$$



Manifold Embedding, or similarity of learned semantic vector, R

Word	weapons	medical	companies	define	israel	book
NVDM	guns	medicine	expensive	defined	israeli	books
	weapon	health	industry	definition	arab	reference
	gun	treatment	company	printf	arabs	guide
	militia	disease	market	int	lebanon	writing
	armed	patients	buy	sufficient	lebanese	pages
NADE	weapon	treatment	demand	defined	israeli	reading
	shooting	medecine	commercial	definition	israelis	read
	firearms	patients	agency	refer	arab	books
	assault	process	company	make	palestinian	relevent
	armed	studies	credit	examples	arabs	collection

(b) The five nearest words in the semantic space.

Sorted list of words with association to H in R

Space	*Religion*	*Encryption*	*Sport*	*Policy*
orbit	muslims	rsa	goals	bush
lunar	worship	cryptography	pts	resources
solar	belief	crypto	teams	charles
shuttle	genocide	keys	league	austin
moon	jews	pgp	team	bill
launch	islam	license	players	resolution
fuel	christianity	secure	nhl	mr
nasa	atheists	key	stats	misc
satellite	muslim	escrow	min	piece
japanese	religious	trust	buf	marc

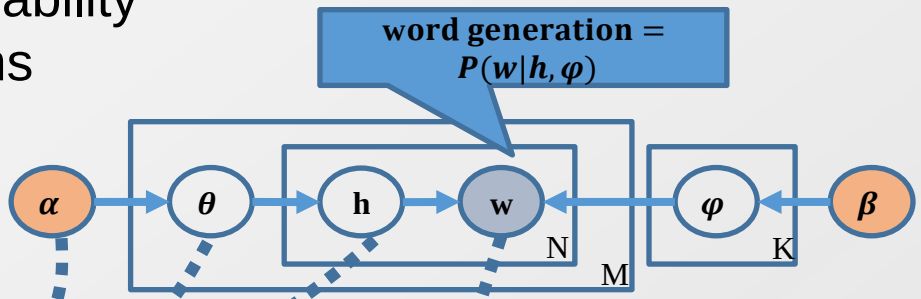
Table 2. The topics learned by NVDM on 20News.

Miao, Yishu, Lei Yu, and Phil Blunsom. "Neural variational inference for text processing." *Proc. ICML*. 2016.

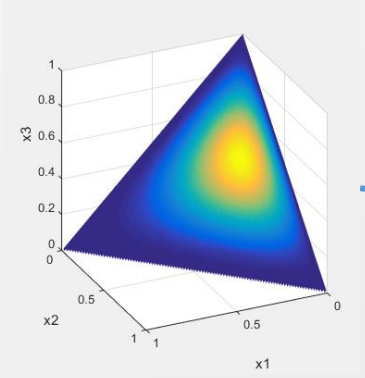
VARIANTS OF VARIATIONAL AUTOENCODER WITH CONDITIONAL PROBABILITY

Detour: Probabilistic Graphical Model

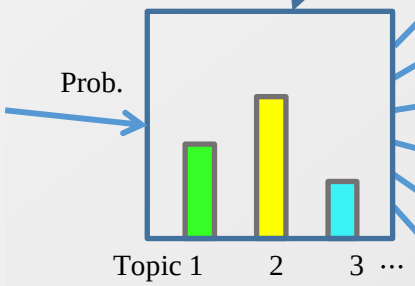
- Probabilistic model with subject matter knowledge
  - Expression from conditional probability
    → Model specification with graphs
- Easier adaptation of expert knowledge and communication



2000~ Probabilistic Graphical Model…



Dirichlet Distribution Prior



Document specific topic proportion

Topic assignment

Estimating Optimized Bidding Price in Virtual Electricity Wholesale Market

Power TAC(Power Trading Agent Competition) is an agent-based simulation for competitions between electricity brokering agents on the smart grid. To win the competition, agents sell their tariff plans to customers and obtain electricity from the power plants. In this operation, a key to success is balancing the demand of the customer and the supply from the

Conditional Variational Autoencoder

- Up to this slide, there was no labels on data instances
  - Only learning the latent representations of given data instances
  - However, there are many cases that we have labels as additional information
- Conditional generation of data instance
  - Original VAE : $\mathcal{L} = -D_{KL}\left(q_\phi(h|e)||p_\theta(h)\right) + \mathbb{E}_{q_\phi(h|e)}[\log p_\theta(e|h)]$
  - Conditional VAE : $\mathcal{L} = -D_{KL}\left(q_\phi(h|e,y)||p_\theta(h|y)\right) + \mathbb{E}_{q_\phi(h|e,y)}[\log p_\theta(e|y,h)]$
    - $\log p(e|y) = \log \sum_h q(h|e,y)\frac{p(e,h|y)}{q(h|e,y)} \geq \sum_h q(h|e,y)\left[\log p(e,h|y) - \log q(h|e,y)\right]$

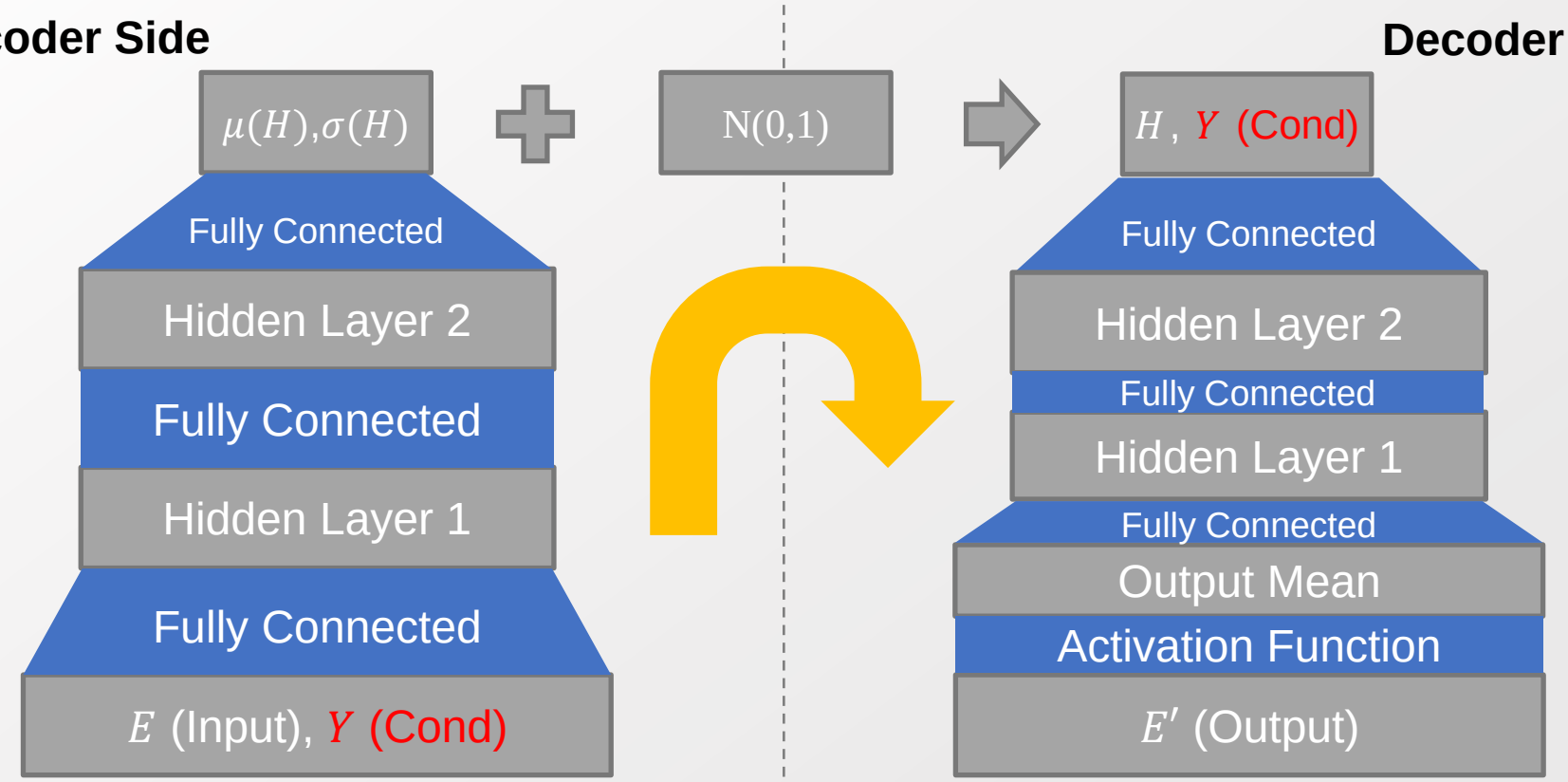
      $= \sum_h q(h|e,y)\left[\log\{p(e|h,y)p(h|y)\} - \log q(h|e,y)\right]$

      $= \mathbb{E}_{q(h|e,y)}[\log p(e|h,y)] + \mathbb{E}_{q(h|x,y)}[\log p(h|y) - \log q(h|e,y)]$

      $= \mathbb{E}_{q(h|e,y)}[\log p(e|h,y)] - D_{KL}(\log p(h|y)\,||\,\log q(h|e,y))$
    - Under the assumption that we do not change the prior $p_\theta(h|y)$ by y, there is no change in the ELBO derivation

Sohn, Kihyuk, Honglak Lee, and Xinchen Yan. "Learning structured output representation using deep conditional generative models." *Advances in neural information processing systems*. 2015.

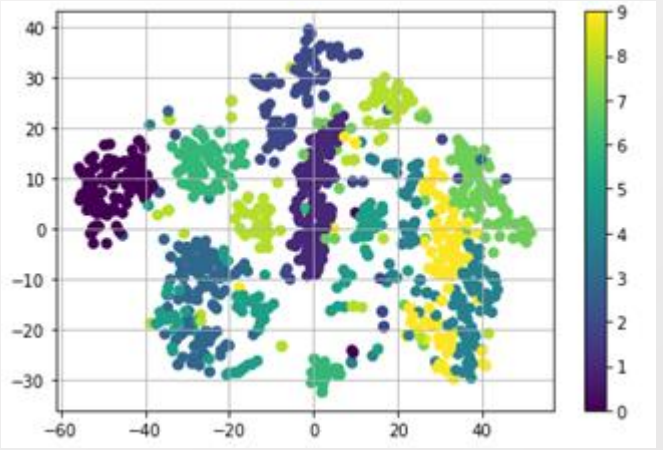
Feed-Forward Neural Net. Of Cond. VAE

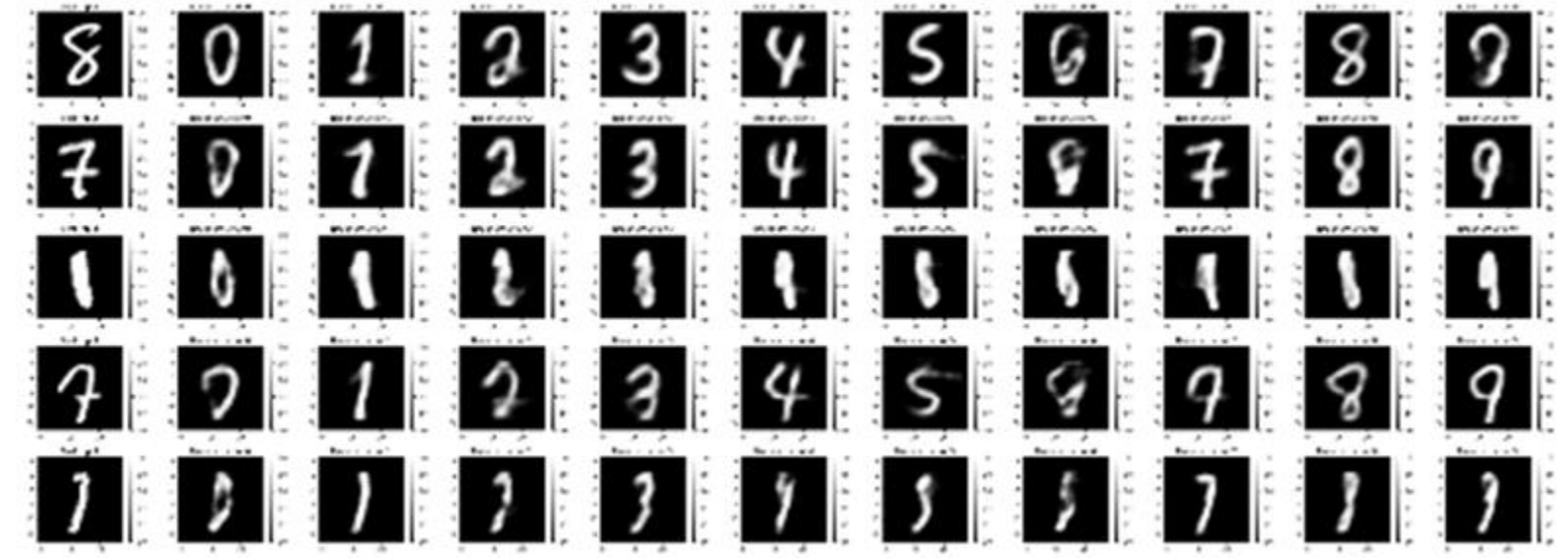


- Original VAE : $\mathcal{L} = -D_{KL}\left(q_\phi(h|e)||p_\theta(h)\right) + \mathbb{E}_{q_\phi(h|e)}[\log p_\theta(e|h)]$
- Conditional VAE : $\mathcal{L} = -D_{KL}\left(q_\phi(h|e,y)||p_\theta(h|y)\right) + \mathbb{E}_{q_\phi(h|e,y)}[\log p_\theta(e|y,h)]$
  - y is concatenated to $q_\phi(h|e,y)$ and $p_\theta(e|y,h)$
  - Under the same prior assumption, $p_\theta(h|y) \sim N(0,1)$ regardless of y

Results of Conditional VAE

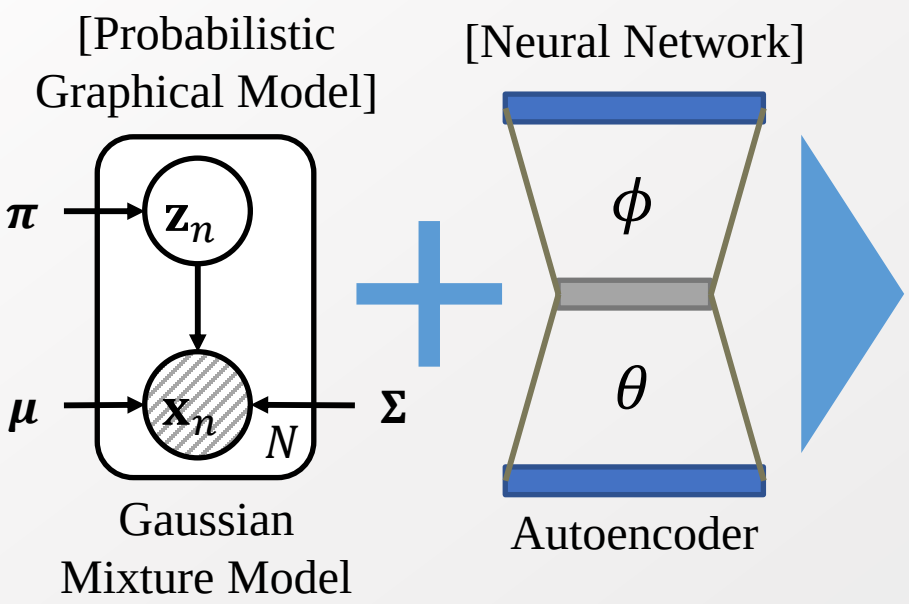
- Conditional reconstruction
  - $\mathcal{L} = -D_{KL}\left(q_\phi(h|e,y)||p_\theta(h|y)\right) + \mathbb{E}_{q_\phi(h|e,y)}[\log p_\theta(e|y,h)]$
    - $p_\theta(e|y,h)$: Decoding structure of CVAE
      - Depending on the latent variable h and the condition variable y
    - $q_\phi(h|e,y)$: Encoding structure of CVAE
      - Depending on the observation variable e and the condition variable y

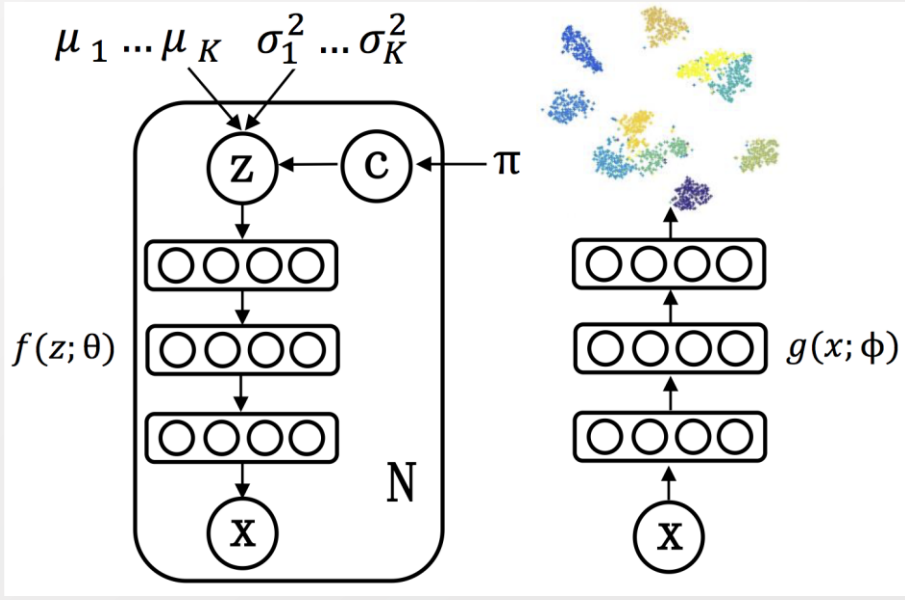




Variational Deep Embedding (VaDE)

Zhuxi, Jiang, et al. "Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering" *IJCAI* (2017).





- The Generative Process of VADE
  1. Choose a cluster $c \sim Cat(\boldsymbol{\pi})$
  2. Choose a latent vector $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2 \boldsymbol{I})$
  3. Choose a sample $\mathbf{e}$
     - If $\mathbf{e}$ is binary, $\mathbf{e} \sim Ber(\boldsymbol{\mu}_e)$
     - If $\mathbf{e}$ is real-valued, $\mathbf{e} \sim \mathcal{N}(\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2 \boldsymbol{I})$

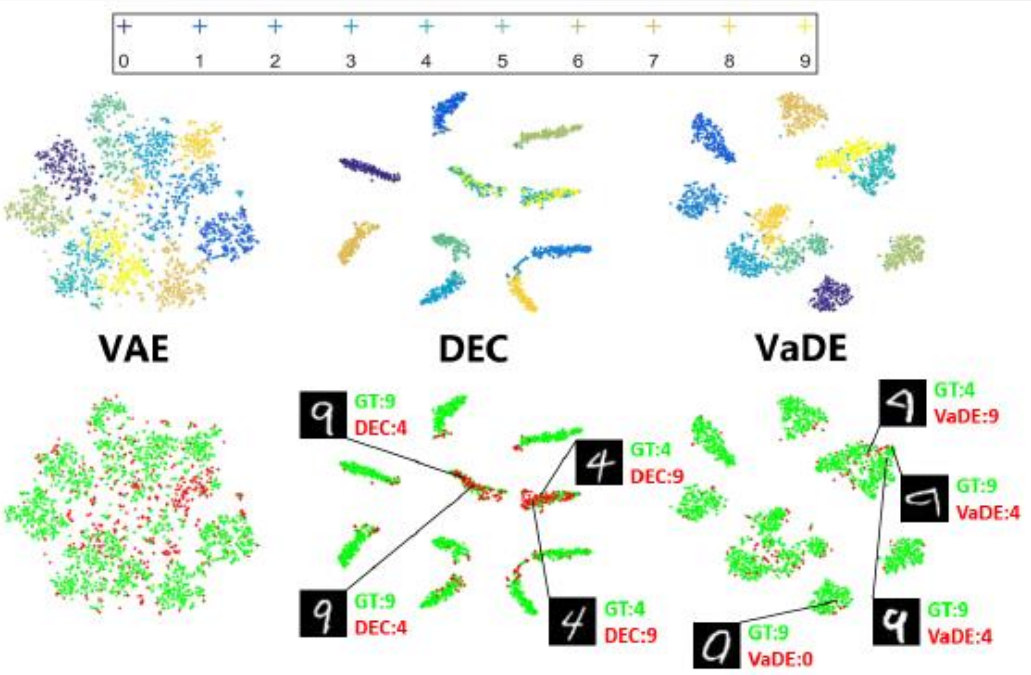
- $p(\mathbf{e}, \mathbf{h}, c) = p(\mathbf{e}|\mathbf{h})p(\mathbf{h}|c)p(c)$
  - $p(c) = Cat(c|\boldsymbol{\pi})$
  - $p(\mathbf{h}|c) = \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2 \boldsymbol{I})$
  - $p(\mathbf{e}|\mathbf{h}) = Ber(\mathbf{e}|\boldsymbol{\mu}_e)$ *or* $\mathcal{N}(e|\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2 \boldsymbol{I})$
- $\approx q_\phi(\mathbf{h}, c|\boldsymbol{e})$

Key Idea on Probabilistic Modeling

- In Variational AutoEncoders,
  - Posterior: $q_\phi(\boldsymbol{h}|\boldsymbol{e}) = \mathcal{N}(\boldsymbol{h}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$
    $\rightarrow h = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
- In Conditional VAE
  - ELBO : $\mathcal{L} = -D_{KL}\left(q_\phi(h|e,y)||p_\theta(h|y)\right) + \mathbb{E}_{q_\phi(h|e,y)}[\log p_\theta(e|y,h)]$
- In VaDE,
  - Posterior: $q_\phi(\boldsymbol{h}, c|\mathbf{e})$
    - $q_\phi(\boldsymbol{h}|\boldsymbol{e}) = \sum_c q_\phi(\boldsymbol{h}, c|\mathbf{e}) = \sum_k \boldsymbol{\pi}_k \mathcal{N}\left(\boldsymbol{h}; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2\boldsymbol{I}\right)$ → Reparameterization
      - $q_\phi(\boldsymbol{h}, c|\mathbf{x}) = q_\phi(\boldsymbol{h}|\mathbf{e})q_\phi(c|\mathbf{e}) = \mathcal{N}(\boldsymbol{h}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2\boldsymbol{I})$
        - Mean field assumption
  - ELBO derivation:
    - $\log p(\boldsymbol{e}) = \log \int_h \sum_c p(\mathbf{e}, \mathbf{h}, c)\, d\mathrm{h} \geq \mathbb{E}_{q(\mathrm{h},c|\mathrm{e})}\left[\log \frac{p(\mathrm{e},\mathrm{h},c)}{q(\mathrm{h},c|\mathrm{e})}\right] = \mathcal{L}_{ELBO}$
    - $\mathcal{L}_{ELBO}(\boldsymbol{e}) = \mathbb{E}_{q(\mathrm{h},c|\mathrm{e})}\left[\log \frac{p(\mathrm{e},\mathrm{h},c)}{q(\mathrm{h},c|\mathrm{e})}\right] = \mathbb{E}_{q(\mathrm{h},c|e)}[\log p(\mathrm{e},\mathrm{h},c) - \log q(\mathrm{h},c|e)]$
      $= \mathbb{E}_{q(\mathrm{h},c|\mathrm{e})}[\log p(\mathrm{e}|h) + \log p(\mathrm{h}|c) + \log p(c) - \log q(\mathrm{h}|\mathrm{e}) - \log q(c|e)]$
- $p(\mathbf{e}, \mathbf{h}, c) = p(\mathbf{e}|\mathbf{h})p(\mathbf{h}|c)p(c)$
  - $p(c) = Cat(c|\boldsymbol{\pi})$
  - $p(\mathbf{h}|c) = \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2\boldsymbol{I})$
  - $p(\mathbf{e}|\mathbf{h}) = Ber(\mathbf{e}|\boldsymbol{\mu}_e)$ *or* $\mathcal{N}(e|\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e^2\boldsymbol{I})$
- $\approx q_\phi(\mathbf{h}, c|\boldsymbol{e})$

VaDE Experimental Result

▼ #Clust.=10



▲ #Clust.=7

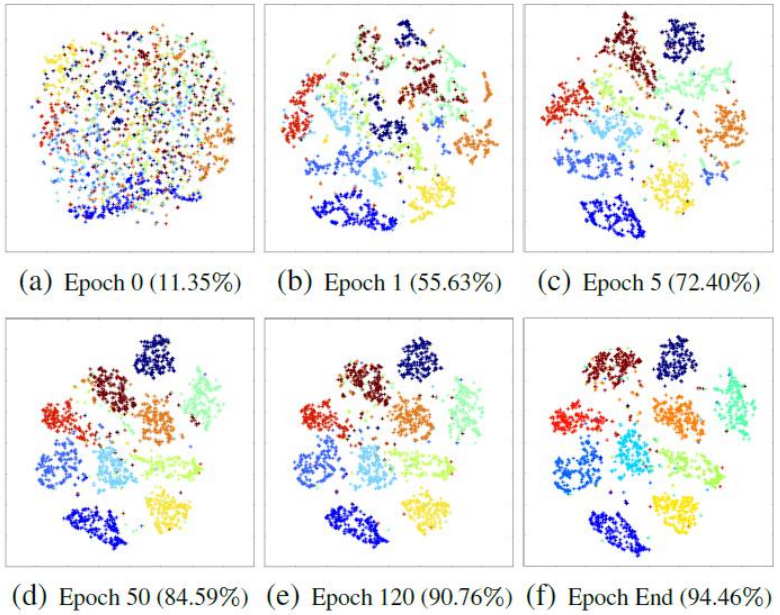


Figure 3: The illustration about how data is clustered in the latent space learned by VaDE during training on MNIST. Different colors

Method	MNIST	HHAR	REUTERS-10K	REUTERS	STL-10
GMM	53.73	60.34	54.72	55.81	72.44
AE+GMM	82.18	77.67	70.13	70.98	79.83
VAE+GMM	72.94	68.02	69.56	60.89	78.86
LDMGI	84.09†	63.43	65.62	N/A	79.22
AAE	83.48	83.77	69.82	75.12	80.01
DEC	84.30†	79.86	74.32	75.63†	80.62
VaDE	**94.46**	**84.46**	**79.83**	**79.38**	**84.45**

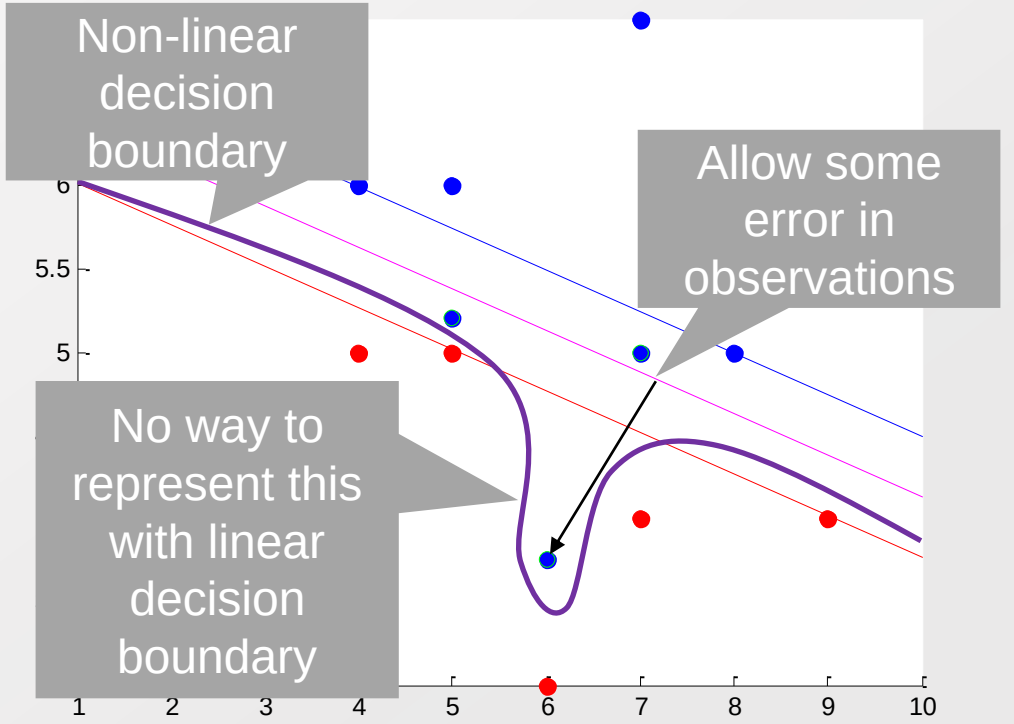
†: Taken from [Xie *et al.*, 2016].

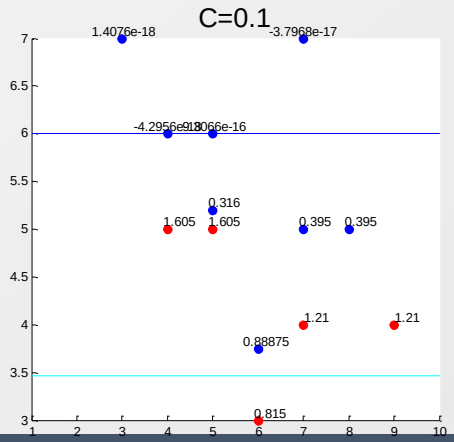
Table 2: Clustering accuracy (%) performance comparison on all datasets.

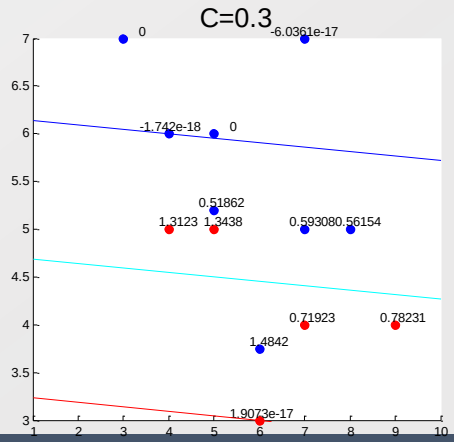
VARIANTS OF VARIATIONAL AUTOENCODER WITH ELABORATED LOSSES

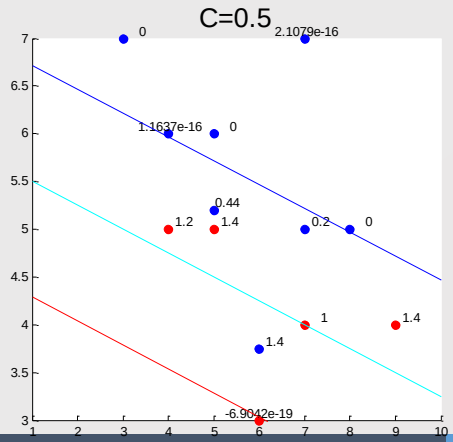
Detour:"Error" Cases in SVM

- Data points that are
  - Impossible to classify with a linear decision boundary
- So called, "error" cases…
- How to manage these?
  - $\min_{w,b} \|w\| + C \sum_j \xi_j$
  - $s.t. \left(we_j + b\right) y_j \geq 1 - \xi_j, \forall j$
  - $\xi_j \geq 0, \forall j$





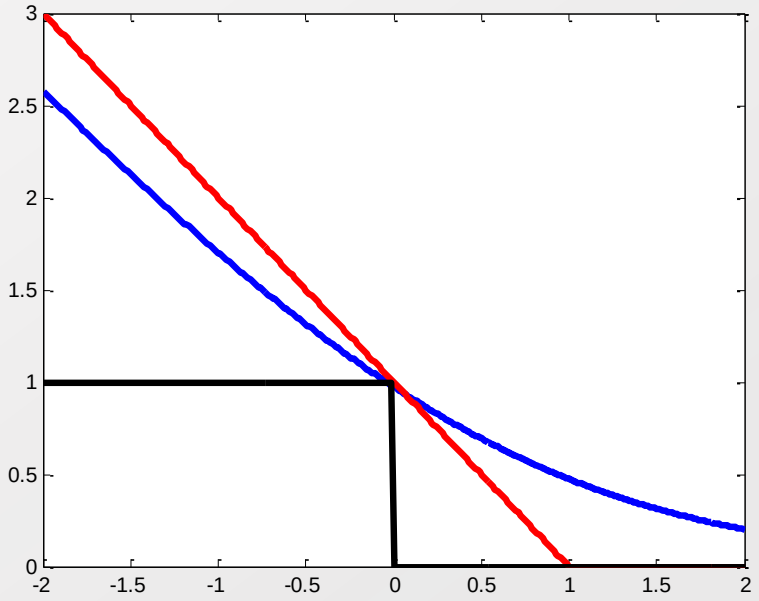


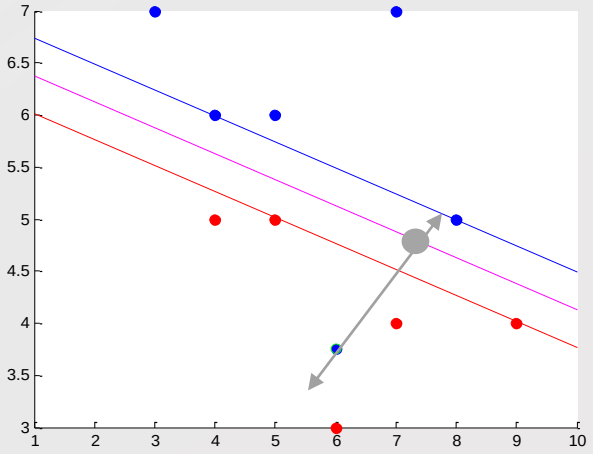


Detour: Comparison to Logistic Regression

- Loss function
  - $\xi_j = loss(f(e_j), y_j)$
- SVM loss function: Hinge Loss
  - $\xi_j = (1 - (we_j + b)y_j)_+$
- Logistic Regression
  loss function: Log Loss
  - $\hat{\theta} = \underset{\theta}{\text{argmax}} \sum_{1 \le i \le N} log(P(Y_i|E_i;\theta))$

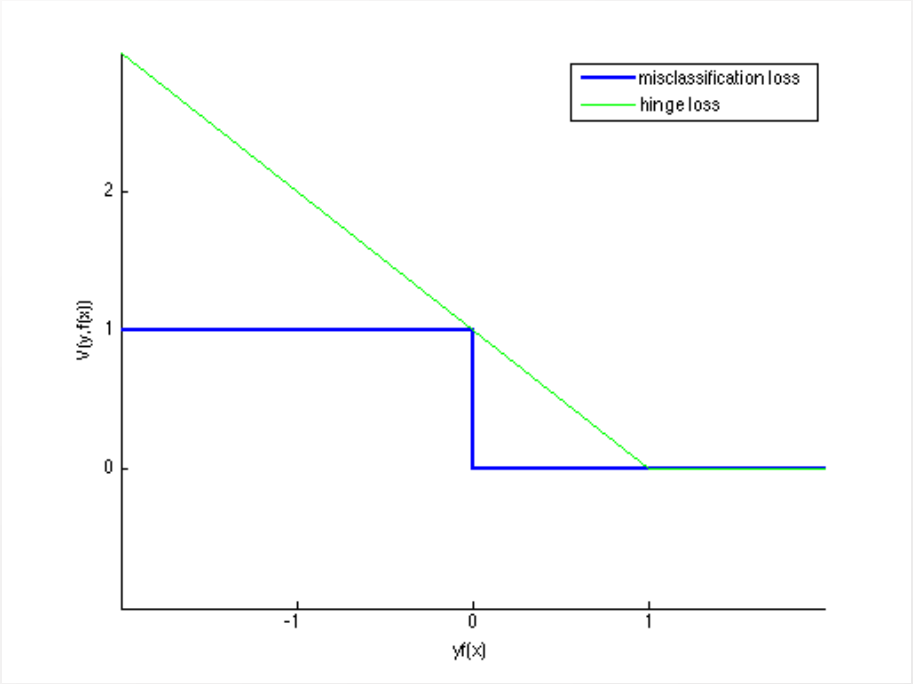
    $= \underset{\theta}{\text{argmax}} \sum_{1 \le i \le N} \{Y_i E_i \theta - \log(1 + e^{E_i\theta})\}$
  - $\xi_j = -\log\left(P(Y_j|E_j, w, b)\right) = \log\left(1 + e^{(wE_j+b)y_j}\right)$
- Which loss function is preferable?
  - Around the decision boundary?
  - Overall place?





Detour: Regularization and SVM

$$f = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} V\left(y_i, f(x_i)\right) + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$



$$V\left(y_i, f(x_i)\right) = \left(1 - yf(x)\right)_+$$

$$(s)_+ = \max(s, 0)$$

$$f = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(1 - yf(x)\right)_+ + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

$$f = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ C \sum_{i=1}^{n} \left(1 - yf(x)\right)_+ + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right\}$$

$$C = \frac{1}{2\lambda n}$$

- Support vector is a special case of regularization with the hinge loss

β-Variational Autoencoder

$$\min_{w,b} \|w\| + C\sum_j \xi_j$$
$$s.t.\,(we_j + b)y_j \geq 1 - \xi_j, \forall j$$
$$\xi_j \geq 0, \forall j$$

- From the generative modeling perspective, ELBO is derived
  - from the Jensen's inequality
  - with an assumed variational distribution
  - for a latent variable model
    - $\ln P(E) = \ln \sum_H P(H,E) = \ln \sum_H Q(H|E)\frac{P(H,E)}{Q(H|E)}$

      $\geq \sum_H Q(H|E) \ln \left[\frac{P(H,E)}{Q(H|E)}\right] = -D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) + \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)]$
- From the original autoencoder's perspective,
  - Autoencoder with a regularization imposed by a prior distribution
  - Regularization method from the Lagrange method and the soft margin of SVM
    - $max_{\phi,\theta}\mathbb{E}_{E\sim D}[\mathbb{E}_{H\sim q_\phi(H|E)}[\log p_\theta(E|H)]]$, Subject to $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) < \epsilon$
    - $F(\phi,\theta,\beta;E,H) = \mathbb{E}_{H\sim q_\phi(H|E)}[\log p_\theta(E|H)] - \beta\left(D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) - \epsilon\right)$

      $\geq \mathbb{E}_{H\sim q_\phi(H|E)}[\log p_\theta(E|H)] - \beta D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right)$
  - β becomes the hyperparameter of the regularization
    - β = 1 → Original VAE
    - $\beta > 1$ → Stronger regularization with $p_\theta(H)\sim N(0,1^2)$
      - The covariance structure of 1^2 impose the independence between the latent dimensions of H
      - Called, Latent disentanglemnt

Evidence Lower Bound with Neural Net.

- Given the evidence lower bound, we can perform derivations on multiple directions
  - $\mathcal{L} = -D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) + \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)]$

    $= \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)] + \sum_H q_\phi(H|E)[\log p_\theta(H) - \log q_\phi(H|E)]$

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

    $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) = -\sum_H q_\phi(H|E)\log\frac{p_\theta(H)}{q_\phi(H|E)}$
  - $\mathcal{L} = \mathbb{E}_{h\sim q(H|E,\phi)}[\log p(E,H|\theta) - \log q(H|E,\phi)] = \mathbb{E}_{h\sim q(H|E,\phi)}\left[\log\frac{p(E,H|\theta)}{q(H|E,\phi)}\right]$

    $= \mathbb{E}_{\epsilon\sim N(0,I)}\left[\log\frac{p(e,h(\epsilon,e,\theta)|\theta)}{q(h(\epsilon,e,\theta)|e,\theta)}\right]$
- Approximate expectation by generating k samples with ϵ
  - $\mathcal{L} = \mathbb{E}_{\epsilon\sim N(0,I)}\left[\log\frac{p(e,h(\epsilon,e,\theta)|\theta)}{q(h(\epsilon,e,\theta)|e,\theta)}\right] = \frac{1}{k}\sum_{i=1..k}\log\frac{p(e,h(\epsilon_i,e,\theta)|\theta)}{q(h(\epsilon_i,e,\theta)|e,\theta)}$

    $= \frac{1}{k}\sum_{i=1..k}\log w(e,h(\epsilon_i,e,\theta),\theta) = \frac{1}{k}\sum_{i=1..k} w_i$
    - $w(e,h(\epsilon,e,\theta),\theta) = \frac{p(e,h(\epsilon,e,\theta)|\theta)}{q(h(\epsilon,e,\theta)|e,\theta)}$: Unnormalized Importance weight
- We can sample just one ϵ or multiple ϵs
  - $\mathcal{L}_k = \mathbb{E}_{\epsilon\sim N(0,I)}\left[\log\frac{1}{k}\sum_{i=1}^k w_i\right] \le \log\mathbb{E}_{\epsilon\sim N(0,I)}\left[\frac{1}{k}\sum_{i=1}^k w_i\right] = \log p(e)$
  - Original VAE : $k = 1$
  - More samples → Tighter ELBO

Detour: Importance Sampling

- Huge waste from the rejection
- Is generating the PDF the end goal?
  - No… Usually, the question follows
    - Calculating the expectation of PDF
    - Calculating a certain probability
- Let's use the wasted sample to answer the questions

$$\mathcal{L} = \mathbb{E}_{\epsilon=<\epsilon_1,\ldots,\epsilon_L>\sim N(0,I)}\left[\log \frac{p(e,h(\epsilon,e,w)|w)}{q(h(\epsilon,e,w)|e,w)}\right]$$

  - $E(f) = \int f(h)p(h)dh = \int f(h)\frac{p(h)}{q(h)}q(h)dh \cong \frac{1}{L}\sum_{l=1}^{L}\frac{P(h^l)}{q(h^l)}f(h^l)$
    - L = # of samples, h^l=a sample of H
    - Here, the importance weight plays the role
      - $r^l = \frac{P(h^l)}{q(h^l)}$
    - What if $P(h^l)$ and $q(h^l)$ is not normalized, as they should be as probability distributions
    - $E(f) \cong \frac{1}{L}\sum_{l=1}^{L}\frac{P(h^l)}{q(h^l)}f(h^l) = \frac{1}{L}\frac{H_q}{H_p}\sum_{l=1}^{L}\frac{\tilde{P}(h^l)}{\tilde{q}(h^l)}f(h^l)$
  - $\mathrm{P(H>1)} = \int_1^{\infty} 1_{h>1}p(h)dh = \int_1^{\infty} 1_{h>1}\frac{p(h)}{q(h)}q(h)dh \cong \frac{1}{L}\sum_{l=1}^{L}\frac{P(h^l)}{q(h^l)}1_{h^l>1}$

Learning on Importance Weighted VAE

- Learning on $\mathcal{L}_k = \mathbb{E}_{\epsilon\sim N(0,I)}\left[\log\frac{1}{k}\sum_{i=1}^{k}\frac{p(e,h(\epsilon_i,e,\theta)|\theta)}{q(h(\epsilon_i,e,\theta)|e,\theta)}\right]$

$$w(e,h(\epsilon,e,\theta),\theta) = \frac{p(e,h(\epsilon,e,\theta)|\theta)}{q(h(\epsilon,e,\theta)|e,\theta)}$$

- $\nabla_\theta\mathcal{L}_k = \nabla_\theta\mathbb{E}_{\epsilon_1\dots\epsilon_k}\left[\log\frac{1}{k}\sum_{i=1}^{k}\frac{p(e,h(\epsilon_i,e,\theta)|\theta)}{q(h(\epsilon_i,e,\theta)|e,\theta)}\right]$

$$E(f) = \int f(h)\frac{p(h)}{q(h)}q(h)dh \cong \frac{1}{L}\sum_{l=1}^{L}\frac{P(h^l)}{q(h^l)}f(h^l)$$

$$= \mathbb{E}_{\epsilon_1\dots\epsilon_k}\left[\nabla_\theta\log\frac{1}{k}\sum_{i=1}^{k}w(e,h(\epsilon_i,e,\theta),\theta)\right]$$

$$\nabla_\theta w = w\nabla_\theta\log w, \quad \because \frac{d}{d\theta}\log f(\theta) = \frac{\nabla_\theta f(\theta)}{f(\theta)}$$

$$= \mathbb{E}_{\epsilon_1\dots\epsilon_k}\left[\frac{\frac{1}{k}\sum_{i=1}^{k}\nabla_\theta w(e,h(\epsilon_i,e,\theta),\theta)}{\frac{1}{k}\sum_{j=1}^{k}w\left(e,h\left(\epsilon_j,e,\theta\right),\theta\right)}\right] = \mathbb{E}_{\epsilon_1\dots\epsilon_k}\left[\sum_{i=1}^{k}\frac{\nabla_\theta w_i}{\sum_{j=1}^{k}w_j}\right]$$

$$= \mathbb{E}_{\epsilon_1\dots\epsilon_k}\left[\sum_{i=1}^{k}\frac{w_i\nabla_\theta\log w_i}{\sum_{j=1}^{k}w_j}\right] = \mathbb{E}_{\epsilon_1\dots\epsilon_k}\left[\sum_{i=1}^{k}\frac{w_i}{\sum_{j=1}^{k}w_j}\nabla_\theta\log w(e,h(\epsilon_i,e,\theta),\theta)\right]$$

- Here, we use the unbiased estimate of gradient.
  - The stochastic gradient $\nabla f_i(x)$ is the unbiased estimate of gradient $\nabla f(x)$
    - $Bias_\theta[\bar{\theta}] = E_{x|\theta}[\bar{\theta}] - \theta$, Unbiased estimator: $Bias_\theta[\bar{\theta}] = 0$
    - $E_i[\nabla f_i(x)] = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) = \nabla f(x)$

Derivation of Implicit Distribution from ELBO

- Optimal prior : $p_\lambda^*(h) = \int p_D(e) q_\phi(\mathrm{h}|\mathrm{e}) d\mathrm{e} \equiv q_\phi(\mathrm{h})$
  - A.k.a. aggregated prior
  - Original VAE ELBO
    - $\mathcal{L}(e;\phi,\theta) = \mathbb{E}_{q_\phi(h|e)}[\log p_\theta(e|h)] - D_{KL}\left(q_\phi(h|e)||p_\theta(h)\right)$
    - $D_{KL}\left(q_\phi(h|e)||p_\theta(h)\right)$ can be calculated as a closed form
  - Optimal prior ELBO
    - $\mathcal{L}(e;\phi,\theta) = \mathbb{E}_{q_\phi(h|e)}[\log p_\theta(e|h)] - D_{KL}\left(q_\phi(h|e)||q_\phi(h)\right)$
    - $D_{KL}\left(q_\phi(h|e)||q_\phi(h)\right)$ is difficult to be calculated in the closed form
    - $D_{KL}\left(q_\phi(h|e)||q_\phi(h)\right) = E_{q_\phi(h|e)}\left[\ln \frac{q_\phi(h|e)}{q_\phi(h)}\right]$

      $= \int q_\phi(h|e) \ln \frac{q_\phi(h|e)p(h)}{q_\phi(h)p(h)} dh = \int q_\phi(h|e) \ln \frac{q_\phi(h|e)}{p(h)} dh + \int q_\phi(h|e) \ln \frac{p(h)}{q_\phi(h)} dh$

      $= D_{KL}\left(q_\phi(h|e)||p(h)\right) - \mathbb{E}_{q_\phi(h|e)}\left[\ln \frac{q_\phi(h)}{p(h)}\right]$
  - $D_{KL}\left(q_\phi(h|e)||p(h)\right)$ is the KL divergence from the original VAE
  - $\mathbb{E}_{q_\phi(h|e)}\left[\ln \frac{q_\phi(h)}{p(h)}\right]$ is the expected density ratio
    - between the model and the variational distribution

Learning in Implicit Models

- $\mathbb{E}_{q_\phi(h|e)}\left[\ln \frac{q_\phi(h)}{p(h)}\right]$ is the expected density ratio
  - The density ratio can be computed by building a classifier to distinguish observed data from that generated by the model.
    - $p(h)$ is used for a set of n samples $E_p = \{e_1^{(p)}, \dots, e_n^{(p)}\}$
    - $q_\phi(h)$ is used for a set of n samples $E_q = \{e_1^{(q)}, \dots, e_n^{(q)}\}$
      - Through the ancestral sampling of $q_\phi(h|e)$ by randomly selecting e
  - A random variable y that assigns a label $y = 1$ to all samples in E_p and $y = 0$ to all samples in E_q.
    - $p(h) = p(h|y = 0)$ and $q_\phi(h) = p(h|y = 1)$
    - $p^*(h|y) \equiv \begin{cases} q_\phi(h), y = 1 \\ p(h), y = 0 \end{cases}$
- By applying Bayes' rule, we can compute the ratio $r(\boldsymbol{h}) = \ln \frac{q_\phi(h)}{p(h)}$ as:
  - $\frac{q_\phi(h)}{p(h)} = \frac{p^*(h|y=1)}{p^*(h|y=0)} = \frac{\frac{p^*(h,y=1)}{p^*(y=1)}}{\frac{p^*(h,y=0)}{p^*(y=0)}} = \frac{\frac{p^*(y=1|h)p^*(h)}{p^*(y=1)}}{\frac{p^*(y=0|h)p^*(h)}{p^*(y=0)}} = \frac{p^*(y=1|h)}{p^*(y=0|h)} \cdot \frac{\pi}{1-\pi} = \frac{p^*(y=1|\boldsymbol{h})}{p^*(y=0|\boldsymbol{h})} = \frac{D(h)}{1-D(h)}$
    - $D(h) = p^*(y = 1|h)$
    - Which indicates density ratio estimation = class probability estimation.
  - The problem is reduced to computing the probability $p(y = 1|h)$
    - Discriminative modeling can be applied, as a discriminator

ELBO Learning with Neural Net. Implicit Prior

- $D(h) = p^*(y=1|\mathbf{h})$, $\mathbb{E}_{q_\phi(h|e)}\left[\ln \frac{q_\phi(h)}{p(h)}\right]$
  - Consider a neural network classifier, $D(h) = \sigma(T_\psi(h))$
    - Learning objective of $T_\psi(h)$
      - $T^*(h) = max_\psi \mathbb{E}_{q_\phi(h)}\left[\ln\left(\sigma\left(T_\psi(h)\right)\right)\right] + \mathbb{E}_{p(h)}\left[\ln\left(1-\sigma\left(T_\psi(h)\right)\right)\right]$
  - $\frac{q_\phi(h)}{p(h)} = \frac{D(h)}{1-D(h)} = \frac{\sigma(T^*(h))}{1-\sigma(T^*(h))}$
    - $q_\phi(h) = \sigma\left(T^*(h)\right)\left(p(h)hq_\phi(h)\right) \rightarrow \sigma\left(T^*(h)\right) = \frac{1}{1+\exp(-T^*(h))} = \frac{q_\phi(h)}{p(h)+q_\phi(h)}$

      $\rightarrow p(h) = q_\phi(h)\exp\left(-T^*(h)\right) \rightarrow T^*(h) = \ln \frac{q_\phi(h)}{p(h)}$
- Optimal prior ELBO
  - $\mathcal{L}(e;\phi,\theta) = \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(e|h)] - D_{KL}\left(q_\phi(h|e)||q_\phi(h)\right)$

    $= \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(e|h)] - D_{KL}\left(q_\phi(h|e)||p(h)\right) + \mathbb{E}_{q_\phi(H|E)}\left[\ln \frac{q_\phi(h)}{p(h)}\right]$

    $= \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(e|h)] - D_{KL}\left(q_\phi(h|e)||p(h)\right) + \mathbb{E}_{q_\phi(H|E)}[T^*(h)]$

    $= \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(e|h) + T^*(h)] - D_{KL}\left(q_\phi(h|e)||p(h)\right)$
  - Expectation can be the Monte-Carlo estimation with the reparametrization trick
- Iterating objectives by learning ϕ, θ and ψ alternatively

detour) β-Variational Autoencoder

$$\min_{w,b} \|w\| + C\sum_j \xi_j$$
$$s.t.\,(we_j + b)y_j \geq 1 - \xi_j, \forall j$$
$$\xi_j \geq 0, \forall j$$

- From the generative modeling perspective, ELBO is derived
  - from the Jensen's inequality
  - with an assumed variational distribution
  - for a latent variable model
    - $\ln P(E) = \ln \sum_H P(H,E) = \ln \sum_H Q(H|E)\frac{P(H,E)}{Q(H|E)}$

      $\geq \sum_H Q(H|E) \ln \left[\frac{P(H,E)}{Q(H|E)}\right] = -D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) + \mathbb{E}_{q_\phi(H|E)}[\log p_\theta(E|H)]$
- From the original autoencoder's perspective,
  - Autoencoder with a regularization imposed by a prior distribution
  - Regularization method from the Lagrange method and the soft margin of SVM
    - $max_{\phi,\theta}\mathbb{E}_{E\sim D}[\mathbb{E}_{H\sim q_\phi(H|E)}[\log p_\theta(E|H)]]$, Subject to $D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) < \epsilon$
    - $F(\phi,\theta,\beta;E,H) = \mathbb{E}_{H\sim q_\phi(H|E)}[\log p_\theta(E|H)] - \beta\left(D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right) - \epsilon\right)$

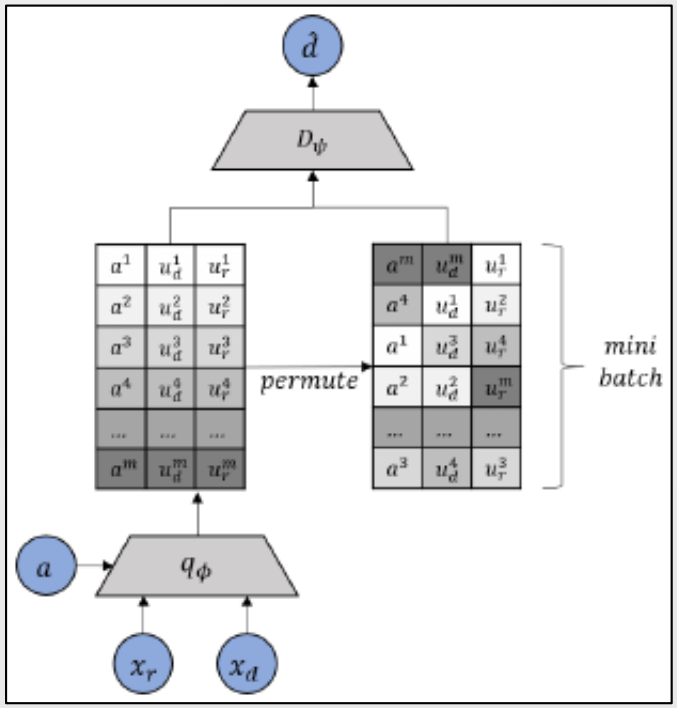
      $\geq \mathbb{E}_{H\sim q_\phi(H|E)}[\log p_\theta(E|H)] - \beta D_{KL}\left(q_\phi(H|E)||p_\theta(H)\right)$
  - β becomes the hyperparameter of the regularization
    - β = 1 → Original VAE
    - $\beta > 1$ → Stronger regularization with $p_\theta(H)\sim N(0,1^2)$
      - The covariance structure of 1^2 impose the independence between the latent dimensions of H
      - Called, Latent disentanglemnt

Factorizing Z dimensions: FactorVAE

- Obvious disadvantage of β-VAE : over-regularization!
- Disentanglement requires a restriction → often, regularization
  - It is unwise to regularize the latent variable to simple $N(0, I)$
    - Here, regularize ≈ prior modeling
- The restriction that we want is reducing the correlation across z
  - When we assume $\bar{q}(z) = \bar{q}(z_1, \dots, z_d) \equiv \prod_{j=1}^{d} q(z_j)$
  - Total correlation: $KL(q(z)||\bar{q}(z))$
    - $q(z|x^{(i)})$: samples will deliver $q(z)$
    - $\bar{q}(z|x^{(i)})$
      - Generate from samples
        - so the dimensional distribution is maintained
      - Permute a dimension of z_d by dimension-wise
        - so the cross-dimensional correlation can be broken
  - Finally, how to calculate $KL(q(z)||\bar{q}(z))$
  - $TC(z) = KL(q(z)||\bar{q}(z)) = E_{q(z)}\left[\log \frac{q(z)}{\bar{q}(z)}\right]$

$$\approx E_{q(z)}\left[\log \frac{D(z)}{1 - D(z)}\right]$$

    - $D(z)$ is a discriminator to identify whether z is coming from the original q or the permuted $\bar{q}$
- Eventually, this is an alternative optimization



Variational Inference and Implicit Models

- Traditional VI requires conjugacy and tractable likelihood.
  - VAE resort the conjugacy issue by forming the inference networks for variational distribution.
    - VAE still requires an explicit likelihood function.

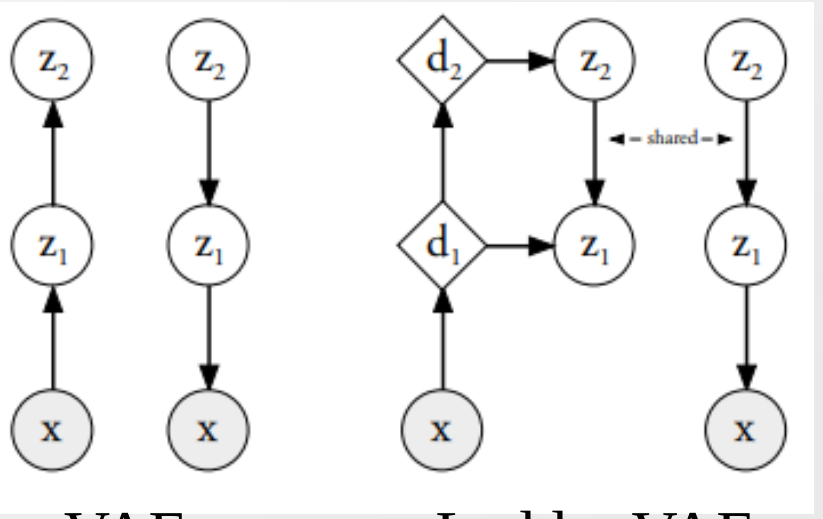
$$q^*(\boldsymbol{h}) = \underset{q \in Q}{\operatorname{argmin}}\, KL(q_\phi(\mathbf{h}|\boldsymbol{e})||p(\boldsymbol{h}|\boldsymbol{e}))$$

$$q_\phi(\boldsymbol{h}|\boldsymbol{e}) = \prod_{n=1}^{N} q(h_n; \lambda_n = \mathrm{NN}(e_n; \mathrm{NN}(e_n|\phi));\ \mathrm{q}(.) = \mathrm{Normal}$$

- What if we combine the methods of “learning in implicit models” with VI?
  - We can use an implicit form of q
    - More expressive than explicit forms
  - We also can use an implicit form of p
    - GAN, simulator...
  - Of course we can use both p and q in an implicit form.

Structuring Multiple Latent Variables (1)

- Maximize the likelihood of
  $p_\theta(x) = \int p_\theta(x,z)dz$
  - $p_\theta(x,z) = p_\theta(x|z)p_\theta(z)$
- Let's assume that z has a layered structure
  - $p_\theta(z) = p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})$
  - $p_\theta(z_i|z_{i+1}) = N\left(z_i|\mu_{p,i}(z_{i+1}), \sigma^2_{p,i}(z_{i+1})\right)$
  - $p_\theta(x|z_1) = N\left(x|\mu_{p,0}(z_1), \sigma^2_{p,0}(z_1)\right)$
    - or can follow the Bernoulli distribution
    - $\mu_{p,i}(z_{i+1}), \sigma^2_{p,i}(z_{i+1})$: amortized inferenece on the model distribution for generations



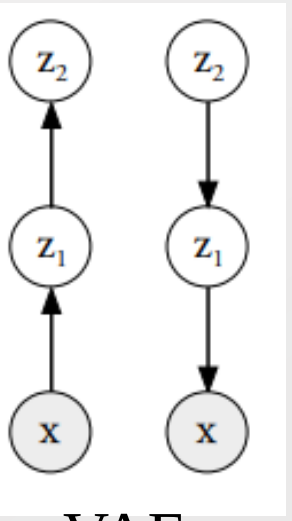
VAE Ladder VAE

- Eventually, the ELBO structure will be
  - $\log p(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = L(\theta,\phi;x)$
  - $L(\theta,\phi;x) = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}\right]$
    $= E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}{q_\phi(z|x)}\right]$

Structuring Multiple Latent Variables (2)

- Eventually, the ELBO structure will be
  - $\log p(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = L(\theta,\phi;x)$

$$L(\theta,\phi;x) = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}{q_\phi(z|x)}\right]$$



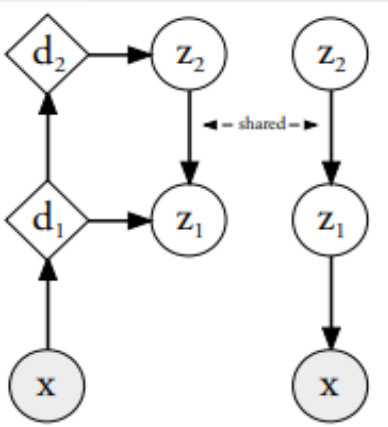
VAE

- How to structure the variational distribution?
  - by following the Bayesian inference of the posterior definition
  - $q_\phi(z|x) = q_\phi(z_1|x)\prod_{i=2}^{L} q_\phi(z_i|z_{i-1})$
  - $q_\phi(z_1|x) = N\left(z_1|\mu_{q,1}(x), \sigma_{q,1}^2(x)\right)$
  - $q_\phi(z_i|z_{i-1}) = N\left(z_i|\mu_{q,i}(z_{i-1}), \sigma_{q,i}^2(z_{i-1})\right), i = 2 \ldots L$
- Any potential problem?
  - Hierarchical random variable → hierarchically increasing variance
    - More over, this is a single forward path of the encoder
  - Need to correct the latent variable inference in multiple directions

Structuring Multiple Latent Variables (3)

- Eventually, the ELBO structure will be
  - $\log p(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = L(\theta,\phi;x)$

$$L(\theta,\phi;x) = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}{q_\phi(z|x)}\right]$$



Ladder VAE

- How to structure the variational distribution?
  - What was the best variational distribution?
    - The mirror of the model distribution → **sharing the inference parameters**
  - $q_\phi(z|x) = q_\phi(z_L|x)\prod_{i=1}^{L-1} q_\phi(z_i|z_{i+1})$
  - $q_\phi(z_L|x) = N(z_L|\mu_{q,L},\sigma_{q,L}^2)$
  - $q_\phi(z_i|z_{i+1}) = N\left(z_i|\mu_{q,i},\sigma_{q,i}^2\right), i = 1 \ldots L-1$
  - $\mu_{q,i} = \frac{\hat{\mu}_{q,i}\,\hat{\sigma}_{q,i}^{-2} + \hat{\mu}_{p,i}\hat{\sigma}_{p,i}^{-2}}{\hat{\sigma}_{q,i}^{-2} + \hat{\sigma}_{p,i}^{-2}}, \sigma_{q,i} = \frac{1}{\hat{\sigma}_{q,i}^{-2} + \hat{\sigma}_{p,i}^{-2}}$
  - $d_n = MLP(d_{n-1})$
  - $\hat{\mu}_{q,i} = Linear(d_i), \hat{\sigma}_{q,i}^2 = Softplus\left(Linear(d_i)\right), i = 1 \ldots L$

Structuring Multiple Latent Variables (4)

- Eventually, the ELBO structure will be
  - $\log p(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = L(\theta,\phi;x)$

$$L(\theta,\phi;x) = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}{q_\phi(z|x)}\right]$$

$$= E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}{q_\phi(z_L|x)\prod_{i=1}^{L-1} q_\phi(z_i|z_{i+1})}\right]$$

$$= E_{q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z_L|x)\prod_{i=1}^{L-1} q_\phi(z_i|z_{i+1})}{p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}\right]$$

$$= E_{q_\phi(z|x)}[\log p_\theta(x|z)] + E_{q_\phi(z|x)}\left[\log \frac{q_\phi(z_L|x)}{p_\theta(z_L)}\right] + \sum_{i=1}^{L-1} E_{q_\phi(z|x)}\left[\log \frac{\prod_{i=1}^{L-1} q_\phi(z_i|z_{i+1})}{\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})}\right]$$

- $L(\theta,\phi;x)$

$$= E_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL(q_\phi(z_L|x)||p_\theta(z_L)) + \sum_{i=1}^{L-1} KL(q_\phi(z_i|z_{i+1})||p_\theta(z_i|z_{i+1}))$$

- What is added?
  - The recursive regularization of KL divergence on the latent variables
  - If we select the Gaussain distribution, the KL divergence calculation can be the closed-form solution
- What is gained?
  - The direct chaining of latent random variable inferred by the amortized variational inference
  - This is a specific structure of VAE of chaining
- Ladder VAE!

- Some slides are generated by Hyemi Kim, AAILab, KAIST