CMPUT 651 (Fall 2019)

Sentence Generation

Lili Mou
lmou@ualberta.ca
lili-mou.github.io

CMPUT 651 (Fall 2019)

Roapmap

- Motivation and examples
- Techniques
  - Generation from latent space
  - Generation from word space

Motivation

- Sentence generation
  - Dialogue systems
  - Paraphrase generation
  - Machine translation
- A seq2seq model may not suffice
  - No input
  - Constructing new information
  - Diversity needed
- Probabilistic sentence generation
  - Prior sampling, posterior sampling

CMPUT 651 (Fall 2019)

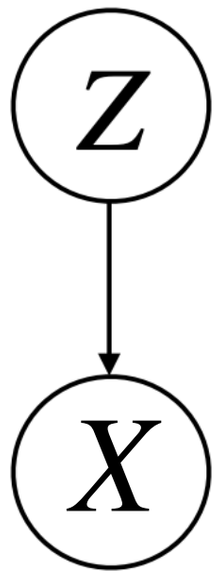
Latent Space Sampling

Variational Autoencoder

- Humans' sentence generation involves two steps
  - First, we have some "vague" idea of the sentence
  - Then, we flesh it out by words

- A sentence $\boldsymbol{x} = (x_1, \cdots, x_T)$ is subject to some latent representation $\boldsymbol{z}$

$$p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z})p(\boldsymbol{x} \,|\, \boldsymbol{z})$$



Kingma, D. P., & Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Variational Autoencoder

- Humans' sentence generation involves two steps
  - First, we have some "vague" idea of the sentence
  - Then, we flesh it out by words

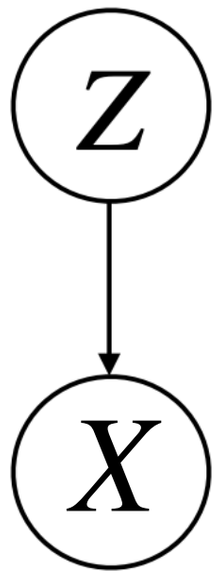
- A sentence $\boldsymbol{x} = (x_1, \cdots, x_T)$ is subject to some latent representation $\boldsymbol{z}$

$$p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z})$$

How can we learn a model with latent variables?

E-step: $p(z \mid x) = \dfrac{p(z)p(x \mid z)}{p(x)} = \dfrac{p(z)p(x \mid z)}{\int p(z')p(x \mid z')dz'}$

M-step: maximize $\mathbb{E}_{z \sim p(z|x)} \log p(z, x)$



Variational Autoencoder

- Humans' sentence generation involves two steps
  - First, we have some "vague" idea of the sentence
  - Then, we flesh it out by words

- A sentence $\boldsymbol{x} = (x_1, \cdots, x_T)$ is subject to some latent representation $\boldsymbol{z}$

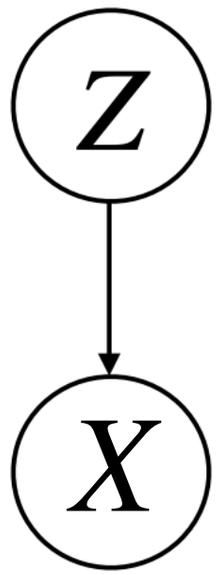
$$p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z})$$

How can we learn a model with latent variables?

E-step: $p(z \mid x) = \dfrac{p(z)p(x \mid z)}{p(x)} = \dfrac{p(z)p(x \mid z)}{\int p(z')p(x \mid z')dz'}$

M-step: maximize $\mathbb{E}_{z \sim p(z|x)} \log p(z, x)$

Intractable

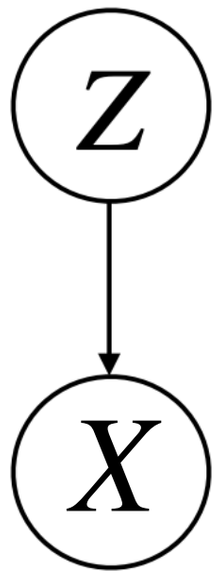


Variational Autoencoder

- Humans' sentence generation involves two steps
  - First, we have some "vague" idea of the sentence
  - Then, we flesh it out by words

- A sentence $\boldsymbol{x} = (x_1, \cdots, x_T)$ is subject to some latent representation $\boldsymbol{z}$

$$p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z})$$



How can we learn a model with latent variables?

Recognization

E-step: $p(z \mid x) = \dfrac{p(z)p(x \mid z)}{p(x)} = \dfrac{p(z)p(x \mid z)}{\int p(z')p(x \mid z')dz'}$

Reconstruction

M-step: maximize $\mathbb{E}_{z \sim p(z|x)} \log p(z, x)$

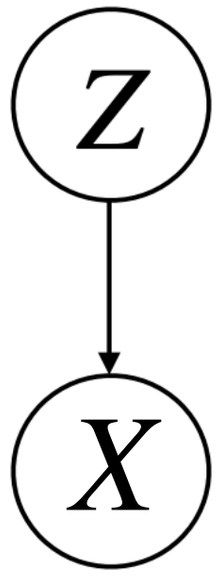
(in a more general sense)

Variational Inference

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = \log\left(\int_z p(\boldsymbol{x}, z;\boldsymbol{\theta})dz\right)$$

$$= \underbrace{\int q(z|\boldsymbol{x}) \log \frac{p(\boldsymbol{y}, z;\boldsymbol{\theta})}{q(z|\boldsymbol{x})}dz}_{L(q,\boldsymbol{\theta})} + \underbrace{\int q(z|\boldsymbol{x}) \log \frac{q(z|\boldsymbol{x})}{p(z|\boldsymbol{x};\boldsymbol{\theta})}dz}_{\mathrm{KL}(q(Z|\boldsymbol{x})\|p(Z|\boldsymbol{x}))}$$

Variational inference	vs	**EM**
Variational family		q can be any distribution
$q \in Q$		True posterior is the best
Ignore the KL-term		KL=0 after E step



Variational Inference

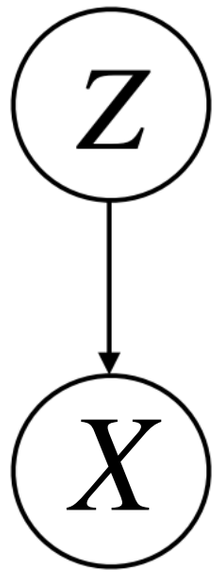
$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = \log\left(\int_z p(\boldsymbol{x}, z;\boldsymbol{\theta})dz\right)$$

$$= \underbrace{\int q(z|\boldsymbol{x}) \log \frac{p(\boldsymbol{y}, z;\boldsymbol{\theta})}{q(z|\boldsymbol{x})}dz}_{L(q,\boldsymbol{\theta})} + \underbrace{\int q(z|\boldsymbol{x}) \log \frac{q(z|\boldsymbol{x})}{p(z|\boldsymbol{x};\boldsymbol{\theta})}dz}_{\mathrm{KL}(q(Z|\boldsymbol{x})\|p(Z|\boldsymbol{x}))}$$

- Two extremes
  - Q = any function ⇒ EM

    ⇒ powerful model; optimization intractable
  - Q = {a fixed distribution}

    ⇒ degenerated model; optimization easy



Variational Inference

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = \log\left(\int_z p(\boldsymbol{x}, z;\boldsymbol{\theta})dz\right)$$

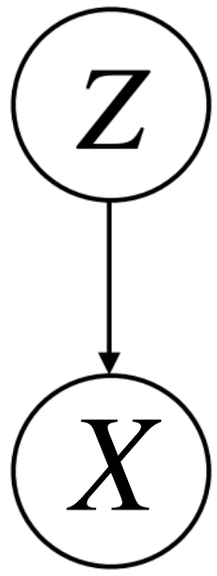
$$= \underbrace{\int q(z|\boldsymbol{x}) \log \frac{p(\boldsymbol{y}, z;\boldsymbol{\theta})}{q(z|\boldsymbol{x})}dz}_{L(q,\boldsymbol{\theta})} + \underbrace{\int q(z|\boldsymbol{x}) \log \frac{q(z|\boldsymbol{x})}{p(z|\boldsymbol{x};\boldsymbol{\theta})}dz}_{\mathrm{KL}(q(Z|\boldsymbol{x})\|p(Z|\boldsymbol{x}))}$$

- Two extremes
  - Q = any function ⇒ EM

    ⇒ powerful model; optimization intractable
  - Q = {a fixed distribution}

    ⇒ degenerated model; optimization easy

Trade-off, e.g.,
- Independent assumption
- Gaussian assumption



Example

[PRML]

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0,(\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0,b_0) \end{aligned}$$

Variational family: factorized distribution

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$$

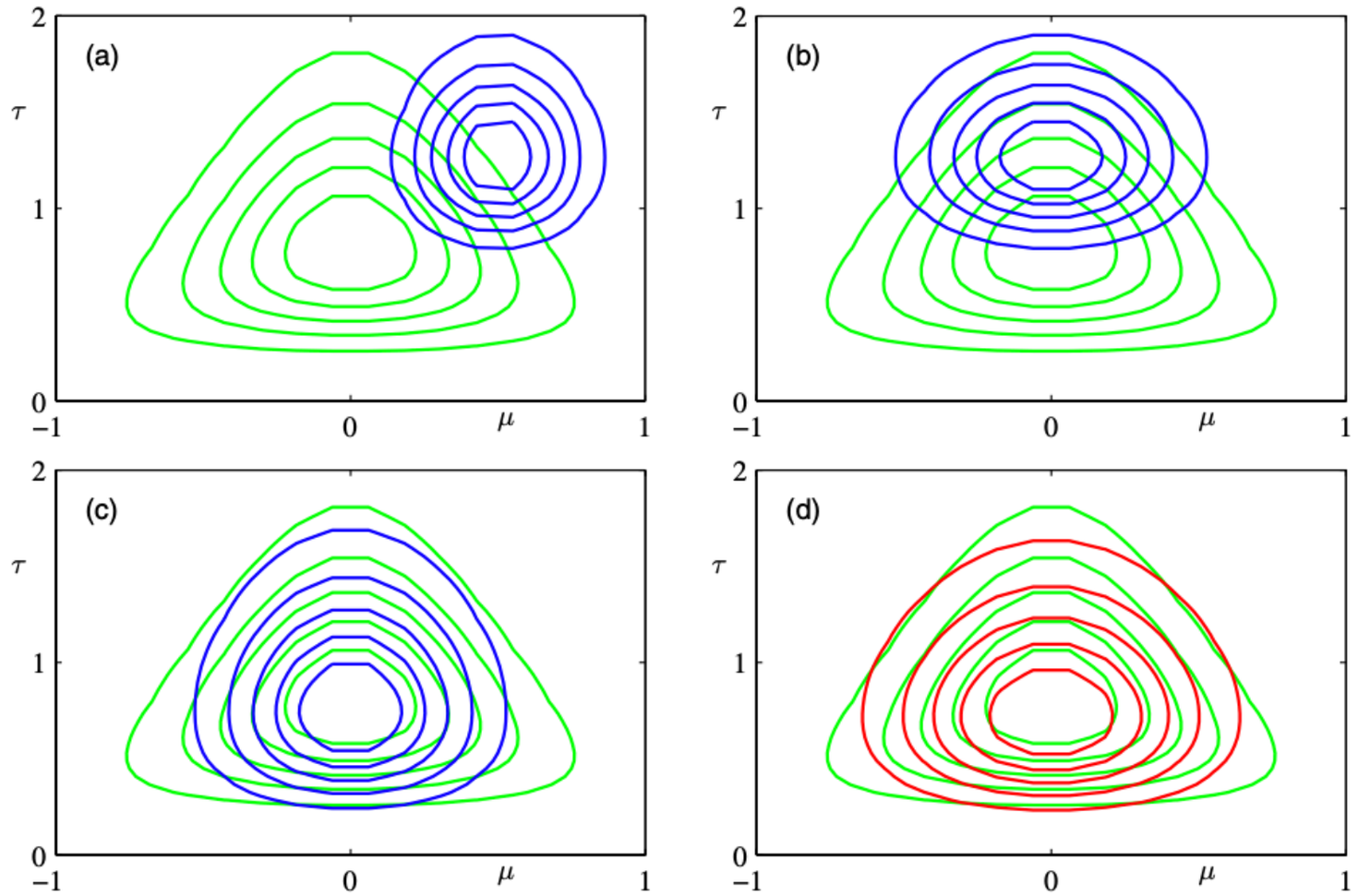
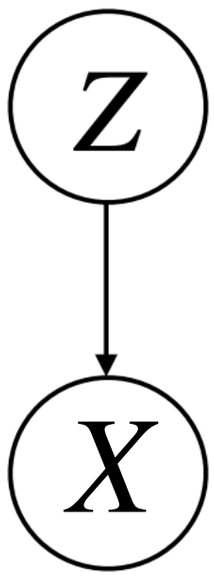


Figure 10.4 Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu,\tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

UNIVERSITY OF ALBERTA

Variational Autoencoder

- Variational autoencder
  - Variational family: $Q = \{\mathcal{N}(\boldsymbol{\mu}, \operatorname{diag} \boldsymbol{\sigma}^2) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\sigma} \in \mathbb{R}^d_{++}\}$
  - Recognizing $\boldsymbol{\mu}, \boldsymbol{\sigma}$ by NN
  - Modeling $\boldsymbol{x}$ also by NN (need a little bit more efforts)



Variational Autoencoder

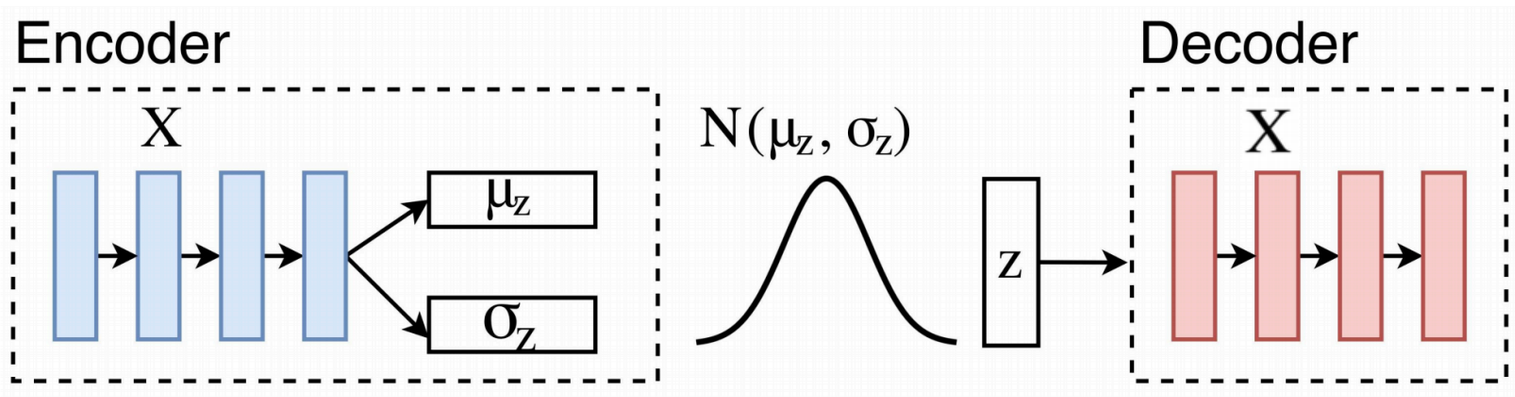
$$\log p_\theta(x) = \log\left(\int_z p_\theta(x,z)dz\right)$$

$$= \int q_\phi(z\,|\,x)\,\log\frac{p_\theta(y,z)}{q_\phi(z\,|\,x)}dz + \int q_\phi(z\,|\,x)\,\log\frac{q_\phi(z\,|\,x)}{p_\theta(z\,|\,x)}dz$$

$$\geq \int q_\phi(z\,|\,x)\,\log\frac{p_\theta(y,z;\theta)}{q_\phi(z\,|\,x)}dz$$

$$= \int q_\phi(z\,|\,x)\,\log p_\theta(y\,|\,z)dz + \int q_\phi(z\,|\,x)\log\frac{p_\theta(z)}{q_\phi(z\,|\,x)}dz$$

$$= \mathbb{E}_{z\sim q_\phi(z|x)}\log p_\theta(y\,|\,z) - \mathrm{KL}(q_\phi(z\,|\,x)\|p_\theta(z))$$



UNIVERSITY OF ALBERTA

CMPUT 651 (Fall 2019)

Formula Zoo

$$\mathbb{E}_{x\sim p_{data}(x)}[\log p(x)] \geq \underbrace{\mathbb{E}_{x\sim p_{data}(x)}[\mathbb{E}_{q(z|x)}[\log p(x|z)]]}_{\text{-VAE Reconstruction}} - \underbrace{\mathbb{E}_{x\sim p_{data}(x)}[KL(q(z|x)\,\|\,p(z))]}_{\text{VAE Regularization}} \quad (1)$$

$$= \underbrace{\mathbb{E}_{x\sim p_{data}(x)}[\mathbb{E}_{q(z|x)}[\log p(x|z)]]}_{\text{-AVB Reconstruction}} - \underbrace{KL(q(z,x)\,\|\,p(z)\,p_{data}(x))}_{\text{AVB Regularization}} \quad (2)$$

$$- \underbrace{\mathbb{E}_{x\sim p_{data}(x)}[\mathbb{E}_{q(z|x)}[\log p(x|z)]]}_{\text{-AAE Reconstruction}} \underbrace{-KL(q(z)\,\|\,p(z))}_{\text{AAE Regularization}} \underbrace{- I(z;x)}_{\text{Mutual info.}} \quad (3)$$

$$= \underbrace{-\mathbb{E}_{z\sim q(z)}[KL(q(x|z)\,\|\,p(x|z))]}_{\text{IAE Reconstruction}} \underbrace{-KL(q(z)\,\|\,p(z))}_{\text{IAE Regularization}} \underbrace{- H_{data}(x)}_{\text{Entropy of data}} \quad (4)$$

$$= \underbrace{-KL(q(x,z)\,\|\,r(x,z))}_{\text{IAE Reconstruction}} \underbrace{- KL(q(z)\,\|\,p(z))}_{\text{IAE Regularization}} \underbrace{- H_{data}(x)}_{\text{Entropy of data}} \quad (5)$$

$$= \underbrace{-KL(q(x,z)\,\|\,p(x,z))}_{\text{ALI/BiGAN Cost}} \underbrace{- H_{data}(x)}_{\text{Entropy of data}} \quad (6)$$

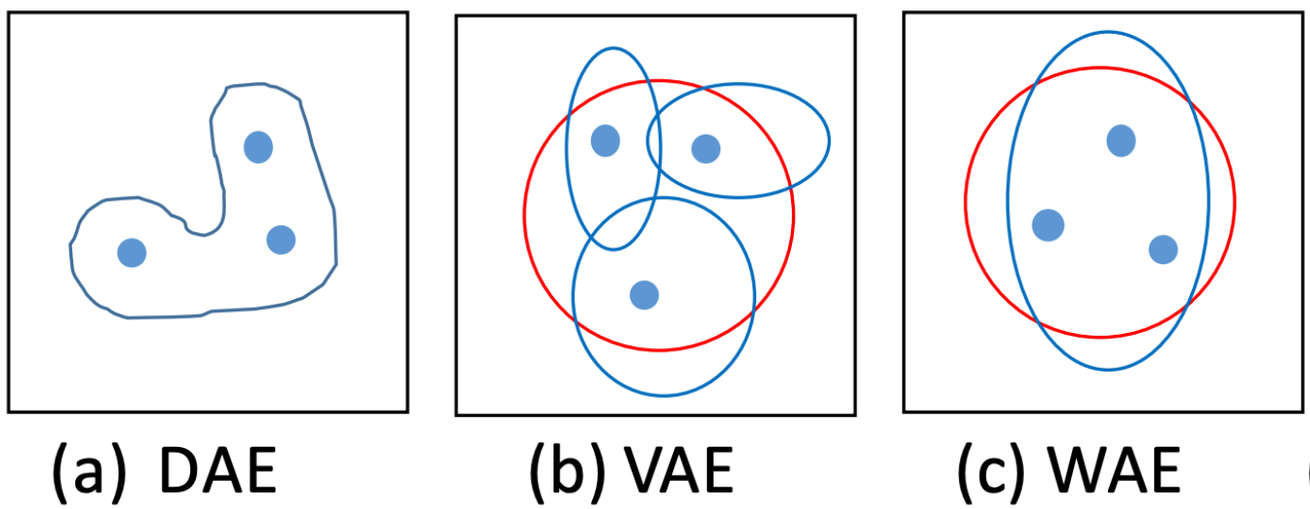
UNIVERSITY OF ALBERTA

Adversarial/Wasserstein Autoencoder

- VAE: $q(z \mid x) \rightarrow p(z)$

- WAE: $q(z) = \int p_{\mathscr{D}}(x) q(z \mid x) dx \xrightarrow{close} p(z)$

$$J = \mathbb{E}_{x \in p_{\mathscr{D}}(x)} \mathbb{E}_{z \in q(z|x)} \log p(z \mid x) + \mathbb{D}(q(z), p(z))$$

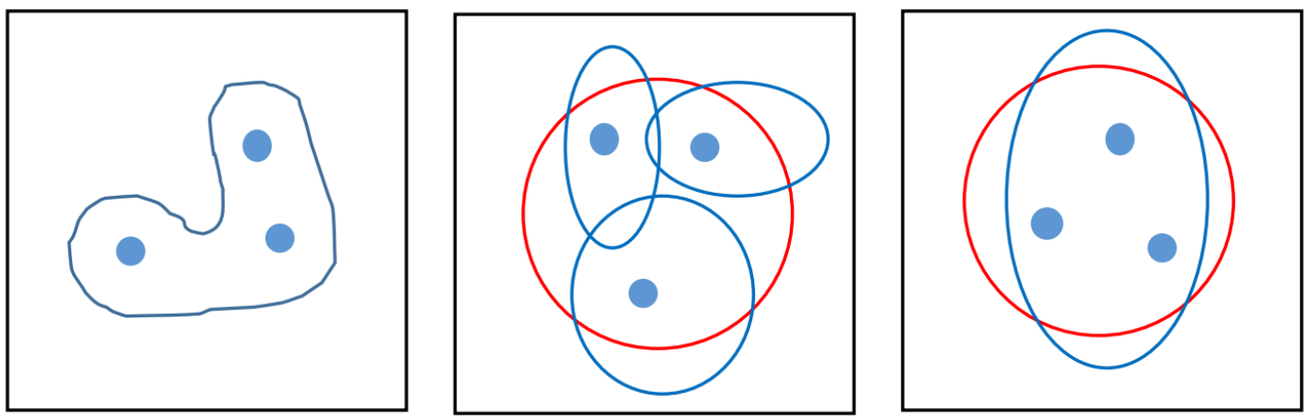


Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. arXiv preprint arXiv:1511.05644. 2015

Implicit Distributions

- We penalize some distance between $q(z)$ and $p(x)$
- We do not have an explicit form or $q(z)$
- But samples from $q(z) := \int p_{\mathscr{D}}(x) q(z \mid x) dx$

$$x^{(i)} \sim p_{\mathscr{D}}(x), \quad z \sim q(z \mid x^{(i)})$$



(a) DAE (b) VAE (c) WAE

Adversarial Training

- We penalize some distance between $q(z)$ and $p(z)$
- We deliberately introduce a classifier (discriminator/adversary) to distinguish samples from $q(z)$ and $p(z)$
- We train the model to fool the discriminator
  - Flipping gradient
  - Maximizing predicted entropy

Algorithm

foreach *mini-batch* **do**

  minimize $J_{\text{dis}}\left(\boldsymbol{\theta}_{\text{dis}}\right)$ w.r.t. $\boldsymbol{\theta}_{\text{dis}}$

  minimize J_{ovr} w.r.t. $\boldsymbol{\theta}_{\text{E}}, \boldsymbol{\theta}_{\text{D}}$

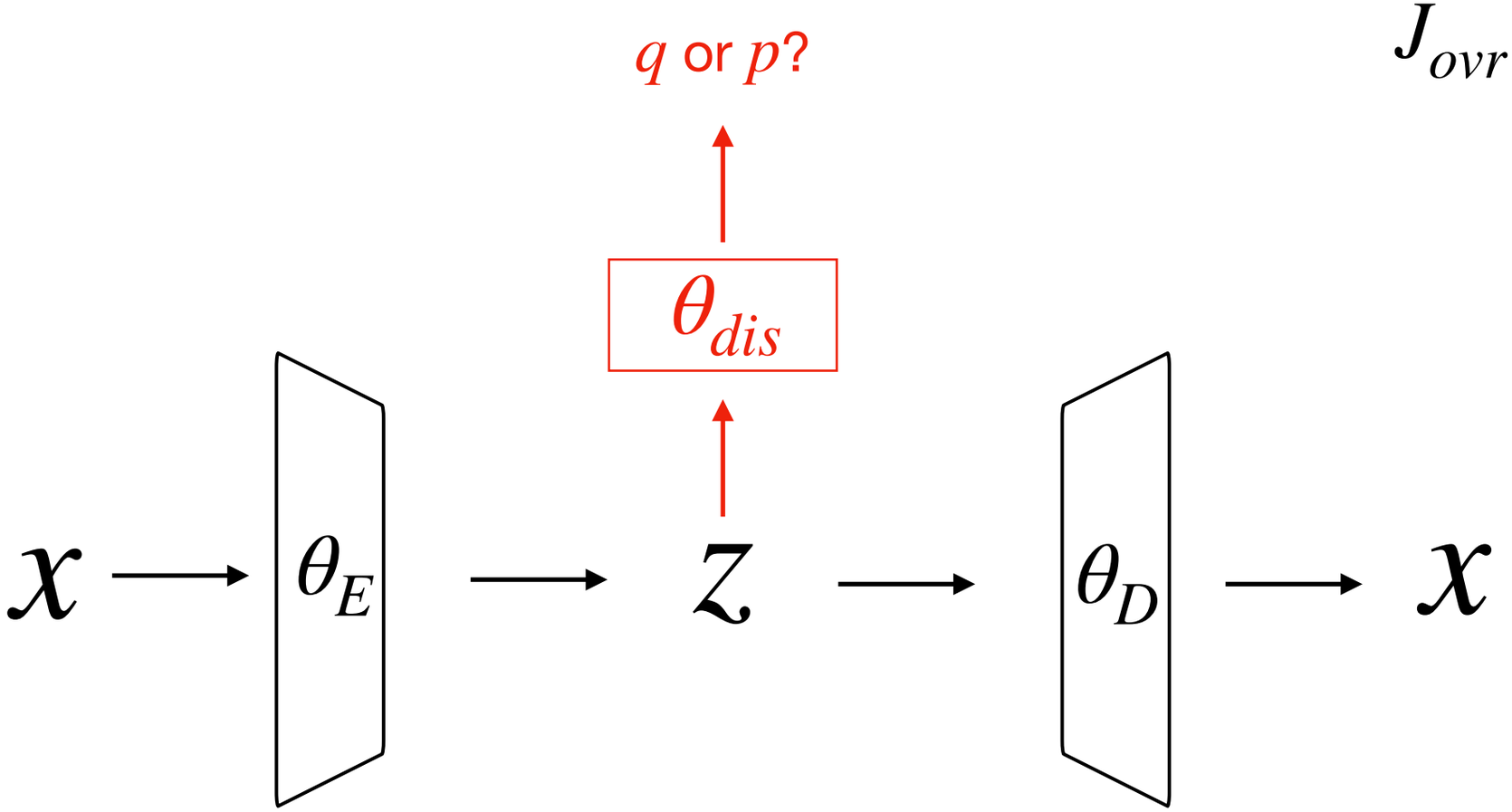
end

- Modeling J_{ovr}
  - Flipping gradient

$$J_{ovr} = J_{rec} - J_{dis}$$

  - Maximizing predicted entropy

$$J_{ovr} = J_{rec} - \mathscr{H}(\mathbf{y}_{dis})$$



Applications of Adv Training

- Adversarial/Wasserstein autoencoder

$$q(z) \quad vs \quad p(z)$$

- Generative adversarial network

$$p_{gen}(z) \quad vs \quad p_{\mathscr{D}}(z)$$

- Domain adaptation

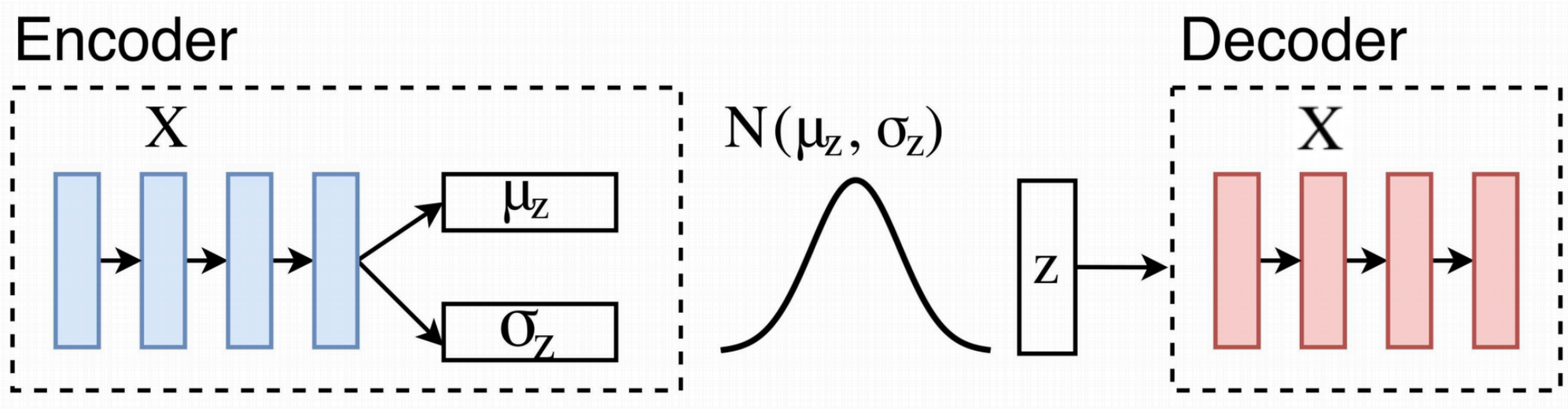
$$p_{D1}(z) \quad vs \quad p_{D2}(z)$$

A Few Fundamental Questions

- VAE: KL collapse

$$J = \mathbb{E}_{z\sim q(z|x)}[\log p(x\,|\,z)] + \mathrm{KL}(q(z\,|\,x)\|p(z))$$

  - WAE alleviates this problem

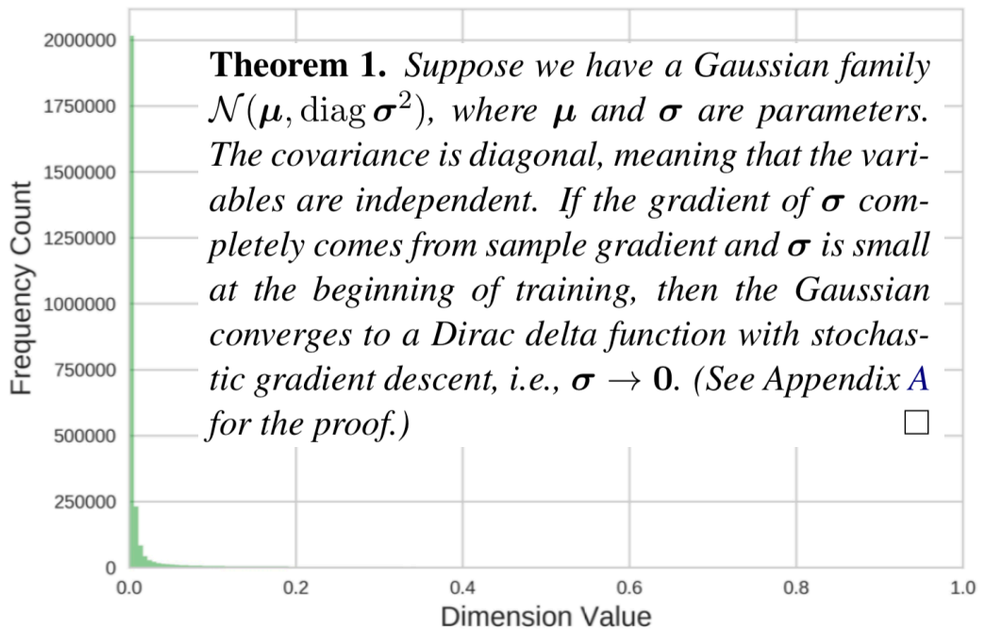


Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R, Bengio S. Generating sentences from a continuous space. In *CoNLL*, 2016.

A Few Fundamental Questions

- WAE: Stochasticity collapse

$$J = \mathbb{E}_{z\sim q(z|x)}[\log p(x\,|\,z)] + \mathbb{D}(q(z), p(z))$$



Theorem 1. *Suppose we have a Gaussian family* $\mathcal{N}(\boldsymbol{\mu}, \operatorname{diag}\boldsymbol{\sigma}^2)$*, where* $\boldsymbol{\mu}$ *and* $\boldsymbol{\sigma}$ *are parameters. The covariance is diagonal, meaning that the variables are independent. If the gradient of* $\boldsymbol{\sigma}$ *completely comes from sample gradient and* $\boldsymbol{\sigma}$ *is small at the beginning of training, then the Gaussian converges to a Dirac delta function with stochastic gradient descent, i.e.,* $\boldsymbol{\sigma} \to \mathbf{0}$*. (See Appendix A for the proof.)* □

(a) $\lambda_{\mathrm{KL}} = 0$

- If
  - Gaussian encoder
  - Gradient comes from samples
  - Sampling var << Data var
- Then
  - $\sigma^2 \to 0$ by SGD

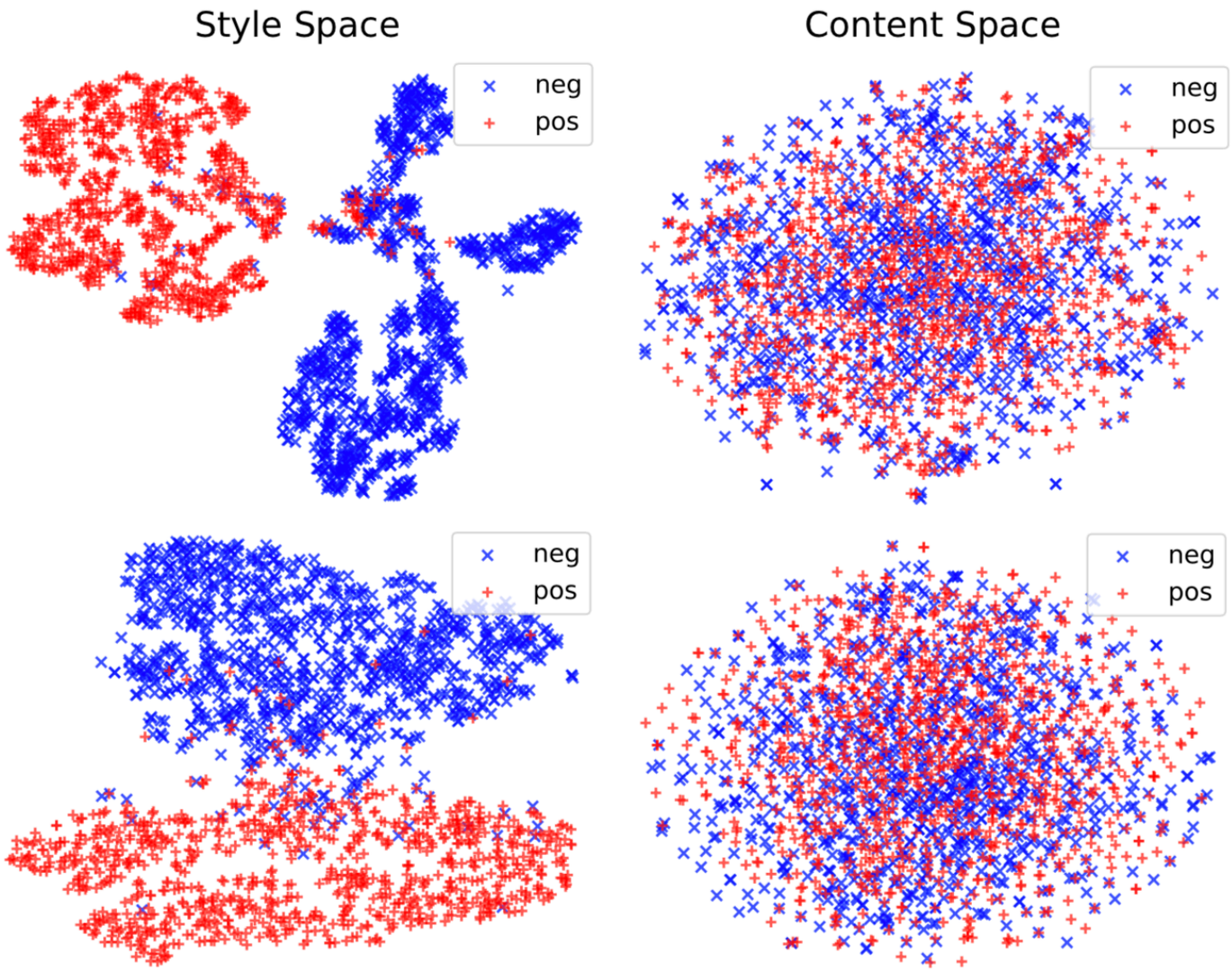
Bahuleyan, Mou, Zhou, and Vechtomova. Stochastic Wasserstein autoencoder for probabilistic sentence generation. In *NAACL-HLT*, 2019

UNIVERSITY OF ALBERTA

Applications of *AEs

- Regularization

Especially good for interpolation



Applications of *AEs

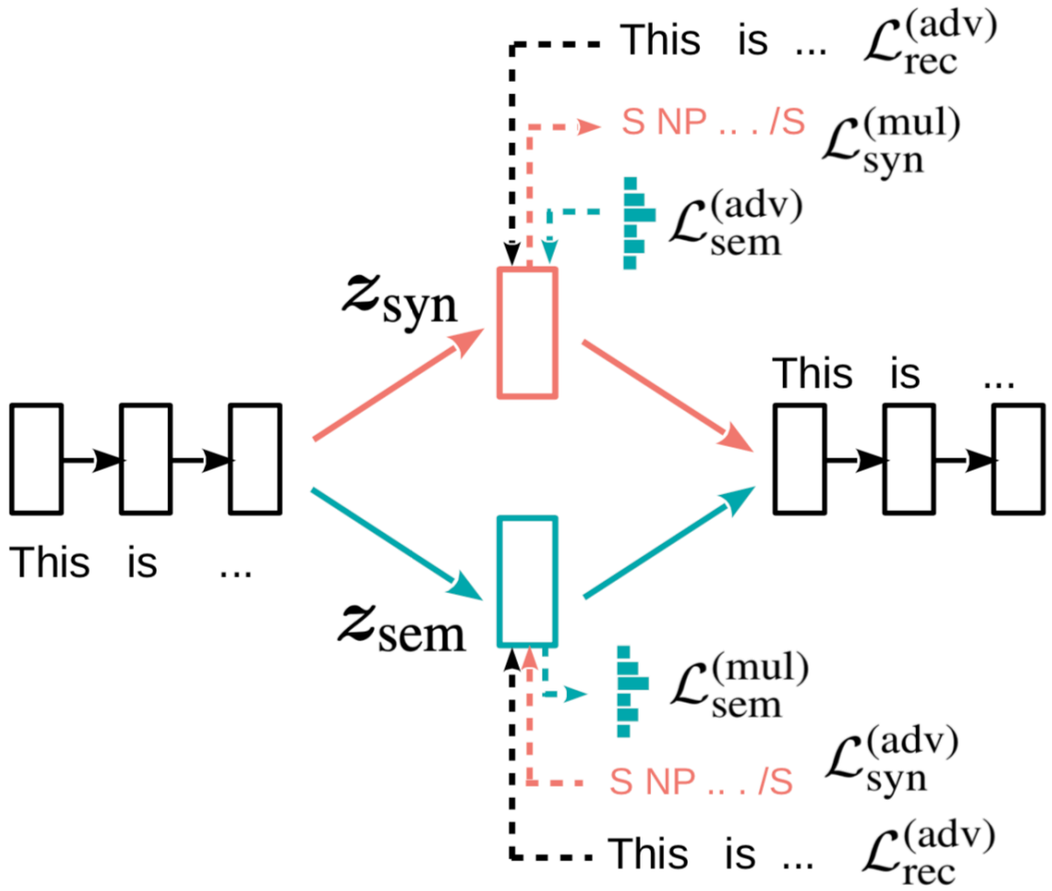
- Prior sampling

WAE-D ($\lambda_{\text{WAE}} = 10$)
the lone man is working .
the group of men is using ice at the sunset .
a family is outside in the background .
two women are standing on a busy street outside a fair
a tourists is having fun on a sunny day
WAE-S ($\lambda_{\text{WAE}} = 10$, $\lambda_{\text{KL}} = 0.01$)
an asian man is dancing in a highland house .
a person wearing a purple snowsuit jumps over the tree .
the vocalist is at the music and dancing with a microphone .
a young man is dressed in a white shirt cleaning clothes .
three children lie together and a woman falls in a plane .

Training Samples
a mother and her child are outdoors.
the people are opening presents.
the girls are looking toward the water.
a small boy walks down a wooden path in the woods.
a person in a green jacket it surfing while holding on to a line.
DAE
two families walking in a towel down alaska sands a cot .
a blade is rolling its nose furiously paper .
a woman in blue shirts is passing by a some beach
transporting his child are wearing overalls .
a guys are blowing on professional thinks the horse .
VAE without Annealing
a man is playing a guitar .
a man is playing with a dog .
a man is playing with a dog .
a man is playing a guitar .
a man is playing with a dog .
VAE with Annealing
the band is sitting on the main street .
couple dance on stage in a crowded room .
two people run alone in an empty field .
the group of people have gathered in a picture .
a cruise ship is docking a boat ship .

Applications of *AEs

- Posterior sampling
  - Paraphrase generation
  - Style-transfer generation



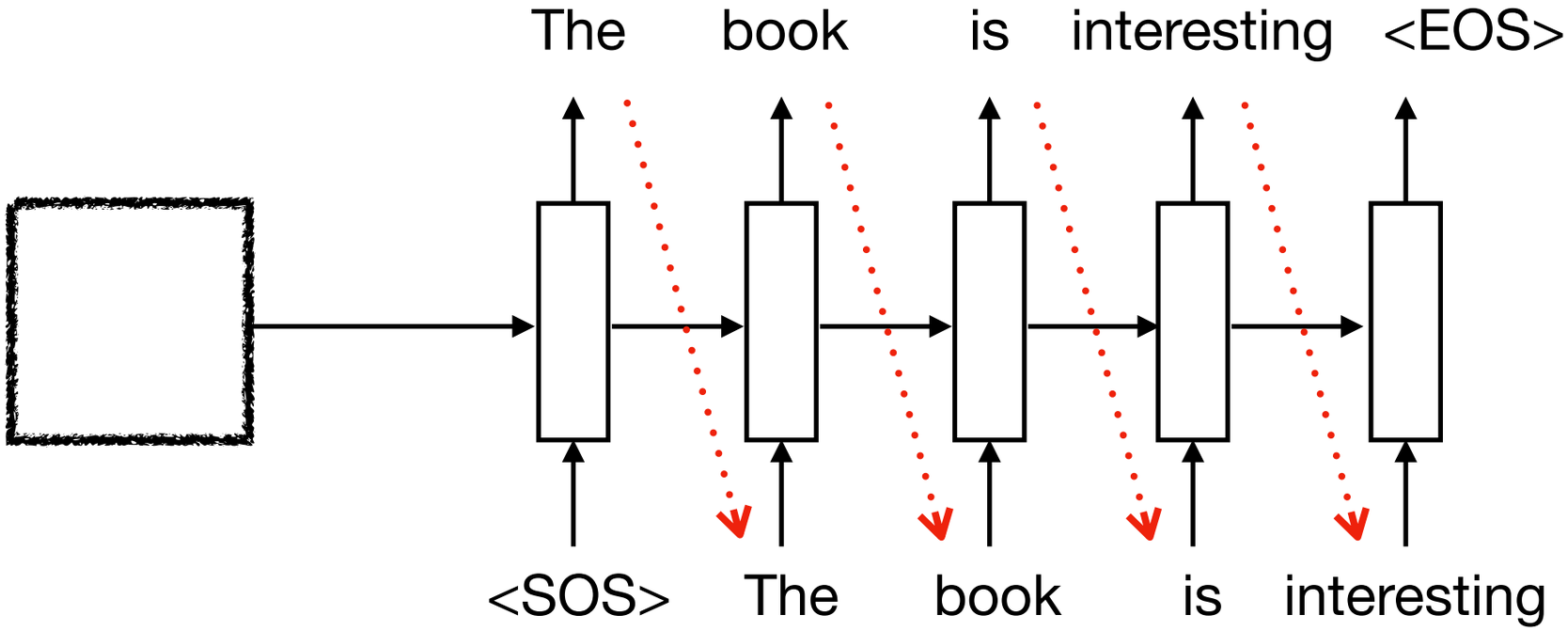
Semantic and Syntactic Providers		Syntax-Transfer Output	
Ref$_{\text{syn}}$:	There is an apple on the table.	**VAE**:	The man is in the kitchen.
Ref$_{\text{sem}}$:	The airplane is in the sky.	**DSS-VAE**:	There is a airplane in the sky.
Ref$_{\text{syn}}$:	The shellfish was cooked in a wok.	**VAE**:	The man was filled with people.
Ref$_{\text{sem}}$:	The stadium was packed with people.	**DSS-VAE**:	The stadium was packed with people.
Ref$_{\text{syn}}$:	The child is playing in the garden.	**VAE**:	There is a person in the garden.
Ref$_{\text{sem}}$:	There is a dog behind the door.	**DSS-VAE**:	A dog is walking behind the door.

Word Space Sampling

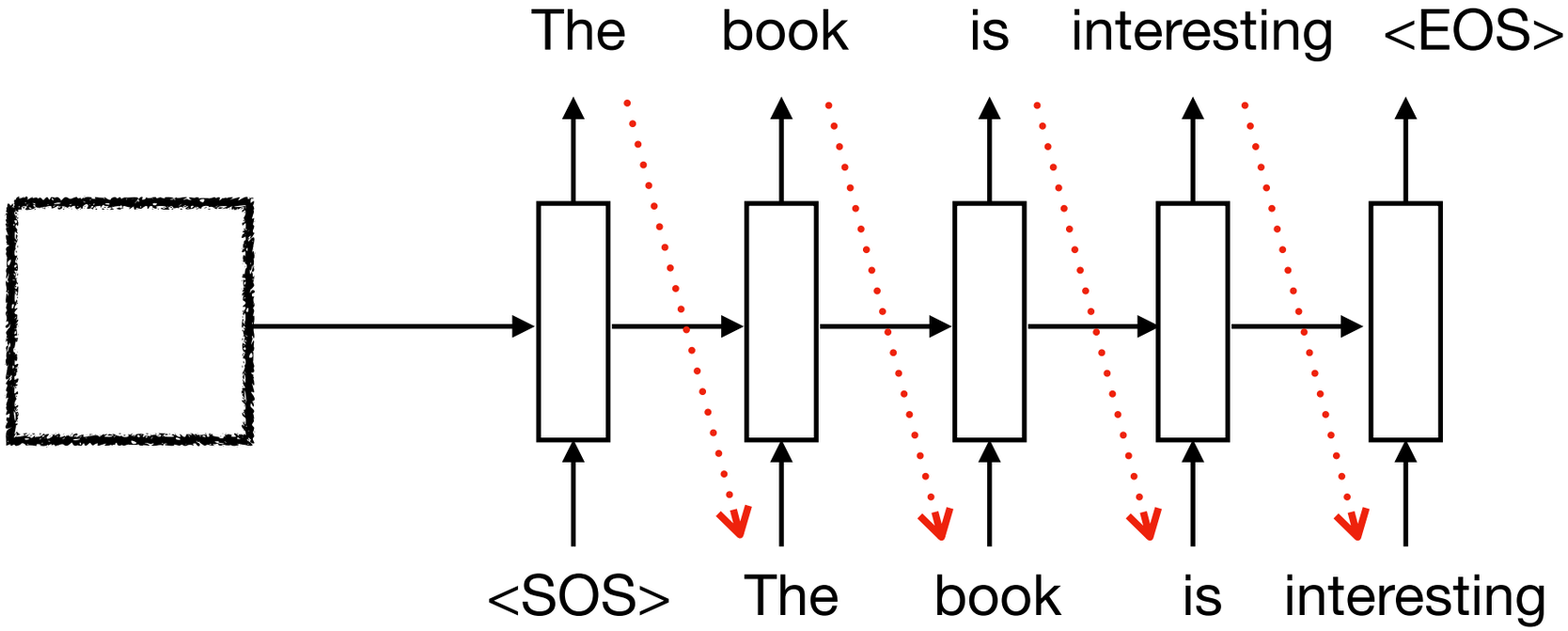
Miao, Zhou, Mou, Yan, Li. CGMH: Constrained sentence generation by Metropolis-Hastings sampling. In *AAAI*, 2019.

UNIVERSITY OF ALBERTA

RNN Generation



RNN Generation

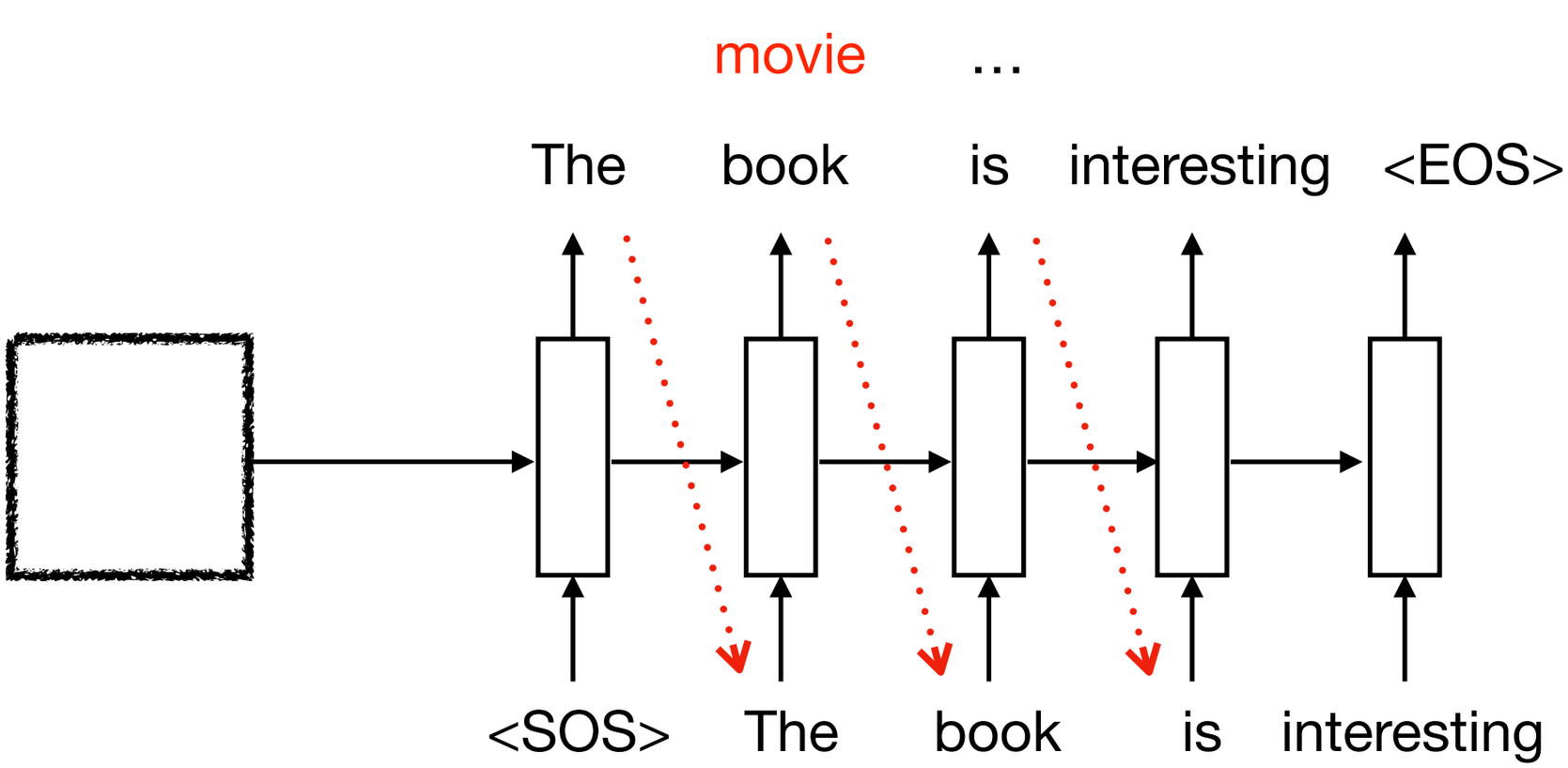


Question: Can we generate a sentence right-to-left?

UNIVERSITY OF ALBERTA

Issues with Single Directional Generation

- Information bottleneck
- Error cumulation
  - Due to sampling or incompetency of the RNN



Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

~~The~~ book is interesting <EOS>

This

Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

~~The~~ book is interesting <EOS>

This

quite

Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

fascinating

~~The~~ book is ~~interesting~~ <EOS>

This

quite

Applications

- Paraphrase generation
  - “Sample” a sentence with similar semantics but different wordings
- Summarization
  - “Sample” a sentence with similar semantics
- Grammatical error correction
  - “Sample” a more likely sentence with the same semantics

UNIVERSITY OF ALBERTA

Sampling Methods

Independent Sampling

- Sampling from CDF
  - Probabilistic density function (PDF)

$$\Pr[a \le x \le b] = \int_a^b f(x)\,dx$$

  - Cumulative density function (CDF)

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \Pr[u \le x]$$

  - Sampling procedure

$$u \sim \mathrm{U}[0,1]; \quad x = \mathrm{CDF}^{-1}(u)$$

- Problems
  - CDF not analytic
  - Especially, the conditional CDF in multivariate cases

Independent Sampling

- Rejection sampling
  - To sample from $p(x) = \frac{1}{Z}\widetilde{p}(x)$
  - We instead sample $x \sim q(x)$
  - Accept the sample x with probability $\frac{\widetilde{p}(x)}{k \cdot q(x)}$

    where k is a constant s.t. $kq(x) \geq \widetilde{p}(x), \forall x$
  - Reject x w.p. $1 - \frac{\widetilde{p}(x)}{k \cdot q(x)}$
- Many other sampling methods

Dependent Sampling

- Goal: Sample from $p(x)$
- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} \mid x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} \mid x^{(1)}), \quad \cdots$
  - Hope $p(x^{(n)}) \to p(x)$ as $n \to \infty$

Markov Chain

- States: $S = \{s_1, s_2, \cdots\}$
- Initial distribution $\pi^{(0)}$
- Transition probability: $\mathcal{T}_{i \to j} = p(x^{(t+1)} = s_j \,|\, x^{(t)} = s_i)$
  - $x^{(t+1)}$ is independent of $x^{(t-1)}$, given $x^{(t)}$
  - $\mathcal{T}_{i \to j}$ works for all time steps t
- **Thm**: Starting from an arbitrary initial distribution, a Markov Chain converges to a **unique** stationary distribution (under mild assumptions).

UNIVERSITY OF ALBERTA

Markov Chain Monte Carlo

- Goal: Sample from $p(x)$
- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} \mid x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} \mid x^{(1)}), \quad \cdots$
  - Hope $p(x^{(n)}) \to p(x)$ as $n \to \infty$

Markov Chain Monte Carlo

- Goal: Sample from $p(x)$
- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} | x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} | x^{(1)}), \quad \cdots$

    by following a carefully designed Markov chain
  - ~~Hope~~ $p(x^{(n)}) \to p(x)$ as $n \to \infty$

Guaranteed that

UNIVERSITY OF ALBERTA

Metropolis—Hastings Sampler

- **Input**
  - An arbitrary desired distribution $p(x)$
- **Output**
  - An unbiased sample $x \sim p(x)$
- **Algorithm**
  - Start from an arbitrary initial state $x^{(0)}$
  - For every step t

    $g(x'|x)$: arbitrary proposal distribution

    Propose a new state $x' \sim g(x'|x^{(t)})$

    Accept x' w.p. $A(x'|x) = \min\left\{1, \frac{p(x')g(x^{(t)}|x')}{p(x)g(x'|x^{(t)})}\right\}$, i.e., $x^{(t+1)} = x'$

    Reject x' otherwise, i.e., $x^{(t+1)} = x^{(t)}$
  - Return $x^{(t)}$ with a large t

UNIVERSITY OF ALBERTA

Proof Sketch

- Detailed balance property = > Stationary distribution

If

$$\forall x, y, \qquad \pi(x) \cdot \mathcal{T}_{x \to y} = \pi(y) \cdot \mathcal{T}_{y \to x}$$

Then

$\pi(x)$ is a stationary distribution

Because

$$\forall x, \qquad \pi(x) = \sum_y \pi(y) \mathcal{T}_{y \to x} = \sum_y \pi(x) \mathcal{T}_{x \to y} = \pi(x)$$

Proof Sketch (Cont.)

- MH Sampler satisfies detailed balance
  - $\forall x, y,$ if $x \neq y$, $p(x) \cdot \mathcal{T}_{x \to y} = p(x) \cdot g(y \mid x) \cdot \min\left\{1, \frac{p(y)g(x \mid y)}{p(x)g(y \mid x)}\right\}$ (1)

$$p(y) \cdot \mathcal{T}_{y \to x} = p(y) \cdot g(x \mid y) \cdot \min\left\{1, \frac{p(x)g(y \mid x)}{p(y)g(x \mid y)}\right\} \quad (2)$$

  - W.L.O.G., we assume $p(x)g(y \mid x) \geq p(y)g(x \mid y)$

$$(1) = p(y) \cdot g(x \mid y)$$

$$(2) = p(y) \cdot g(x \mid y)$$

  - $\forall x, y,$ if $x = y$, $p(x)\mathcal{T}_{x \to y} = p(y)\mathcal{T}_{y \to x}$ also holds

Gibbs Sampler

- Suppose $\mathbf{x} = (x_1, x_2, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n)$
- If the proposal distribution is $x_i' \sim p(x_i \,|\, \mathbf{x}_{-i})$
- Then, the acceptance rate is $A(\mathbf{x}' \,|\, \mathbf{x}) = \min\left\{1, \frac{p(\mathbf{x}')g(\mathbf{x} \,|\, \mathbf{x}')}{p(\mathbf{x})g(\mathbf{x}' \,|\, \mathbf{x})}\right\}$
  - Notice that $\mathbf{x}' = (x_1, x_2, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)$
  - Thus, $\frac{p(\mathbf{x}')g(\mathbf{x} \,|\, \mathbf{x}')}{p(\mathbf{x})g(\mathbf{x}' \,|\, \mathbf{x})} = \frac{p(\mathbf{x}_{-i})p(x_i' \,|\, \mathbf{x}_{-i}) \cdot p(x_i \,|\, \mathbf{x}_{-i})}{p(\mathbf{x}_{-i})p(x_i \,|\, \mathbf{x}_{-i}) \cdot p(x_i' \,|\, \mathbf{x}_{-i})} = 1$

=> Gibbs step is a special case of an MH step, with AC rate = 1

Applying MH to Sentence Generation

MH Components

- State: Every sentence
- Target distribution: Depend on the task
- Proposal distribution
  - Task agnostic, or task specific
- Compute acceptance rate
  - We can’t do anything here

Target distribution

- General formula
  - $p(\mathbf{x}) \propto p_{\mathrm{LM}}(\mathbf{x}) \cdot s_1(\mathbf{x}) \cdots s_n(\mathbf{x})$
  - $s_i(\mathbf{x})$: scoring functions specific to the task

Target distribution

- General formula
  - $p(\mathbf{x}) \propto p_{\mathrm{LM}}(\mathbf{x}) \cdot s_1(\mathbf{x}) \cdots s_n(\mathbf{x})$
  - $s_i(\mathbf{x})$: scoring functions specific to the task

- Keywords-to-sentence generation

$$s(\mathbf{x}) = \begin{cases} 1, & \text{if keywords in } \mathbf{x} \\ 0, & \text{otherwise} \end{cases}$$

- Paraphrase generation/Grammatical error correction
  - $s(\mathbf{x}) = \mathrm{sim}_{\mathrm{semantic}}(\mathbf{x}, \mathbf{x}_0) + \mathrm{diff}_{\mathrm{word}}(\mathbf{x}, \mathbf{x}_0)$

Proposal Distribution

- Replace

$$g_{\text{replace}}(\mathrm{x}'|\mathrm{x}) = \pi(w_m^* = w^c|\mathrm{x}_{-m}) = \frac{\pi(w_1, \cdots, w_{m-1}, w^c, w_{m+1}, \cdots, w_n)}{\sum_{w \in \mathcal{V}} \pi(w_1, \cdots, w_{m-1}, w, w_{m+1}, \cdots, w_n)}$$

- Delete
- Insert
  - Also sample from posterior

Examples: Keywords-to-Sentence

Keyword(s)	**Generated Sentences**
friends	My good **friends** were in danger .
project	The first **project** of the scheme .
have, trip	But many people **have** never made the **trip** .
lottery, scholarships	But the **lottery** has provided **scholarships** .
decision, build, home	The **decision** is to **build** a new **home** .
attempt, copy, painting, denounced	The first **attempt** to **copy** the **painting** was **denounced** .

Examples: Paraphrase Generation

Model	BLEU-ref	BLEU-ori	NLL
Origin Sentence	30.49	100.00	7.73
VAE-SVG (100k)	22.50	-	-
VAE-SVG-eq (100k)	**22.90**	-	-
VAE-SVG (50k)	17.10	-	-
VAE-SVG-eq (50k)	17.40	-	-
Seq2seq (100k)	22.79	33.83	6.37
Seq2seq (50k)	20.18	27.59	**6.71**
Seq2seq (20k)	16.77	22.44	6.67
VAE (unsupervised)	9.25	27.23	7.74
CGMH *w/o matching*	18.85	50.28	7.52
w/ KW	20.17	53.15	7.57
w/ KW + WVA	20.41	53.64	7.57
w/ KW + WVM	20.89	54.96	7.46
w/ KW + ST	20.70	54.50	7.78

Type	Examples
Ori	what 's the best plan to lose weight
Ref	what is a good diet to lose weight
Gen	what 's the best way to slim down quickly
Ori	how should i control my emotion
Ref	how do i control anger and impulsive emotions
Gen	how do i control my anger
Ori	why do my dogs love to eat tuna fish
Ref	why do my dogs love eating tuna fish
Gen	why do some dogs like to eat raw tuna and raw fish

UNIVERSITY OF ALBERTA

Examples: Paraphrase Generation

Model	#parallel data	GLEU
AMU	2.3M	44.85
CAMB-14	155k	46.04
MLE	720k	52.75
NRL	720k	**53.98**
CGMH	0	45.5

Ori	Even if **we are failed** , We have to try to get **a new things** .
Ref	Even if we all failed , we have to try to get new things .
Gen	Even if we are failing , We have to try to get some new things .

Ori	In the world **oil price very high** right now .
Ref	In today 's world , oil prices are very high right now .
Gen	In the world , oil prices are very high right now .

UNIVERSITY OF ALBERTA

Analysis

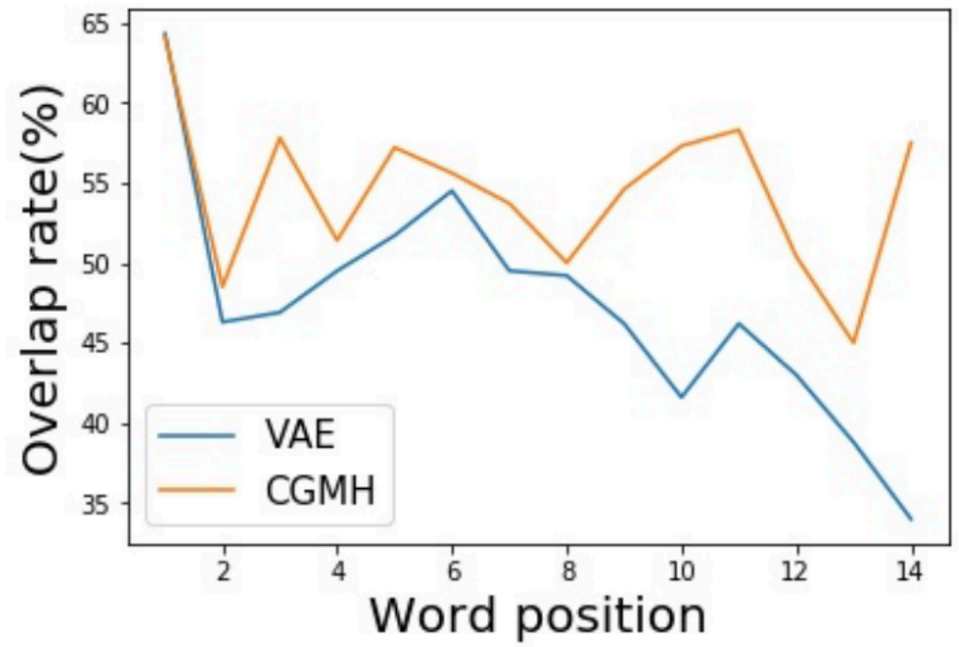
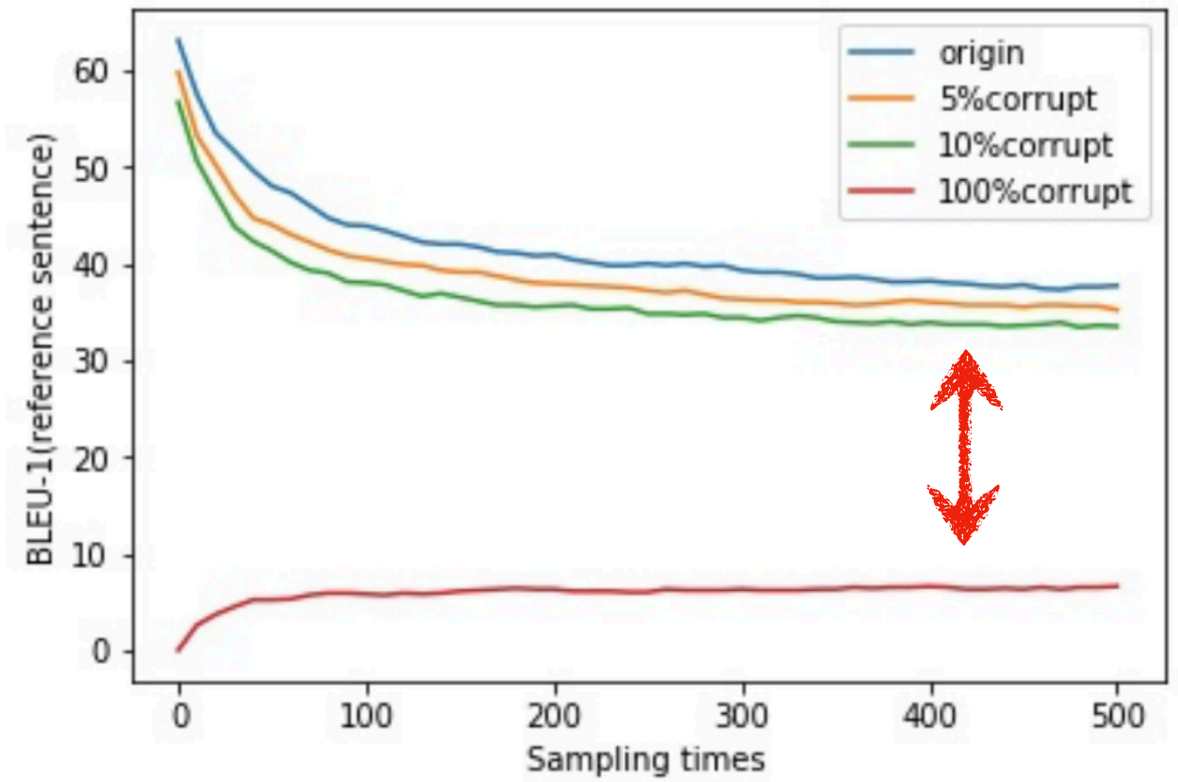


Figure 3: Overlap rates of CGMH and VAE for each word position of sentences.

UNIVERSITY OF ALBERTA

Analysis (Cont.)



The Markov Chain never mixes.
We mainly use MH as SA.

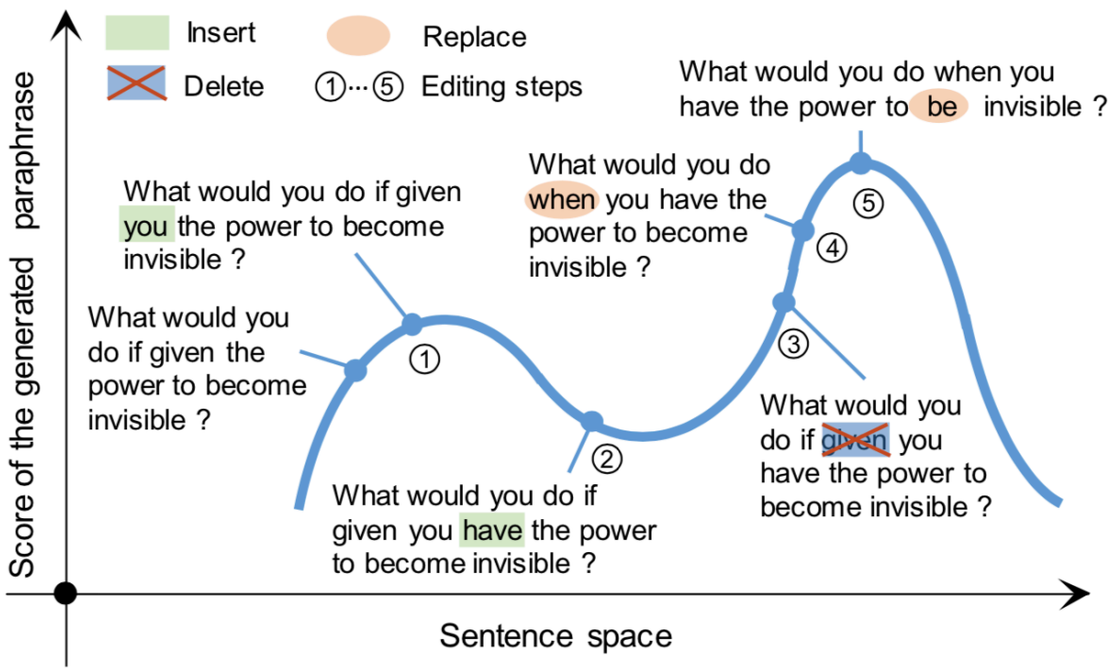
Figure 2: Generation quality with corrupted initial states. At each situation, 0/5%/10%/100% of the words in initial sentences are randomly replaced with other words.

UNIVERSITY OF ALBERTA

Simulated Annealing

- MH: Sampling from $\propto \exp\{s(x)\}$

- SA: Searching the optimum of $s(x)$
  - Define $p_\tau(x) \propto \exp\{s(x)/\tau\}$
  - Start from high temperature, but cool it down gradually
  - With $\tau \to 0$,
    $p_\tau(x) = 1$ if $x = \text{argmax}\, s(x)$, or 0 otherwise



$$p(\text{accept}|\mathrm{x}_*, \mathrm{x}_t, T) = \min\left(1, e^{\frac{f(\mathrm{x}_*) - f(\mathrm{x}_t)}{T}}\right)$$

$$T = \max(0, T_{\text{init}} - C \cdot t)$$

Liu, Mou, Meng, Zhou, Zhou, Song. Unsupervised Paraphrasing by Simulated Annealing. arXiv preprint arXiv:1909.03588, 2019.

References

- Kingma, D. P., & Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.
- Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. arXiv preprint arXiv:1511.05644. 2015.
- Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R, Bengio S. Generating sentences from a continuous space. In *CoNLL*, 2016.
- Bahuleyan, Mou, Zhou, and Vechtomova. Stochastic Wasserstein autoencoder for probabilistic sentence generation. In *NAACL-HLT*, 2019.
- Miao, Zhou, Mou, Yan, Li. CGMH: Constrained sentence generation by Metropolis-Hastings sampling. In *AAAI*, 2019.
- Liu, Mou, Meng, Zhou, Zhou, Song. Unsupervised Paraphrasing by Simulated Annealing. arXiv preprint arXiv:1909.03588, 2019.

CMPUT 651 (Fall 2019)

Thank you!

Q&A

UNIVERSITY OF ALBERTA