Journal Club

Dec 01, 2020

Term-project: Latent Dirichlet Allocation and Wasserstein LDA

Jaewon Bae and Sangheon Lee

Korea Advanced Institute of Science and Technology, Daejeon, South Korea
Email: {duckgoose, buaaaaang}@kaist.ac.kr

CONTENTS

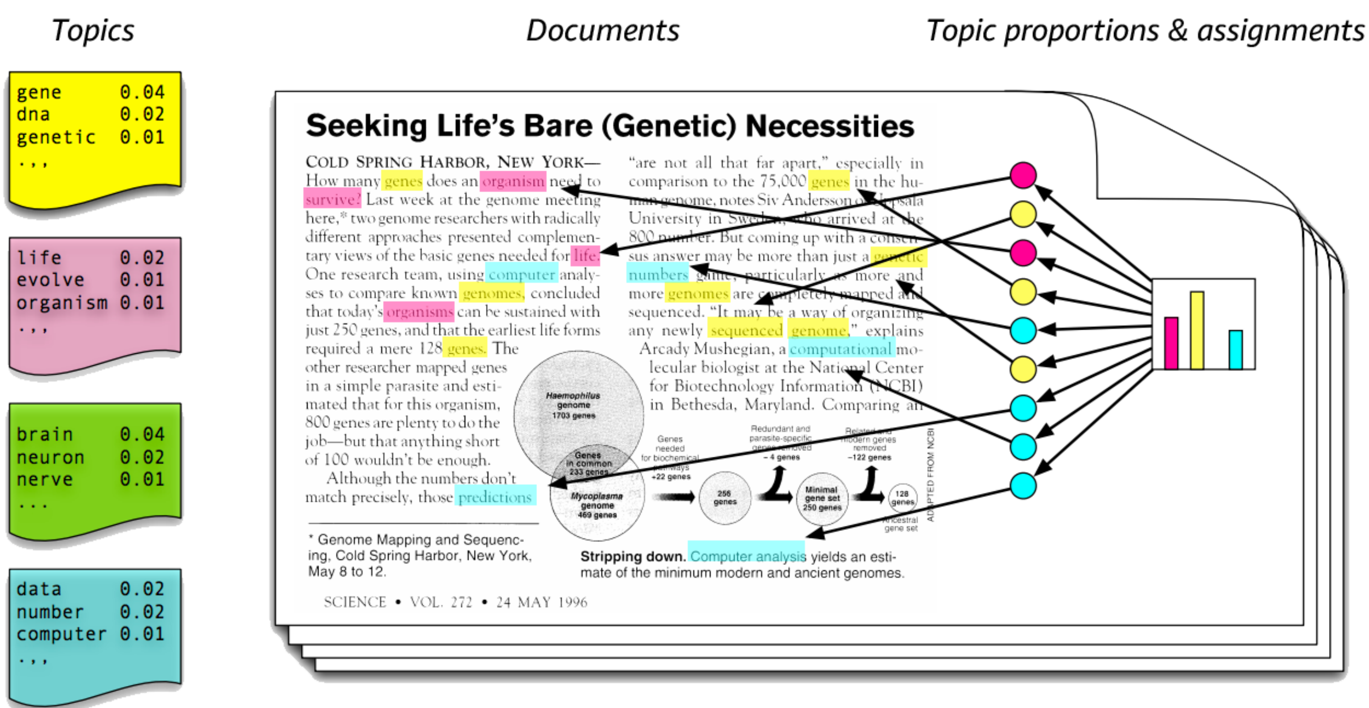
CONTENTS

Contents – LDA

- Dirichlet Distribution
- Latent Dirichlet Allocation
  - Process
  - Example
  - LDA objectives
- Comparison with previous models
  - Motivation of LDA
- LDA and exchangeability
  - De Finetti Theorem

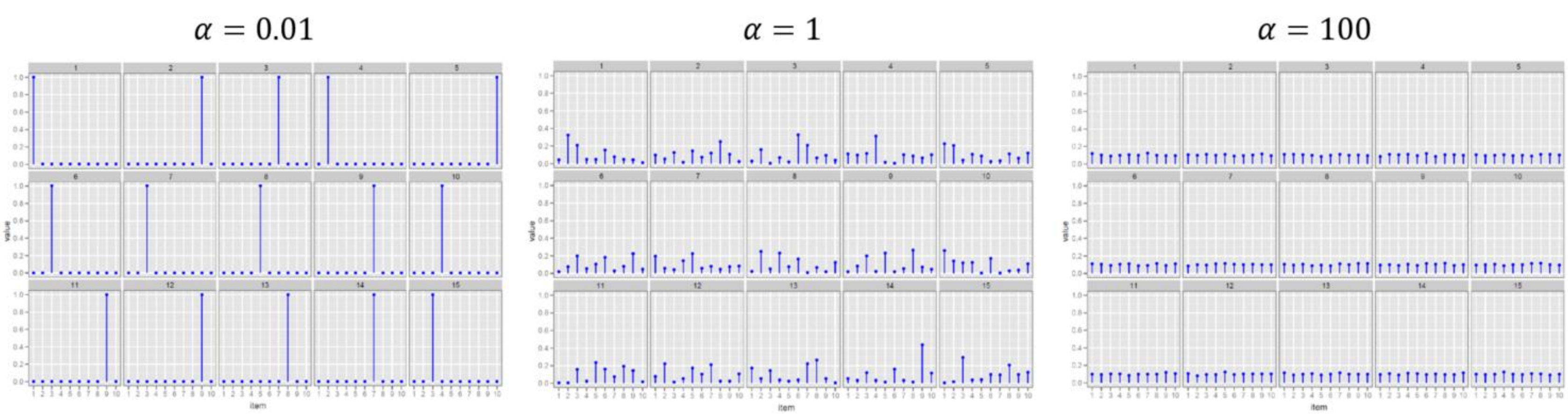
Latent Dirichlet Allocation

- Basic ideas
  - Documents are represented as random mixtures over latent topics
  - Each topic is characterized by a distribution over words.



Dirichlet Distribution

- Dirichlet Distribution
  - Sum to one
  - Captures the intuition: Document typically belongs to a sparse subset of topics
  - Not belongs to the location family



LATENT DIRICHLET ALLOCATION PROCESS

- LDA notation
  - Word: a basic unit, v$^{\text{th}}$ word: V-vector w with $w^v = 1$ and $w^u = 0$ for $u \neq v$
  - Document: sequence of N words, $\mathbf{w} = (w_1, \ldots w_N)$
  - Corpus: collection of M documents, $\mathrm{D} = (\mathbf{w_1}, \ldots \mathbf{w_M})$

- LDA Process for each document
  1. Choose N
  2. Choose $\theta \sim \mathrm{Dir}(\alpha)$, where $\alpha = (\alpha_1, \ldots \alpha_k)$
  3. For each of the N words w_n:
     a. Choose a topic $z_n \sim \mathrm{Multi}(\theta)$
     b. Choose a word $w_n \sim p(w_n | z_n, \beta)$, Multinomial probability given topic z_n

     where $\beta \in \mathrm{Mat}_{k \times V}$, $\beta_{\mathrm{ij}} = p\left(w^j = 1 \middle| z^i = 1\right)$

LDA Process Example

Consider 3rd document and 1st word

1. Choose $N = 10$

2. Choose $\theta \sim \mathrm{Dir}(\alpha)$, where $\alpha = (\alpha_1, \ldots \alpha_3)$

Docs	Topic 1	Topic 2	Topic 3
Doc 1	0.400	0.000	0.600
Doc 2	0.000	0.600	0.400
Doc 3	0.375	0.625	0.000
Doc 4	0.000	0.375	0.625
Doc 5	0.500	0.000	0.500
Doc 6	0.500	0.500	0.000

3.a. Choose a topic $z_1 \sim \mathrm{Multi}(\theta)$ as Topic 2

Note.

$p(z_n|\theta) = \theta_i$ for a unique i such that $z_n^i = 1$

The pmf: $\frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}$ with $n = \sum x_i = 1$

3.b. Choose a word $w_1 \sim p(w_1|z_1, \beta)$ given z_1: Topic 2

Terms	Topic 1	Topic 2	Topic 3
Baseball	0.000	0.000	0.200
Basketball	0.000	0.000	0.267
Boxing	0.000	0.000	0.133
Money	0.231	0.313	0.400
Interest	0.000	0.312	0.000
Rate	0.000	0.312	0.000
Democrat	0.269	0.000	0.000
Republican	0.115	0.000	0.000
Cocus	0.192	0.000	0.000
President	0.192	0.063	0.000

LDA Process Example

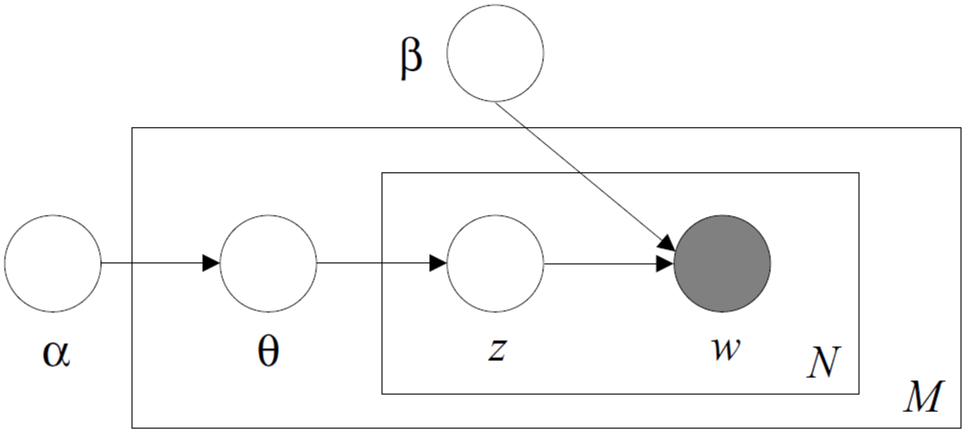


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

LDA AND EXCHANGEABILITY

- Basic ideas
  - Documents are represented as **random mixtures** over latent topics
  - Each topic is characterized by a distribution over words.
- Exchangeability: order can be neglected, i.e., $p(x_1, \dots, x_n) = p\left(x_{\sigma(1)}, \dots, x_{\sigma(n)}\right)$
- De Finetti Theorem:
  - Infinitely exchangeable ⇔ Conditionally independent given some r.v.

$$p(w, z) = \int p(\theta)\{\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n)\}d\theta$$

Theorem 2 (De Finetti, 1930s). *A sequence of random variables $(x_1, x_2, \ldots)$ is infinitely exchangeable iff, for all n,*

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta)P(d\theta),$$

for some measure P on θ.

LATENT DIRICHLET ALLOCATION OBJECTIVES

- Distributions
  - Dirichlet pdf: $p(\theta|\alpha) = \frac{\Gamma(\sum\alpha_i)}{\prod\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$, $\theta_i \geq 0, \sum\theta_i = 1$
  - Multinomial pmf: $\frac{n!}{x_1!\cdots x_k!}\theta_1^{x_1}\cdots\theta_k^{x_k}$ with $n = \sum x_i$
- Probabilities
  - $p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta)$
  - $p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta)\right) d\theta$
  - $p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N} p\left(z_{d,n}\middle|\theta_d\right) p\left(w_{d,n}\middle|z_{d,n}, \beta\right)\right) d\theta_d$

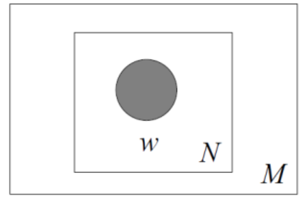
LDA vs Other Latent Variable Models

- Unigram model
  - A single topic for all documents.
  - $w_n \sim \text{Multi}(\theta)$ for a given θ. $p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$
- Mixture of unigrams
  - A single topic for each document.
  - $w_n \sim p(w_n|z)$. $p(\mathbf{w}) = \sum p(z) \prod_{n=1}^{N} p(w_n|z)$
- pLSI (probability Latent Semantic Indexing)
  - For each document, $z \sim \text{Multi}(d)$. Multiple topics for a single document
  - $p(d, w_n) = p(d) \sum p(w_n|z) p(z|d)$

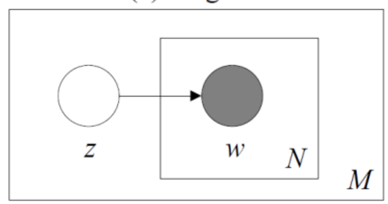
LDA VS OTHER LATENT VARIABLE MODELS

- Limitations
  - Unigram and mixture of unigrams
    - Do not consider multiple topics for a document
  - pLSI
    - $p(z|d)$: only for trained sets → cannot assign probability to new one
      - pLSI: z~Multi(d) where d from document labels
      - LDA: z_n~Multi(θ) where θ~Dir(α)
    - Number of parameters: pLSI depends on the number of documents
      - pLSI: $kM + kV$ → may cause overfitting
      - LDA: $k + kV$

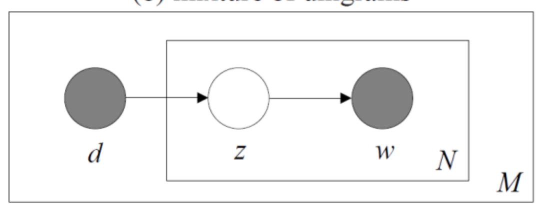
LDA AND OTHER MODEL FIGURES



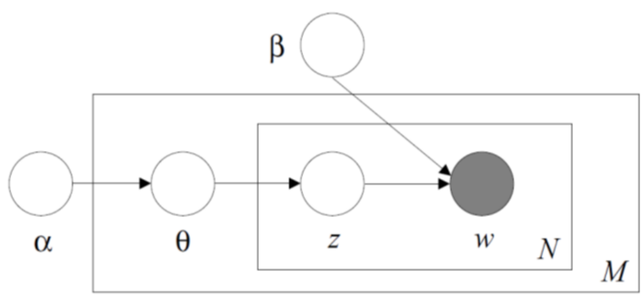
(a) unigram



(b) mixture of unigrams



(c) pLSI/aspect model



NMF AND PLSI

- Setting
  - Document: sequence of N words, $\mathbf{w} = (w_1, \dots w_N)$
  - Corpus: collection of M documents, $\mathrm{D} = (\mathbf{w_1}, \dots \mathbf{w_M})$
  - Word-to-document matrix: $F \in M_{N\times M}$, $\left(F_{ij}\right) = F\left(w_i, d_j\right)$ with normalization
- NMF (Non-negative Matrix Factorization): $F = CH^T$
  - Minimize $J_{\mathrm{NMF}} = \sum_{i=1}^{N} \sum_{j=1}^{M} F_{ij} \log \frac{F_{ij}}{(CH^T)_{ij}} - F_{ij} + (CH^T)_{ij}$
- pLSI
  - Maximize $J_{\mathrm{pLSI}} = \sum_{i=1}^{N} \sum_{j=1}^{M} F_{ij} \log p\left(w_i, d_j\right)$

Inference of LDA: Intractability

- Intractable posterior: $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$
  - $p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta$
  - $p(\theta|\alpha) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)}\theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$ and $p(z_n|\theta) = \theta_i$ for a unique i with $z_n^i = 1$
  - $p(z_n|\theta)p(w_n|z_n, \beta) = \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i\, \beta_{ij})^{w_n^j}$: coupling of θ and β
- Approximation of the posterior
  - Variational inference
  - Collapsed Gibbs sampling
  - MCMC

Variational Inference: Based on Convexity

- Decoupling of θ and β: $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) \approx q(\theta, \mathbf{z}|\gamma, \phi)$
    - $q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$
    - Dirichlet parameter γ and Multinomial parameters $\{\phi_n\}_1^N$
- Aim: minimize $D_{KL}\big(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)\big)$
    - $\log\big(p(\mathbf{w}|\alpha, \beta)\big) = L(\gamma, \phi; \alpha, \beta) + D_{KL}\big(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)\big)$
    - $L(\gamma, \phi; \alpha, \beta) = \mathrm{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - \mathrm{E}_q[\log q(\theta, \mathbf{z})]$
- Equivalent with Aim: maximize $L(\gamma, \phi; \alpha, \beta)$

VARIATIONAL INFERENCE: BASED ON CONVEXITY

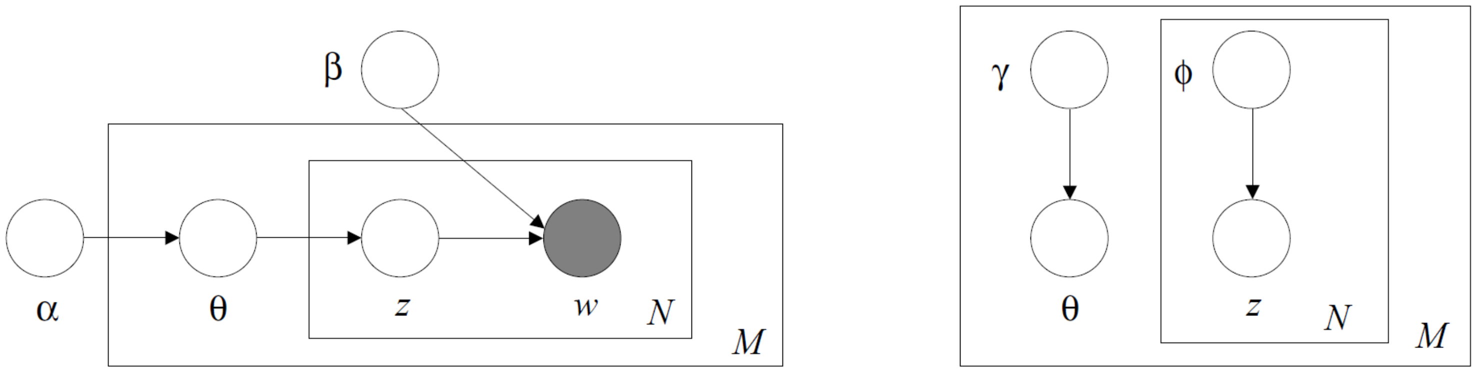


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

MAXIMIZE A LOWER BOUND OF $\log\big(p(\mathbf{w}|\alpha,\beta)\big)$

- $L(\gamma,\phi;\alpha,\beta) = \mathrm{E}_q[\log p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)] - \mathrm{E}_q[\log q(\theta,\mathbf{z})]$
  - $\mathrm{E}_q[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi\big(\sum\gamma_j\big)$

$$\begin{aligned} L(\gamma,\phi;\alpha,\beta) = {} & \mathrm{E}_q[\log p(\theta\,|\,\alpha)] + \mathrm{E}_q[\log p(\mathbf{z}\,|\,\theta)] + \mathrm{E}_q[\log p(\mathbf{w}\,|\,\mathbf{z},\beta)] \\ & - \mathrm{E}_q[\log q(\theta)] - \mathrm{E}_q[\log q(\mathbf{z})]. \end{aligned}$$

$$\begin{aligned} L(\gamma,\phi;\alpha,\beta) = {} & \log\Gamma\left(\textstyle\sum_{j=1}^k \alpha_j\right) - \sum_{i=1}^k \log\Gamma(\alpha_i) + \sum_{i=1}^k (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right) \\ & + \sum_{n=1}^N \sum_{i=1}^k \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right) \\ & + \sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^V \phi_{ni} w_n^j \log\beta_{ij} \\ & - \log\Gamma\left(\textstyle\sum_{j=1}^k \gamma_j\right) + \sum_{i=1}^k \log\Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right) \\ & - \sum_{n=1}^N \sum_{i=1}^k \phi_{ni}\log\phi_{ni}, \end{aligned}$$

CONTENTS

CONTENTS – WASSERSTEIN LDA

- Introduction: Motivation and Mission
- Encoder-Decoder pair
  - Deterministic Encoder & Decoder
  - Objective
  - Distribution Matching
  - WAE objective: WAE vs VAE
  - Measures for topic diversity and coherence
- Theoretical Part
  - Objective function of WAE
  - Maximum Mean Discrepancy

Motivation of W-LDA

- Topic Modeling
  - Popular: LDA (Variational Bayesian)
  - Recent: deep neural network used (ex. VAE)

- Topic Modeling used VAE
  - Inference is carried out easily without expensive iterative sampling as in VB
  - Location family assumption
  - All samples be forced to match the prior – ELBO term

MISSION OF W-LDA

- Dirichlet Distribution
  - Not belongs to the location family
  - Captures the intuition: Document typically belongs to a sparse subset of topics

- Mission
  - Prior follows Dirichlet distribution – LDA
  - Suitable objective – Two proposed kernels
  - Encoder output should be dependent of the input – Aggregated posterior

WASSERSTEIN LDA (W-LDA)

- Wasserstein LDA
  - Prior is assumed to be followed Dirichlet distribution
  - Apply suitable kernel for distribution matching
  - It performs better than GAN in high dimensional Dirichlet distribution
  - Produce diverse topics – high TU and NPMI scores

Wasserstein LDA Objective

$$\inf_{Q(\theta|\mathbf{w})} \mathbb{E}_{P_{\mathbf{w}}} \mathbb{E}_{Q(\theta|\mathbf{w})} [c(\mathbf{w}, \mathrm{dec}(\theta))] + \lambda \cdot \mathcal{D}_{\Theta}(Q_{\Theta}, P_{\Theta})$$

$D_{\mathrm{WAE}}(P_{\mathbf{w}}, P_{\mathrm{dec}}) = W_c(P_{\mathbf{w}}, P_{\mathrm{dec}}) + \lambda \cdot D_{\theta}(Q_{\theta}, P_{\theta})$ with $W_c(P_{\mathbf{w}}, P_{\mathrm{dec}}) \coloneqq \inf \mathrm{E}_{(X,Y)\sim\Gamma} [c(X, Y)]$.

where $\Gamma \in P(X \sim P_{\mathbf{w}}, Y \sim P_{\mathrm{dec}})$, $p_{\mathrm{dec}}(\mathbf{w}) \coloneqq \int_{\theta} p_{\mathrm{dec}}(\mathbf{w}|\theta) p(\theta) d\theta$, and c is any measurable cost function.

Theorem 1 *For P_G as defined above with deterministic $P_G(X|Z)$ and any function $G \colon \mathcal{Z} \to \mathcal{X}$*

$$\inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma} \left[c(X, Y) \right] = \inf_{Q \colon Q_Z = P_Z} \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)} \left[c(X, G(Z)) \right],$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

DETERMINISTIC W-LDA ENCODER

- MLP mapping $\mathbf{w}$ to an output layer of K units $\theta \in S^{K-1}$
- Inference: $Q(\theta|\mathbf{w}) \approx p(\theta|\mathbf{w})$ (in this case, $Q(\theta|\mathbf{w})$ is a Dirac Delta function)
- Deterministic encoder: $\theta = \mathrm{enc}(\mathbf{w})$ (in contrast, random encoder in VAE)
- Two purposes
  - Distribution matching $Q_\theta \approx P_\theta$: minimize regularization term
  - θ are informative enough for reconstruction at the decoder

DETERMINISTIC W-LDA DECODER

- Single layer NN mapping θ to an output layer of V units $\hat{\mathbf{w}} \in S^{V-1}$
- $\hat{\mathbf{w}} = (\widehat{w_1}, \ldots, \widehat{w_V}), \widehat{w_i} = \frac{\exp h_i}{\sum \exp h_i}$ where $\mathbf{h} = [\beta_1 \cdots \beta_K][\theta_1, \ldots, \theta_K]^{\mathrm{t}} + \mathrm{b}$
- A cost function is simply the negative cross-entropy loss between **w** and $\hat{\mathbf{w}}$
- Deterministic decoder: $\hat{\mathbf{w}} = \mathrm{dec}(\theta)$ (in contrast, random encoder in VAE)
- If decoder is non-deterministic, THM1 gives an upper bound

Theorem 1 *For P_G as defined above with deterministic $P_G(X|Z)$ and any function $G\colon \mathcal{Z} \to \mathcal{X}$*

$$\inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}\big[c(X,Y)\big] = \inf_{Q\colon Q_Z = P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big],$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

DISTRIBUTION MATCHING OF W-LDA

$$\text{MMD}_k(P_\theta, Q_\theta) = \left\| \int_\Theta k(\theta,\cdot) dP_\Theta(\theta) - \int_\Theta k(\theta,\cdot) dQ_\Theta(\theta) \right\|_{\mathcal{H}_k}$$

- Geodesic distance: $d(\theta, \theta') = 2 \arccos\left(\sum \sqrt{\theta_i \theta_i'}\right)$.
- Information diffusion kernel using the distance: $k(\theta, \theta') = \exp\left(-\frac{d^2(\theta,\theta')}{4}\right)$
  - Much more sensitive to points near the boundary of the simplex
- Significance of distribution matching
  - Encoder: stuck in bad local minima: only one dimension θ is nonzero
  - Decoder: fails to produce all meaningful topics

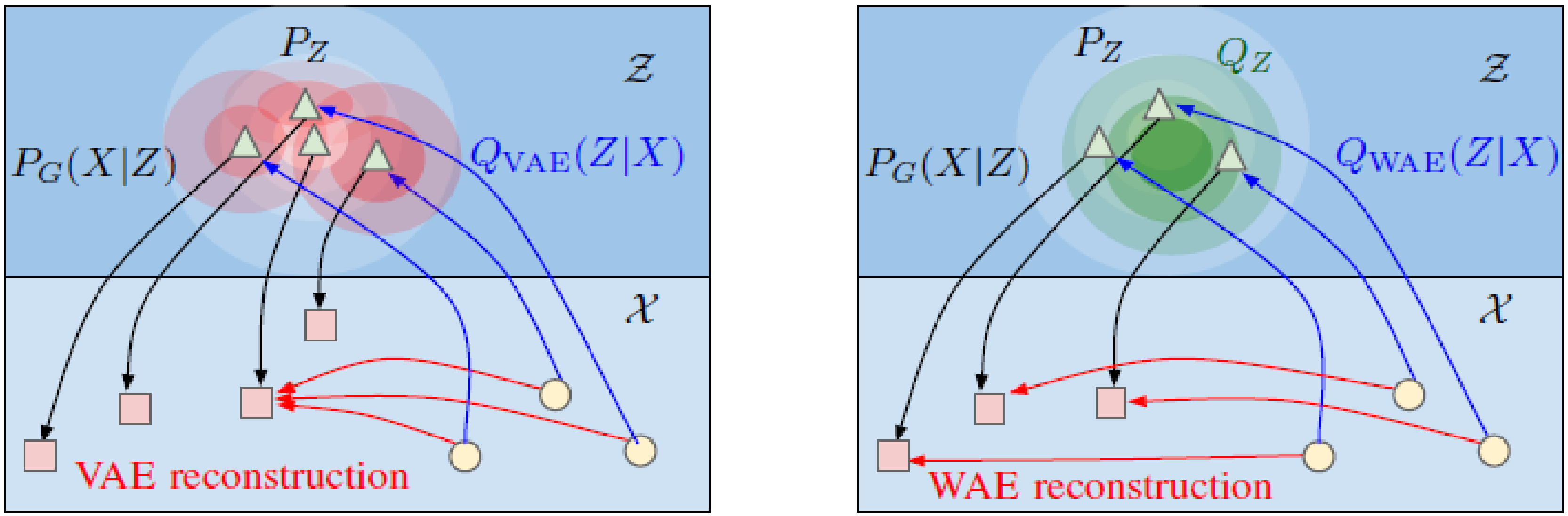
WASSERSTEIN AUTO-ENCODER VS VAE

- Wasserstein Auto-encoder
  - A new family of regularized auto-encoder
  - Shares many of the properties of VAEs
  - Minimize the optimal transport between $P_{\mathbf{w}}$ and $P_{\mathbf{dec}}$ (weaker topology)

- VAE vs WAE
  - VAE: minimize ELBO; regularization term: $Q(\theta|\mathbf{w})$ matches to P_θ
  - WAE: minimize objective; regularization term: $Q(\theta)$ matches to P_θ where $Q(\theta)$ is the aggregated posterior, i.e., $Q(\theta) \coloneqq \mathrm{E}_{P_{\mathbf{w}}}\left(Q(\theta|\mathbf{w})\right)$

WASSERSTEIN AUTO-ENCODER VS VAE

(a) VAE (b) WAE



- VAE: minimize ELBO; regularization term: $Q(\theta|\mathbf{w})$ matches to P_θ
- WAE: minimize objective; regularization term: $Q(\theta)$ matches to P_θ where $Q(\theta)$ is the aggregated posterior, i.e., $Q(\theta) \coloneqq \mathrm{E}_{P_\mathbf{w}}\big(Q(\theta|\mathbf{w})\big)$

REGULARIZATION PENALTY OF WAE

- Two different divergences
  - GAN-based: $D_\theta(P_\theta, Q_\theta) = D_{JS}(P_\theta, Q_\theta)$
  - MMD-based: $\mathrm{MMD}_k(P_\theta, Q_\theta) = \left\| \int_\Theta k(\theta,\cdot) dP_\Theta(\theta) - \int_\Theta k(\theta,\cdot) dQ_\Theta(\theta) \right\|_{\mathcal{H}_k}$
- GAN-based:
  - Vanishing gradient problem occurs
  - The encoder fails to update for distribution matching
- MMD-based:
  - Use SGD methods thanks to un-biasness and U-statistic property
  - performs well when matching high-dimensional standard normal

WAE-GAN ALGORITHM

ALGORITHM 1 Wasserstein Auto-Encoder with GAN-based penalty (WAE-GAN).

Require: Regularization coefficient $\lambda > 0$.

Initialize the parameters of the encoder Q_ϕ, decoder G_θ, and latent discriminator D_γ.

while (ϕ, θ) not converged **do**

Sample $\{x_1, \ldots, x_n\}$ from the training set

Sample $\{z_1, \ldots, z_n\}$ from the prior P_Z

Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \ldots, n$

Update D_γ by ascending:

$$\frac{\lambda}{n} \sum_{i=1}^{n} \log D_\gamma(z_i) + \log\big(1 - D_\gamma(\tilde{z}_i)\big)$$

Update Q_ϕ and G_θ by descending:

$$\frac{1}{n} \sum_{i=1}^{n} c\big(x_i, G_\theta(\tilde{z}_i)\big) - \lambda \cdot \log D_\gamma(\tilde{z}_i)$$

end while

$$\min_G \max_D V(D,G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] .$$

$$\min_G \max_D V(D,G) = \qquad + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] .$$

WAE-MMD ALGORITHM

ALGORITHM 2 Wasserstein Auto-Encoder with MMD-based penalty (WAE-MMD).

Require: Regularization coefficient $\lambda > 0$, characteristic positive-definite kernel k.

Initialize the parameters of the encoder Q_ϕ, decoder G_θ, and latent discriminator D_γ.

while (ϕ, θ) not converged **do**

Sample $\{x_1, \dots, x_n\}$ from the training set

Sample $\{z_1, \dots, z_n\}$ from the prior P_Z

Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \dots, n$

Update Q_ϕ and G_θ by descending:

$$\frac{1}{n}\sum_{i=1}^{n} c\big(x_i, G_\theta(\tilde{z}_i)\big) + \frac{\lambda}{n(n-1)}\sum_{\ell \neq j} k(z_\ell, z_j) + \frac{\lambda}{n(n-1)}\sum_{\ell \neq j} k(\tilde{z}_\ell, \tilde{z}_j) - \frac{2\lambda}{n^2}\sum_{\ell, j} k(z_\ell, \tilde{z}_j)$$

end while

$$\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, y_j).$$

ENCODED OUTPUT VS PRIOR: 2D DIRICHLET

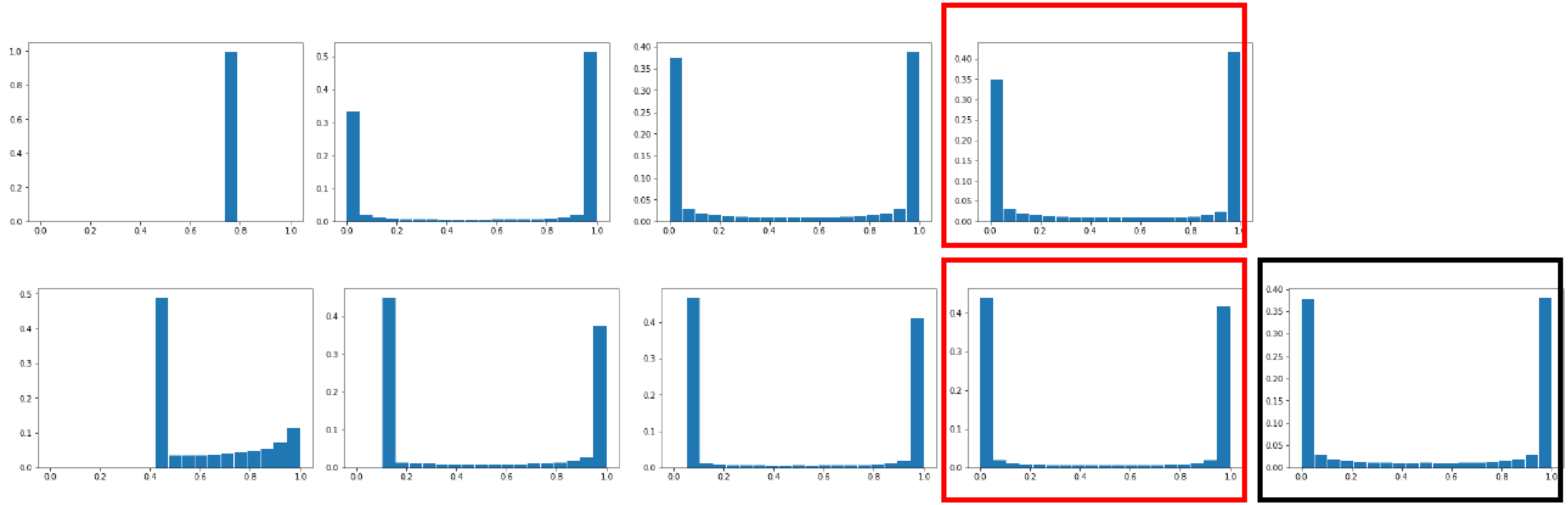


Figure 4: Histogram for the encoded latent distribution over epochs. First row corresponds to epochs 0, 10, 20 and 50 of GAN training; second row corresponds to epochs 0, 10, 20 and 50 of MMD training; the right most figure on the second row corresponds to the histogram of the prior distribution: 2D Dirichlet of parameter 0.1

ENCODED OUTPUT VS PRIOR: 50D DIRICHLET

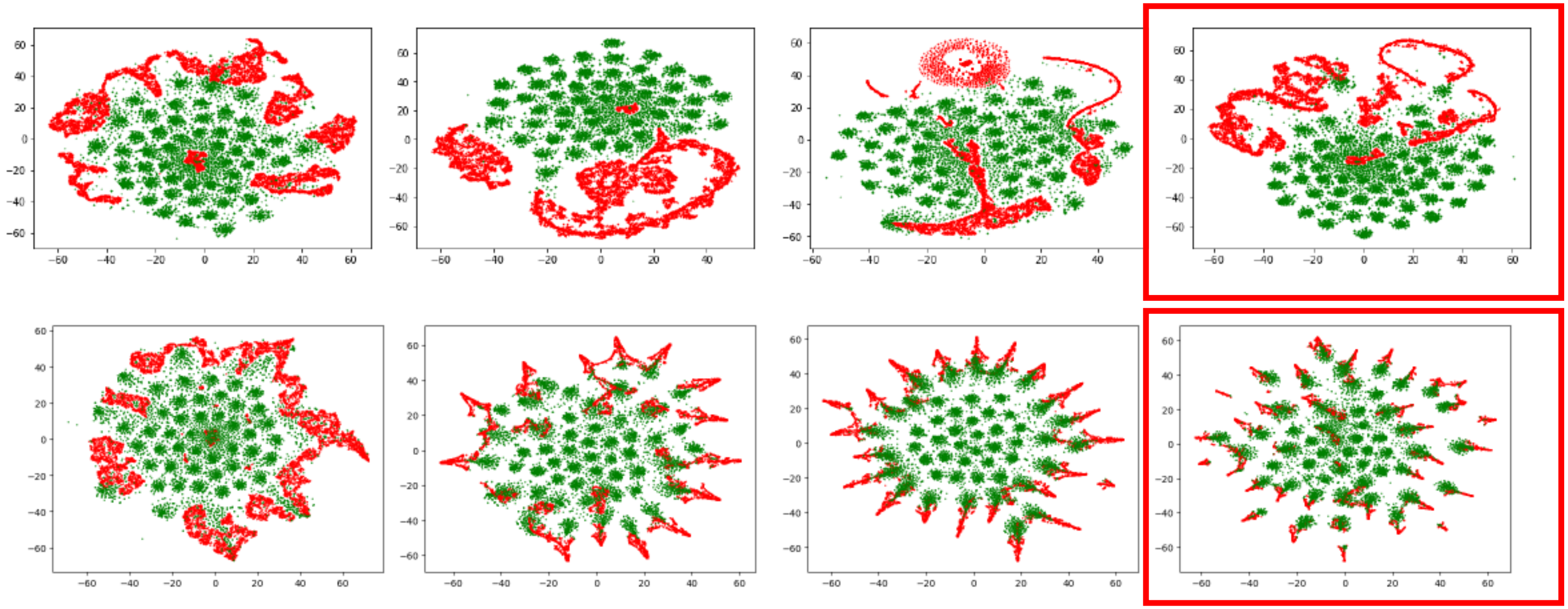


Figure 5: t-SNE plot of encoder output vectors (red) and samples from the Dirichlet prior (green) over epochs. First row corresponds to epochs 0,10,30,99 of GAN training; second row corresponds to those of MMD training

THEORETICAL PART OF WAE: DETERMINISTIC

Theorem 1 *For P_G as defined above with deterministic $P_G(X|Z)$ and any function $G\colon \mathcal{Z} \to \mathcal{X}$*

$$\inf_{\Gamma \in \mathcal{P}(X\sim P_X, Y\sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}\big[c\big(X,Y\big)\big] = \inf_{Q\colon Q_Z = P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X,G(Z)\big)\big],$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

Lemma 1 $\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X, P_G)$ *with identity if*[6] $P_G(Y|Z=z)$ *are Dirac for all* $z \in \mathcal{Z}$.

Proof The first assertion is obvious. To prove the identity, note that when Y is a deterministic function of Z, for any A in the sigma-algebra induced by Y we have $\mathbb{E}\big[\mathbb{1}_{[Y\in A]}|X,Z\big] = \mathbb{E}\big[\mathbb{1}_{[Y\in A]}|Z\big]$. This implies $(Y \perp\!\!\!\perp X)|Z$ and concludes the proof. ∎

$$\begin{aligned}
W_c^\dagger(P_X,P_G) &= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{(X,Y,Z)\sim P}\big[c(X,Y)\big] \\
&= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\mathbb{E}_{Y\sim P(Y|Z)}\big[c(X,Y)\big] \\
&= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\big[c\big(X,G(Z)\big)\big] \\
&= \inf_{P\in\mathcal{P}_{X,Z}} \mathbb{E}_{(X,Z)\sim P}\big[c\big(X,G(Z)\big)\big].
\end{aligned}$$

MAXIMUM MEAN DISCREPANCY

- Motivation of MMD

Problem 1 *Let x and y be random variables defined on a topological space $\mathcal{X}$, with respective Borel probability measures p and q . Given observations $X := \{x_1, \dots, x_m\}$ and $Y := \{y_1, \dots, y_n\}$, independently and identically distributed (i.i.d.) from p and q, respectively, can we decide whether $p \neq q$?*

Definition 2 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and let p, q, x, y, X, Y be defined as above. We define the maximum mean discrepancy (MMD) as*

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)] \right). \tag{1}$$

RKHS AND RIESZ THEOREM

- RKHS (Reproducing Kernel Hilbert Space)

  We say that a Hilbert space of real-valued function of $\mathcal{X}$, $\mathcal{H}$ is a **reproducing kernel Hilbert space** if, for all $x \in \mathcal{X}$, L_x is continuous at any $f \in \mathcal{H}$ or equivalently, if L_x is a bounded operator on $\mathcal{H}$, i.e. there exists an M such that,

$$|L_x(f)| = |f(x)| \leq M||f||_{\mathcal{H}} \;\; \forall f \in \mathcal{H}$$

- Riesz Representation THM

  Let $\mathcal{H}$ be a Hilbert space over $\mathbb{R}$. If $T \in \mathcal{H}^*$, then there exists a unique vector u in $\mathcal{H}$ such that

$$T(v) = \langle v, u \rangle_{\mathcal{H}} \text{ for all } v \in \mathcal{H}$$

$$L_x(f) = \langle f, K_x \rangle_{\mathcal{H}} \text{ for all } f \in \mathcal{H}$$

$$K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}} = K_y(x) = K_x(y)$$

Existence of mean Embedding

$$\text{Reproducing kernel Hilbert space: } \mathcal{H} \xrightarrow{\text{Riesz representation thm}} \text{Symmetric, positive-definite kernel: } K$$

$$\text{Reproducing kernel Hilbert space: } \mathcal{H} \xleftarrow{\text{Moore–Aronszajn thm}} \text{Symmetric, positive-definite kernel: } K$$

- $\exists \phi(x)$ such that $f(x) =< f, \phi(x) >_{\mathcal{H}}$ where $\phi(x) = k(x,\cdot)$

- Mean embedding: $\mu_p \in \mathcal{H}, s.t. E_x f =< f, \mu_p >_{\mathcal{H}}$

Lemma 3 *If $k(\cdot,\cdot)$ is measurable and $\mathbf{E}_x \sqrt{k(x,x)} < \infty$ then $\mu_p \in \mathcal{H}$.*

Proof The linear operator $T_p f := \mathbf{E}_x f$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$|T_p f| = |\mathbf{E}_x f| \leq \mathbf{E}_x |f| = \mathbf{E}_x |\langle f, \phi(x) \rangle_{\mathcal{H}}| \leq \mathbf{E}_x \left(\sqrt{k(x,x)} \|f\|_{\mathcal{H}} \right).$$

Hence by the Riesz representer theorem, there exists a $\mu_p \in \mathcal{H}$ such that $T_p f = \langle f, \mu_p \rangle_{\mathcal{H}}$. If we set $f = \phi(t) = k(t,\cdot)$, we obtain $\boxed{\mu_p(t) = \langle \mu_p, k(t,\cdot) \rangle_{\mathcal{H}} = \mathbf{E}_x k(t,x)}$: in other words, the mean embedding of the distribution p is the expectation under p of the canonical feature map. ■

RE-WRITTEN MMD

- MMD in terms of mean embeddings

Lemma 4 *Assume the condition in Lemma 3 for the existence of the mean embeddings μ_p, μ_q is satisfied. Then*

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \left\|\mu_p - \mu_q\right\|_{\mathcal{H}}^2.$$

Proof

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{F}, p, q] &= \left[\sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbf{E}_x\left[f(x)\right] - \mathbf{E}_y\left[f(y)\right]\right)\right]^2 \\
&= \left[\sup_{\|f\|_{\mathcal{H}} \leq 1} \left\langle \mu_p - \mu_q, f\right\rangle_{\mathcal{H}}\right]^2 \\
&= \left\|\mu_p - \mu_q\right\|_{\mathcal{H}}^2.
\end{aligned}$$

CALCULATION OF MMD WITH SAMPLES

- The condition of MMD vanishing

Theorem 5 *Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathcal{X}$, with associated continuous kernel $k(\cdot,\cdot)$. Then* $\mathrm{MMD}[\mathcal{F},p,q]=0$ *if and only if $p=q$.*

- Mean embedding: $\mu_p \in \mathcal{H}, s.t. E_x f = < f, \mu_p >_{\mathcal{H}}$

Lemma 6 *Given x and x' independent random variables with distribution p, and y and y' independent random variables with distribution q, the squared population* MMD *is*

$$\mathrm{MMD}^2[\mathcal{F},p,q] = \mathbf{E}_{x,x'}\left[k(x,x')\right] - 2\mathbf{E}_{x,y}\left[k(x,y)\right] + \mathbf{E}_{y,y'}\left[k(y,y')\right],$$

Proof Starting from the expression for $\mathrm{MMD}^2[\mathcal{F},p,q]$ in Lemma 4,

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{F},p,q] &= \left\|\mu_p-\mu_q\right\|_{\mathcal{H}}^2\\ &= \langle\mu_p,\mu_p\rangle_{\mathcal{H}} + \langle\mu_q,\mu_q\rangle_{\mathcal{H}} - 2\langle\mu_p,\mu_q\rangle_{\mathcal{H}}\\ &= \mathbf{E}_{x,x'}\langle\phi(x),\phi(x')\rangle_{\mathcal{H}} + \mathbf{E}_{y,y'}\langle\phi(y),\phi(y')\rangle_{\mathcal{H}} - 2\mathbf{E}_{x,y}\langle\phi(x),\phi(y)\rangle_{\mathcal{H}},\end{aligned}$$

$$\widehat{\mathrm{MMD}}_{\mathbf{k}}(Q_\Theta,P_\Theta) = \frac{1}{m(m-1)}\sum_{i\neq j}\mathbf{k}(\theta_i,\theta_j) + \frac{1}{m(m-1)}\sum_{i\neq j}\mathbf{k}(\theta'_i,\theta'_j) - \frac{2}{m^2}\sum_{i,j}\mathbf{k}(\theta_i,\theta'_j).$$

Some examples: If $f(x)=x$ the U-statistic $f_n(x)=\bar{x}_n=(x_1+\cdots+x_n)/n$ is the sample mean.

If $f(x_1,x_2)=|x_1-x_2|$, the U-statistic is the mean pairwise deviation $f_n(x_1,\ldots,x_n)=2/(n(n-1))\sum_{i>j}|x_i-x_j|$, defined for $n\geq 2$.

Measures: Topic Diversity and Coherence

- NMPI (normalized point mutual information) : topic coherence
  - $\text{NMPI}(x, y) \coloneqq \log \frac{p(x,y)}{p(x)p(y)} / -\log p(x, y) \in [-1,1]$
  - NMPI = $+1$ (-1) complete (no) co-occurrences, NMPI = 0: independent

- TU (topic uniqueness) : topic diversity
  - $\text{TU}(k) \coloneqq \frac{1}{L} \sum_{l=1}^{L} \frac{1}{\text{count}(l,k)}$: TU of kth topic,
  - $\text{count}(l, k)$ is the total number of times the lth top word in topic k appears in the top words across all topics
  - The higher TU value, the more diverse topics be produced

Contents

- Backgrounds on topic modeling
  - Latent Dirichlet Allocation (LDA)
  - Wasserstein Latent Dirichlet Allocation (W-LDA)
- Term-project
  - Simulation results
- Discussion

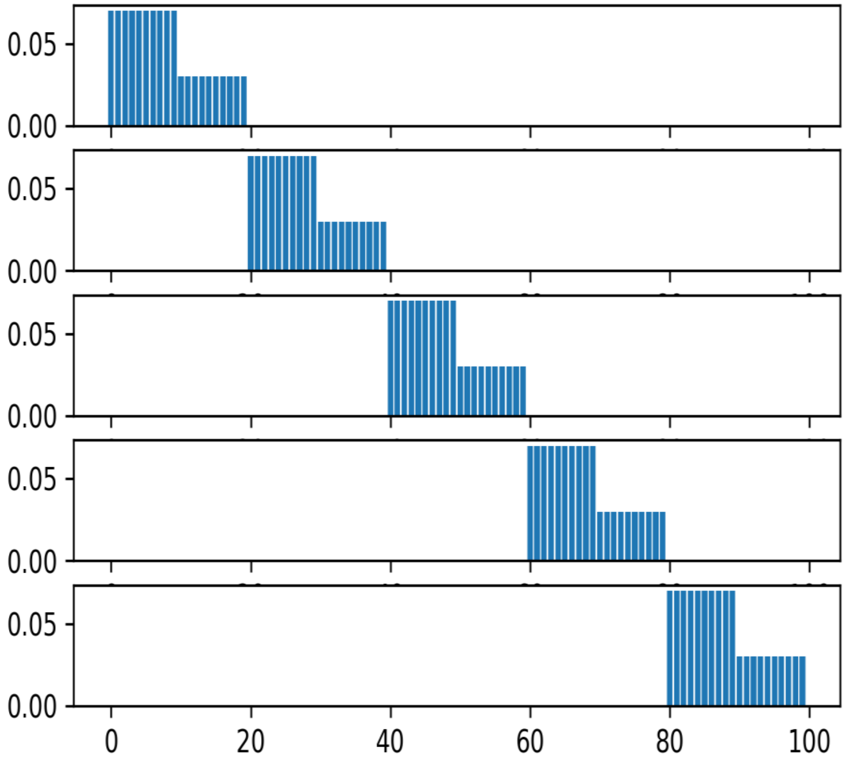
CONTENTS

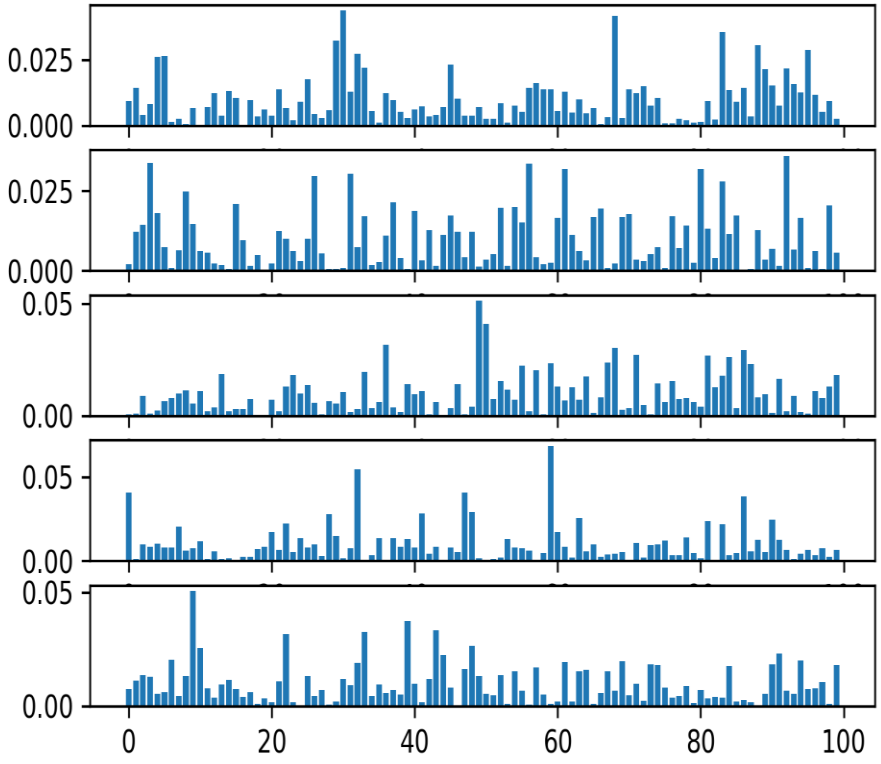
TERM-PROJECT: SIMULATION SETTINGS

- Model
  - Wasserstein Auto-Encoder (used softmax activation)
  - Modified W-LDA (non-negativity constrain and relu activation)
- Simulation parameters
  - Corpus: not real, made by Tensorflow keras
  - Beta: Two types – simple(0.7 and 0.3) and complex(from exponential)

TERM-PROJECT: SIMULATION RESULTS

- W-LDA implementation with tensorflow keras
- Application on generated corpus: simple and complex





TERM-PROJECT: SIMULATION RESULTS

- W-LDA implementation on paper
- Can’t show beta

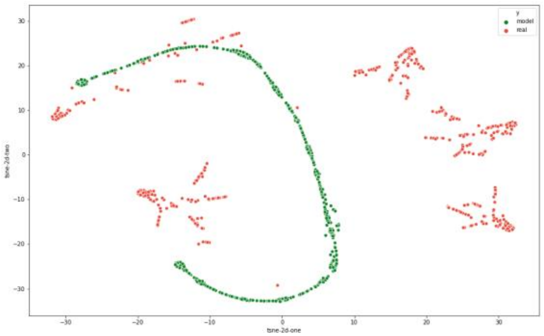
```
topic 1
real:   [0 1 2 3 4 5 6 7 8 9]
model:  [8 5 2 7 1 6 0 9 4 3]
topic 2
real:   [24 27 25 28 23 22 21 20 29 26]
model:  [20 26 22 28 29 24 21 27 38 25]
topic 3
real:   [49 40 42 43 44 45 46 47 48 41]
model:  [42 46 44 49 48 45 43 41 40 47]
topic 4
real:   [69 60 62 63 64 65 66 67 68 61]
model:  [68 65 62 67 61 69 66 60 63 64]
topic 5
real:   [80 88 81 82 83 84 85 86 87 89]
model:  [86 83 85 82 89 88 84 80 81 87]
```

```
topic 1
real:   [30 68 83 29 88 95 32  5  4 45]
model:  [30 29 95 68 88  5  4 89 58 72]
topic 2
real:   [92  3 56 80 61 31 26 83  8 37]
model:  [80 56 92  3 31 26 61 65  8 66]
topic 3
real:   [49 50 36 68 86 71 81 84 67 59]
model:  [50 49 87 36 86 68 67 13 71 23]
topic 4
real:   [59 32 47  0 86 48 41 28 63 90]
model:  [59  0 47 28 32 86 41 20 48  7]
topic 5
real:   [ 9 39 43 33 22 48 10 91 44  6]
model:  [39  9 43  6 44 91 22 10 33 73]
```

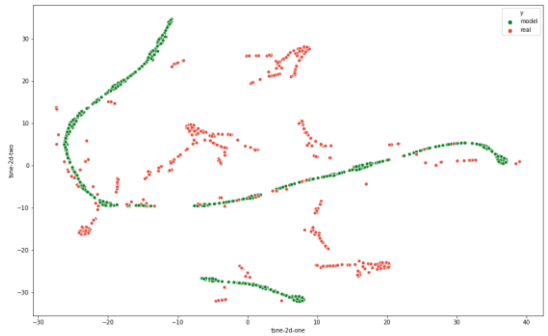
TERM-PROJECT: SIMULATION RESULTS

- T-SNE plot of how model is being educated

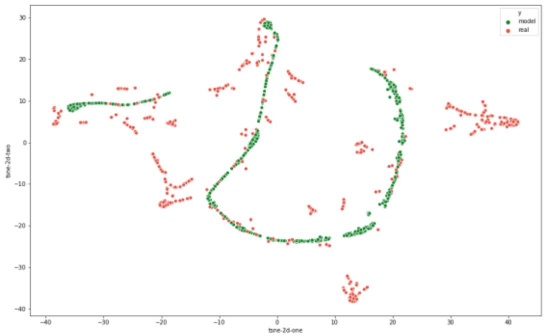
Epoch 10



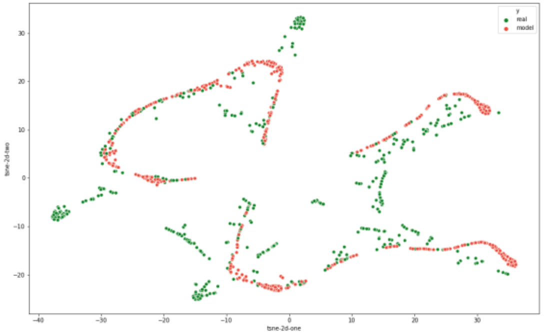
Epoch 20



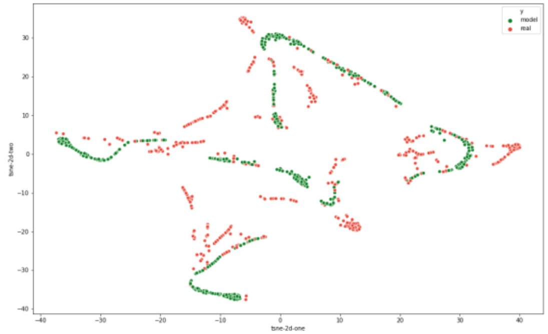
Epoch 30



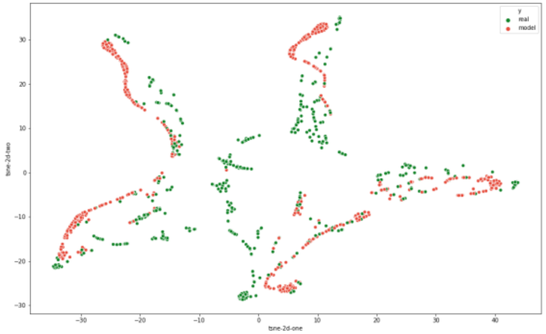
Epoch 40



Epoch 50



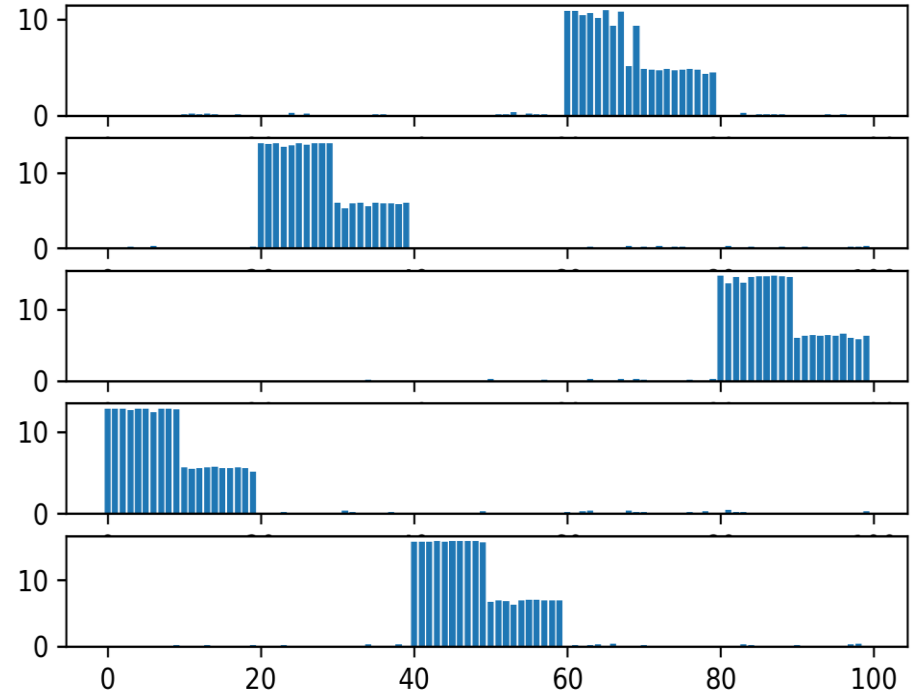
Epoch 60



Term-project: Simulation Results

- Modified W-LDA implementation – nonnegative decoder, linear activation
- application on generated corpus with simple beta

```
topic 1
real:   [0 1 2 3 4 5 6 7 8 9]
model:  [4 7 8 0 5 1 2 9 3 6]
topic 2
real:   [24 27 25 28 23 22 21 20 29 26]
model:  [29 22 28 25 27 20 21 26 24 23]
topic 3
real:   [49 40 42 43 44 45 46 47 48 41]
model:  [43 47 46 45 48 44 42 41 40 49]
topic 4
real:   [69 60 62 63 64 65 66 67 68 61]
model:  [65 61 60 67 63 62 64 66 69 68]
topic 5
real:   [80 88 81 82 83 84 85 86 87 89]
model:  [80 87 88 85 86 84 89 82 83 81]
```



TERM-PROJECT: SIMULATION RESULTS

- Modified W-LDA implementation - – nonnegative decoder, linear activation
- Better work on complicated corpus!

```
topic 1
real:   [30 68 83 29 88 95 32  5  4 45]
model:  [30 68 83 29 88 95 32  4  5 33]
topic 2
real:   [92  3 56 80 61 31 26 83  8 37]
model:  [92  3 56 80 31 83 61 26  8 15]
topic 3
real:   [49 50 36 68 86 71 81 84 67 59]
model:  [49 50 36 86 68 81 84 71 67 59]
topic 4
real:   [59 32 47  0 86 48 41 28 63 90]
model:  [59 32 47  0 86 28 41 48 90 81]
topic 5
real:   [ 9 39 43 33 22 48 10 91 44  6]
model:  [ 9 39 33 43 22 10 48 91 44  6]
```

Discussion

- Edit encoder in W-LDA
  - Deterministic property of Encoder is too strong
  - Release the condition while maintain Dirichlet condition
- Further work
  - Apply W-LDA to scRNA sequencing data
  - Apply modified W-LDA to scRNA sequencing data

REFERENCES

[1] Tolstikhin, Ilya & Bousquet, Olivier & Gelly, Sylvain & Schölkopf, Bernhard. (2017). Wasserstein Auto-Encoders.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

[3] Nan, Feng & Ding, Ran & Nallapati, Ramesh & Xiang, Bing. (2019). Topic Modeling with Wasserstein Autoencoders.

[4] Kingma, Diederik & Welling, Max. (2013). Auto-Encoding Variational Bayes. ICLR.

[5] Bousquet, Olivier & Gelly, Sylvain & Tolstikhin, Ilya & Simon-Gabriel, Carl-Johann & Schölkopf, Bernhard. (2017). From optimal transport to generative modeling: the VEGAN cookbook.

[6] Gretton, A & Borgwardt, K. & Rasch, Malte & Schölkopf, Bernhard & Smola, AJ. (2012). A Kernel Two-Sample Test. The Journal of Machine Learning Research. 13. 723-773.

[7] Michael I. Jordan, Stat260: Bayesian Modeling and Inference, Lecture 1: History and De Finetti's Theorem

[8] O. Kallenberg, Probabilistic symmetries and invariance principles, Springer, New York, 2005

[9] Ding, Chris & Li, Tao & Peng, Wei. (2006). Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence Chi-Square Statistic, and a Hybrid Method.

[10] https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/.

THANK YOU FOR YOUR LISTENING

ANY QUESTIONS OR COMMENTS

Sangheon Lee and Jaewon Bae
Email: {buaaaaang, duckgoose}@kaist.ac.kr