TOPIC MODELING & LATENT DIRICHLET ALLOCATION (LDA)

Hugh Nguyen

Data Scientist
RBC DNA - Texonomy

2021.01.29

AGENDA

1. The Problem Overview
2. The Tools
3. LDA Details
4. Discussion

THE PROBLEM

PROBLEM OVERVIEW

Topic 1

Sports

Topic 2

Economics

Topic 3

Games

Topic 4

Games

Sports

PROBLEM OVERVIEW

1. What is LDA?

2. How is LDA trained to effectively discover the hidden topics from the observed texts?

   - Gibbs Sampling

   - **Variational E-M Algorithm**

THE TOOLS

BAYES' THEOREM

Let θ be model's parameter and X be the observed data, the posterior probability distribution of θ **given** the observed data X is defined as follow:

$$P(\theta \mid X) = \frac{\text{joint probability distribution}}{\text{evidence}}$$

$$= \frac{(\text{likelihood})(\text{prior})}{\text{evidence}}$$

$$= \frac{P(X, \theta)}{P(X)}$$

$$= \frac{P(X|\theta)P(\theta)}{P(X)}$$

BAYES' THEOREM

General Conditional Probability Rules:

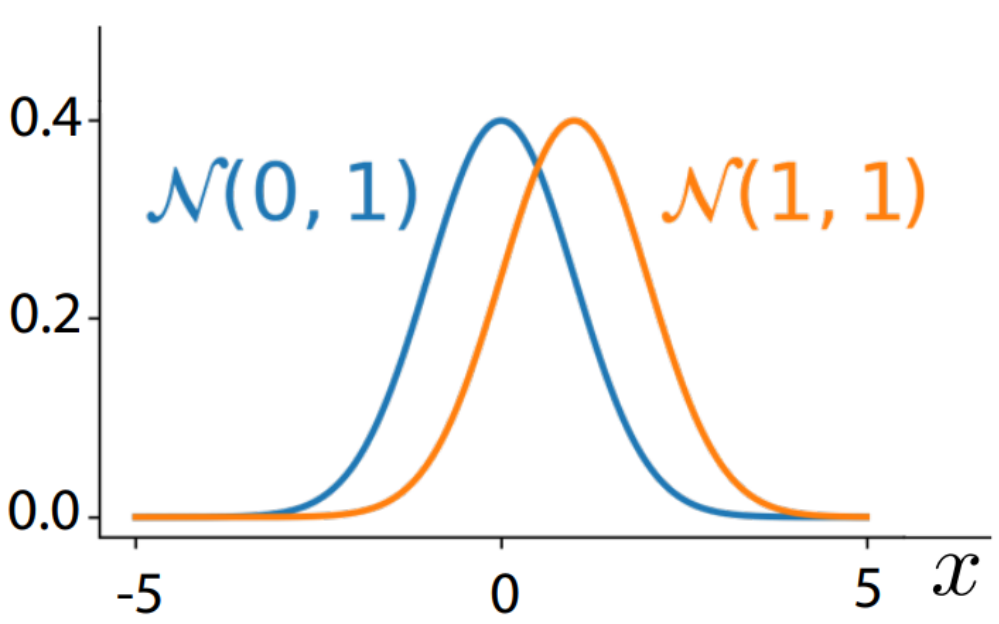
$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

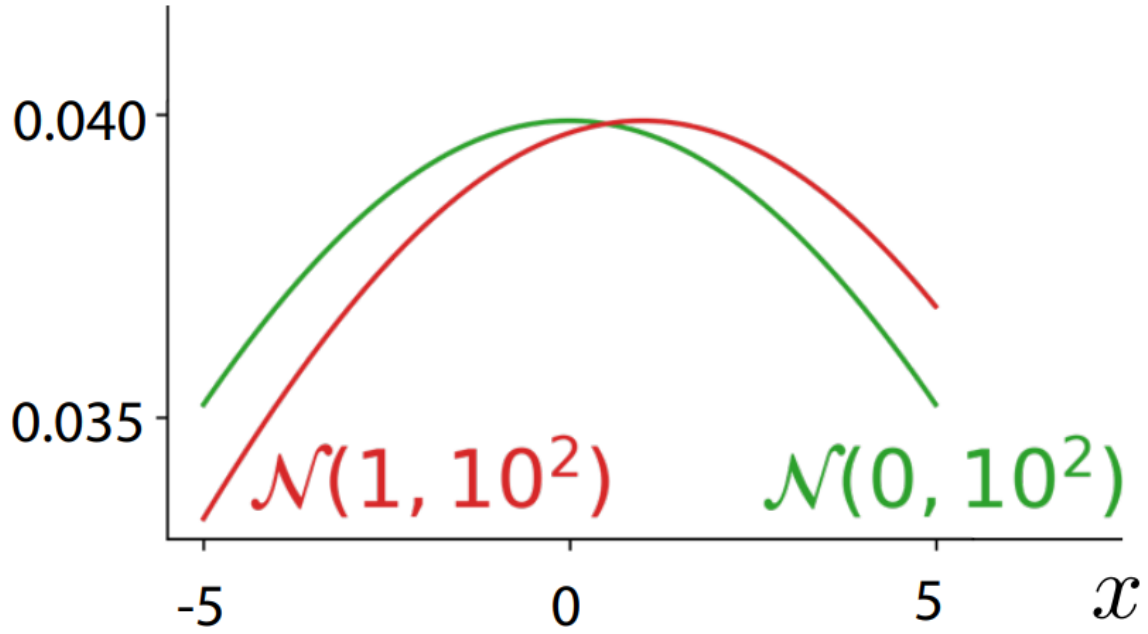
$$P(X_1,...,X_N) = \prod_{i=1}^{N} P(X_i|X1,...,X_{i-1})$$

KULLBACK LEIBLER (KL) DIVERGENCE

KL[N(0,1) || N(1,1)] = 0.5



KL[N(0,100) || N(1,100)] = 0.005



KULLBACK LEIBLER (KL) DIVERGENCE

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

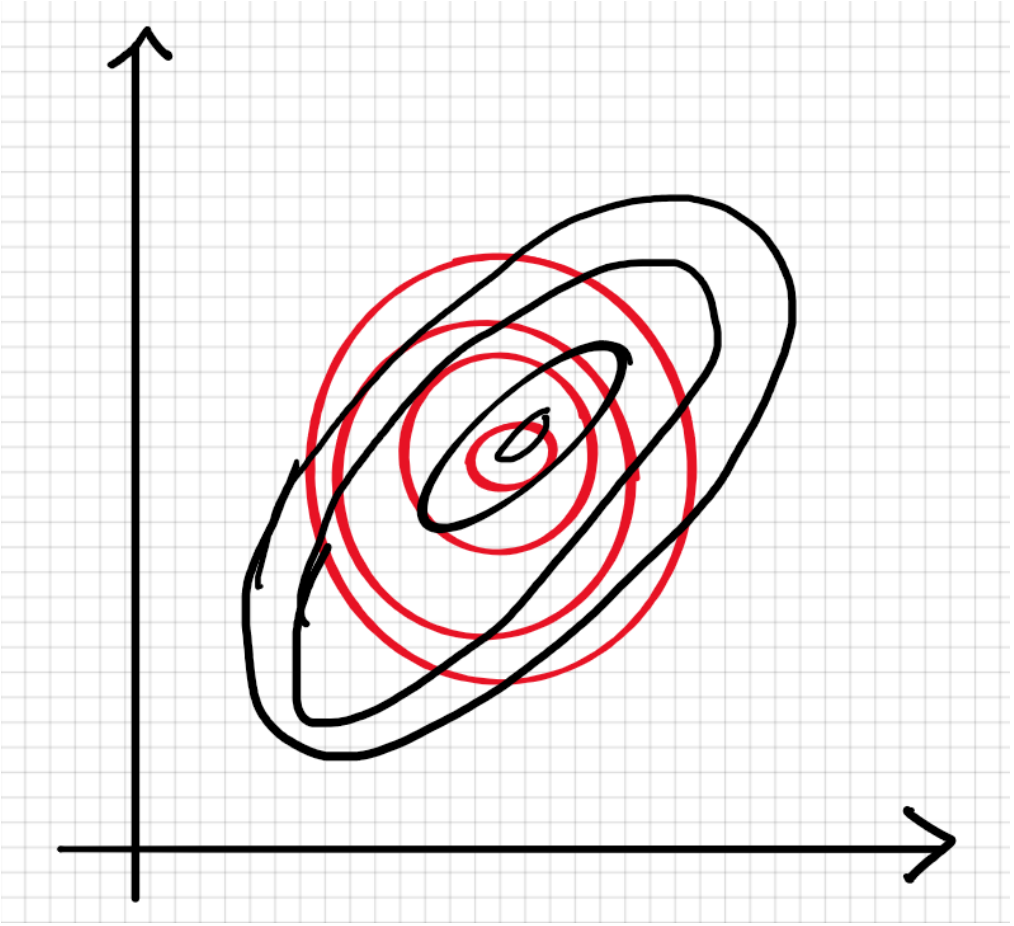
VARIATIONAL INFERENCE

Example:

$$p(Z_1, Z_2) \approx q(Z_1)q(Z_2)$$

$$p(Z_1, Z_2) \sim N(0, \Sigma)$$

$$q(Z_1)q(Z_2) \sim N(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix})$$



VARIATIONAL INFERENCE

Use to approximate the full posterior distribution

1. Select a famlily of distributions Q: $Q = \{q \,|\, q(Z) = \prod_{i=1}^{d} q_i(Z_i)\}$; i.e:

$$Q \sim N(\mu, \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \\ \vdots & \vdots & \vdots & \dots & \sigma_{d-1}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_d^2 \end{pmatrix}$$

2. Next we will try to approximate the full posterior $p(\mathbf{Z})$ with some variational distribution $q(\mathbf{Z})$:

$$\mathcal{KL}[q(Z) \,||\, p(Z)] \rightarrow min_{q \in Q}$$

VARIATIONAL INFERENCE

Then apply coordinate descend until convergence:

$$\mathcal{KL}[q(Z)\,||\,p(Z)] \rightarrow min_{q_1}$$

$$\mathcal{KL}[q(Z)\,||\,p(Z)] \rightarrow min_{q_2}$$

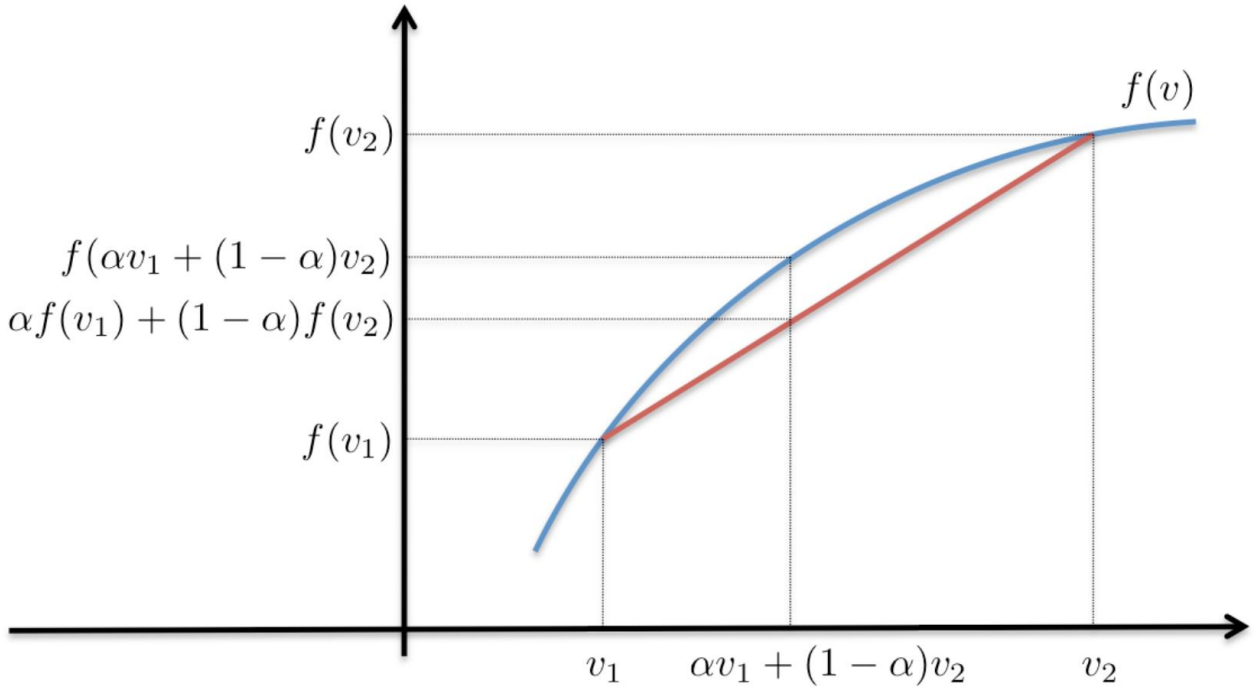
$$\dots$$

$$\mathcal{KL}[q(Z)\,||\,p(Z)] \rightarrow min_{q_d}$$

JENSEN'S INEQUALITY

If $f(x)$ is concave (i.e. $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$), then for any random variable X with corresponding probability density function (pdf) $f(X)$:

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$



E-M ALGORITHM

Use to solve parameter estimation problems, especially when some data is missing or not observable (latent variables)

- $X = \{x_1, ..., x_n\}$
- $Z|X \sim p_\theta, \theta \in \Theta$
- p_θ belongs to the exponential family (i.e. Gaussian, Beta, Dirichlet, Exponential, Gamma, etc.), which is the **primary** assumption for EM algo to work well (idk why this is so).
- Main objective: $\hat{\theta}_{EM} \in argmax_\theta \, p_\theta(X)$

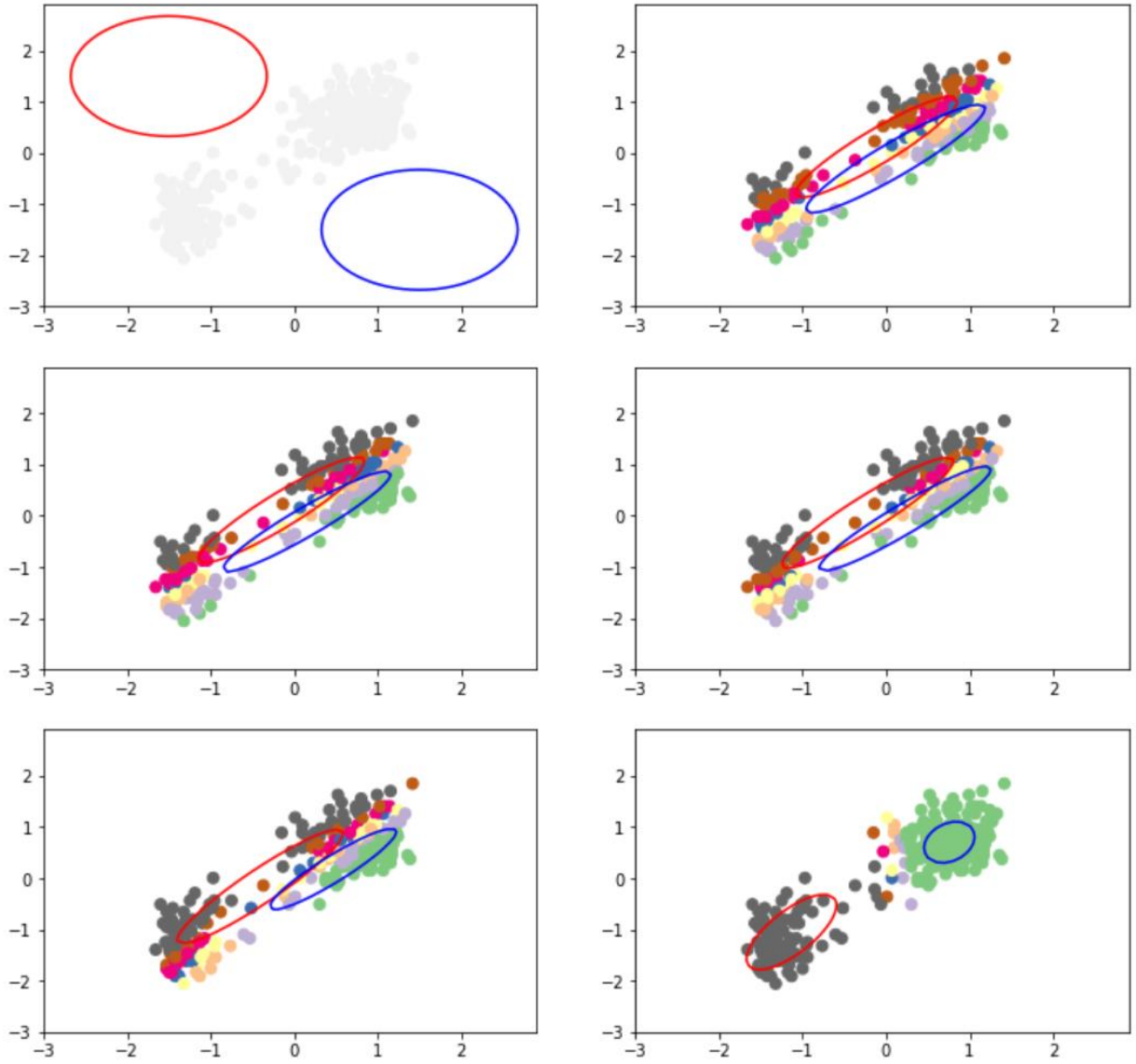
E-M ALGORTHIM

Ground truth:

- $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ = [-1,1,0.8,0.5]
- $\pi = [\pi_1, \pi_2]$ = [0.4,0.6]

EM Init:

- $\widehat{\mu_1}, \widehat{\mu_2}, \widehat{\sigma_1}, \widehat{\sigma_2}$ = [-1.5,1.5,1,1]
- $\pi = [\pi_1, \pi_2]$ = [0.5,0.5]
- Plot every 10 iterations
- Converge at the 60th iteration



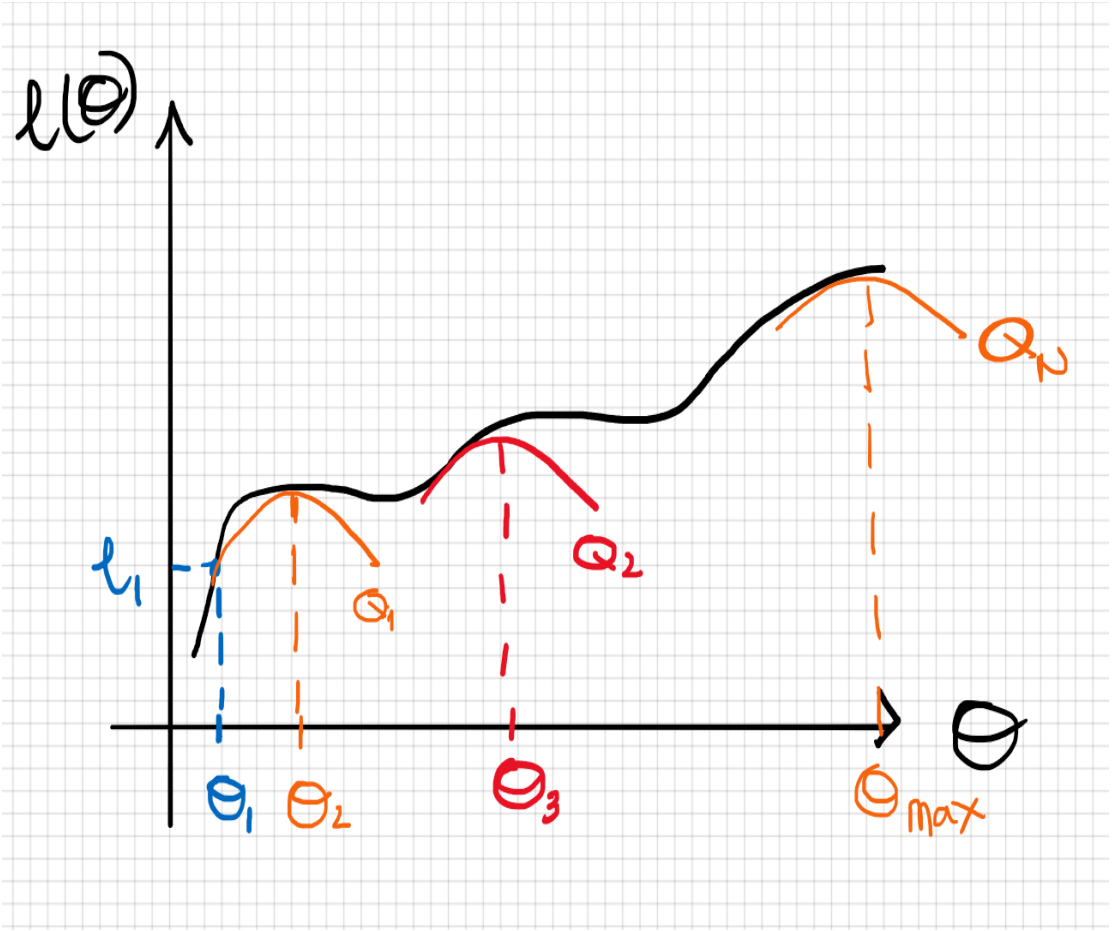
E-M ALGORTHIM

While not converge:

- E-Step: find Q_i using current θ_i
- M-step: find θ_{i+1} by max Q_i

Formally:

- the loop: for i in range(0,I)
- init $\theta_0 \in \Theta$
- E-step: $Q(\theta, \theta_i) = E_{\theta_i}[log\, p_\theta(X, Z|X)]$
- M-step: $\theta_{i+1} \in argmax_\theta Q(\theta, \theta_i)$



E-M ALGORTHIM

Pros:

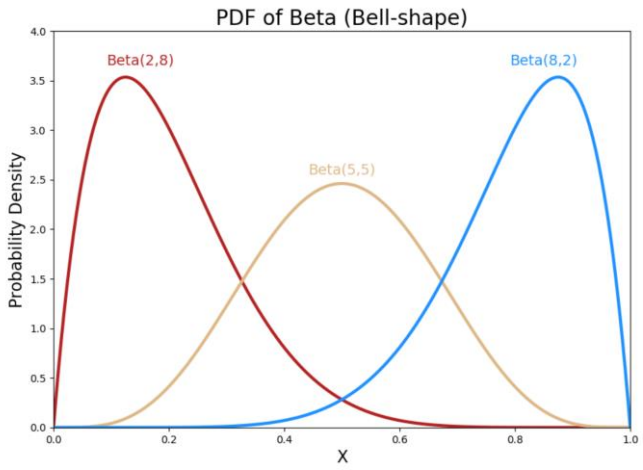
- Works well even when the derivative of the likelihood function wrt to θ is hard to compute

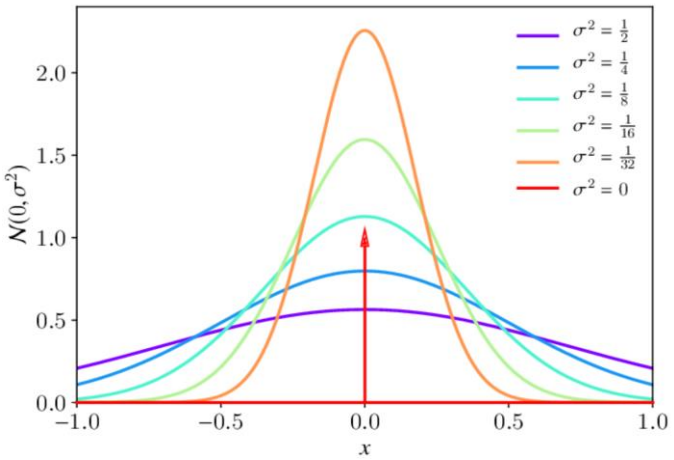
Cons:

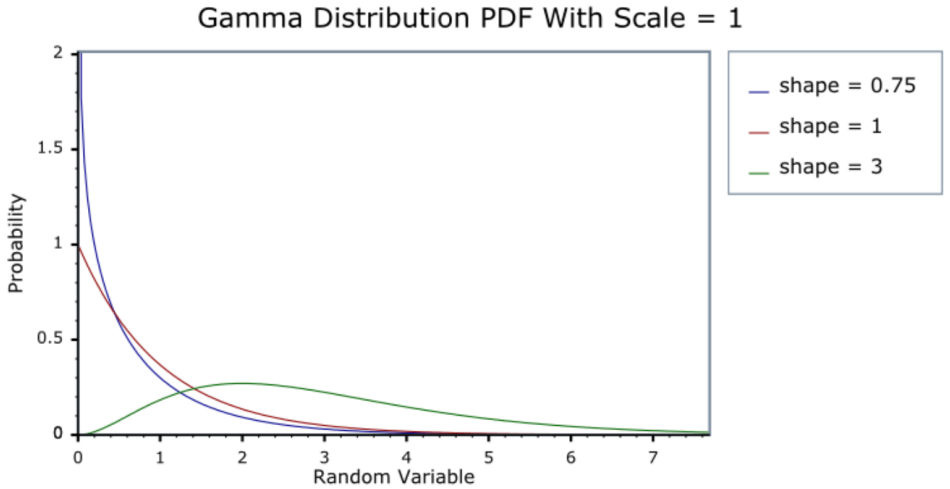
- Not guarantee to reach global maximum
- Computational expensive so convergence rate is slow
- Works well only when p_θ belong to exponential family

DIRICHLET DISTRIBUTION

Dirichlet distribution belongs to the Exponential families





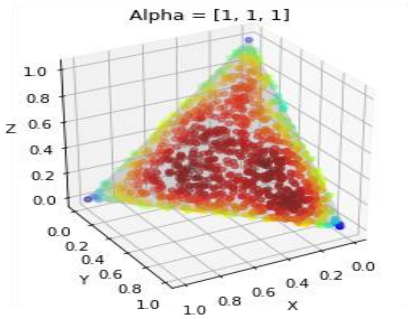


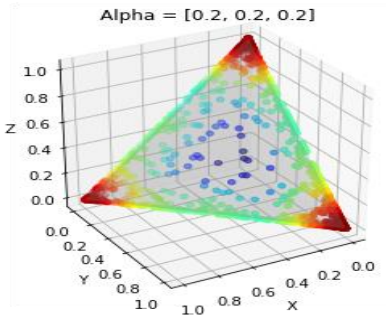
etc

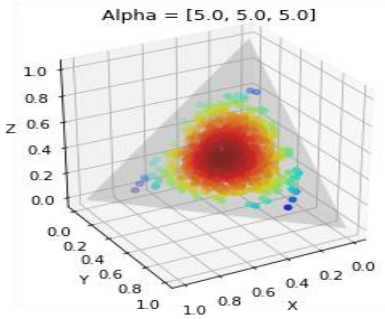
STATISTICS & GEOMETRY

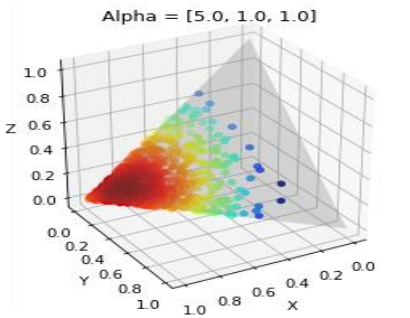
PDF: $\theta \sim \mathbf{Dir}(\theta \mid \alpha)$

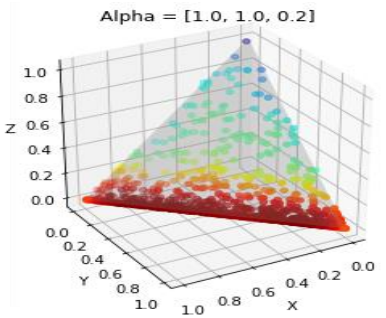
$$p(\theta \mid \alpha) = \frac{1}{C(\alpha)} \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$$

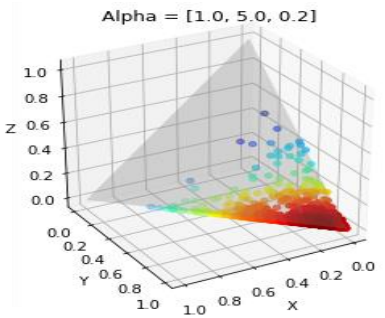










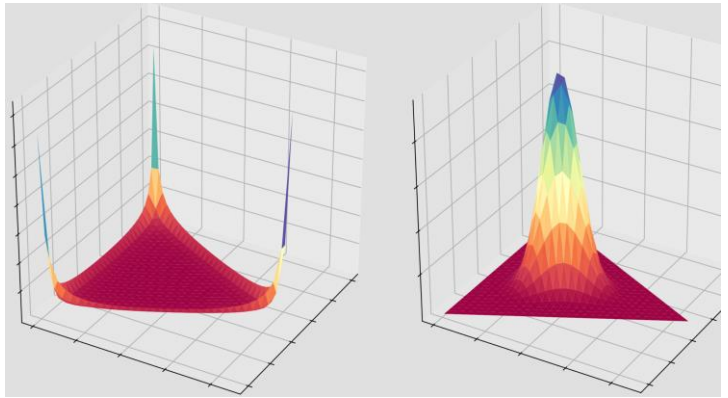


Where:

- $\theta_t \geq 0$ and $\sum_T \theta_t = 1$
- Model param $\alpha_t > 0$

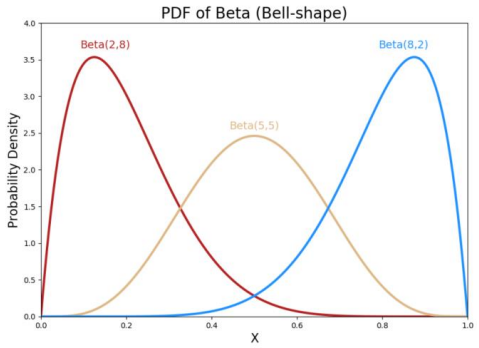
Statistics:

- Let $\alpha_0 = \sum_T \alpha_t$ where $t \in \{1, 2, ..., T\}$
- $\mathbb{E}[\theta_t] = \frac{\alpha_t}{\alpha_0}$
- $Cov(\theta_t, \theta_j) = \frac{\alpha_t \alpha_0 [i=j] - \alpha_t \alpha_j}{\alpha_0^2(\alpha_0+1)}$

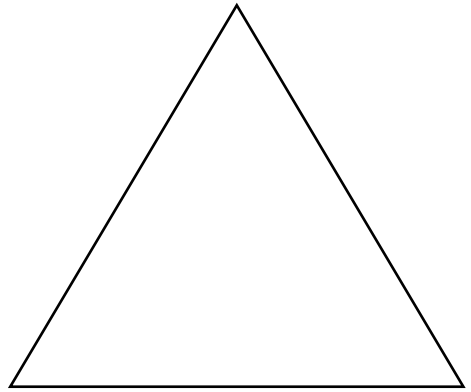


Alpha = (0.1,0.1,0.1) Alpha = (10,10,10)

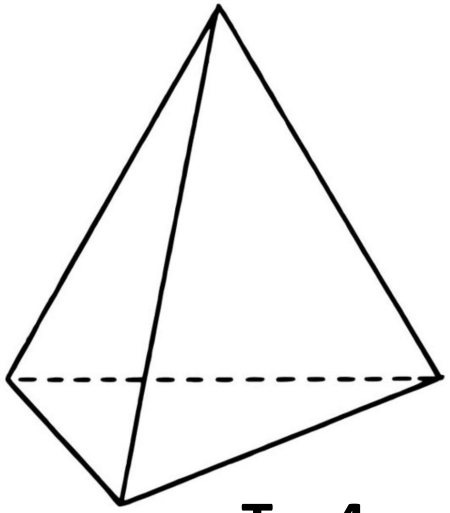
STATISTICS & GEOMETRY



T = 2



T = 3



T = 4

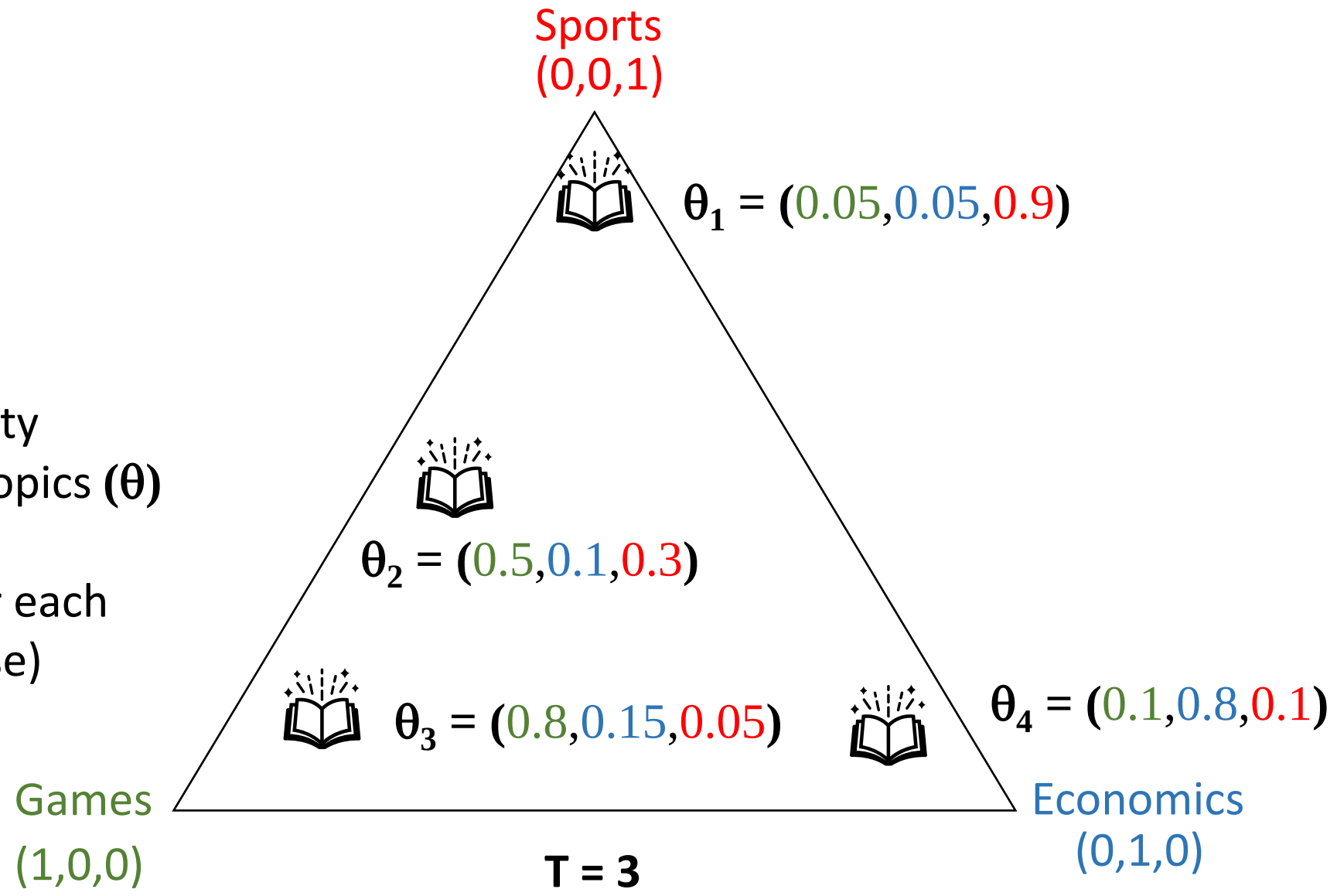
T = N --> N-dim simplex

LATENT DIRICHLET ALLOCATION (LDA)

DIR & TOPIC MODELING

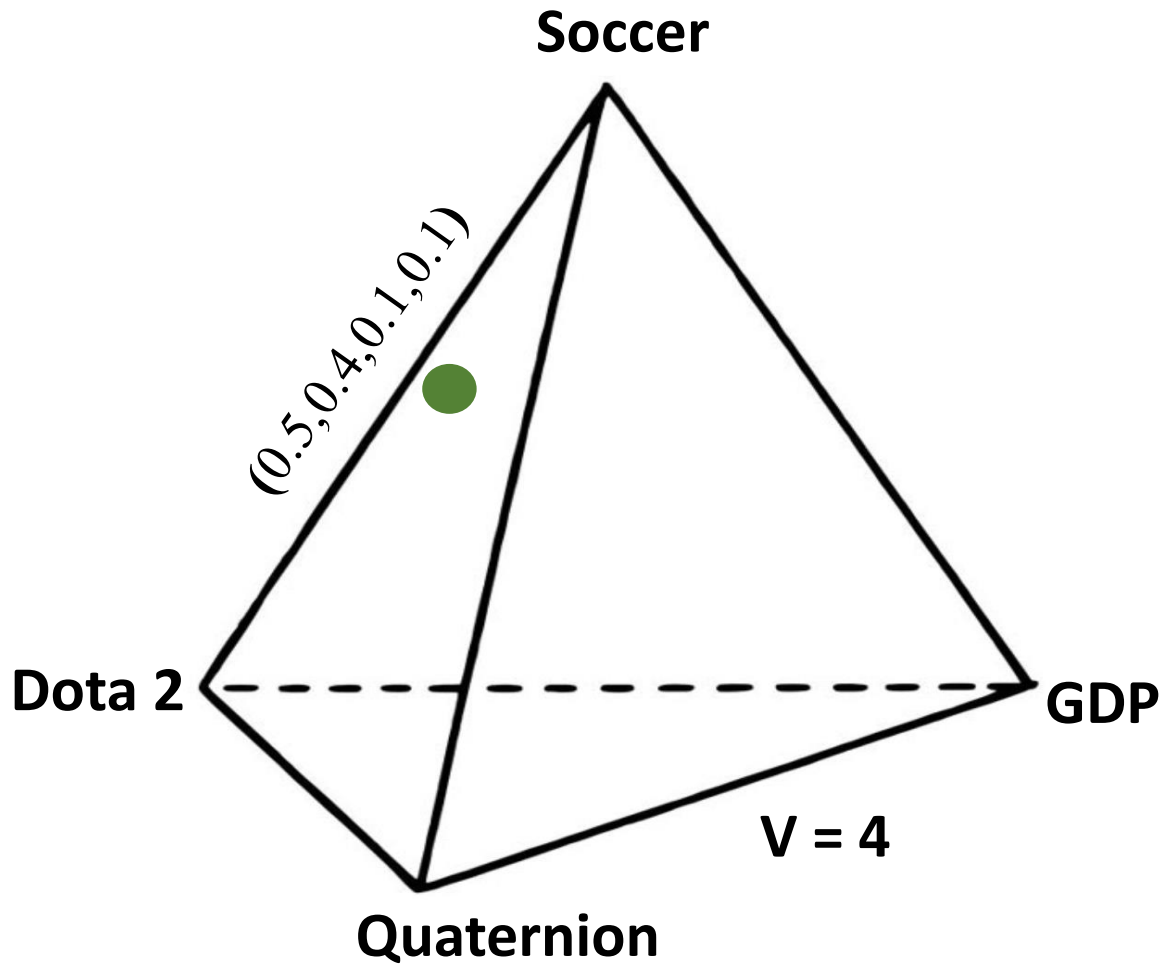
Document: a probability distribution over the topics **(θ)**

High α: more topic for each document (more sparse)



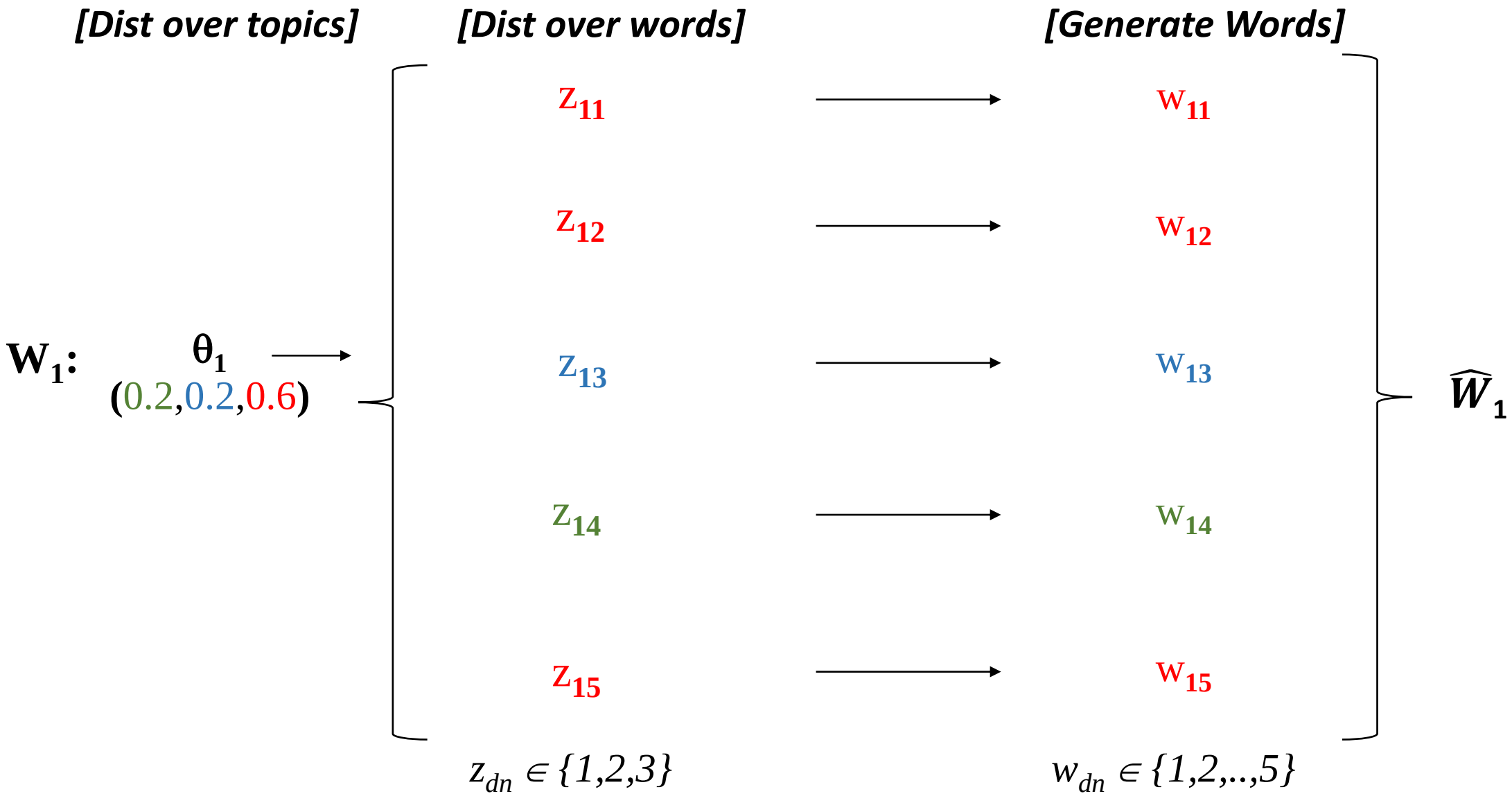
DIR & TOPIC MODELING

Topic: a probability distribution over the words

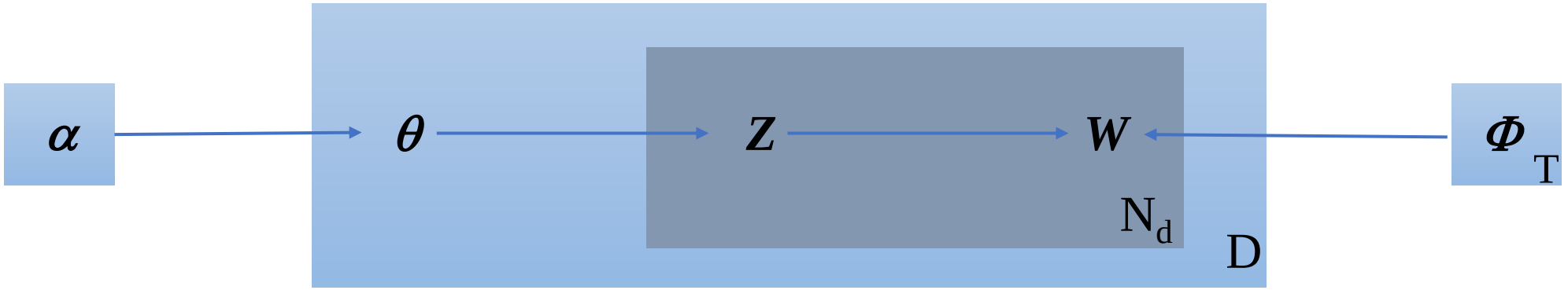


Games = 0.5(Dota2) + 0.4(Soccer) + 0.1(GDP) + 0.1(Quaternion)

GENERATIVE PROCESS



BLUEPRINT



$$p(\theta, \mathbf{Z} \mid \mathbf{W}; \alpha, \Phi) \propto p(\theta, \mathbf{Z}, \mathbf{W} \mid \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}; \phi_{tw})$$

For each d in D:

- Generate topic probability distribution of θ_d using $p(\theta_d \mid \alpha)$
- For each word in document *d,* assign a topic z_{dn} by sampling according to $p(z_{dn} \mid \theta_d)$
- For each assigned topic z_{dn}, sample a word w_{dn} according to $p(w_{dn} \mid z_{dn}; \phi_{tw})$

BLUEPRINT

$$p(\theta, \mathbf{Z} \mid \mathbf{W}; \alpha, \Phi) \propto p(\theta, \mathbf{Z}, \mathbf{W} \mid \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}; \phi_{tw})$$

- ***W:*** collection of documents, each of which has N_d words
- ***Z:*** latent variable which represents the distribution of topics over the words
- $\boldsymbol{\theta}$: latent variable which represents the distribution over the topics for each document
- $\boldsymbol{\Phi}$ **:** model parameters which represents the distribution over the words for each selected topic – i.e a word probability matrix for each topic (row) and each word (column)

BLUEPRINT

$$p(\theta, \mathbf{Z} \mid \mathbf{W}; \alpha, \Phi) \propto p(\theta, \mathbf{Z}, \mathbf{W} \mid \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}; \phi_{tw})$$

θ_d: the probability distribution over topics for document d

w_{dn}: the nth word in document d; $w_{dn} \in \{1, ..., N_d\}$

z_{dn}: topic of the nth word in document d; $z_{dn} \in \{1, ..., T\}$

$p(\theta_d) \sim Dir(\theta_d|\alpha) = \frac{1}{C(\alpha)} \prod_{t=1}^{T} \theta_{dt}^{\alpha_t - 1}$

$p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$

$p(w_{dn}|z_{dn}) = \Phi_{z_{dn} w_{dn}}$:

Where: $\phi_{tw} \geq 0$ and $\sum_{w} \phi_{tw} = 1$

LOG JOINT LIKELIHOOD FUNCTION

$$p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}; \phi)$$

$$\iff p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi) = \prod_{d=1}^{D} \{ \frac{1}{C(\alpha)} \prod_{t=1}^{T} \theta_{dt}^{\alpha_t - 1} \} \prod_{n=1}^{N_d} \{ \prod_{t=1}^{T} [z_{dn} = t] \theta_{dt} \phi_{t w_{dn}} \} \quad \text{(Substitution)}$$

$$\Rightarrow log(p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi)) = \sum_{d=1}^{D} [\sum_{t=1}^{T} (\alpha_t - 1) log(\theta_{dt}) + \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] (log\theta_{dt} + log\phi_{t w_{dn}})] + const \quad \text{(Take log both side)}$$

OPTIMIZATION & INFERENCE

MAXIMIZE LIKELIHOOD FUNCTION

$$log(p(\mathbf{W} \mid \alpha, \Phi)) = log \sum_{\mathbf{Z},\theta} p(\mathbf{W}, Z, \theta \mid \alpha, \Phi) \quad \text{(Marginalize over latent variables)}$$

$$= log \sum_{\mathbf{Z},\theta} \frac{q(Z)q(\theta)}{q(Z)q(\theta)} p(W, Z, \theta \mid \alpha, \Phi) \quad \text{(multiply by 1)}$$

$$= log \sum_{\mathbf{Z},\theta} q(Z)q(\theta) \frac{p(W, Z, \theta \mid \Phi)}{q(Z)q(\theta)}$$

$$= log\, \mathbb{E}_{q(\mathbf{Z})q(\theta)} [\frac{p(W, Z, \theta \mid \alpha, \Phi)}{q(Z)q(\theta)}]$$

$$\geq \mathcal{L}(q, \alpha, \Phi) \quad \text{(Jensen's inequality)}$$

$$= \mathbb{E}_{q(\theta)q(\mathbf{Z})} log [\frac{p(W, Z, \theta \mid \alpha, \Phi)}{q(Z)q(\theta)}]$$

$$\Rightarrow log(p(\mathbf{W} \mid \alpha, \Phi)) = \mathbb{E}_{q(\theta)q(\mathbf{Z})} log [\frac{p(W, Z, \theta \mid \alpha, \Phi)}{q(Z)q(\theta)}] + C$$

MAXIMIZE LIKELIHOOD FUNCTION

There is a very neat proof for the following:

$$C = log(p(\mathbf{W} \mid \alpha, \Phi)) - \mathbb{E}_{q(\theta)q(\mathbf{Z})} log[\frac{p(\mathbf{W}, \mathbf{Z}, \theta \mid \alpha, \Phi)}{q(\mathbf{Z})q(\theta)}]$$

$$= \mathcal{KL}(q(\theta)q(\mathbf{Z}) \,||\, p(\theta, \mathbf{Z}|\mathbf{W}))$$

Hence , we now have:

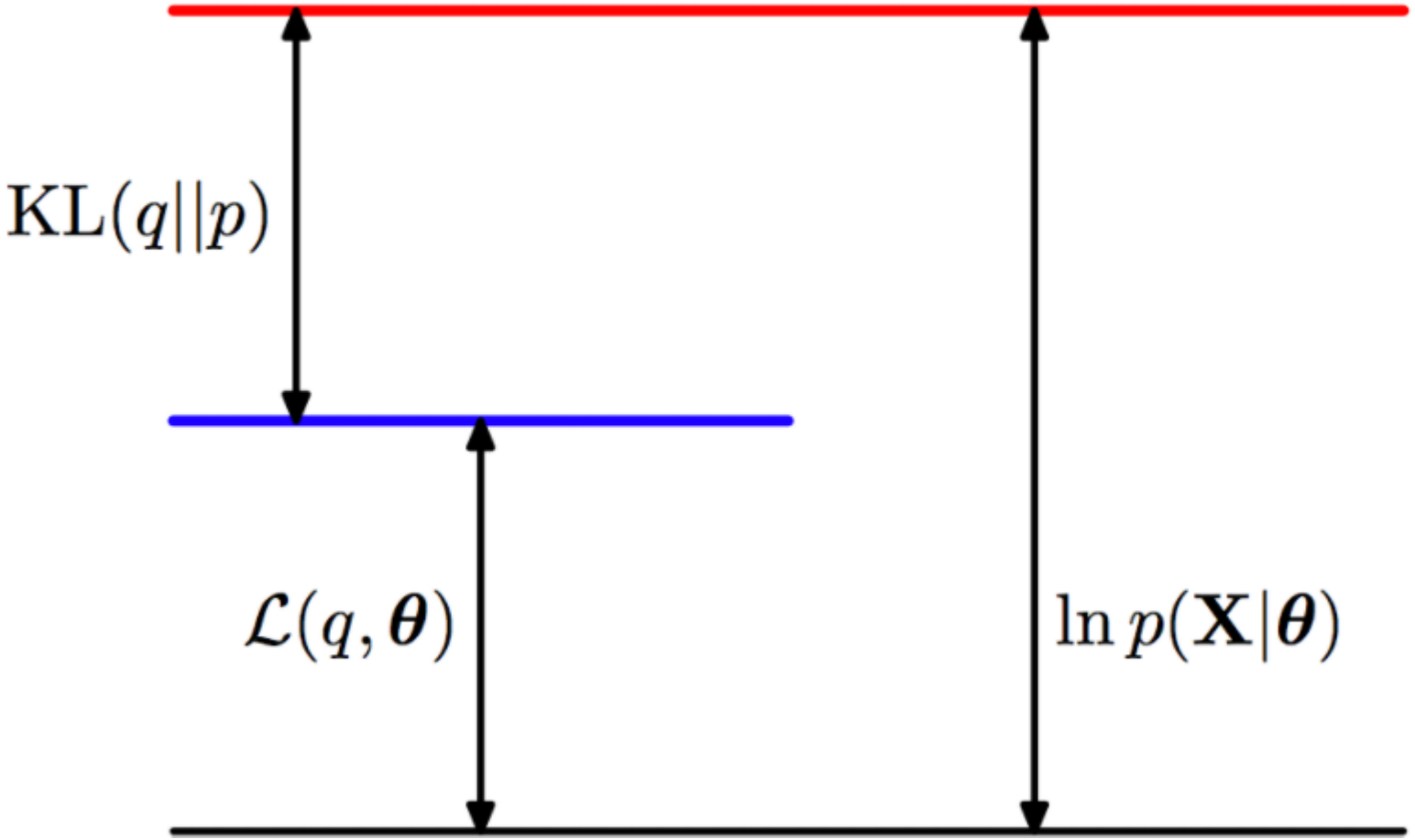
$$log(p(\mathbf{W} \mid \alpha, \Phi)) = \mathbb{E}_{q(\theta)q(\mathbf{Z})} log[\frac{p(\mathbf{W}, \mathbf{Z}, \theta \mid \alpha, \Phi)}{q(\mathbf{Z})q(\theta)}] + \mathcal{KL}(q(\theta)q(\mathbf{Z}) \,||\, p(\theta, \mathbf{Z}|\mathbf{W}))$$

$$= \mathcal{L}(q, \alpha, \Phi) + \mathcal{KL}(q(\theta)q(\mathbf{Z}) \,||\, p(\theta, \mathbf{Z}|\mathbf{W}))$$

Where:

- $q(\mathbf{Z})$ and $q(\theta)$ are some known lower bound distributions over the latent variables, i.e. **variational Gaussian distribution**

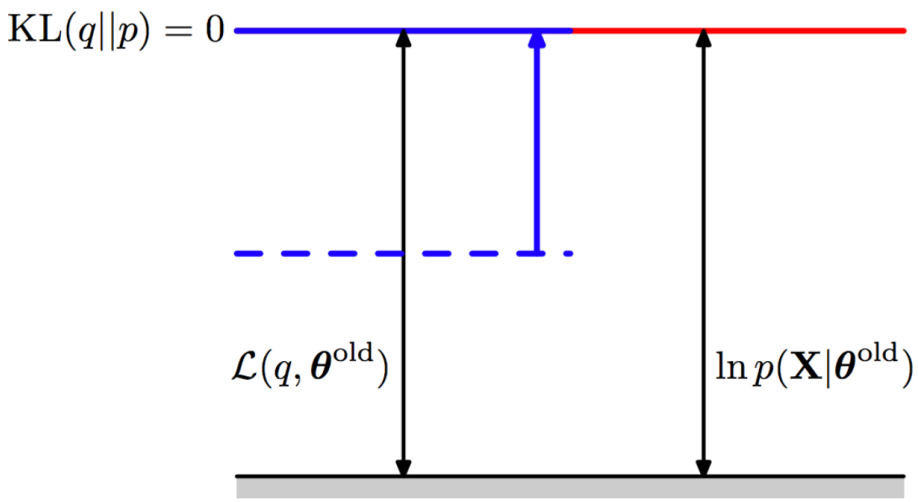
MAXIMIZE LIKELIHOOD FUNCTION



E-STEP

E-Step: θ and $\mathbf{Z}$

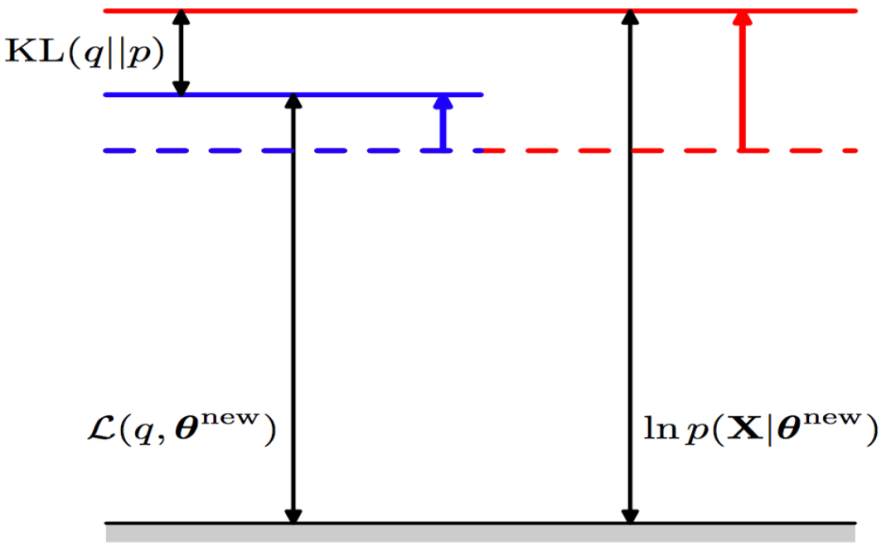
Objective: $\mathcal{KL}(q(\mathbf{Z})q(\theta)\,||\,p(\theta,\mathbf{Z}|\mathbf{W})) \rightarrow \min_{q(\theta)q(\mathbf{Z})}$



M-STEP

Objective: $\mathbb{E}_{q(\theta)q(z)}[log(p(W, Z, \theta))] \rightarrow \max_{q(\phi)}$ such that:

- since ϕ_{tw} is a probability: $\phi_{tw} \geq 0\,, \forall t, w$. This is already been satisfied because it is under the log function as we see above.
- The probability distribution should sum up to 1 over vocal size V: $\sum_{w=1}^{V} \phi_{tw} = 1\,, \forall t$



E-STEP: FIND q(θ)

Consider the joint log likelihood function:

$$log\, p(\mathbf{W}, \theta, \mathbf{Z} \mid \alpha, \Phi) = \sum_{d=1}^{D}[\sum_{t=1}^{T}(\alpha_t - 1)log(\theta_{dt}) + \sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn} = t](log\theta_{dt} + log\phi_{tw_{dn}})] + const \qquad \textbf{(1)}$$

Note that, w.r.t $q(\theta)$, $\sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn} = t]log\phi_{tw_{dn}}$ in **(1)** is a constant so we can ignore it the below derivation of $log[q(\theta)]$.

$$log[q(\theta)] = \mathbb{E}_{q(\mathbf{Z})} log\,[p(Z, \theta, \mathbf{W})] + C_1$$

$$= \mathbb{E}_{q(\mathbf{Z})} \sum_{d=1}^{D}[\sum_{t=1}^{T}(\alpha_t - 1)log(\theta_{dt}) + \sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn} = t]log\theta_{dt}] + C_1$$

$$= \sum_{d=1}^{D}[\sum_{t=1}^{T}(\alpha_t - 1)log(\theta_{dt}) + \sum_{n=1}^{N_d}\sum_{t=1}^{T}\mathbb{E}_{q(\mathbf{z_{dn}})}\{[z_{dn} = t]\}log\theta_{dt}] + C_1$$

Let $\gamma_{dn} = \mathbb{E}_{q(\mathbf{z_{dn}})}\{[z_{dn} = t]\} \Rightarrow q(z_{dn} = t) = \gamma_{dn}^{t}$

$$= \sum_{d=1}^{D}\sum_{t=1}^{T}[(\alpha_t - 1) + \sum_{n=1}^{N_d}\gamma_{dn}]\, log\theta_{dt} + C_1$$

E-STEP: FIND q(θ)

Now take the *exp* both side to get $q(\theta)$:

$$q(\theta) \propto \prod_{d=1}^{D} \prod_{t=1}^{T} \theta_{dt}^{[(\alpha_t - 1) + \sum_{n=1}^{N_d} \gamma_{dn}]}$$

$$= \prod_{d=1}^{D} \{ \prod_{t=1}^{T} \theta_{dt}^{[(\alpha_t + \sum_{n=1}^{N_d} \gamma_{dn}) - 1]} \}$$

$$\Rightarrow q(\theta) = \prod_{d=1}^{D} q(\theta_d), \text{where } q(\theta_d) \sim \mathbf{Dir}(\theta_d | \alpha_t + \sum_{n=1}^{N_d} \gamma_{dn}) \qquad \textbf{(2)}$$

Note that the update rules for $q(\theta)$ depends on γ_{dn} which depends on z_{dn}. So to compete the E-step we will also need to find updating rule for $q(\mathbf{Z})$

E-STEP: FIND q(Z)

Similarly, W.r.t $q(\mathbf{Z})$ $\sum_{t=1}^{T}(\alpha_t - 1)log(\theta_{dt})$ in **(1)** is a constant so we can ignore it the below derivation.

$$log[q(\mathbf{Z})] = \mathbb{E}_{q(\theta)} log\,[p(Z, \theta, \mathbf{W})] + C_2$$

$$= \mathbb{E}_{q(\theta)} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t](log\theta_{dt} + log\phi_{tw_{dn}}) + C_2 \quad \text{(Substitution from (1))}$$

$$= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \{\sum_{t=1}^{T} [z_{dn} = t](\mathbb{E}_{q(\theta)}[log\theta_{dt}] + log\phi_{tw_{dn}})\} + C_2; \quad (z_{dn} \text{ sums up to 1 over T and } \{*\} \text{ is dist over } z_{dn})$$

E-STEP: FIND q(Z)

Take *exp* both sides we have the product over independent distribution as following:

$$q(\mathbf{Z}) \propto \prod_{d=1}^{D} \prod_{n=1}^{N_d} q(z_{dn})$$

Where:

$$q(z_{dn} = t) \propto exp\{\mathbb{E}_{q(\theta)}[log\theta_{dt}] + log\phi_{tw_{dn}}\}$$

$$= \frac{\phi_{tw_{dn}} exp(\mathbb{E}_{q(\theta_{dt})}[log\theta_{dt}])}{\text{sum over all possible values of t}}$$

$$= \frac{\phi_{tw_{dn}} exp(\mathbb{E}_{q(\theta_{dt})}[log\theta_{dt}])}{\sum_{t'=1}^{T} \phi_{t'w_{dn}} exp(\mathbb{E}_{q(\theta_{dt'})}[log\theta_{dt'}])}$$

$$= \gamma_{dn}^{t}, \quad (\text{since } \gamma_{dn} = \mathbb{E}_{q(\mathbf{z_{dn}})}\{[z_{dn} = t]\}) \qquad \mathbf{(3)}$$

M-STEP

Consider the joint log likelihood function:

$$log\, p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi) = \sum_{d=1}^{D}[\sum_{t=1}^{T}(\alpha_t - 1)log(\theta_{dt}) + \sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn} = t](log\theta_{dt} + log\phi_{tw_{dn}})] + const \qquad \textbf{(1)}$$

M-Step: Φ

Objective: $\mathbb{E}_{q(\theta)q(z)}[log(p(W, Z, \theta))] \rightarrow \max_{q(\phi)}$ such that:

- since ϕ_{tw} is a probability: $\phi_{tw} \geq 0\,, \forall t, w$. This is already been satisfied because it is under the log function as we see above.
- The probability distribution should sum up to 1 over vocal size V: $\sum_{w=1}^{V} \phi_{tw} = 1\,, \forall t$

M-STEP

M-STEP

$$L = \mathbb{E}_{q(\theta)q(z)}[log(p(W, Z, \theta))] + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1)$$

$$= \mathbb{E}_{q(\theta)q(z)} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] log\phi_{tw_{dn}} + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1) \quad \text{(Substitution)}$$

$$= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} (\mathbb{E}_{q(\theta)}[z_{dn} = t]) log\phi_{tw_{dn}} + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1)$$

$$= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} \gamma_{dn}^{t} log\phi_{tw_{dn}} + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1)$$

M-STEP

Taking derivative w.r.t ϕ_{tw} then set it to 0 to solve for ϕ_{tw}, we get:

$$\phi_{tw} = \frac{\sum_{d,n} \gamma_{dn}^{t}[w_{dn} = w]}{\text{sum over all possible words in our data } \mathbf{W}}$$

$$= \frac{\sum_{d,n} \gamma_{dn}^{t}[w_{dn} = w]}{\sum_{w',d,n} \gamma_{dn}^{t}[w_{dn} = w']} \tag{4}$$

SUMMARY

- E-Step: keep Φ fixed, we iteratively update θ and $\mathbf{Z}$ until converges using $\mathcal{KL}(q(\mathbf{Z})q(\theta)\,||\,p(\theta,\mathbf{Z}|\mathbf{W}))$
- M-step: keep θ and $\mathbf{Z}$ fixed, we iteratively update Φ by maximizing $\mathbb{E}_{q(\theta)q(z)}[log(p(W,Z,\theta))]$

MAKING PREDICTION

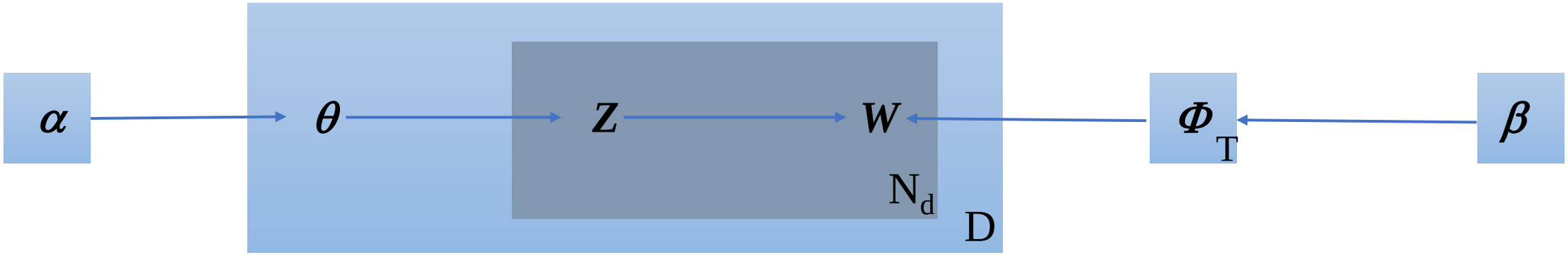
Prediction

Objective: $\mathcal{KL}(q(\theta_{d^*})q(\mathbf{Z}_{d^*}) \,||\, p(\theta_{d^*}, \mathbf{Z}_{d^*}|W; \alpha, \Phi)) \rightarrow \min_{q(\theta_{d^*})q(z_{d^*})}$

For the new document, we wanna predict:

- the new value for $\mathbf{Z}$, i.e. assign a topic for each words
- the new value for θ, i.e. the global topic distribution of the document

EXTENSION



This time we consider Φ as a random variable which follows $DIR(\beta)$. Now the joing probability distribution becomes:

$$p(\theta, \mathbf{Z}, \mathbf{W} \mid \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} p(z_{dn}|\theta_d) p(\phi_{tw}|\beta) p(w_{dn}|z_{dn}, \phi_{tw})$$

High β: there are less words in a topic (sparse distribution of topics over words)

REFERENCES

0. D. Blei, Andrew Ng, Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003.
1. Reed C. Latent Dirichlet Allocation: Towards a Deeper Understanding. 2012.
2. Coursera: Bayesian Methods in Machine Learning. National Research University - Higher School of Economics. 2020.
3. Chenouri S. Stat440/840: Computational Inference. University of Waterloo. 2019.
4. Lysy M. Stat946: Advanced Computational Inference. University of Waterloo. 2019.
5. Struthers C. Stat450/850: Estimation and Hypothesis Testing. University of Waterloo. 2018.
6. Bishop C. Pattern Recognition and Machine Learning. Information Science and Statistics. 2006.
7. Wolf W. Deriving Expectation-Maximization. Blog. 2018.
8. Serrano L. Latent Dirichlet Distribution. Youtube. 2020.


THANK YOU!