LDA In Slow Motion
- Journey of a Parametric Topic Model

Pradipto Das
CSE Department, SUNY Buffalo

What is a topic?

- "I call the same thing *element* and *topos*; for an element or a topos is a heading under which many enthymemes fall." - Aristotle
- An **enthymeme,** in its modern sense, is an informally stated syllogism (a three-part deductive argument) with an unstated assumption that must be true for the premises to lead to the conclusion
  - Aristotle. *Topica.* Translated by E. S. Forster. Loeb Classical Library. Cambridge: Harvard University Press, 1989
    - en.wikipedia.org/wiki/Topics_(Aristotle)
- The Topics contains and relies upon Aristotle's definition of reasoning; *a verbal expression in which, certain things having been laid down, other things necessarily follow from these*

What is a topic? – Aristotle's view

Case 1:

- Major premise: All men are mortal
- Minor premise: All Greeks are men
- Conclusion: All Greeks are mortal

Case 2:

- Major premise: No homework is fun
- Minor premise: Some reading is homework
- Conclusion: Some reading is not fun

What is a topic? – LDA's view

Case 1:

- men, mortal co-occur
- Greeks, men co-occur
- Conclusion: men, mortal and Greeks must be related under some notion

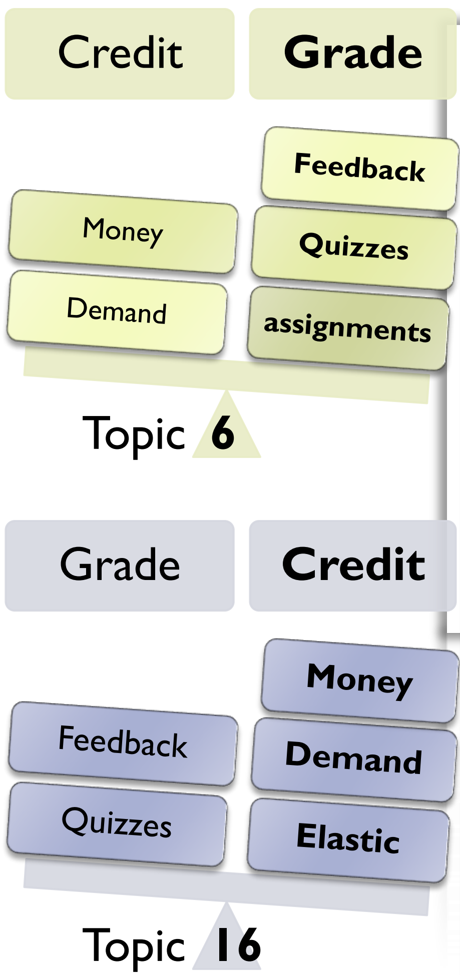
Case 2:

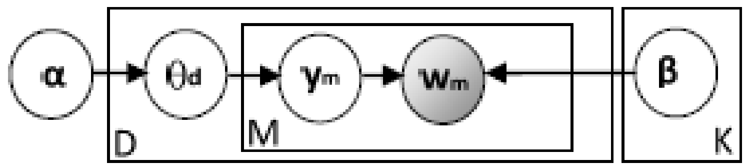
- homework, fun co-occur
- reading, homework co-occur
- Conclusion: homework, fun and reading must be related under some *other* notion – not the one involving Greeks and mortal

What is a topic? – LDA's view

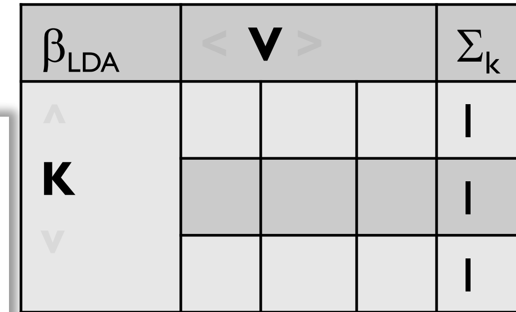
- Repeated syllogisms referring to the same "theme" introduces word co-occurrence
- That words under a "theme" are related is a direct outcome of the way text presents itself in English
- We will hardly see a syllogism like
  - Major premise: All men are mortal
  - Minor premise: Some reading is homework
  - Lack of Conclusion: All men are mortal AND Some reading is homework? (When I write my first draft of a paper after prolonged coding, it reads like this – all I can think of is "Our model performs better..." and "for (k=0; k < …)")

LDA





(a) LDA model

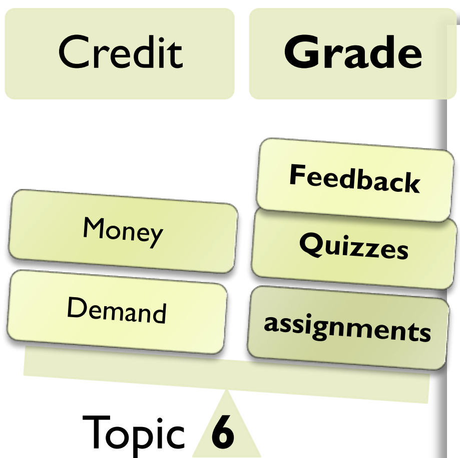


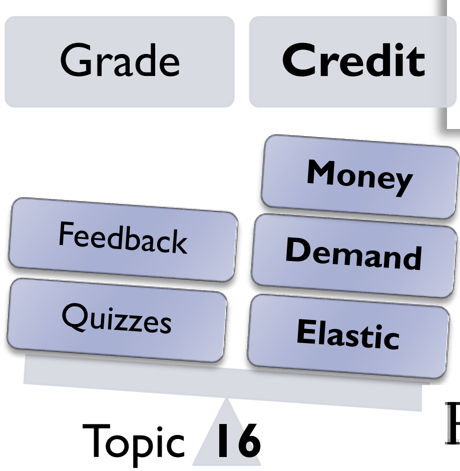
- D = number of documents
- M = number of word positions in the document
- K = number of latent dimensions a.k.a topics/clusters
- V = total number of words in corpus

- w_m = observed word in document d ∈ D
- y_m = hidden topic i ∈ K for word w_m in document d
- θ_d = document d's hidden distribution over K topics
  - These 3 grow with the data (think about D)

- α = parameter of LDA
- β = parameter of LDA (that we finally like to visualize)
  - These 2 does not grow with data (independent of D)

LDA and TagLDA





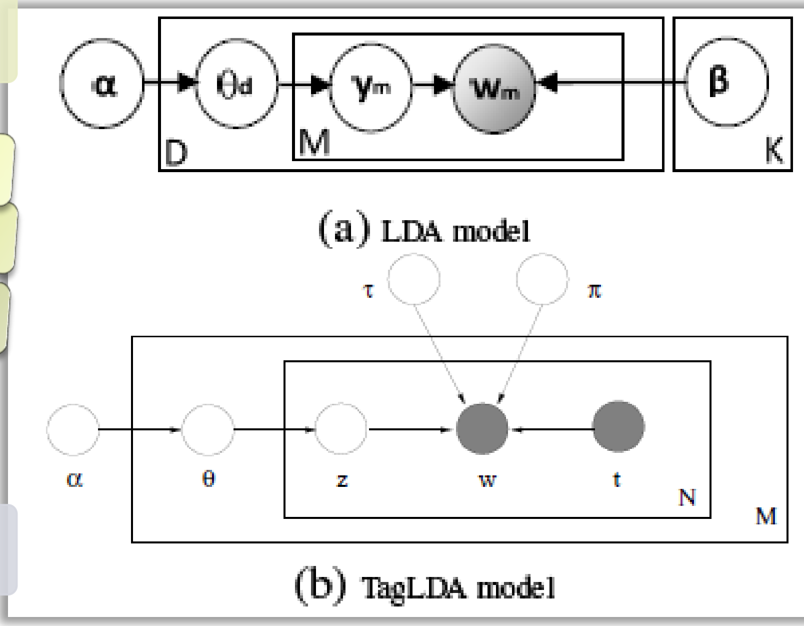
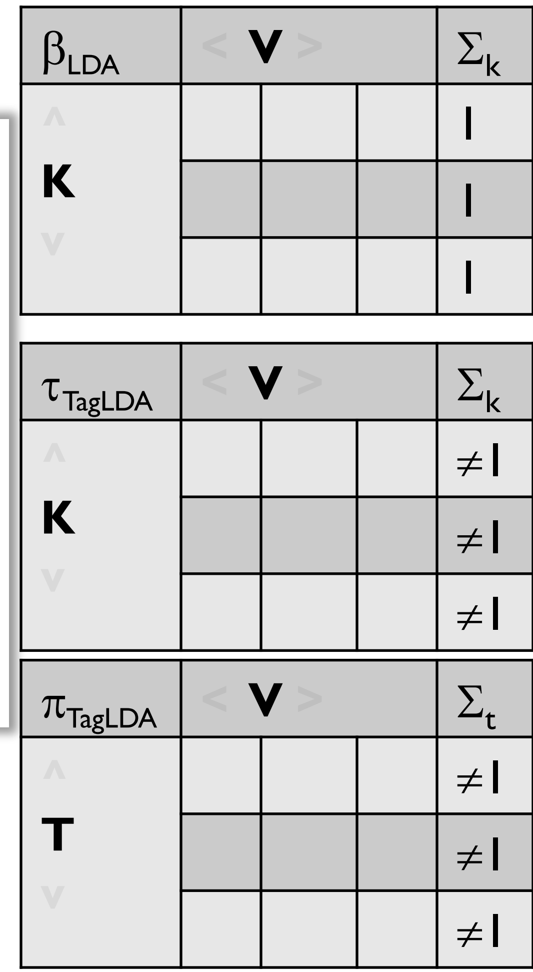


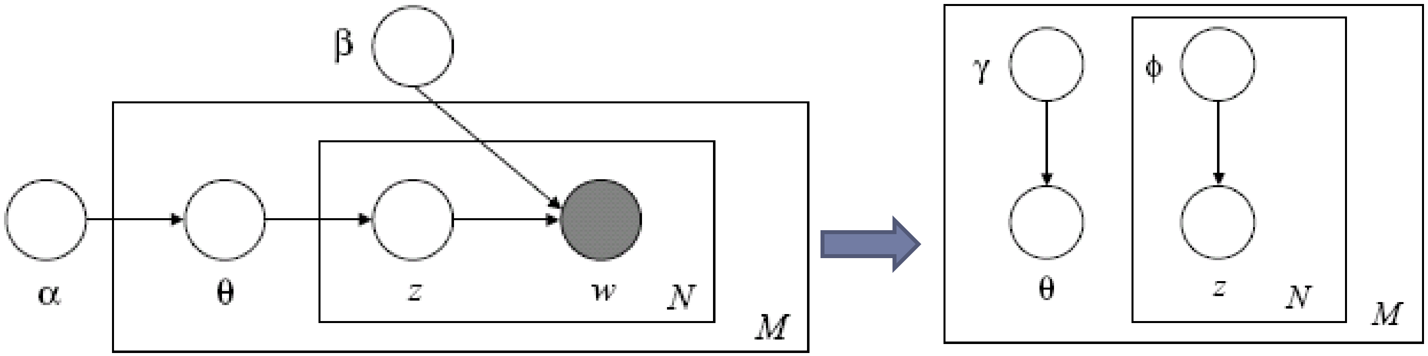
Figure 1: Graphical model representations of LDA and TagLDA models



Parameters of LDA

- β is a set of distributions (topics)
- For each dataset, we only see it's values for the words present in the dataset
- β is row normalized and is represented in log-space since we calculate log likelihoods
- α_i is a prior for θ_i encoding the belief that a document is "about a topic *i*." In basic LDA, all α_i's are equal to α (symmetric)
  - However, β needs to be updated as each document is encountered
  - β_sufficient-statistics is a SINGLE auxiliary storage for β that contains un-normalized entries for β
  - β_sufficient-statistics has the same dimensions as β
  - α_sufficient-statistics similarly is an auxiliary storage for α

Variational Inference in LDA

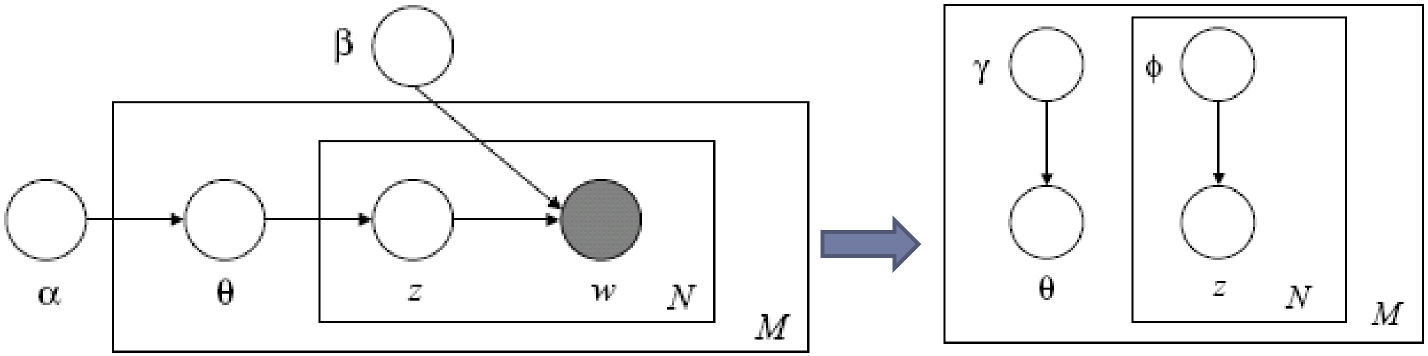


$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) d\theta,$$

- Normalizing constant is intractable due to the coupling between θ and β in the summation over latent topics
- Key motivation - compute the posterior distribution of the hidden variables (z and θ) given a document (word and document distribution over topics)

Variational Inference in LDA



- Find tightest possible lower bound to the value of the expression, by using optimization of variational parameters
  - A simple way to obtain a tractable family of lower bounds is to consider *simple modifications* of the original graphical model in which some of the edges and nodes are removed
- In particular, in the LDA model shown above, the coupling between θ and β arises due to the edges between θ, **z**, and ***w***
  - By dropping *these edges* and the ***w** node*, and using free variational parameters in the resulting simplified graphical model, we obtain a *family* of distributions (for ϕ and γ) on the latent variables θ *and* **z**

General EM algorithm

- **Local** optimization algorithm
- Given a set of observed (visible) variables V , a set of unobserved (hidden/ latent/missing) variables H, and model parameters θ, optimize the log probability of the observed data

$$\mathcal{L}(\theta) = log p(V|\theta) = \log \int dH\, p(H,V|\theta)$$

- Using Jensen's inequality for any distribution of hidden states q(H) [Note: log is concave]

$$\mathcal{L}(\theta) = \log \int dH\, q(H) \frac{p(H,V|\theta)}{q(H)} \geq \int dH\, q(H) log \frac{p(H,V|\theta)}{q(H)} = \mathrm{F}(q,\theta)$$

- Normally we are given only V, not (H,V)
- In the EM algorithm, we alternately optimize F(q,θ) w.r.t q and θ

Acknowledgements: Matthew Beal's lecture notes

General EM algorithm

- The lower bound on the log likelihood

$$\mathcal{F}(q,\theta) = \int dH\, q(H) log \frac{p(H,V|\theta)}{q(H)} = \int dH\, q(H) \log p(H,V|\theta) + \mathcal{H}(q)$$

- E step: optimize F(q,θ) w.r.t the distribution over hidden variables given the parameters

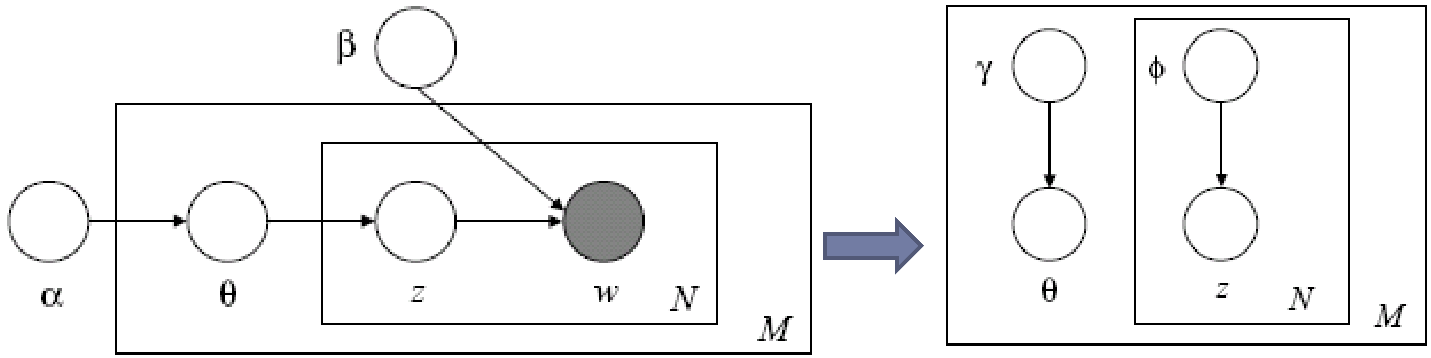
$$q^{(k)}(H) := \underset{q(H)}{\operatorname{argmax}} \mathcal{F}(q(H), \theta^{(k-1)})$$

- M step: maximize F(q,θ) w.r.t the parameters given the hidden distribution

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \mathcal{F}\left(q^{(k)}(H), \theta\right) = \underset{\theta}{\operatorname{argmax}} \int dH q^{(k)}(H) \log p(H,V|\theta)$$

- Take Home Message:
  - E step: fill in values for the hidden variables according to their posterior probabilities
  - M step: learn model as if hidden variables were not hidden

Variational Inference (VBEM) in LDA



- Note: γ (Dirichlet) and $(\phi_1, \phi_2, \ldots, \phi_N)$ (Multinomials) are *free parameters*

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

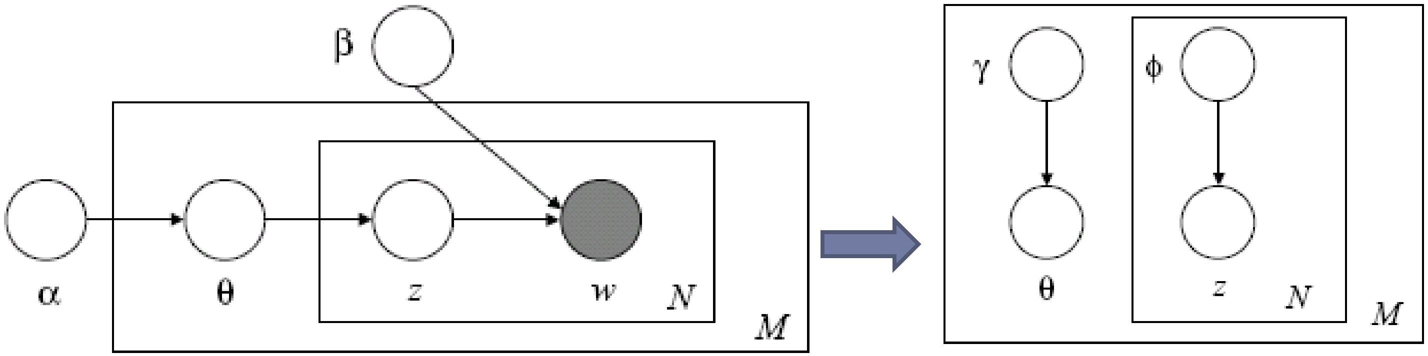
$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta))$$

$$\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) | \gamma]\}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

$$\mathrm{E}_q[\log(\theta_i) \mid \gamma] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)$$

Variational Inference (VBEM) in LDA



$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*{}_{dni} w^j{}_{dn}$$

$$\alpha_{\text{new}} = \alpha_{\text{old}} - H(\alpha_{\text{old}})^{-1} g(\alpha_{\text{old}})$$ See original paper for α optimization

- Take Home Message:
  - Training: Initialize α and β to initial values → compute γ, ϕ from observed documents → again compute α and β from γ, ϕ
  - Testing: Compute γ, ϕ for unseen documents from learned α and β for training documents. Update α and β accordingly
    - Note that testing is implicit in training (why?)

Intuitions for VBEM in LDA

- Rarely, if ever, does a document contain all words belonging to the same topic
  - Represent a document as a distribution over topics
- Given any decently sized document set, it is not possible to exactly estimate the values of the distribution over all combinations of observations in our lifetime
- Assume simpler distributions that tally with our beliefs and compute **tractable expectations rather than exact estimations**
- Find as tight a lower bound to the absolute truth – the probability of the data given the model

Expectation Maximization (EM) for LDA

- compute_initial_guesses() // for beta (corpus distribution)

```
prev_ELBO = 0;
while (converged > tolerance)
{
  e-step(); // zero initialize sufficient statistics; initialize and update
per document distributions from corpus distributions; sufficient
statistics depend on per document distributions
  m-step(); // update corpus distributions from sufficient statistics
  curr_ELBO = compute_likelihood(); // sums up the individual
doc_likelihoods from compute_doc_likelihood()s in doc-e-step()s
  converged = (prev_ELBO-curr_ELBO)/prev_ELBO;
  prev_ELBO = curr_ELBO;
}
```

Computing Likelihood for LDA

- Variational Inference
  - Recall the variational distribution we had before

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^{N} q(z_n | \phi_n)$$

$$\begin{aligned} log p(\mathbf{w}|\alpha,\beta) &= log \int \sum_{z} p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta) d\theta \\ &= log \int \sum_{z} \frac{p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) q(\theta,\mathbf{z})}{q(\theta,\mathbf{z})} d\theta \\ &\geq \int \sum_{z} q(\theta,z) log p(\theta,z,w|\alpha,\beta) d\theta - \int \sum_{z} q(\theta,z) log q(\theta,z) d\theta \\ &= Eq[\log p(\theta,z,w|\alpha,\beta)] - Eq[log q(\theta,z)] \end{aligned}$$

$$log p(w|\alpha,\beta) = L(\gamma,\phi;\alpha,\beta) + D(q(\theta,z|\gamma,\phi)||p(\theta,z|w,\alpha,\beta))$$

Where, $L(\gamma,\phi;\alpha,\beta) = E_q[log p(\theta,z,w|\alpha,\beta)] - E_q[log q(\theta,z)]$

Computing Likelihood for LDA

- Variational Inference
  - We now expand the L(γ,ϕ; α,β) into 5 core expressions

$$L(\gamma,\phi;\alpha,\beta) = E_q[log p(\theta|\alpha)] + E_q[log p(z|\theta)] + E_q[log p(\mathbf{w}|\mathbf{z},\beta)] - E_q[log q(\theta)] - E_q[log q(z)]$$

  - We expand the expression in terms of the model parameters (α,β) and the variational parameters (γ,ϕ)

compute_doc_likelihood()

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

Expectation Maximization (EM) for LDA

doc-e-step() { $\alpha^{(suffstats)} = 0; \beta^{(suffstats)} = 0$

initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*

initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*

prev_ELBO = 0; t = 0;

repeat

for *n* = *1* to *N* {

for *i* = *1* to K

$$\phi^{t+1}_{n,i} = \beta_{i,w_n} \exp(\psi(\gamma^t_i) - \psi(\sum_{j=1}^{K} \gamma^t_j));$$

normalize $\phi^{t+1}_{n,i}$ so that row-sum=1

}

$$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$$

curr_ELBO = compute_doc_likelihood();

converged = (prev_ELBO – curr_ELBO)/prev_ELBO;

until converged > tolerance or t < Max_T for e-step

for *n* = *1* to *N*

for *i* = *1* to *K*

$$\beta^{(suffstats)}_{i,w_{d,n}} += \phi_{n,i} w_{d,n};$$

for *i* = *1* to *K*

$$\alpha^{(suffstats)} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$$

}

This expression is not needed since it is same for all *n* in *N_d*

Store final value of curr_ELBO as the doc_likelihood for this document

Precompute this term in the alpha update first

Synthetic Example for LDA

- model.input =
  - 5 0:1:1 1:1:1 1:1:1 3:1:1 3:1:1 //wordID:count:tagID (count is always 1)
  - 6 1:1:1 2:1:1 2:1:1 2:1:1 3:1:1 3:1:1 // tagID not used in LDA but
  - 4 4:1:1 5:1:1 6:1:1 0:1:1 // used in TagLDA (Later slides)
  - 6 4:1:1 5:1:1 4:1:1 5:1:1 4:1:1 5:1:1 // 6 = num. of positions
- term_to_id.txt=
  - Gilchrist=0 wicket=1 keeper=3 captain=2 Jones=4 data=5 architect=6
- model.input + term_to_id.txt =
  - doc0: Gilchrist wicket wicket keeper keeper
  - doc1: wicket captain captain captain keeper keeper
  - doc2: Jones data architect Gilchrist
  - doc3: Jones data Jones data Jones data

Synthetic Example for LDA

- What should we expect from model.input?
  - doc0: Gilchrist wicket wicket keeper keeper
  - doc1: wicket captain captain captain keeper keeper
  - doc2: Jones data architect Gilchrist
  - doc3: Jones data Jones data Jones data
- Given number of hidden topics, K=2
- The documents should be more about one topic than another (based on our prior belief)
- One topic should be about **sport** and the other about **software** and most words should be more focused on one topic than another
- Gilchrist may belong to both topics as suggested by data (although we know that Gilchrist was a sportsman)

Synthetic Example for LDA

- **compute_initial_guesses()**
  - Randomly initializing β_sufficient-statistics in this case and obtaining row-normalized log(β) as: [Note: log(β) = {log($\beta_{i,j}$)}]

log beta =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	-1.67219	-1.49886	-2.05742	-2.92477	-3.356	-1.50465	-1.89479
Topic1:	-3.11147	-2.12028	-2.81038	-1.41226	-1.54886	-1.42601	-2.53912

alpha = 0.5

Synthetic Example for LDA

- **E-Step begin**

Processing document 0: Gilchrist(0/0) wicket(1/1) wicket(1/2) keeper(3/3) keeper(3/4)

termID Position

gamma initial = 3 3

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5

doc-e-step() { $\alpha^{(suffstats)} = 0; \beta^{(suffstats)} = 0$

initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*

initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*

prev_ELBO = 0; t = 0;

repeat

for *n* = *1* to *N* {

for *i* = *1* to *K*

$\phi^{t+1}_{n,i} = \beta_{i,w_n} \exp(\psi(\gamma^t_i) - \psi(\sum_{j=1}^{K} \gamma^t_j))$;

normalize $\phi^{t+1}_{n,i}$ so that row-sum=1

}

$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$

curr_ELBO = compute_doc_likelihood();

converged = (prev_ELBO – curr_ELBO)/prev_ELBO;

until converged > tolerance or t < Max_T for e-step

for *n* = *1* to *N*

for *i* = *1* to *K*

$\beta^{(suffstats)}_{i,w_{d,n}} += \phi_{n,i} w_{d,n}$;

for *i* = *1* to *K*

$\alpha^{(suffstats)} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$

}

gamma =	2.90867	3.09133
Phi =	0.79803	0.20196
	0.63522	0.36477
	0.63477	0.36522
	0.17038	0.82961
	0.17025	0.82974

Beta suff stats =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.79803	1.27	0	0.34063	0	0	0
Topic1:	0.20196	0.7300	0	1.65936	0	0	0

alpha suff stats = -1.56795

Synthetic Example for LDA

E-Step continue

Processing document 1: wicket(1/0) captain(2/1) captain(2/2) captain(2/3) keeper(3/4) keeper(3/5)

gamma initial = 3.5 3.5

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;
0.5 0.5; 0.5 0.5;

gamma =	3.74424	3.25576
Phi =	0.68789	0.31210
	0.71509	0.28490
	0.71480	0.28519
	0.71454	0.28545
	0.20602	0.79397
	0.20587	0.79412

doc-e-step() { $\alpha^{(suffstats)} = 0; \beta^{(suffstats)} = 0$
initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
for *n* = *1* to *N* {
for *i* = *1* to *K*
$\phi^{t+1}_{n,i} = \beta_{i,w_n} \exp(\psi(\gamma^t_i) - \psi(\sum_{j=1}^{K} \gamma^t_j))$;
normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
}
$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n* = *1* to *N*
for *i* = *1* to *K*
$\beta^{(suffstats)}_{i,w_{d,n}} += \phi_{n,i} w_{d,n}$;
for *i* = *1* to *K*
$\alpha^{(suffstats)} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

Beta suff stats =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.798034	1.95789	2.14445	0.752533	0	0	0
Topic1:	0.201966	1.04211	0.855551	3.24747	0	0	0

alpha suff stats = -3.11368

Synthetic Example for LDA

E-Step continue

Processing document 2: Jones(4/0) data(5/1) architect(6/2) Gilchrist(0/3)

gamma initial = 2.5 2.5
Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;

gamma =	2.80352	2.19648
Phi =	0.18017	0.81982
	0.55360	0.44639
	0.71931	0.28068
	0.85043	0.14957

doc-e-step() { $\alpha^{(suffstats)} = 0; \beta^{(suffstats)} = 0$
initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
for *n* = *1* to *N* {
for *i* = *1* to *K*
$\phi^{t+1}_{n,i} = \beta_{i,w_n} \exp(\psi(\gamma^t_i) - \psi(\sum_{j=1}^{K} \gamma^t_j))$;
normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
}
$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n* = *1* to *N*
for *i* = *1* to *K*
$\beta^{(suffstats)}_{i,w_{d,n}} += \phi_{n,i} w_{d,n}$;
for *i* = *1* to *K*
$\alpha^{(suffstats)} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

Beta suff stats =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	1.6484	1.9578	2.1444	0.7525	0.1801	0.5536	0.7193
Topic1:	0.3515	1.0421	0.8555	3.2474	0.8198	0.4463	0.2806

alpha suff stats = -4.74159

Synthetic Example for LDA

E-Step continue

Processing document 3: Jones(4/0) data(5/1) Jones(4/2) data(5/3) Jones(4/4) data(5/5)

gamma initial = 3.5 3.5

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5

gamma =	0.62764	6.37235
Phi =	0.00661	0.99338
	0.03610	0.96389
	0.00658	0.99342
	0.03594	0.96405
	0.00655	0.99344
	0.03584	0.96415

doc-e-step() { $\alpha^{(suffstats)} = 0; \beta^{(suffstats)} = 0$

initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*

initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*

prev_ELBO = 0; t = 0;

repeat

for *n* = *1* to *N* {

for *i* = *1* to *K*

$\phi^{t+1}_{n,i} = \beta_{i,w_n} \exp(\psi(\gamma^t_i) - \psi(\sum_{j=1}^{K} \gamma^t_j))$;

normalize $\phi^{t+1}_{n,i}$ so that row-sum=1

}

$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$

curr_ELBO = compute_doc_likelihood();

converged = (prev_ELBO – curr_ELBO)/prev_ELBO;

until converged > tolerance or t < Max_T for e-step

for *n* = *1* to *N*

for *i* = *1* to *K*

$\beta^{(suffstats)}_{i,w_{d,n}} += \phi_{n,i} w_{d,n}$;

for *i* = *1* to *K*

$\alpha^{(suffstats)} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$

}

Beta suff stats =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	1.6484	1.9578	2.1444	0.7525	0.1999	0.6614	0.7193
Topic1:	0.3515	1.0421	0.8555	3.2474	3.8000	3.3385	0.2806

alpha suff stats = -8.15944

Synthetic Example for LDA

- **E-Step end**
- **M-Step begin**

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^{*}_{dni} w^{j}_{dn}.$$

Log Beta =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	-1.5900	-1.4180	-1.3270	-2.3742	-3.6997	-2.503	-2.41936
Topic1:	-3.6039	-2.5172	-2.7145	-1.3806	-1.2234	-1.3525	-3.82897

alpha = 0.947154

log_likelihood= -43.55398 : converged = inf em_steps = 1 (out of 50)

TopicHeap0: captain wicket Gilchrist keeper architect data Jones

TopicHeap1: Jones data keeper wicket captain Gilchrist architect

- **M-Step end**

Synthetic Example for LDA

- **Iteration 2**
- **E-Step begin … E-Step end**
- **M-Step begin**

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}.$$

Log Beta =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	-1.7347	-1.3735	-1.3185	-1.6982	-4.6176	-3.4413	-2.55193
Topic1:	-4.0856	-3.3162	-3.7959	-1.6210	-1.0081	-1.0685	-3.99843

alpha = 1.13942

log_likelihood= -39.944901 : converged = 0.08286 em_steps = 2 (out of 50)

TopicHeap0:	captain	wicket	keeper	Gilchrist	architect	data	Jones
TopicHeap1:	Jones	data	keeper	wicket	captain	architect	Gilchrist

- **M-Step end**

Synthetic Example for LDA

- **Iteration 3**
- **E-Step begin … E-Step end**
- **M-Step begin**

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}.$$

Log Beta =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	-1.8527	-1.4130	-1.3982	-1.2863	-5.9505	-4.7532	-2.72934
Topic1:	-4.3027	-4.6651	-5.3798	-2.5688	-0.8214	-0.8397	-3.72119

alpha = 0.885771

log_likelihood= -37.33781618 : converged = 0.06527 em_steps = 3 (out of 50)

TopicHeap0: keeper captain wicket Gilchrist architect data Jones

TopicHeap1: Jones data keeper wicket captain architect Gilchrist

- **M-Step end**

Synthetic Example for LDA

- **Iteration 4**
- **E-Step begin … E-Step end**
- **M-Step begin**

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}.$$

Log Beta =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	-1.8896	-1.4260	-1.4240	-1.1591	-7.7516	-6.5435	-2.95167
Topic1:	-4.2712	-6.6943	-7.5649	-4.4922	-0.7617	-0.7649	-3.19709

alpha = 0.605925

log_likelihood= -34.94676348 : converged = 0.06404 em_steps = 4 (out of 50)

TopicHeap0: keeper captain wicket Gilchrist architect data Jones

TopicHeap1: Jones data architect Gilchrist keeper wicket captain

- **M-Step end**

Synthetic Example for LDA

- **Iteration 46**
- **E-Step begin … E-Step end**
- **M-Step begin**

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^{*}_{dni} w^{j}_{dn}$$

Log Beta =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	-2.3979	-1.2992	-1.2992	-1.0116	-337.72	-337.72	-336.77
Topic1:	-2.3025	-337.14	-337.55	-337.23	-0.9162	-0.9162	-2.3025

alpha = 0.0123227

log_likelihood= -29.05400653 : converged = 0.00010 em_steps = 46 (out of 50)

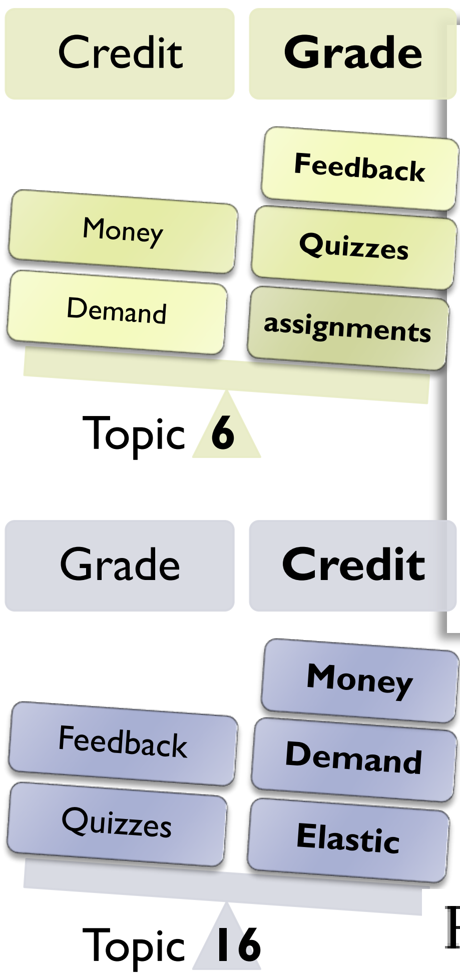
TopicHeap0: keeper wicket captain Gilchrist architect Jones data

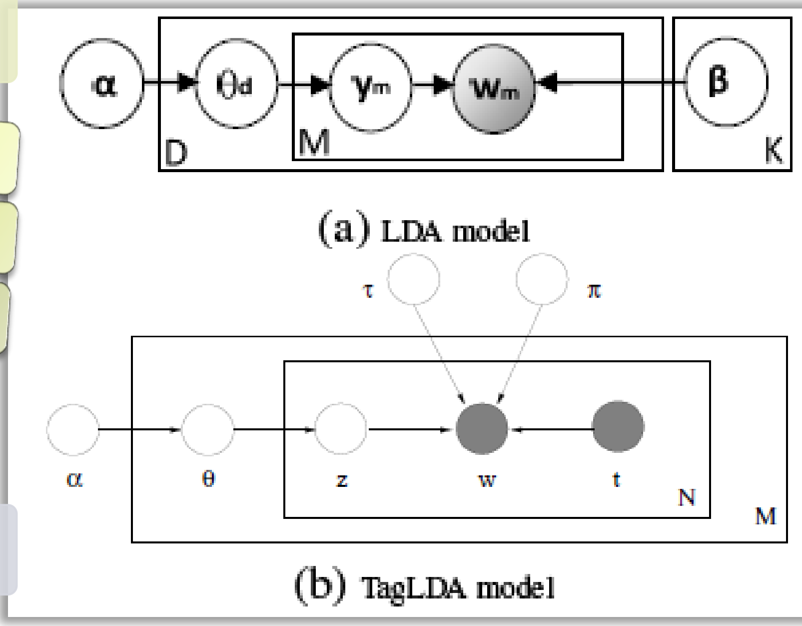
TopicHeap1: Jones data Gilchrist architect wicket keeper captain

Ties broken arbitrarily

- **M-Step end**

LDA and TagLDA





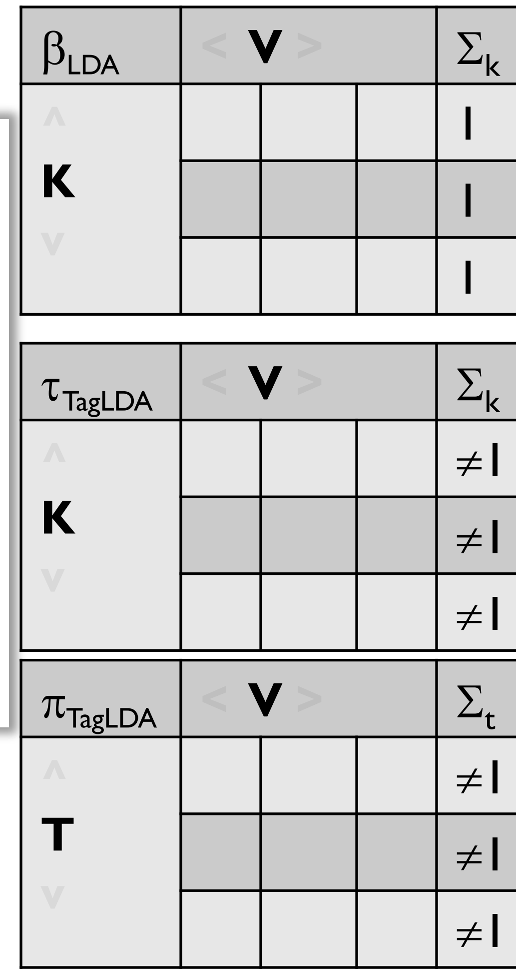
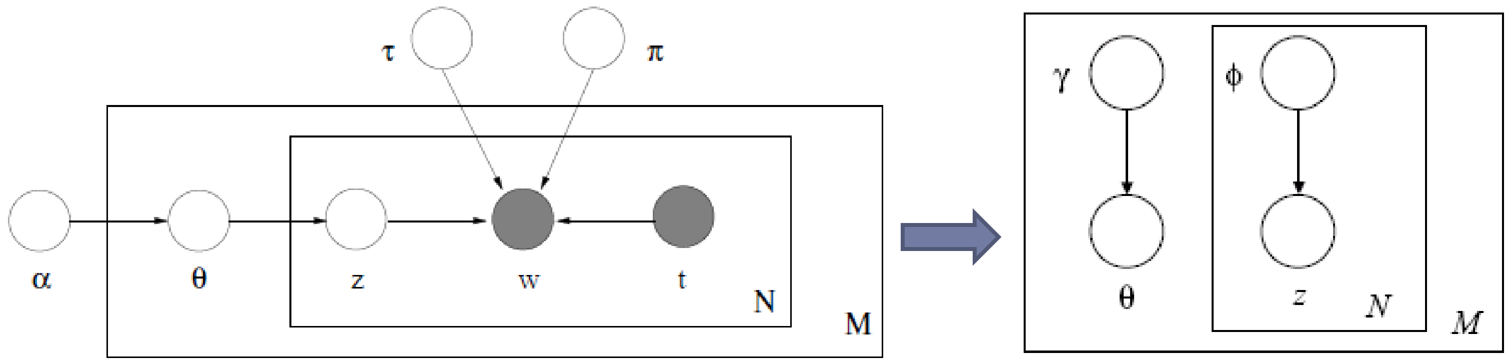


Figure 1: Graphical model representations of LDA and TagLDA models

Parameters of TagLDA

- τ is a set of distributions (topics over words) [Not a multinomial]
- π is a set of distributions (tags over words) [Not a multinomial]
- For each dataset, we only see their values for the words and tags present in the dataset
- How much a topic explains the data vs. how much a tag explains the data
- α_i is a prior for θ_i encoding the belief that a document is "about a topic *i*." In basic LDA, all α_i's are equal to α
  - To update τ in the m-step() there needs to be TWO τ_marginal-statistics matrices which are updated per document in the e-step
  - To update π in the m-step() there needs to be TWO π_marginal-statistics matrices which are updated per document in the e-step
  - α_sufficient-statistics similarly is an auxiliary storage for α

Properties of TagLDA



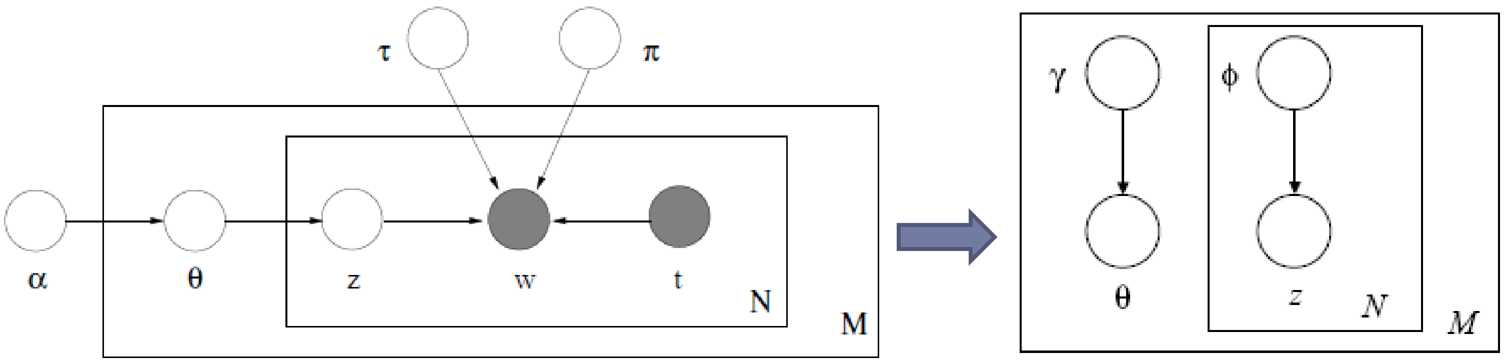
$$P(w_n = v|z_n, t_n, \tau, \pi) \propto \exp(\tau_{z_n,v} + \tau_{t_n,v})$$

$$p(w|t,\alpha,\tau,\pi) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n=1}^{k} p(z_n|\theta)p(w_n|z_n,t_n,\tau,\pi)\right)d\theta$$

$$\log p(w|t,\alpha,\tau,\pi) \geq \int_\theta \sum_z q(\theta,z|\gamma,\phi)(logp(w,\theta,z|t,\alpha,\tau,\pi) - logq(\theta,z|\gamma,\phi))d\theta$$
$$\equiv L(\gamma,\phi;\alpha,\tau,\pi)$$

$$L(\gamma,\phi;\alpha,\tau,\pi)$$
$$= Eq[logp(w,\theta,z|t,\alpha,\tau,\pi)] - E_q[logq(\theta,z|\gamma,\phi)]$$
$$= Eq[logp(\theta|\alpha)] + E_q[logp(z|\theta)]\ E_q[logp(w|z,\tau,\pi)]$$
$$-Eq[logq(\theta)] - Eq[logq(z)]$$

Properties of TagLDA



$$E_q[logp(w|z,\tau,\pi)] = E_q[\sum_{n=1}^{N} log \frac{\exp(\tau_{z_n,v} + \pi_{t_n,v})}{\sum_{v=1}^{V} \exp(\tau_{z_n,v} + \pi_{t_n,v})}]$$

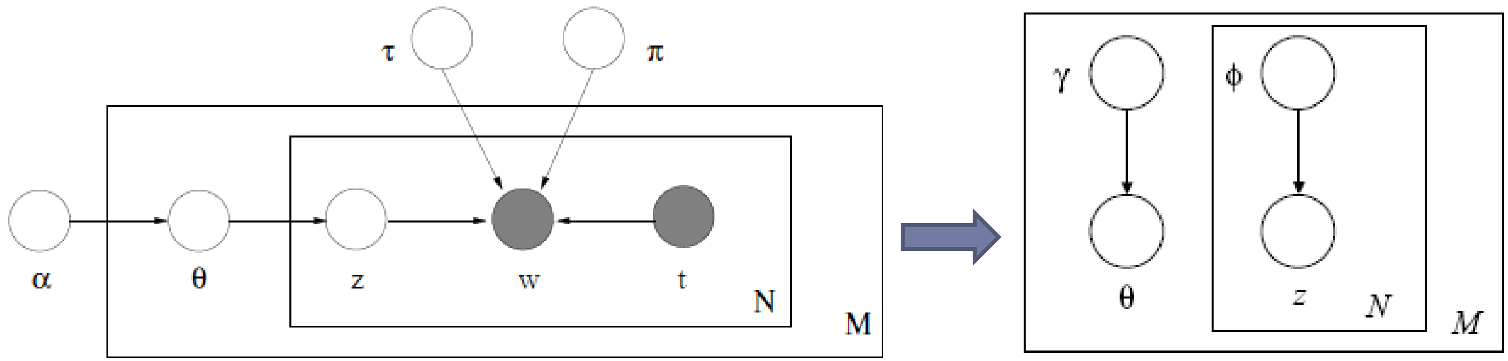
$$\log(x) \leq \zeta^{-1}x + \log(\zeta) - 1, \forall\zeta > 0$$

$$\sum_{n=1}^{N} E_q[log \sum_{v=1}^{V} \exp(\tau_{z_n,v} + \pi_{t_n,v})] \leq \sum_{n=1}^{N} [\zeta_n{}^{-1} \left(\sum_{v=1}^{V} \sum_{i=1}^{k} \phi_{ni} \exp(\tau_{z_n,v} + \pi_{t_n,v}) \right) + log\zeta_n - 1]$$

(ϕ_{ni} ↑ $\mathrm{E}_{q(z_n)}$)

NOTE: $\tau_{i,v}$ and $\pi_{t,v}$ are in log-space all throughout TagLDA discussions

Variational Inference in TagLDA



- Note: γ (Dirichlet), $(\phi_1, \phi_2, \ldots, \phi_N)$ (Multinomials) and ξ are *free parameters*

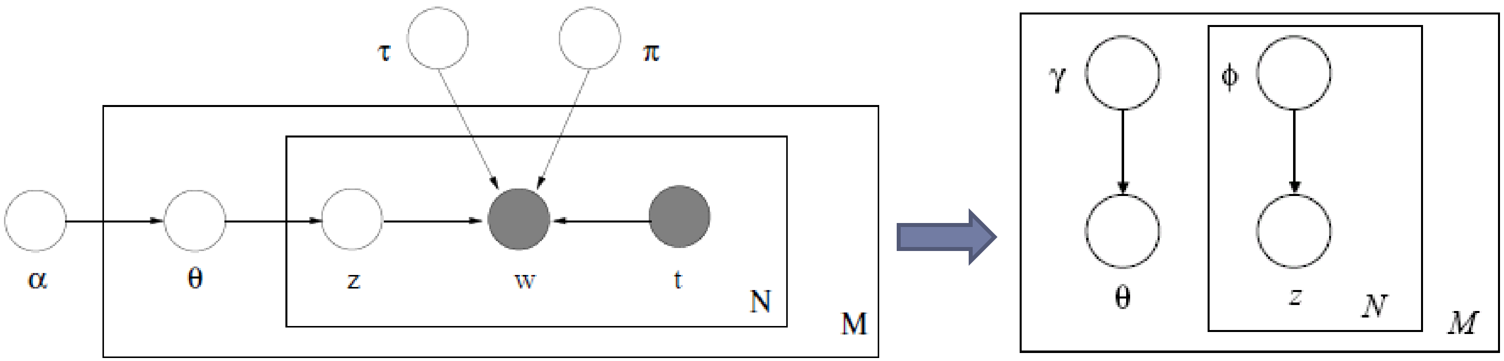
Precompute a KxT matrix before every E-step

$$\hat{\xi}_n = \sum_{v=1}^{V} \sum_{i=1}^{k} \phi_{ni} \exp(\tau_{i,v} + \pi_{t_n,v})$$ Equ. TL- A

$$\hat{\phi}_{ni} \propto \exp\left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{k} \gamma_j\right) + \left(\tau_{i,w_n} + \pi_{t_n,w_n}\right) - \sum_{v=1}^{V} \xi_n^{-1} \exp\left(\tau_{i,v} + \pi_{t_n,v}\right)\right)$$ Equ. TL- B

$$\hat{\gamma}_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$ Equ. TL- C

Variational Inference in TagLDA



$$\tau_{i,v} = \log\left(\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dn,i}\delta(w_{dn},v)\right) - \log(\sum_{d=1}^{M}\sum_{n=1}^{N_d}\xi_{dn}{}^{-1}\phi_{dn,i}exp(\pi_{t_{dn},v}))$$

Equu.
TL- D

$$\pi_{t,v} = \log\left(\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dn,i}\delta(t_{dn},t)\delta(w_{dn},v)\right) - \log(\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\xi_{dn}{}^{-1}\phi_{dn,i}exp(\tau_{i,v})\delta(t_{dn},t))$$

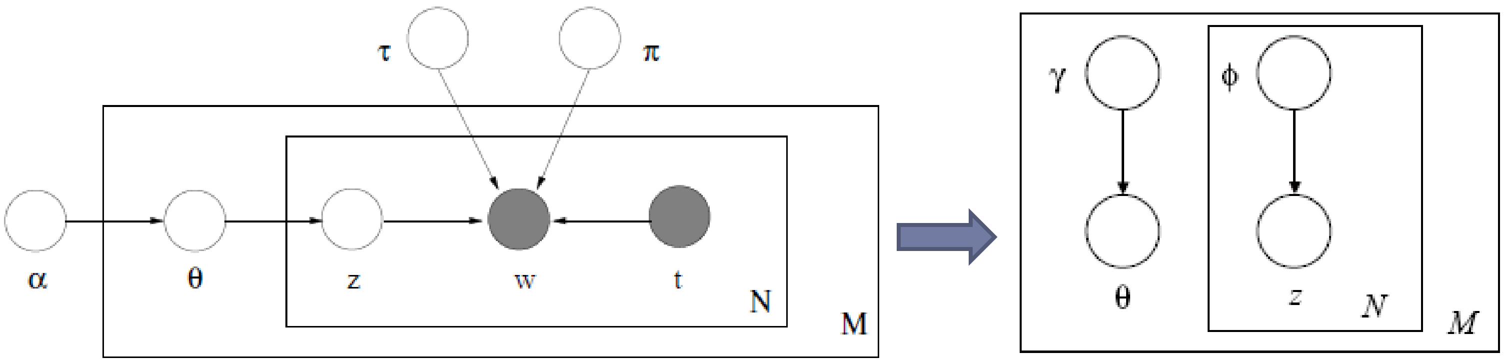
Equ.
TL- E

$$\alpha_{new} = \alpha_{old} - H(\alpha_{old})^{-1}g(\alpha_{old})$$

See original paper for α optimization

Equ.
TL- F

Adding Regularizers in TagLDA



$$\widehat{\mathcal{L}_{(.)}} = \mathcal{L}_{(.)} - \frac{1}{2\sigma^2}\left(\sum_{v=1}^{V}(\exp(\tau_{i,v}))^2\right) - \frac{1}{2\sigma^2}\left(\sum_{v=1}^{V}(\exp(\pi_{t,v}))^2\right)$$

0-mean, σ-stddev Gaussian regularizer for every $\tau_{i,v}$ and $\pi_{t,v}$

$$\pi_{t,v} = \log\left(\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dn,i}\delta(t_{dn},t)\delta(w_{dn},v)\right) - \log\left(\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\xi_{dn}{}^{-1}\phi_{dn,i}\exp(\tau_{i,v})\delta(t_{dn},t)\right)$$

$$= \log(term_1{}^{\pi}) - \log(term_2{}^{\pi})$$

letting A = exp($\tau_{i,v}$) or exp($\pi_{t,v}$), $2A = -\sigma^2 term_2^{(.)} + \sigma\sqrt{\sigma^2(term_2^{(.)})^2 + 4term_1^{(.)}}$

Update for $\tau_{i,v}$ is exactly the same with $term_1^{\tau}$ and $term_2^{\tau}$

Equ. TL- Er

Expectation Maximization (EM) for TagLDA

- compute_initial_guesses() // for beta (corpus distribution)

  prev_ELBO = 0;
  while (converged > tolerance)
  {
    e-step(); // zero initialize 4 marginal statistic matrices and 1 sufficient statistic for alpha; initialize and update per document distributions from corpus distributions; calls doc-e-step() for all documents
    m-step(); // update corpus distributions from statistics
    curr_ELBO = compute_likelihood(); // sums up the individual doc_likelihoods from compute_doc_likelihood()s in doc-e-step()s
    converged = (prev_ELBO-curr_ELBO)/prev_ELBO;
    prev_ELBO = curr_ELBO;
  }

Computing Likelihood for TagLDA

- Variational Inference
  - We now expand the L(γ,ϕ; α,τ,π) into 5 core expressions

$$L(\gamma, \phi; \alpha, \tau, \pi) = \mathrm{E}_q[\log p(\theta|\alpha)] + \mathrm{E}_q[\log p(\mathbf{z}|\theta)] + \mathrm{E}_q[\log p(\mathbf{w}|\mathbf{z}, \tau, \pi)] \\ -\mathrm{E}_q[\log q(\theta)] - \mathrm{E}_q[\log q(\mathbf{z})],$$

  - So the likelihood in terms of the model parameters (α,τ,π) and the variational parameters (ξ,γ,ϕ)

compute_doc_likelihood()

$$\log\Gamma\left(\sum_{j=1}^{k}\alpha_j\right) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{k}\gamma_j))$$
$$+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\psi(\gamma_i) - \psi(\sum_{j=1}^{k}\gamma_j))$$
$$\sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\tau_{i,wn} + \pi_{t_n,w_n}) - \sum_{n=1}^{N}[\zeta_n^{-1}\left(\sum_{v=1}^{V}\sum_{i=1}^{k}\phi_{ni}\exp(\tau_{z_n,v} + \pi_{t_n,v})\right) + \log\zeta_n - 1]$$
$$- \log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) - \sum_{i=1}^{k}\log\Gamma(\gamma_i) + \sum_{i=1}^{k}(\gamma_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{k}\gamma_j))$$
$$- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni} \equiv L2(\gamma, \phi, \zeta; \alpha, \tau, \pi)$$

Expectation Maximization (EM) for TagLDA

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$

initialize $\phi^{0}_{n,i} = 1/K$ for all *i* and *k*

initialize $\gamma^{0}_{i} = \alpha_{i} + N/K$ for all *i*

prev_ELBO = 0; t = 0;

repeat

for *n* = *1* to *N* {

for *i* = *1* to *K*

Update $\phi^{t+1}_{n,i}$ according to Equ.TL-B;

normalize $\phi^{t+1}_{n,i}$ so that row-sum=1

}

$$\gamma^{t+1}_{i} = \gamma^{t}_{i} + \left(\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^{t}_{n,i}\right)$$

curr_ELBO = compute_doc_likelihood();

converged = (prev_ELBO – curr_ELBO)/prev_ELBO;

until converged > tolerance or t < Max_T for e-step

for *n* = *1* to *N*

Update $\pi^{marginal_{1,2}}_{t_{d,n}, w_{d,n}}$ according to Equ.TL-Er

for *i* = *1* to *K*

Update $\tau^{marginal_{1,2}}_{i, w_{d,n}}$ according to Equ.TL-Er;

for *i* = *1* to *K*

$$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi\left(\sum_{j=1}^{K} \gamma_{d,j}\right)$$

}

Synthetic Example for TagLDA

- model.input =
  - 5 0:1:0 1:1:1 1:1:1 3:1:1 3:1:1 //wordID:count:tagID
  - 6 1:1:1 2:1:0 2:1:0 2:1:0 3:1:1 3:1:1 // tagID = 0 means "un-important"
  - 4 4:1:0 5:1:1 6:1:1 0:1:0 // tagID = 1 means "important"
  - 6 4:1:0 5:1:1 4:1:0 5:1:1 4:1:0 5:1:1 // count is always 1 in this format
- term_tag_to_id.txt =
  - 0=un-important 1=important
- term_to_id.txt=
  - Gilchrist=0 wicket=1 keeper=3 captain=2 Jones=4 data=5 architect=6
- model.input + term_to_id.txt + tern_tag_to_id.txt =
  - doc0: Gilchrist#0 wicket#1 wicket#1 keeper#1 keeper#1
  - doc1: wicket#1 captain#0 captain#0 captain#0 keeper#1 keeper#1
  - doc2: Jones#0 data#1 architect#1 Gilchrist#0
  - doc3: Jones#0 data#1 Jones#0 data#1 Jones#0 data#1

Synthetic Example for TagLDA

- What should we expect from model.input?
  - doc0: Gilchrist#0 wicket#1 wicket#1 keeper#1 keeper#1
  - doc1: wicket#1 captain#0 captain#0 captain#0 keeper#1 keeper#1
  - doc2: Jones#0 data#1 architect#1 Gilchrist#0
  - doc3: Jones#0 data#1 Jones#0 data#1 Jones#0 data#1
- Given number of hidden topics, K=2
- The documents should be more about one topic than another (based on our prior belief)
- One topic should be about **sport** and the other about **software** and most words should be more focused on one topic than another
- Gilchrist may belong to both topics as suggested by data (although we know that Gilchrist was a sportsman)

Synthetic Example for TagLDA

- What should we expect from model.input?
  - doc0: Gilchrist#0 wicket#1 wicket#1 keeper#1 keeper#1
  - doc1: wicket#1 captain#0 captain#0 captain#0 keeper#1 keeper#1
  - doc2: Jones#0 data#1 architect#1 Gilchrist#0
  - doc3: Jones#0 data#1 Jones#0 data#1 Jones#0 data#1
- Given number of observed tags, T=2
- We know wicket, keeper, data, architect are important words (Word level domain knowledge)
- If Gilchrist were to be "important" based on the dominance of important words on his topic would he be more or less "important" than Jones?
- Jones should remain less important because his topic is conditioned on "importance" less than that of Gilchrist's

Synthetic Example for TagLDA

- **compute_initial_guesses()**
  - Randomly initializing τ_marginal-statistics in this case and obtaining $\log(\tau)$ as: [Note: $\log(\tau) = \{\log(\tau_{i,j})\}$]
  - Similarly for $\log(\pi)$

Log tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.1189	0.3112	-1.2743	0.4446	0.2119	0.3041	0.5725
Topic1:	0.1181	0.5153	-0.0278	0.2744	0.3097	0.0091	-0.4654

Log pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	-1.67219	-1.49886	-2.05742	-2.92477	-3.356	-1.50465	-1.89479
Tag1:	-3.11147	-2.12028	-2.81038	-1.41226	-1.54886	-1.42601	-2.53912

alpha = 0.5

Synthetic Example for TagLDA

- ## E-Step begin

Processing document 0: Gilchrist(0/0/0) wicket(1/1/1) wicket(1/2/1) keeper(3/3/1) keeper(3/4/1)

termID Position Tag

gamma initial = 3 3

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$

initialize $\phi^{0}_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^{0}_{i} = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
  for *n = 1* to *N* {
    for *i = 1* to *K*
      Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;
    normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
  }
  $\gamma^{t+1}_{i} = \gamma^{t}_{i} + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^{t}_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n = 1* to *N*
  Update $\pi^{marginal_{1,2}}_{t_{d,n}, w_{d,n}}$ according to Equ. TL-Er
  for *i = 1* to *K*
    Update $\tau^{marginal_{1,2}}_{i, w_{d,n}}$ according to Equ. TL-Er;
for *i = 1* to *K*
  $\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

gamma =	1.54437	4.45563
Phi =	0.20887	0.79112
	0.18003	0.81997
	0.17886	0.82113
	0.23915	0.76084
	0.23743	0.76256

Tau suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.2088	0.3588	0	0.4765	0	0	0
Topic1:	0.7911	1.6411	0	1.52341	0	0	0

Tau suff stats 2=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.0122	0.0456	0	0.0290	0	0	0
Topic1:	0.0463	0.2087	0	0.0928	0	0	0

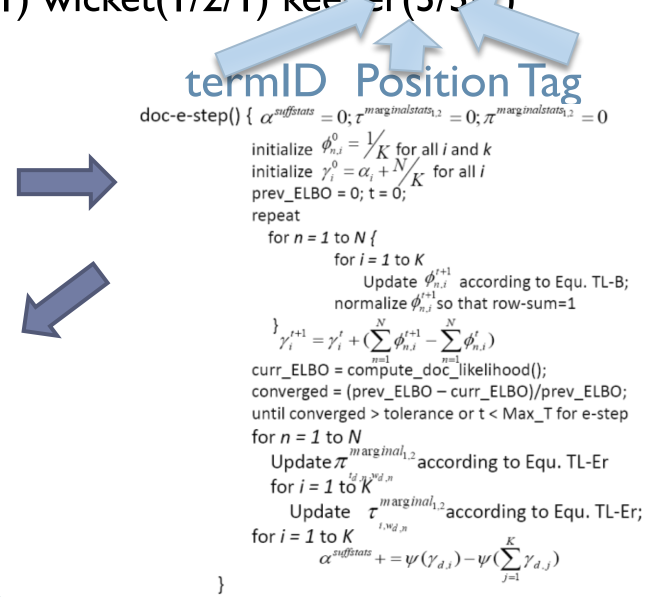
Synthetic Example for TagLDA

- **E-Step begin**

Processing document 0: Gilchrist(0/0/0) wicket(1/1/1) wicket(1/2/1) keeper(3/3/1) keeper(3/4/1)

gamma initial = 3 3

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5



gamma =	1.54437	4.45563
Phi =	0.20887	0.79112
	0.18003	0.81997
	0.17886	0.82113
	0.23915	0.76084
	0.23743	0.76256

Pi suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	1	0	0	0	0	0	0
Tag1:	0	2	0	2	0	0	0

Pi suff stats 2=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.1133	0	0	0	0	0	0
Tag1:	0	0.4076	0	0.3450	0	0	0

alpha suff stats = -1.95729

Synthetic Example for TagLDA

E-Step continue

Processing document 1: wicket(1/0/1) captain(2/1/0) captain(2/2/0) captain(2/3/0) keeper(3/4/1) keeper(3/5/1)

gamma initial = 3.5 3.5

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;
0.5 0.5; 0.5 0.5;

gamma =	0.67885	6.32115
Phi =	0.03696	0.96303
	0.01283	0.98716
	0.01280	0.98719
	0.01276	0.98723
	0.05197	0.94802
	0.05151	0.94848

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$
initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
for *n = 1* to *N* {
for *i = 1* to *K*
Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;
normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
}
$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n = 1* to *N*
Update $\pi^{marginal_{1,2}}_{t_{d,n},w_{d,n}}$ according to Equ. TL-Er
for *i = 1* to *K*
Update $\tau^{marginal_{1,2}}_{i,w_{d,n}}$ according to Equ. TL-Er;
for *i = 1* to *K*
$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

Tau suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.2088	0.3958	0.0384	0.5800	0	0	0
Topic1:	0.7911	2.6041	2.9616	3.4199	0	0	0

Tau suff stats 2=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.0122	0.0503	0.0060	0.0354	0	0	0
Topic1:	0.0463	0.3320	0.4657	0.2094	0	0	0

Synthetic Example for TagLDA

E-Step continue

Processing document 1: wicket(1/0/1) captain(2/1/0) captain(2/2/0) captain(2/3/0) keeper(3/4/1) keeper(3/5/1)

gamma initial = 3.5 3.5

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;

gamma =	0.67885	6.32115
Phi =	0.03696	0.96303
	0.01283	0.98716
	0.01280	0.98719
	0.01276	0.98723
	0.05197	0.94802
	0.05151	0.94848

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$

initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*

initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*

prev_ELBO = 0; t = 0;

repeat

for *n = 1* to *N* {

for *i = 1* to *K*

Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;

normalize $\phi^{t+1}_{n,i}$ so that row-sum=1

}

$$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$$

curr_ELBO = compute_doc_likelihood();

converged = (prev_ELBO – curr_ELBO)/prev_ELBO;

until converged > tolerance or t < Max_T for e-step

for *n = 1* to *N*

Update $\pi^{marginal_{1,2}}_{t_{d,n}, w_{d,n}}$ according to Equ. TL-Er

for *i = 1* to *K*

Update $\tau^{marginal_{1,2}}_{i, w_{d,n}}$ according to Equ. TL-Er;

for *i = 1* to *K*

$$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$$

}

Pi suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	1	0	3	0	0	0	0
Tag1:	0	3	0	4	0	0	0

Pi suff stats 2=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.1133	0	0.2957	0	0	0	0
Tag1:	0	0.6183	0	0.6815	0	0	0

alpha suff stats = -5.22157

Synthetic Example for TagLDA

E-Step continue

Processing document 2: Jones(4/0/0) data(5/1/1) architect(6/2/1) Gilchrist(0/3/0)

gamma initial = 2.5 2.5
Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;

gamma =	4.09844	0.90155
Phi =	0.86101	0.13898
	0.90549	0.09450
	0.95316	0.04683
	0.87875	0.12124

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$

initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
for *n = 1* to *N* {
for *i = 1* to *K*
Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;
normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
}
$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n = 1* to *N*
Update $\pi^{marginal_{1,2}}_{t_{d,n},w_{d,n}}$ according to Equ. TL-Er
for *i = 1* to *K*
Update $\tau^{marginal_{1,2}}_{i,w_{d,n}}$ according to Equ. TL-Er;
for *i = 1* to *K*
$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

Tau suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	1.0876	0.3958	0.0384	0.5800	0.8610	0.9054	0.9531
Topic1:	0.9123	2.6041	2.9616	3.4199	0.1389	0.0945	0.0468

Tau suff stats 2=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.0875	0.0503	0.0060	0.0354	0.0953	0.1961	0.08919
Topic1:	0.0567	0.3320	0.4657	0.2094	0.0153	0.0204	0.00438

Synthetic Example for TagLDA

E-Step continue

Processing document 2: Jones(4/0/0) data(5/1/1) architect(6/2/1) Gilchrist(0/3/0)

gamma initial = 2.5 2.5
Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;

gamma =	4.09844	0.90155
Phi =	0.86101	0.13898
	0.90549	0.09450
	0.95316	0.04683
	0.87875	0.12124

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$
initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
for *n* = *1* to *N* {
for *i* = *1* to *K*
Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;
normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
}
$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n* = *1* to *N*
Update $\pi^{marginal_{1,2}}_{t_{d,n}, w_{d,n}}$ according to Equ. TL-Er
for *i* = *1* to *K*
Update $\tau^{marginal_{1,2}}_{i, w_{d,n}}$ according to Equ. TL-Er;
for *i* = *1* to *K*
$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

Pi suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	1	0	3	0	1	0	0
Tag1:	1	3	0	4	0	1	1

Pi suff stats 2=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.1133	0	0.2957	0	0.1205	0	0
Tag1:	0.1374	0.6183	0	0.6815	0	0.1612	0.2091

alpha suff stats = -7.70207

Synthetic Example for TagLDA

- **E-Step continue**

Processing document 3: Jones(4/0/0) data(5/1/1) Jones(4/2/0) data(5/3/1) Jones(4/4/0) data(5/5/1)

gamma initial = 3.5 3.5

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5

gamma =	4.6866	2.3134
Phi =	0.65012	0.34987
	0.74024	0.25975
	0.65295	0.34704
	0.74261	0.25738
	0.65572	0.34427
	0.74493	0.25506

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$

initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*

initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*

prev_ELBO = 0; t = 0;

repeat

for *n* = 1 to *N* {

for *i* = 1 to *K*

Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;

normalize $\phi^{t+1}_{n,i}$ so that row-sum=1

}

$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$

curr_ELBO = compute_doc_likelihood();

converged = (prev_ELBO – curr_ELBO)/prev_ELBO;

until converged > tolerance or t < Max_T for e-step

for *n* = 1 to *N*

Update $\pi^{marginal_{1,2}}_{t_{d,n},w_{d,n}}$ according to Equ. TL-Er

for *i* = 1 to *K*

Update $\tau^{marginal_{1,2}}_{i,w_{d,n}}$ according to Equ. TL-Er;

for *i* = 1 to *K*

$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$

}

Tau suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	1.0876	0.3958	0.0384	0.5800	2.8198	3.1333	0.9531
Topic1:	0.9123	2.6041	2.9616	3.4199	1.1801	0.8667	0.0468

Tau suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.0875	0.0503	0.0060	0.0354	0.3155	0.6823	0.08919
Topic1:	0.0567	0.3320	0.4657	0.2094	0.1324	0.1890	0.00438

Synthetic Example for TagLDA

E-Step continue

Processing document 3: Jones(4/0/0) data(5/1/1) Jones(4/2/0) data(5/3/1) Jones(4/4/0) data(5/5/1)

gamma initial = 3.5 3.5

Phi initial = 0.5 0.5; 0.5 0.5; 0.5 0.5; 0.5 0.5;
0.5 0.5; 0.5 0.5

gamma =	4.6866	2.3134
Phi =	0.65012	0.34987
	0.74024	0.25975
	0.65295	0.34704
	0.74261	0.25738
	0.65572	0.34427
	0.74493	0.25506

doc-e-step() { $\alpha^{suffstats} = 0; \tau^{marginalstats_{1,2}} = 0; \pi^{marginalstats_{1,2}} = 0$
initialize $\phi^0_{n,i} = 1/K$ for all *i* and *k*
initialize $\gamma^0_i = \alpha_i + N/K$ for all *i*
prev_ELBO = 0; t = 0;
repeat
for *n = 1* to *N* {
for *i = 1* to *K*
Update $\phi^{t+1}_{n,i}$ according to Equ. TL-B;
normalize $\phi^{t+1}_{n,i}$ so that row-sum=1
}
$\gamma^{t+1}_i = \gamma^t_i + (\sum_{n=1}^{N} \phi^{t+1}_{n,i} - \sum_{n=1}^{N} \phi^t_{n,i})$
curr_ELBO = compute_doc_likelihood();
converged = (prev_ELBO – curr_ELBO)/prev_ELBO;
until converged > tolerance or t < Max_T for e-step
for *n = 1* to *N*
Update $\pi^{marginal_{1,2}}_{t_{d,n},w_{d,n}}$ according to Equ. TL-Er
for *i = 1* to *K*
Update $\tau^{marginal_{1,2}}_{i,w_{d,n}}$ according to Equ. TL-Er;
for *i = 1* to *K*
$\alpha^{suffstats} += \psi(\gamma_{d,i}) - \psi(\sum_{j=1}^{K} \gamma_{d,j})$
}

Pi suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	1	0	3	0	4	0	0
Tag1:	1	3	0	4	0	4	1

Pi suff stats 1=

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.1133	0	0.2957	0	0.4953	0	0
Tag1:	0.1374	0.6183	0	0.6815	0	0.6278	0.2091

alpha suff stats = -9.40611

Synthetic Example for TagLDA

- **E-Step end**
- **M-Step begin**

Log Tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.6932	0.1897	-0.9520	0.3976	1.1176	1.072	0.6235
Topic1:	0.6176	1.0689	1.1010	1.2513	0.7150	0.5202	-0.8475

Log Pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.6364	-100	1.15719	-100	1.26278	-100	-100
Tag1:	0.6244	1.0648	-100	1.2167	-100	1.2299	0.5887

alpha = 0.678869

log_likelihood= -44.89721451 : converged = inf em_steps = 1 (out of 50)

TopicHeap0:	Jones	data	Gilchrist	architect	keeper	wicket	captain
TopicHeap1:	keeper	captain	wicket	Jones	Gilchrist	data	architect
TagHeap0:	Jones	captain	Gilchrist	wicket	keeper	data	architect
TagHeap1:	data	keeper	wicket	Gilchrist	architect	captain	Jones

- **M-Step end**

Synthetic Example for TagLDA

- **Iteration 2**
- **E-Step begin … E-Step end / M-Step begin**

Log Tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.6885	-0.2695	-1.0752	-0.1760	1.1497	1.2356	0.6473
Topic1:	0.6108	1.13506	1.1025	1.2530	-0.2226	-0.2289	-1.2173

Log Pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.6452	-100	1.1153	-100	1.19934	-100	-100
Tag1:	0.6549	1.1609	-100	1.2732	-100	1.2753	0.6575

alpha = 0.52364

log_likelihood= -26.50176340 : converged = 0.40972 em_steps = 2 (out of 50)

TopicHeap0:	data	Jones	Gilchrist	architect	keeper	wicket	captain
TopicHeap1:	keeper	wicket	captain	Gilchrist	Jones	data	architect
TagHeap0:	Jones	captain	Gilchrist	wicket	keeper	data	architect
TagHeap1:	data	keeper	wicket	architect	Gilchrist	captain	Jones

- **M-Step end**

Synthetic Example for TagLDA

- **Iteration 3**
- **E-Step begin … E-Step end / M-Step begin**

Log Tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.6587	-0.9884	-1.5983	-0.8960	1.1676	1.2421	0.6516
Topic1:	0.6275	1.1430	1.0697	1.2602	-0.9764	-1.0936	-1.7940

Log Pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.6329	-100	1.0739	-100	1.1829	-100	-100
Tag1:	0.6541	1.1516	-100	1.2682	-100	1.2514	0.6556

alpha = 0.35718

log_likelihood= -23.45330186 : converged = 0.11503 em_steps = 3 (out of 50)

TopicHeap0:	data	Jones	Gilchrist	architect	keeper	wicket	captain
TopicHeap1:	keeper	wicket	captain	Gilchrist	Jones	data	architect
TagHeap0:	Jones	captain	Gilchrist	wicket	keeper	data	architect
TagHeap1:	keeper	data	wicket	architect	Gilchrist	captain	Jones

- **M-Step end**

Synthetic Example for TagLDA

- **Iteration 4**
- **E-Step begin … E-Step end / M-Step begin**

Log Tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.6495	-1.9324	-2.4608	-1.8273	1.1657	1.228	0.6491
Topic1:	0.6228	1.1423	1.0526	1.2568	-1.8753	-2.1385	-2.7003

Log Pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.6228	-100	1.0537	-100	1.1699	-100	-100
Tag1:	0.6497	1.1441	-100	1.2587	-100	1.2307	0.6499

alpha = 0.247851

log_likelihood= -21.21241265 : converged = 0.09555 em_steps = 4 (out of 50)

TopicHeap0:	data	Jones	Gilchrist	architect	keeper	wicket	captain
TopicHeap1:	keeper	wicket	captain	Gilchrist	Jones	data	architect
TagHeap0:	Jones	captain	Gilchrist	wicket	keeper	data	architect
TagHeap1:	keeper	data	wicket	architect	Gilchrist	captain	Jones

- **M-Step end**

Synthetic Example for TagLDA

- **Iteration 5**
- **E-Step begin … E-Step end / M-Step begin**

Log Tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.6453	-3.0959	-3.5984	-2.9767	1.1604	1.2163	0.6453
Topic1:	0.6168	1.1392	1.0399	1.2528	-2.9986	-3.3860	-3.9017

Log Pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.6167	-100	1.0402	-100	1.1614	-100	-100
Tag1:	0.6454	1.1395	-100	1.2531	-100	1.2167	0.6454

alpha = 0.181814

log_likelihood= -19.68089582 : converged = 0.07220 em_steps = 5 (out of 50)

TopicHeap0:	data	Jones	architect	Gilchrist	keeper	wicket	captain
TopicHeap1:	keeper	wicket	captain	Gilchrist	Jones	data	architect
TagHeap0:	Jones	captain	Gilchrist	wicket	keeper	data	architect
TagHeap1:	keeper	data	wicket	Gilchrist	architect	captain	Jones

- **M-Step end**

Synthetic Example for TagLDA

- **Iteration 27**
- **E-Step begin … E-Step end / M-Step begin**

Log Tau =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Topic0:	0.6414	-44.2895	-44.8044	-44.1473	1.1544	1.2054	0.6414
Topic1:	0.6110	1.1363	1.0281	1.2492	-44.1279	-44.7341	-45.2408

Log Pi =

	Gilchrist	wicket	captain	keeper	Jones	data	architect
Tag0:	0.6110	-100	1.0281	-100	1.1544	-100	-100
Tag1:	0.6414	1.1363	-100	1.2492	-100	1.2054	0.6414

alpha = 0.020894

log_likelihood = -17.49778259 : converged = 0.00047 em_steps = 27 (out of 50)

Ties broken arbitrarily

TopicHeap0:	data	Jones	Gilchrist	architect	keeper	wicket	captain
TopicHeap1:	keeper	wicket	captain	Gilchrist	Jones	data	architect
TagHeap0:	Jones	captain	Gilchrist	wicket	keeper	data	architect
TagHeap1:	keeper	data	wicket	Gilchrist	architect	captain	Jones

- **M-Step end**

References

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. JMLR, 3:993–1022, 2003
- Xiaojin Zhu, David Blei, and John Lafferty. Taglda: Bringing document structure knowledge into topic models. UWisc Technical Report TR-1533, 2006