

# Вероятностные тематические модели

## Лекция 6. Байесовский вывод для модели LDA

К. В. Воронцов  
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса  
<http://www.MachineLearning.ru/wiki>  
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ – ФИЦ ИУ РАН • 16 октября 2019

## 1 Вариационный байесовский вывод

- Байесовский вывод в тематическом моделировании
- Вариационный байесовский вывод для модели LDA
- VB EM-алгоритм для модели LDA

## 2 Сэмплирование Гиббса

- Основная теорема о сэмплировании Гиббса
- Сэмплирование Гиббса для модели LDA
- GS EM-алгоритм для модели LDA

## 3 Замечания о байесовском подходе

- Оптимизация гиперпараметров в LDA
- Графическая нотация
- Сравнение байесовского подхода и ARTM

## Напоминание. Задача тематического моделирования

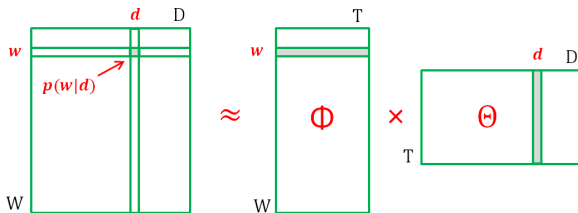
**Дано:** коллекция текстовых документов

- $n_{dw}$  — частоты термов в документах,  $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

**Найти:** параметры тематической модели  $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$

- $\phi_{wt} = p(w|t)$  — вероятности термов  $w$  в каждой теме  $t$
- $\theta_{td} = p(t|d)$  — вероятности тем  $t$  в каждом документе  $d$

Это задача стохастического матричного разложения:



## Напоминание. Тематическая модель LDA

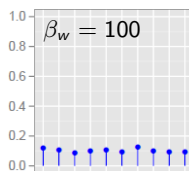
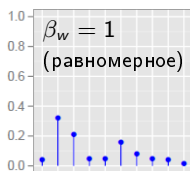
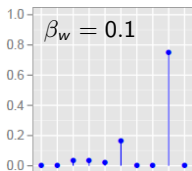
Распределения Дирихле с гиперпараметрами  $\alpha \in \mathbb{R}^T$ ,  $\beta \in \mathbb{R}^W$ :

$$p(\Phi, \Theta | \beta, \alpha) = \prod_{t \in T} \text{Dir}(\phi_t | \beta) \prod_{d \in D} \text{Dir}(\theta_d | \alpha)$$

$$\text{Dir}(\phi_t | \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \quad \beta_t > 0;$$

$$\text{Dir}(\theta_d | \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \quad \alpha_t > 0;$$

**Пример.** Распределение  $\text{Dir}(\phi | \beta)$  при  $|W| = 10$ ,  $\phi, \beta \in \mathbb{R}^{10}$ :



## Свойства распределения Дирихле

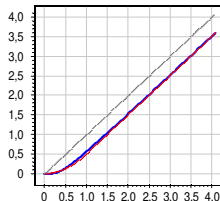
- Математическое ожидание:  

$$E\theta_t = \int \theta_t \text{Dir}(\theta|\alpha) d\theta = \frac{\alpha_t}{\alpha_0} = \text{norm}_t(\alpha_t)$$
- Мода:  $\hat{\theta}_t = \frac{\alpha_t - 1}{\alpha_0 - T} = \text{norm}_t(\alpha_t - 1)$
- Дисперсия:  $D\theta_t = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0 + 1)}$
- Математическое ожидание логарифма:  

$$E \ln \theta_t = \int \ln \theta_t \text{Dir}(\theta|\alpha) d\theta = \psi(\alpha_t) - \psi(\alpha_0),$$
 где  $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$  — дигамма-функция.

Простая, но очень точная аппроксимация  
 экспоненты от дигамма-функции:

$$E(x) = \exp(\psi(x)) = \begin{cases} \frac{x^2}{2}, & 0 \leq x \leq 1 \\ x - \frac{1}{2}, & 1 \leq x \end{cases}$$



## Подходы к оцениванию параметров вероятностных моделей

$X = (d_i, w_i)_{i=1}^n$  — наблюдаемые,  $Z = (t_i)_{i=1}^n$  — скрытые  
 $\Omega = (\Phi, \Theta)$  — параметры,  $\gamma = (\beta, \alpha)$  — гиперпараметры

**Максимизация регуляризованного правдоподобия:**

$$\ln p(X|\Omega) + R(\Omega) = \ln \sum_Z p(X, Z|\Omega) + R(\Omega) \rightarrow \max_{\Omega}$$

**Максимизация апостериорной вероятности:**

$$\ln p(\Omega|X, \gamma) + \text{const} = \ln \sum_Z p(X, Z|\Omega) + \ln p(\Omega|\gamma) \rightarrow \max_{\Omega}$$

**Вариационный байесовский вывод:**

вывести  $p(Z, \Omega|X, \gamma)$

**Сэмплирование Гиббса:**

$Z \sim p(Z|X, \gamma)$

вывести  $p(\Omega|X, Z, \gamma)$

## Основная идея Variational Bayesian inference

$X = (d_i, w_i)_{i=1}^n$  — исходные данные, *наблюдаемые переменные*

$Z = (t_i)_{i=1}^n$  — *скрытые переменные*

$p(\Phi|\beta) = \prod_{t \in T} \text{Dir}(\phi_t|\beta)$  — априорное распределение

$p(\Theta|\alpha) = \prod_{d \in D} \text{Dir}(\theta_d|\alpha)$  — априорное распределение

**Задача:** найти апостериорное распределение  $p(Z, \Phi, \Theta|X, \beta, \alpha)$ .

**Основная идея:** найти его приближение в виде произведения  $n + |T| + |D|$  распределений по блокам переменных  $t_i, \phi_t, \theta_d$ :

$$q(Z, \Phi, \Theta) = \prod_{i=1}^n q_i(t_i) \prod_{t \in T} q_t(\phi_t) \prod_{d \in D} q_d(\theta_d)$$

## Основная теорема вариационного байесовского вывода

**Теорема.** Решение задачи  $\text{KL}(q(Y) \parallel p(Y|X, \Omega)) \rightarrow \min_q$  в семействе факторизованных распределений  $q(Y) = \prod_j q_j(Y_j)$  по переменным  $Y_j, j \in J$ , удовлетворяет системе уравнений

$$\ln q_j(Y_j) = E_{q_{\setminus j}} \ln p(X, Y|\Omega) + \text{const},$$

где  $E_{q_{\setminus j}}$  — матожидание по всем переменным кроме  $Y_j$ ,  
 const — нормировочный множитель распределения  $q_j$ .

Для решения этой системы используют метод простой итерации.

**Идея доказательства:** расписываем  $\text{KL}(\cdot \parallel \cdot)$  и сводим задачу к

$$\sum_{Y_j} q_j(Y_j) \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \ln p(X, Y|\Omega)}_{E_{q_{\setminus j}} \ln p(X, Y|\Omega)} - \sum_{Y_j} q_j(Y_j) \ln q_j(Y_j) \rightarrow \min_q$$



## Доказательство

1. В оптимизационной задаче можно перекидывать  $X$  через условную черту:

$$\sum_Y q(Y) \ln \frac{p(Y|X, \Omega)}{q(Y)} \rightarrow \max_q \Leftrightarrow \sum_Y q(Y) \ln \frac{p(X, Y|\Omega)}{q(Y)} - \sum_Y q(Y) \ln p(X|\Omega) \rightarrow \max_q$$

2. Будем минимизировать KL-дивергенцию поочередно по всем  $Y_j$ .

Применим факторизацию и вынесем слагаемое с  $q_j(Y_j)$  вперёд:

$$\sum_{Y_j} q_j(Y_j) \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \ln p(X, Y|\Omega)}_{E_{q_{\setminus j}} \ln p(X, Y|\Omega)} - \sum_{Y_j} q_j(Y_j) \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \sum_{k \in J} \ln q_k(Y_k)}_{\ln q_j(Y_j) + \text{const}} \rightarrow \max_{q_j}$$

3. Почему вторую фигурную скобку можно заменить на  $\ln q_j(Y_j)$ :

$$\underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \sum_{k \neq j} \ln q_k(Y_k)}_{\text{не зависит от } q_j} + \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \ln q_j(Y_j)}_1$$

4. Введём  $r(Y_j) \propto \exp(E_{q_{\setminus j}} \ln p(X, Y|\Omega))$ , тогда  $\text{KL}(q_j(Y_j) \| r(Y_j)) \rightarrow \min_{q_j}$

5. Точное решение данной задачи  $q_j(Y_j) = r(Y_j)$ , следовательно,

$$\ln q_j(Y_j) = E_{q_{\setminus j}} \ln p(X, Y|\Omega) + \text{const.}$$

## Основная теорема для частного случая модели LDA

Положим  $Y = (Z, \Phi, \Theta)$ ,  $\gamma = (\beta, \alpha)$ ,  $J = \{1, \dots, n\} \sqcup T \sqcup D$ :

$$\ln q_j = E_{q_{\setminus j}} \ln p(X, Z, \Phi, \Theta | \beta, \alpha) + \text{const}$$

Нам предстоит брать матожидания  $E_{q_{\setminus j}}$  по всем распределениям  $q_t(\phi_t)$ ,  $q_d(\theta_d)$ ,  $q_i(t_i)$ , кроме  $j$ -го, от

$$\begin{aligned} \ln p(X, Z, \Phi, \Theta | \beta, \alpha) &= \ln p(X, Z | \Phi, \Theta) p(\Phi | \beta) p(\Theta | \alpha) = \\ &= \ln \prod_{i=1}^n p(d_i, w_i, t_i | \Phi, \Theta) + \ln \prod_{t \in T} \text{Dir}(\phi_t | \beta) + \ln \prod_{d \in D} \text{Dir}(\theta_d | \alpha) = \\ &= \sum_{i=1}^n \ln \phi_{w_i t_i} \theta_{t_i d_i} + \sum_{t, w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d, t} (\alpha_t - 1) \ln \theta_{td} + \text{const}. \end{aligned}$$

**Замечание**, сильно упрощающее выкладки:

если слагаемое  $S$  не зависит от  $j$ -й переменной, то  $E_{q_j} S = \text{const}$ .

## Распределения для блока переменных $q_t(\phi_t)$

Уравнение для распределения переменной  $\phi_t \in \mathbb{R}^W$ :

$$\begin{aligned}
 \ln q_t(\phi_t) &= \sum_{i=1}^n \mathbb{E}_{q_i(t_i)}[t_i = t] \ln \phi_{w_i t_i} + \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \text{const} = \\
 &= \sum_{i=1}^n \sum_{w \in W} [w_i = w] q_i(t) \ln \phi_{wt} + \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \text{const} = \\
 &= \sum_{w \in W} \left( \underbrace{\sum_{i=1}^n [w_i = w] q_i(t)}_{n_{wt}} + \beta_w - 1 \right) \ln \phi_{wt} + \text{const} = \\
 &= \ln \text{Dir}(\phi_t | \tilde{\beta}_t).
 \end{aligned}$$

Это распределение Дирихле с параметрами  $\tilde{\beta}_{wt} = n_{wt} + \beta_w$ ,  
 $n_{wt}$  — оценка числа генераций термина  $w$  из темы  $t$ .

При больших  $n_{wt}$  оно сконцентрировано в точке  $\phi_{wt} = \text{norm}_w(\tilde{\beta}_{wt})$ .

## Распределения для блока переменных $q_d(\theta_d)$

Уравнение для распределения переменной  $\theta_d \in \mathbb{R}^T$ :

$$\begin{aligned}
 \ln q_d(\theta_d) &= \sum_{i=1}^n \mathbb{E}_{q_i(t_i)}[d_i = d] \ln \theta_{t_i d_i} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \text{const} = \\
 &= \sum_{i=1}^n [d_i = d] \sum_{t \in T} q_i(t) \ln \theta_{td} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \text{const} = \\
 &= \sum_{t \in T} \left( \underbrace{\sum_{i=1}^n [d_i = d] q_i(t)}_{n_{td}} + \alpha_t - 1 \right) \ln \theta_{td} + \text{const} = \\
 &= \ln \text{Dir}(\theta_d | \tilde{\alpha}_d).
 \end{aligned}$$

Это распределение Дирихле с параметрами  $\tilde{\alpha}_{td} = n_{td} + \alpha_t$ ,  
 $n_{td}$  — оценка числа термов темы  $t$  в документе  $d$ .

При больших  $n_{td}$  оно сконцентрировано в точке  $\theta_{td} = \text{norm}_t(\tilde{\alpha}_{td})$ .

## Распределения для блока переменных $q_i(t_i)$

Уравнение для распределения переменной  $t_i \in T$ :

$$\begin{aligned} \ln q_i(t) &= E_{q_i}(\ln \phi_{w_i t_i} + \ln \theta_{t_i d_i}) + \text{const} = \\ &= E_{q_t(\phi_t)} \ln \phi_{w_i t} + E_{q_d(\theta_d)} \ln \theta_{t_i d} + \text{const} = \end{aligned}$$

воспользуемся тем, что  $q_t(\phi_t)$  и  $q_d(\theta_d)$  уже найдены:

$$\begin{aligned} &= \psi(n_{w_i t} + \beta_{w_i}) - \psi(\sum_w (n_{w t} + \beta_w)) + \\ &\quad + \psi(n_{t d_i} + \alpha_t) - \psi(\sum_t (n_{t d_i} + \alpha_t)) + \text{const} \end{aligned}$$

Воспользуемся приближением  $\exp(\psi(x)) \approx x - \frac{1}{2}$ :

$$q_i(t) = \text{norm}_{t \in T} \left( \frac{n_{w_i t} + \beta_{w_i} - \frac{1}{2}}{\sum_w (n_{w t} + \beta_w) - \frac{1}{2}} \cdot \frac{n_{t d_i} + \alpha_t - \frac{1}{2}}{\sum_t (n_{t d_i} + \alpha_t) - \frac{1}{2}} \right)$$

Похоже на обычную формулу E-шага  $p(t|d_i, w_i) = \text{norm}_{t \in T}(\phi_{w_i t} \theta_{t d_i})$

## Собираем всё воедино

В итерационном процессе чередуются два шага:

1) распределение термов  $(d_i, w_i)$  по темам,  $E(x) = \exp(\psi(x))$ :

$$q_i(t) = \operatorname{norm}_{t \in T} \left( \frac{E(n_{w_i t} + \beta_{w_i})}{E(\sum_w (n_{wt} + \beta_w))} \cdot \frac{E(n_{td_i} + \alpha_t)}{E(\sum_t (n_{td_i} + \alpha_t))} \right)$$

2) аккумулярование счётчиков  $n_{wt}$  и  $n_{td}$ :

$$n_{wt} = \sum_{i=1}^n [w_i = w] q_i(t) \quad n_{td} = \sum_{i=1}^n [d_i = d] q_i(t)$$

Точечные оценки параметров по матожиданию или моде:

$$\begin{aligned} E\phi_{wt} &= \operatorname{norm}_{w \in W} (n_{wt} + \beta_w) & E\theta_{td} &= \operatorname{norm}_{t \in T} (n_{td} + \alpha_t) \\ \hat{\phi}_{wt} &= \operatorname{norm}_{w \in W} (n_{wt} + \beta_w - 1) & \hat{\theta}_{td} &= \operatorname{norm}_{t \in T} (n_{td} + \alpha_t - 1) \end{aligned}$$

## Промежуточный итог

- Из-за факторизации вариационный байесовский вывод даёт лишь приближённое решение, тем не менее,
- формулы для MAP и VB очень похожи [Asuncion]:
  - при  $n_{wt}, n_{td} \gg 1$  различия неощутимы,
  - при  $n_{wt}, n_{td} \lesssim 1$  тема  $t$  незначима для  $w$  или  $d$ .
- Можно добавить M-шаг для оптимизации  $\beta, \alpha$  [Wallach].
- Некуда добавлять регуляризаторы  $R(\Phi, \Theta)$ .
- Нужны матрицы  $\Phi, \Theta$ , а не распределения  $p(\Phi, \Theta|X)$ .
- Начинает смущать разнообразие оценок... какая лучше?

---

*Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models. Int'l conf. on Uncertainty in Artificial Intelligence, 2009.

*Hanna Wallach, David Mimno, Andrew McCallum.* Rethinking LDA: why priors matter. Neural Information Processing Systems, 2009.

## Сэмплирование Гиббса (Gibbs Sampling)

Основная идея:

- $Z \sim p(Z|X, \beta, \alpha)$  — сэмплировать скрытые переменные
- $p(\Phi, \Theta|X, Z, \beta, \alpha)$  — найти при известных  $X$  и  $Z$

### Основная теорема о сходимости сэмплирования Гиббса

Процесс сэмплирования одномерных случайных величин

$$t_i^{(k+1)} \sim p(t_i|X, Z_{\setminus i}, \Omega) = \frac{p(X, Z|\Omega)}{p(X, Z_{\setminus i}|\Omega)},$$

где  $k$  — номер итерации,  $Z_{\setminus i} = (t_1^{(k+1)}, \dots, t_{i-1}^{(k+1)}, t_{i+1}^{(k)}, \dots, t_n^{(k)})$ ,  
сходится к многомерному распределению  $Z \sim p(Z|X, \Omega)$



## Распределение Дирихле — сопряжённое к мультиномиальному

$p(\Phi, \Theta | \beta, \alpha)$  — априорное распределение Дирихле

$p(\Phi, \Theta | X, Z, \beta, \alpha)$  — апостериорное распределение тоже Дирихле

Вывод апостериорного распределения  $\Phi, \Theta$  при известных  $X, Z$ :

$$p(\Phi, \Theta | X, Z, \beta, \alpha) \propto p(\Phi, \Theta, X, Z | \beta, \alpha) \propto p(X, Z | \Phi, \Theta) p(\Phi, \Theta | \beta, \alpha)$$

$$\propto \prod_{d,w,t} (\phi_{wt} \theta_{td})^{n_{dwt}} \prod_{t \in T} \text{Dir}(\phi_t | \beta) \prod_{d \in D} \text{Dir}(\theta_d | \alpha)$$

$$\propto \prod_{t \in T} \prod_{d,w} \phi_{wt}^{n_{dwt}} \phi_{wt}^{\beta_w - 1} \prod_{d \in D} \prod_{w,t} \theta_{td}^{n_{dwt}} \theta_{td}^{\alpha_t - 1}$$

$$\propto \prod_{t \in T} \prod_w \phi_{wt}^{n_{wt} + \beta_w - 1} \prod_{d \in D} \prod_t \theta_{td}^{n_{td} + \alpha_t - 1}$$

$$\propto \prod_{t \in T} \text{Dir}(\phi_t | \tilde{\beta}_t) \prod_{d \in D} \text{Dir}(\theta_d | \tilde{\alpha}_d), \quad \tilde{\beta}_{wt} = n_{wt} + \beta_w, \quad \tilde{\alpha}_{td} = n_{td} + \alpha_t.$$

## Распределение $p(X, Z|\beta, \alpha)$ для схемы сэмплирования Гиббса

Подынтегральное распределение мы только что вывели, но теперь будем аккуратнее с нормировочными множителями:

$$\begin{aligned}
 p(X, Z|\beta, \alpha) &= \int_{\Phi} \int_{\Theta} p(X, Z|\Phi, \Theta) p(\Phi, \Theta|\beta, \alpha) d\Phi d\Theta = \\
 &= \int_{\Phi} \int_{\Theta} \prod_{w,t} \phi_{wt}^{n_{wt}} \prod_{t,d} \theta_{td}^{n_{td}} \prod_d p_d^{n_d} \prod_{t \in T} \text{Dir}(\phi_t|\beta) \prod_{d \in D} \text{Dir}(\theta_d|\alpha) d\Phi d\Theta = \\
 &= \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \int_{\phi_t} \underbrace{\prod_w \phi_{wt}^{\tilde{\beta}_{wt}-1} d\phi_t}_{\propto \text{Dir}(\phi_t|\tilde{\beta}_t)} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \int_{\theta_d} \underbrace{\prod_t \theta_{td}^{\tilde{\alpha}_{td}-1} d\theta_d}_{\propto \text{Dir}(\theta_d|\tilde{\alpha}_d)} = \\
 &= \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde{\beta}_{wt})}{\Gamma(\sum_w \tilde{\beta}_{wt})} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde{\alpha}_{td})}{\Gamma(\sum_t \tilde{\alpha}_{td})}
 \end{aligned}$$

## Распределение $p(X, Z_{\setminus i} | \beta, \alpha)$ для схемы сэмплирования Гиббса

Итак, мы только что получили распределение

$$p(X, Z | \beta, \alpha) = \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w) \prod_w \Gamma(\tilde{\beta}_{wt})}{\prod_w \Gamma(\beta_w) \Gamma(\sum_w \tilde{\beta}_{wt})} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t) \prod_t \Gamma(\tilde{\alpha}_{td})}{\prod_t \Gamma(\alpha_t) \Gamma(\sum_t \tilde{\alpha}_{td})}$$

Распределение  $p(X, Z_{\setminus i} | \beta, \alpha)$  отличается от него лишь тем, что оно построено по выборке без одной  $i$ -й точки  $(d_i, w_i, t_i)$ :

$$p(X, Z_{\setminus i} | \beta, \alpha) = \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w) \prod_w \Gamma(\tilde{\beta}_{wt} - \delta_{wt}^i)}{\prod_w \Gamma(\beta_w) \Gamma(\sum_w (\tilde{\beta}_{wt} - \delta_{wt}^i))} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t) \prod_t \Gamma(\tilde{\alpha}_{td} - \delta_{td}^i)}{\prod_t \Gamma(\alpha_t) \Gamma(\sum_t (\tilde{\alpha}_{td} - \delta_{td}^i))}$$

где  $\delta_{wt}^i = [w = w_i][t = t_i]$ ,  $\delta_{td}^i = [t = t_i][d = d_i]$

## Ещё чуть-чуть... осталось поделить одно на другое

Для сэмпирования Гиббса нужно одномерное распределение

$$p(t_i|X, Z_{\setminus i}, \beta, \alpha) = \frac{p(X, Z|\beta, \alpha)}{p(X, Z_{\setminus i}|\beta, \alpha)} =$$

В числителе и знаменателе сократятся все множители кроме  $i$ -х:

$$= \frac{\Gamma(n_{w_i t_i} + \beta_{w_i}) \Gamma(\sum_w (n_{wt_i} + \beta_w) - 1) \Gamma(n_{t_i d_i} + \alpha_{t_i}) \Gamma(\sum_t (n_{td_i} + \alpha_t) - 1)}{\Gamma(n_{w_i t_i} + \beta_{w_i} - 1) \Gamma(\sum_w (n_{wt_i} + \beta_w)) \Gamma(n_{t_i d_i} + \alpha_{t_i} - 1) \Gamma(\sum_t (n_{td_i} + \alpha_t))}$$

Воспользуемся свойством гамма-функции  $\frac{\Gamma(x)}{\Gamma(x-1)} = x - 1$ :

$$p(t_i|X, Z_{\setminus i}, \beta, \alpha) = \text{norm}_{t \in T} \left( \frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{wt} + \beta_w) - 1} \cdot \frac{n_{t d_i} + \alpha_t - 1}{\sum_t (n_{td_i} + \alpha_t) - 1} \right)$$

Похоже на обычную формулу E-шага  $p(t|d_i, w_i) = \text{norm}_{t \in T} (\phi_{w_i t} \theta_{t d_i})$

## Собираем всё воедино

Выделены отличия от вариационного алгоритма

1) для каждого  $(d_i, w_i)$ ,  $i = 1, \dots, n$ , сэмпирование темы  $t_i$ :

$$t_i \sim p_i(t) = \operatorname{norm}_{t \in T} \left( \frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{w t} + \beta_w) - 1} \cdot \frac{n_{t d_i} + \alpha_t - 1}{\sum_t (n_{t d_i} + \alpha_t) - 1} \right)$$

2) аккумулярование счётчиков  $n_{wt}$  и  $n_{td}$ :

$$n_{wt} = \sum_{i=1}^n [w_i = w][t_i = t] \quad n_{td} = \sum_{i=1}^n [d_i = d][t_i = t]$$

Точечные оценки параметров по матожиданию или моде:

$$\begin{aligned} E\phi_{wt} &= \operatorname{norm}_{w \in W} (n_{wt} + \beta_w) & E\theta_{td} &= \operatorname{norm}_{t \in T} (n_{td} + \alpha_t) \\ \hat{\phi}_{wt} &= \operatorname{norm}_{w \in W} (n_{wt} + \beta_w - 1) & \hat{\theta}_{td} &= \operatorname{norm}_{t \in T} (n_{td} + \alpha_t - 1) \end{aligned}$$

## Алгоритм сэмплингования Гиббса

**Вход:** коллекция  $D$ , число тем  $|T|$ , параметры  $\alpha, \beta$ ;

**Выход:** распределения  $\Phi$  и  $\Theta$ ;

$n_{wt}, n_{td}, n_t, n_d := 0$  для всех  $d \in D, w \in W, t \in T$ ;

**для всех** итераций  $k := 1, \dots, k_{\max}$

**для всех**  $i = 1, \dots, n$  взять документ  $d := d_i$ , терм  $w := w_i$ ;

**если**  $k \geq 2$  **то**  $t := t_i$ ; -- $n_{wt}$ ; -- $n_{td}$ ; -- $n_t$ ; -- $n_d$ ;

$p(t|d, w) = \operatorname{norm}_{t \in T} \left( \frac{n_{wt} + \beta_w}{n_t + \beta_0} \cdot \frac{n_{td} + \alpha_t}{n_d + \alpha_0} \right)$  для всех  $t \in T$ ;

сэмплировать тему  $t$  из распределения  $p(t|d, w)$ ;

$t_i := t$ ; ++ $n_{wt}$ ; ++ $n_{td}$ ; ++ $n_t$ ; ++ $n_d$ ;

$\phi_{wt} := n_{wt}/n_t$  для всех  $w \in W, t \in T$ ;

$\theta_{td} := n_{td}/n_d$  для всех  $d \in D, t \in T$ ;

## Промежуточный итог

- Похожий алгоритм получится в ARTM, если на E-шаге вместо  $p(t|d, w)$  брать  $\hat{p}(t|d, w) = [t = t_i]$ ,  $t_i \sim p(t|d, w)$ .
- Формулы для MAP, VB и GS очень похожи [Asuncion]:
  - при  $n_{wt}, n_{td} \gg 1$  различия неощутимы,
  - при  $n_{wt}, n_{td} \lesssim 1$  тема  $t$  незначима для  $w$  или  $d$ .
- Необходимость задания априорных распределений:
  - сопряжённые — только распределения Дирихле,
  - не сопряжённые — сильно усложняют задачу.
- VB и GS не имеют удобных механизмов регуляризации, т.к. нет, собственно, и задачи оптимизации по  $(\Phi, \Theta)$
- Проблема неустойчивости даже не ставится.

---

*Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l conf. on Uncertainty in Artificial Intelligence, 2009.*

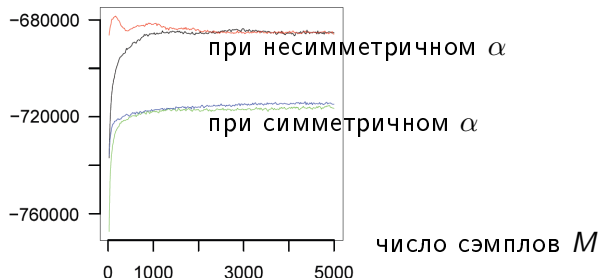
## Проблема выбора гиперпараметров $\alpha$ и $\beta$

Стандартная рекомендация [2004]:  $\alpha_t = 50/|T|$ ,  $\beta_w = 0.01$ .

Выводы по результатам более тонкого исследования [2009]:

- $p(t|d) \sim \text{Dir}(\theta; \alpha)$ , оптимизировать  $\alpha = (\alpha_1, \dots, \alpha_T)$ .
- $p(w|t) \sim \text{Dir}(\phi; \beta)$ , взять симметричное  $\beta_1 = \dots = \beta_T \ll 1$ .

правдоподобие



*H. Wallach, D. Mimno, A. McCallum. Rethinking LDA: why priors matter. NIPS, 2009.*



## Оптимизация гиперпараметра $\alpha$

Обоснованность (evidence) модели на коллекции  $D$ :

$$P(D|\alpha) = \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \rightarrow \max_{\alpha}$$

Метод неподвижной точки [Minka, 2003] — итерационный процесс, встраиваемый между проходами по всей коллекции:

$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

где  $\psi(z) = (\ln \Gamma(z))' = \Gamma'(z)/\Gamma(z)$  — дигамма-функция.

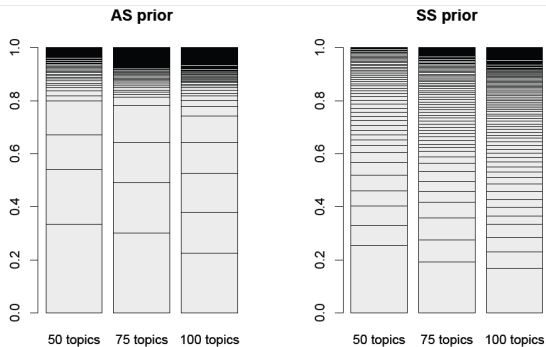
---

*Thomas Minka.* Estimating a Dirichlet distribution. 2003.

*Hanna Wallach.* Structured Topic Models for Language. PhD thesis, University of Cambridge, 2008.

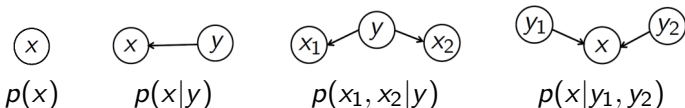
## Преимущество оптимизации гиперпараметра $\alpha$

- Правдоподобие существенно выше.
- Сходимость быстрее.
- Меньшая чувствительность к избыточному  $|T|$ .
- Более естественная несбалансированность тем.

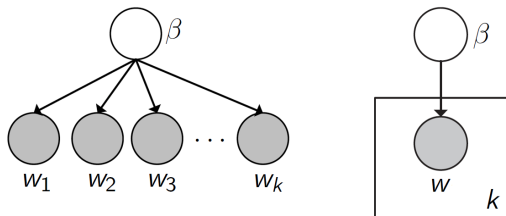


## Графическая нотация «plate notation»

Графическое представление условных зависимостей



Графическое представление выборки  $w_1, \dots, w_k$ , порождаемой распределением  $\beta_w = p(w)$



## Графическая нотация для моделей PLSA и LDA

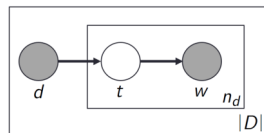
### Модель PLSA:

каждый  $d \in D$  порождает скрытые темы:

$$t_i \sim p(t|d), \quad i = 1, \dots, n_d;$$

каждая тема  $t_i$  порождает слово:

$$w_i \sim p(w|t_i), \quad i = 1, \dots, n_d.$$



### Модель LDA:

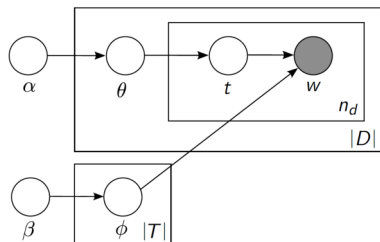
$\alpha$  порождает векторы документов:

$$\theta_d \sim \text{Dir}(\theta|\alpha), \quad d \in D;$$

$\beta$  порождает векторы тем:

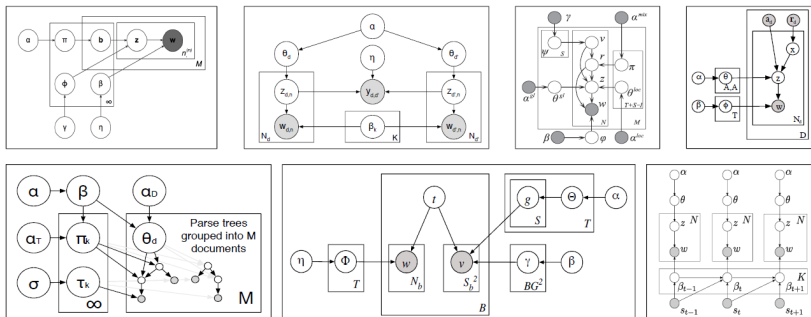
$$\phi_t \sim \text{Dir}(\phi|\beta), \quad t \in T;$$

далее как в PLSA.



## Графическая нотация: представление структуры модели

Большое структурное разнообразие тематических моделей:



## Обсуждение «Stop using Plate Notation»

Единственное достоинство и куча недостатков:

- + хорошо запоминающийся наглядный образ модели
- – множественность вариантов отображения одной модели
- – неполнота и неоднозначность интерпретации
- – не ясен переход от картинки к модели и алгоритму
- – во многих статьях этот переход скрыт или скомкан
- – иллюзия понятности и практическая бесполезность

Один из комментариев:

Every now and then the topic comes up as to why algorithms and procedures are explained in obtuse forms across the entirety of the paper it is described in, usually we just conclude that *it would look too simple if it were explained any other way.*

---

Rob Zinkov. Stop using Plate Notation. 2013-07-28.

<http://zinkov.com/posts/2013-07-28-stop-using-plates>

## Язык псевдокода порождающего процесса (generative story)

**Пример:** вероятностная порождающая модель PLSA

**Вход:**  $p(w|t)$  для всех  $t \in T$ ,  $p(t|d)$  для всех  $d \in D$ ;

**Выход:** коллекция документов;

**для всех** документов  $d \in D$

**для всех** позиций слов  $i = 1, \dots, n_d$  в документе  $d$

        выбрать тему  $t_i$  из  $p(t|d)$ ;

        выбрать слово  $w_i$  из  $p(w|t_i)$ ;

- + легко понимать модель, описание недвусмысленно
- – не ясен переход от модели к алгоритму
- – во многих статьях этот переход скрыт или скомкан

## Язык задач оптимизации: модель $\rightarrow$ критерий $\rightarrow$ алгоритм

1. Вероятностная модель порождения данных:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

2. Постановка задачи оптимизации (ARTM):

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

3. Алгоритм решения — итерационный процесс:

$$\begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \operatorname{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_{w \in W} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$



## Математический язык оптимизационных задач

Один недостаток и куча достоинств:

- – плохо понятен специалистам гуманитарных профессий
- + ясность и однозначность записи модели
- + ясность критерия оптимальности
- + ясность вариантов выбора методов оптимизации
- + переход от модели к алгоритму — это теорема
- + возможность комбинирования моделей
- + унифицированная модульная реализация (BigARTM)

## Байесовское обучение — доминирующий подход в ТМ

Модель LDA предопределила дальнейшее развитие тематических моделей на основе *байесовского вывода*:

$$\text{Posterior}(\Phi, \Theta | \text{data}) \propto \text{Prior}(\Phi, \Theta) P(\text{data} | \Phi, \Theta)$$

Prior и Posterior в модели LDA — распределения Дирихле.

### Проблемы:

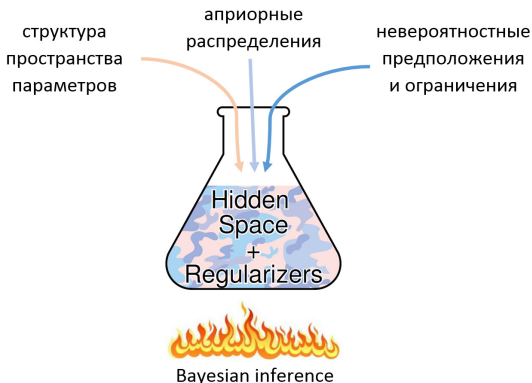
- Нам нужны лишь значения  $\Phi, \Theta$ , а не их распределения
- Prior Дирихле имеет слабые лингвистические обоснования
- Задача сильно усложняется для несопряжённых Prior
- Байесовский вывод уникален для каждой модели
- Технически трудно обобщать и комбинировать модели

---

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

## Алхимия байесовского вывода в тематическом моделировании

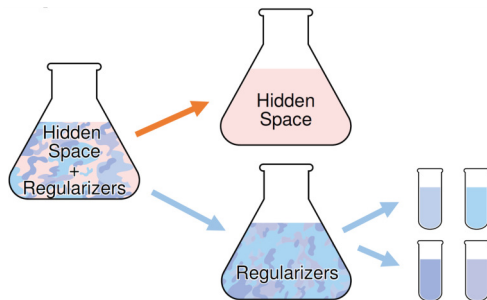
*Вероятностная модель порождения данных объединяет в едином описании структуру пространства параметров, априорные распределения и дополнительные знания о задаче.*



## ARTM — новая алхимия на основе классической регуляризации

Простая *порождающая модель* описывает структуру пространства. Регуляризаторы суммируются с весами, в любых сочетаниях, и каждый описывает только одно дополнительное требование.

**Декомпозиция** — классический способ упрощения задачи



## ARTM: модульный подход к синтезу требуемых моделей


Для построения композитных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».


### Этапы моделирования

### Bayesian TM

### ARTM

	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизуемые этапы

- Максимизация апостериорной вероятности (MAP) даёт точечные оценки  $(\Phi, \Theta)$  и полностью совместима с ARTM.
- *Байесовский вывод* оценивает  $p(\Phi, \Theta|X)$  вместо  $(\Phi, \Theta)$ . Это избыточное усложнение задачи.
- Итерационный процесс в байесовских методах VB и GS не сильно отличается от EM-алгоритма для MAP.
- Байесовский вывод уникален для каждой модели. Нет общего алгоритма. Нет модульной реализации.
- Нет оптимизационной постановки задачи для  $(\Phi, \Theta)$ . Вместо регуляризаторов — априорные распределения.
- Распределение Dir упрощает байесовский вывод, но не имеет убедительных лингвистических обоснований.
- Другие распределения не являются сопряжёнными мультиномиальному, что усложняет байесовский вывод.