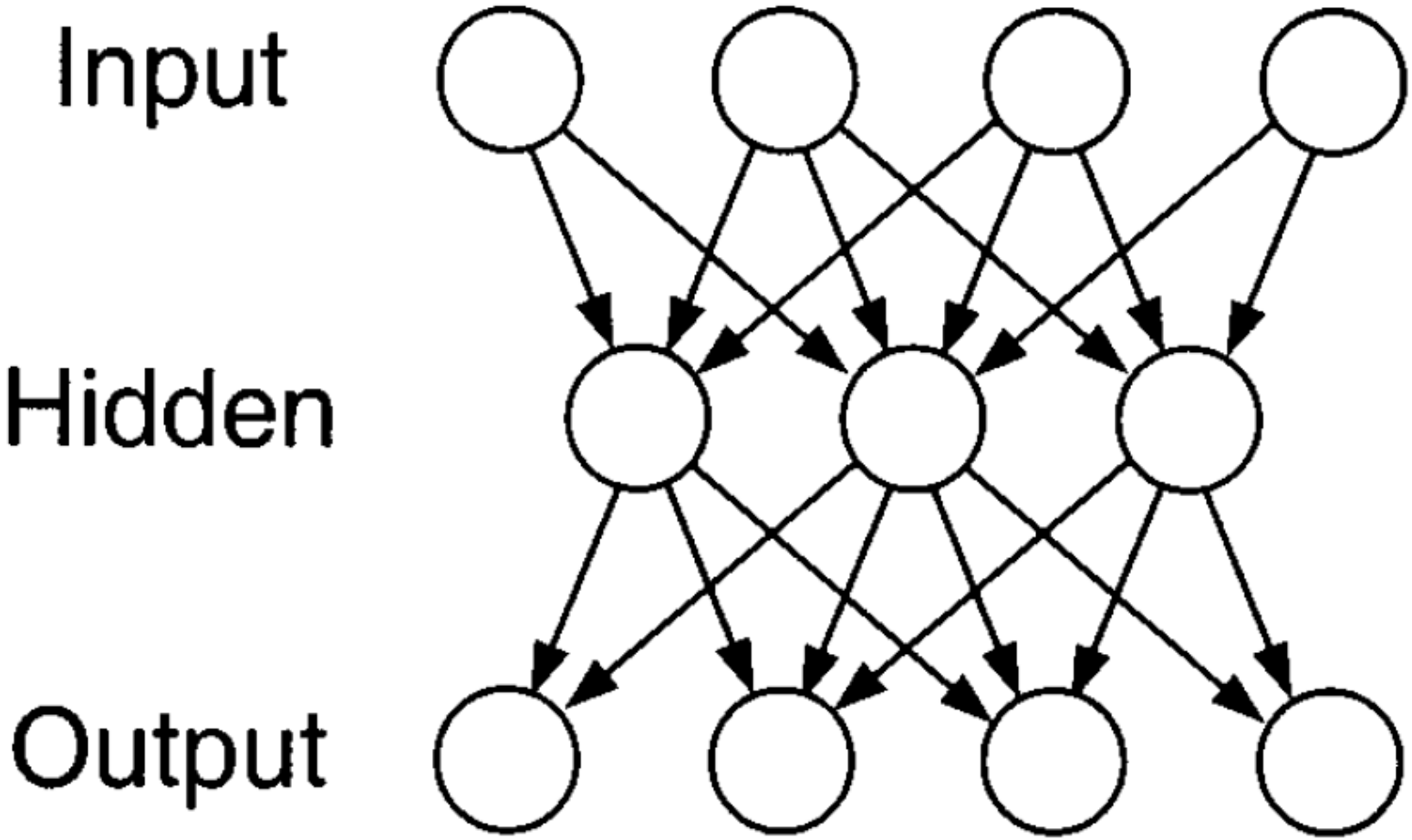
Нейронные сети и LDA

Сергей Николенко

Академический Университет, весенний семестр 2011

Outline

Нейронные сети как графические модели



Нейронные сети

- В узле используем сигмоид-функцию (sigmoid belief networks):

$$p(S_i = 1 \mid S_{\mathrm{pa}(i)}) = 1 / \left(1 + e^{-\sum_{j \in \mathrm{pa}(i)} \theta_{ij} S_j - \theta_{i0}}\right).$$

- Это можно переписать в одну формулу для обоих значений S_i:

$$p(S_i \mid S_{\mathrm{pa}(i)}) = \frac{e^{\left(\sum_{j \in \mathrm{pa}(i)} \theta_{ij} S_j + \theta_{i0}\right) S_i}}{1 + \left(1 + e^{\sum_{j \in \mathrm{pa}(i)} \theta_{ij} S_j + \theta_{i0}}\right)}, \text{ и}$$

$$p(S \mid \theta) = \prod_i \left[\frac{e^{\left(\sum_{j \in \mathrm{pa}(i)} \theta_{ij} S_j + \theta_{i0}\right) S_i}}{1 + \left(1 + e^{\sum_{j \in \mathrm{pa}(i)} \theta_{ij} S_j + \theta_{i0}}\right)} \right].$$

- Мы хотим подсчитать условные вероятности в этом совместном распределении.

- Мы снова применяем метод самосогласованного поля, факторизуем полностью:

$$q(H \mid E, \mu) = \prod_{i \in H} \mu_i^{s_i} (1 - \mu_i)^{1 - s_i}.$$

- Вычисляем KL точно так же, как в случае машин Больцмана, отличие только в $q \ln p$; в итоге получаем:

$$\ln p(E \mid \theta) \geq \sum_{i<j} \theta_{ij}\mu_i\mu_j + \sum_i \theta^c_{i0}\mu_i - \\ - \sum_i \left\langle \ln\left[1 + e^{\sum_{j\in \mathrm{pa}(i)} \theta_{ij}S_j + \theta_{i0}}\right]\right\rangle - \\ - \sum_i \left[\mu_i \ln \mu_i + (1-\mu_i)\ln(1-\mu_i)\right],$$

где $\langle\cdot\rangle$ – ожидание по распределению q.

- Но есть проблема: среднее от $\ln(1 + e^{z_i})$, где $z_i = \sum_{j\in \mathrm{pa}(i)} \theta_{ij}S_j + \theta_{i0}$, не вычислить точно.

- Для этого введём ещё одно вариационное приближение с параметрами ξ_i, из неравенства Йенсена:

$$\langle \ln(1 + e^{z_i}) \rangle = \langle \ln \left[e^{\xi_i z_i} e^{-\xi_i z_i} (1 + e^{z_i}) \right] \rangle =$$

$$\xi_i \langle z_i \rangle + \left\langle \ln \left[e^{-\xi_i z_i} + e^{(1-\xi_i) z_i} \right] \right\rangle \leq \xi_i \langle z_i \rangle + \ln \langle e^{-\xi_i z_i} + e^{(1-\xi_i) z_i} \rangle.$$

- Можно доказать (мы не будем), что в пределе (при большом числе родителей у узла) ξ_i – это ожидание $\sigma(z_i) = 1/(1 + e^{-z})$.

- Для фиксированных ξ_i, дифференцируя KL по μ_i, получаем:

$$\mu_i = \sigma\left(\sum_j \theta_{ij}\mu_j + \theta_{i0} + \sum_j \theta_{ij}(\mu_j - \xi_j) + \sum_j K_{ji}\right),$$

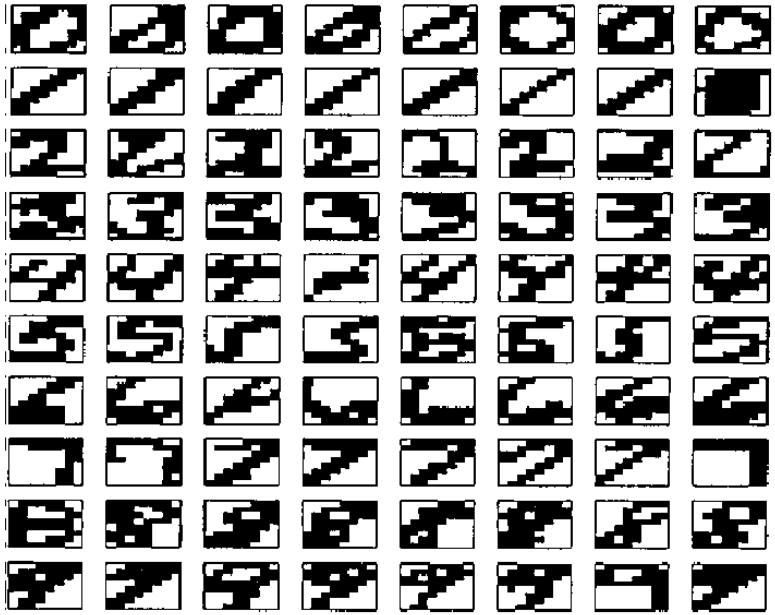
где $K_{ji} = \partial\left(-\ln\left\langle e^{-\xi_i z_i} + e^{(1-\xi_i)z_i}\right\rangle\right)/\partial\mu_i$.

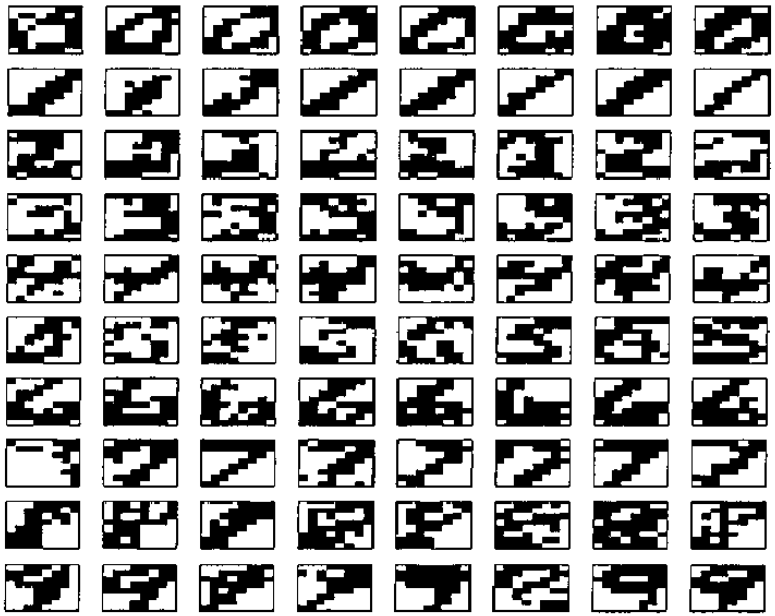
Упражнение. Докажите это.

- Эту производную можно подсчитать из весов узла i, его детей j и других родителей его детей.
- Можно получить уравнения для апдейта ξ_i: [Saul, Jaakkola, Jordan, 1996], [Saul, Jordan, 1999].
- И, наконец, можно получить уравнения для апдейта θ_{ij} при обучении:

$$\Delta\theta_{ij} \propto (\mu_i - \xi_i)\,\mu_j - \theta_{ij}\xi_i\,(1-\xi_i)\,\mu_i\,(1-\mu_i).$$

Пример: распознавание цифр





Outline

LDA

- Ещё один конкретный пример применения – модель LDA (Latent Dirichlet Allocation).
- Задача: смоделировать большую коллекцию текстов (например, для information retrieval или классификации).
- Мы знаем наивный подход: скрытая переменная – тема, слова получаются из темы независимо по дискретному распределению.
- Аналогично работают и подходы, основанные на кластеризации.
- Давайте чуть усложним.

LDA

- Очевидно, что у одного документа может быть несколько тем; подходы, которые кластеризуют документы по темам, никак этого не учитывают.
- Давайте построим иерархическую байесовскую модель:
  - на первом уровне – смесь, компоненты которой соответствуют «темам»;
  - на втором уровне – мультиномиальная переменная с априорным распределением Дирихле, которое задаёт «распределение тем» в документе.

LDA

- Если формально: слова берутся из словаря $\{1, \ldots, V\}$; слово – это вектор w, $w_i \in \{0, 1\}$, где ровно одна компонента равна 1.
- Документ – последовательность из N слов $\mathbf{w}$. Нам дан корпус из M документов $\mathcal{D} = \{\mathbf{w}_d \mid d = 1..M\}$.
- Генеративная модель LDA выглядит так.
  1. Выбрать $N \sim p(N \mid \xi)$.
  2. Выбрать $\theta \sim \mathrm{Di}(\alpha)$.
  3. Для каждого из N слов w_n:
     1. выбрать тему $z_n \sim \mathrm{Mult}(\theta)$;
     2. выбрать слово $w_n \sim p(w_n \mid z_n, \beta)$ по мультиномиальному распределению.

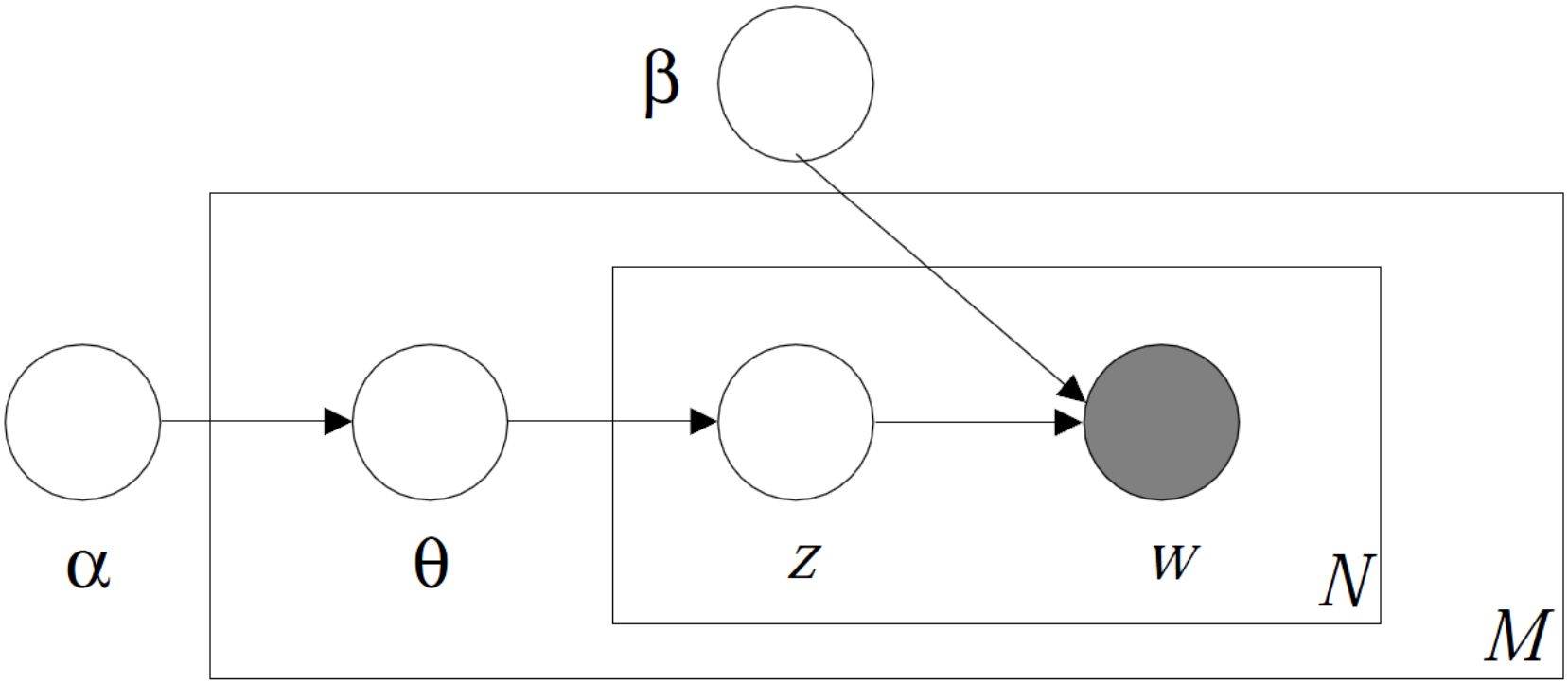
LDA

- Мы пока для простоты фиксируем число тем k, считаем, что β – это просто набор параметров $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, которые нужно оценить, и не беспокоимся о распределении на N.
- Совместное распределение тогда выглядит так:

$$p(\theta, \mathbf{z}, \mathbf{w}, N \mid \alpha, \beta) = p(N \mid \xi) p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta).$$

- В отличие от обычной кластеризации с априорным распределением Дирихле, мы тут не выбираем кластер один раз, а затем накидываем слова из этого кластера, а для каждого слова выбираем по распределению θ, по какой теме оно будет набросано.

LDA: графическая модель



Вывод в LDA

- Рассмотрим задачу байесовского вывода, т.е. оценки апостериорного распределения θ и z после нового документа:

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

- Мы здесь рассматриваем α и β как известные константы; по идее, конечно, их тоже нужно оценивать, но пока для простоты так.
- Знаменатель – правдоподобие – оценивается как

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

и это трудно посчитать, потому что θ и β путаются друг с другом.

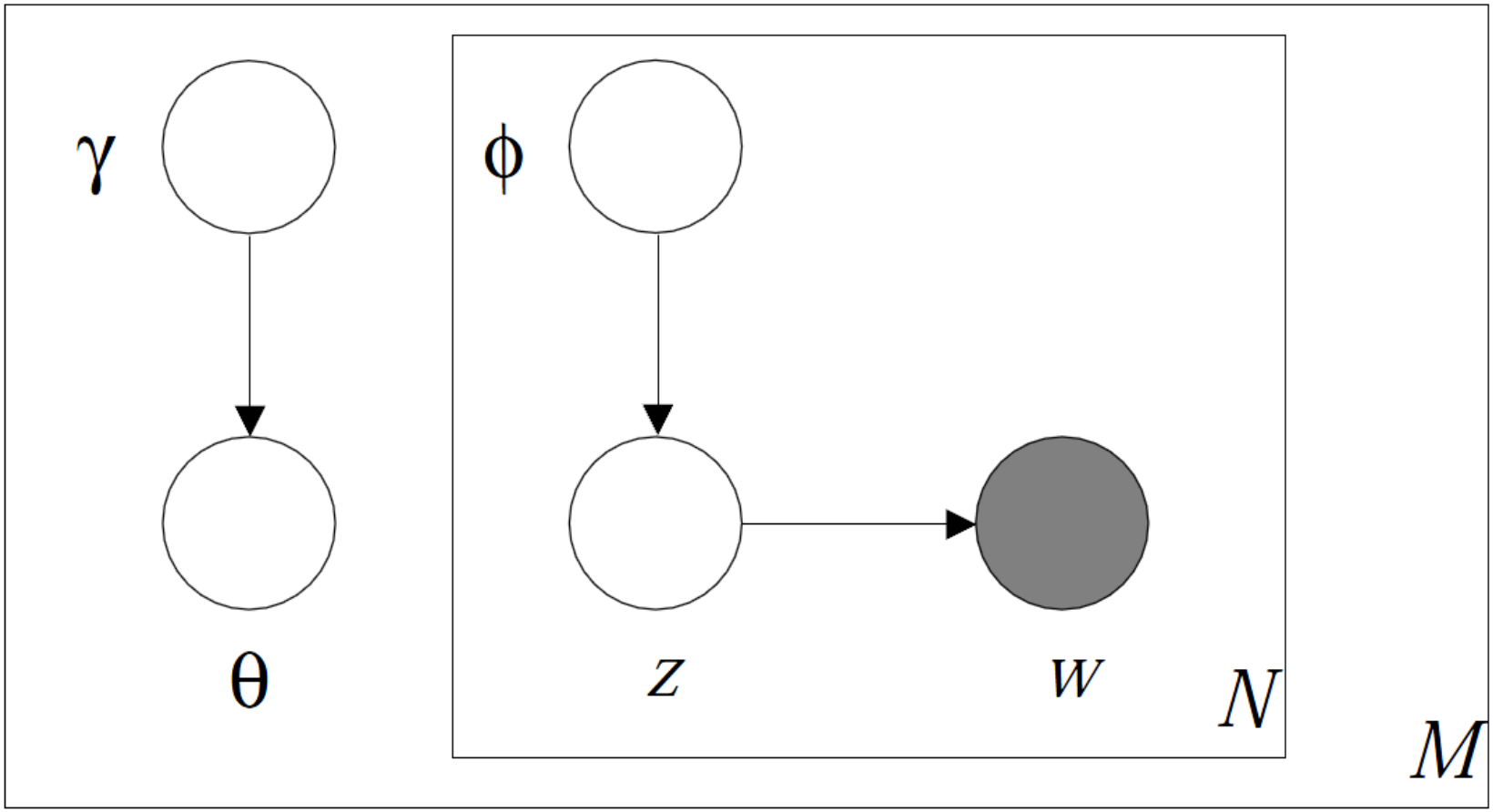
Вывод в LDA

- Вариационное приближение – рассмотрим семейство распределений

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

- Тут всё расщепляется, и мы добавили вариационные параметры γ (Дирихле) и ϕ (мультиномиальный).
- Заметим, что параметры для каждого документа могут быть свои – всё условно по $\mathbf{w}$.

LDA: вариационное приближение



LDA: вариационный вывод

- Теперь можно искать минимум KL-расстояния:

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

- Для этого сначала воспользуемся уже известной оценкой из неравенства Йенсена:

$$\log p(\mathbf{w} \mid \alpha, \beta) = \log \int_\theta \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta =$$

$$= \log \int_\theta \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \geq$$

$$\geq E_q [\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - E_q [\log q(\theta, \mathbf{z})] =: \mathcal{L}(\gamma, \phi; \alpha, \beta).$$

LDA: вариационный вывод

- Распишем произведения:

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

- Свойство распределения Дирихле: если $X \sim \mathrm{Di}(\alpha)$, то

$$E[\log(X_i)] = \Psi(\alpha_i) - \Psi(\sum_i \alpha_i),$$

где $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ – дигамма-функция.

- Теперь можно выписать каждый из пяти членов.

LDA: вариационный вывод

$$\mathcal{L}(\gamma, \phi; \alpha, \beta) = \log \Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1) \left[\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) \right] +$$
$$+ \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \left[\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) \right] +$$
$$+ \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} w_n^j \phi_{ni} \log \beta_{ij} -$$
$$- \log \Gamma(\sum_{i=1}^{k} \gamma_i) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1) \left[\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) \right] -$$
$$- \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}.$$

LDA: вариационный вывод

- Теперь осталось только брать частные производные этого выражения.
- Сначала максимизируем его по ϕ_{ni} (вероятность того, что n-е слово было порождено темой i); надо добавить λ-множители Лагранжа, т.к. $\sum_{j=1}^{k} \phi_{nj} = 1$.
- В итоге получится:

$$\phi_{ni} \propto \beta_{iv} e^{\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)},$$

где v – номер того самого слова, т.е. единственная компонента $w_n^v = 1$.

LDA: вариационный вывод

- Потом максимизируем по γ_i, i-й компоненте апостериорного Дирихле-параметра.
- Получится

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

- Соответственно, для вывода нужно просто пересчитывать ϕ_{ni} и γ_i друг через друга, пока оценка не сойдётся.

LDA: оценка параметров

- Теперь давайте попробуем оценить параметры α и β по корпусу документов $\mathcal{D}$.
- Мы хотим найти α и β, которые максимизируют

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \alpha, \beta).$$

- Подсчитать $p(\mathbf{w}_d \mid \alpha, \beta)$ мы не можем, но у нас есть нижняя оценка $\mathcal{L}(\gamma, \phi; \alpha, \beta)$, т.к.

$$p(\mathbf{w}_d \mid \alpha, \beta) =$$
$$= \mathcal{L}(\gamma, \phi; \alpha, \beta) + \mathrm{KL}(q(\theta, z \mid \mathbf{w}_d, \gamma\phi) \| p(\theta, \mathbf{z} \mid \mathbf{w}_d, \alpha, \beta)).$$

LDA: оценка параметров

- EM-алгоритм:
  1. найти параметры $\{\gamma_d, \phi_d \mid d \in \mathcal{D}\}$, которые оптимизируют оценку (как выше);
  2. зафиксировать их и оптимизировать оценку по α и β.

LDA: оценка параметров

- Для β это тоже делается нехитро:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_n^j.$$

- Для α_i получается система уравнений, которую можно решить методом Ньютона.

Thank you!

Спасибо за внимание!