Вариационные методы II: машины Больцмана

Сергей Николенко

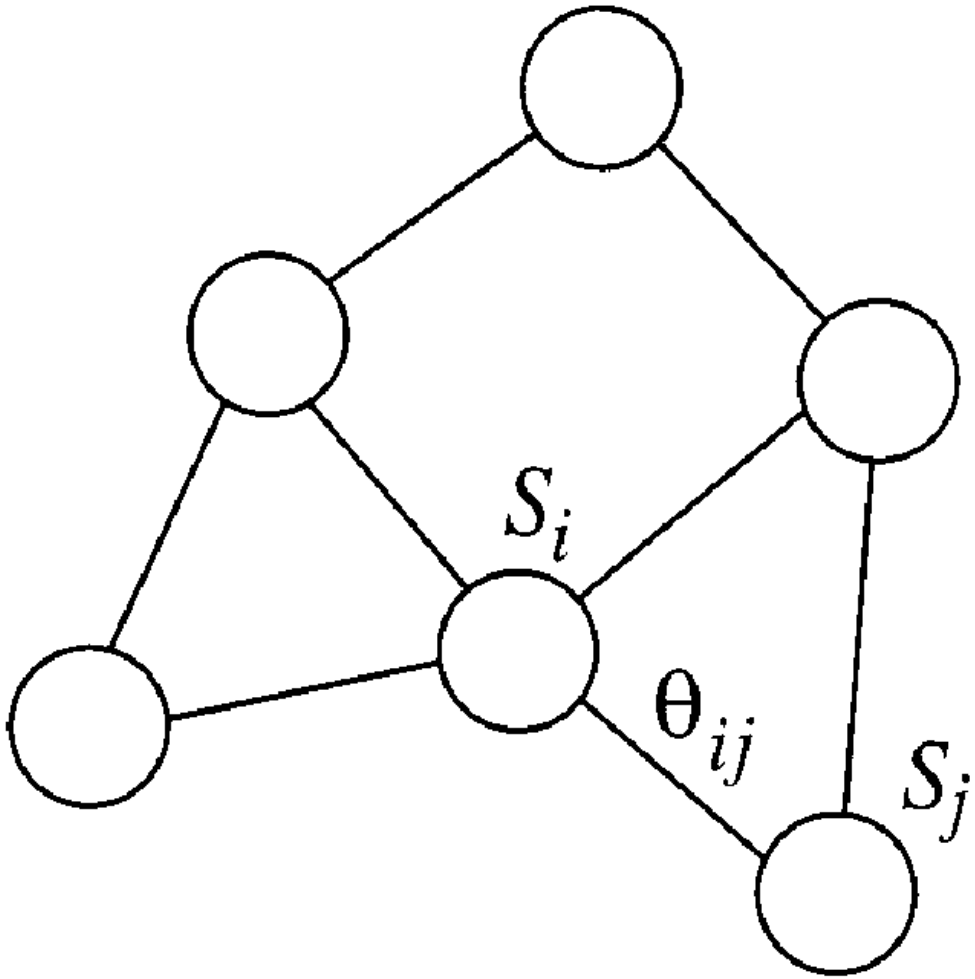
Академический Университет, весенний семестр 2011

Outline

Машина Больцмана

- Машина Больцмана (Boltzmann machine) – это частный случай марковского случайного поля, т.е. ненаправленная графическая модель.
- Вершины – бинарные события S_i, набор функций-потенциалов ограничен.

Машина Больцмана



Машина Больцмана

- Фактор Больцмана (Boltzmann factor) – экспонента от квадратичного выражения от S_i.
- Потенциал каждой клики – произведение факторов, но мы предполагаем, что $e^{\theta_{ij} S_i S_j}$ встречается только в одной из клик.
- Поэтому совместное распределение выглядит как

$$p(S) = \frac{1}{Z} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0} S_i},$$

где $\theta_{ij} = 0$ для несоседних S_i и S_j.

Машина Больцмана

- $E = -\sum_{i<j} \theta_{ij} S_i S_j - \sum_i \theta_{i0} S_i$ называется *энергией*.
- Вообще, совместное распределение $p(E) \sim e^{-\beta E}$ – это *распределение Больцмана* из статистической физики.

Машина Больцмана

- А смысл такой: если мы обучим веса θ_{ij} (это отдельный вопрос) и проведём вывод на каком-нибудь начальном распределении (evidence), то мы получим вероятности других, неизвестных вершин.
- Таким образом, можно дополнять частично известные распределения «по ассоциации».
- Вывод можно вести точно в некоторых частных случаях, но в общем случае он слишком сложен.

Стоящие перед нами задачи

- Мы хотим провести маргинализацию в распределении

$$p(S) = \frac{1}{Z} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0} S_i}.$$

- Для маргинализации вида $p(H) = \sum_{\{S \setminus H\}} p(S)$ нам надо подсчитать сумму экспонент квадратичных функций.
- Для условных вероятностей вида $p(H \mid E) = \frac{p(H,E)}{p(E)}$ нужно подсчитать отношение таких сумм.
- Самая общая такая сумма – это, собственно, $Z = \sum_{\{S\}} p(S)$ (partition function); её и будем искать.

Стоящие перед нами задачи

- Метод такой: будем проводить вариационные преобразования одно за другим, оставаясь в рамках машины Больцмана.
- Один шаг:

$$Z = \sum_{\{S\}} e^{\sum_{j<k} \theta_{jk} S_j S_k + \sum_j \theta_{j0} S_j} =$$

$$= \sum_{\{S \setminus S_i\}} \sum_{S_i \in \{0,1\}} e^{\sum_{j<k} \theta_{jk} S_j S_k + \sum_j \theta_{j0} S_j}.$$

Нижняя оценка

- Легко показать, что внутренняя сумма лог-выпукла. Можно найти вариационную нижнюю оценку:

$$\ln \left[\sum_{S_i \in \{0,1\}} e^{\sum_{j<k} \theta_{ij} S_j S_k + \sum_j \theta_{j0} S_j} \right] =$$

$$= \sum_{\{j<k\} \neq i} \theta_{jk} S_j S_k + \sum_{j \neq i} \theta_{j0} S_j + \ln \left[1 + e^{\sum_{j \neq i} \theta_{ij} S_j + \theta_{i0}} \right] \geq$$

$$\geq \sum_{\{j<k\} \neq i} \theta_{jk} S_j S_k + \sum_{j \neq i} \theta_{j0} S_j + \lambda_i^L \left(\sum_{j \neq i} \theta_{ij} S_j + \theta_{i0} \right) + H(\lambda_i^L),$$

т.к. $\ln(1 + e^{-x}) \geq -\lambda x + H(\lambda)$.

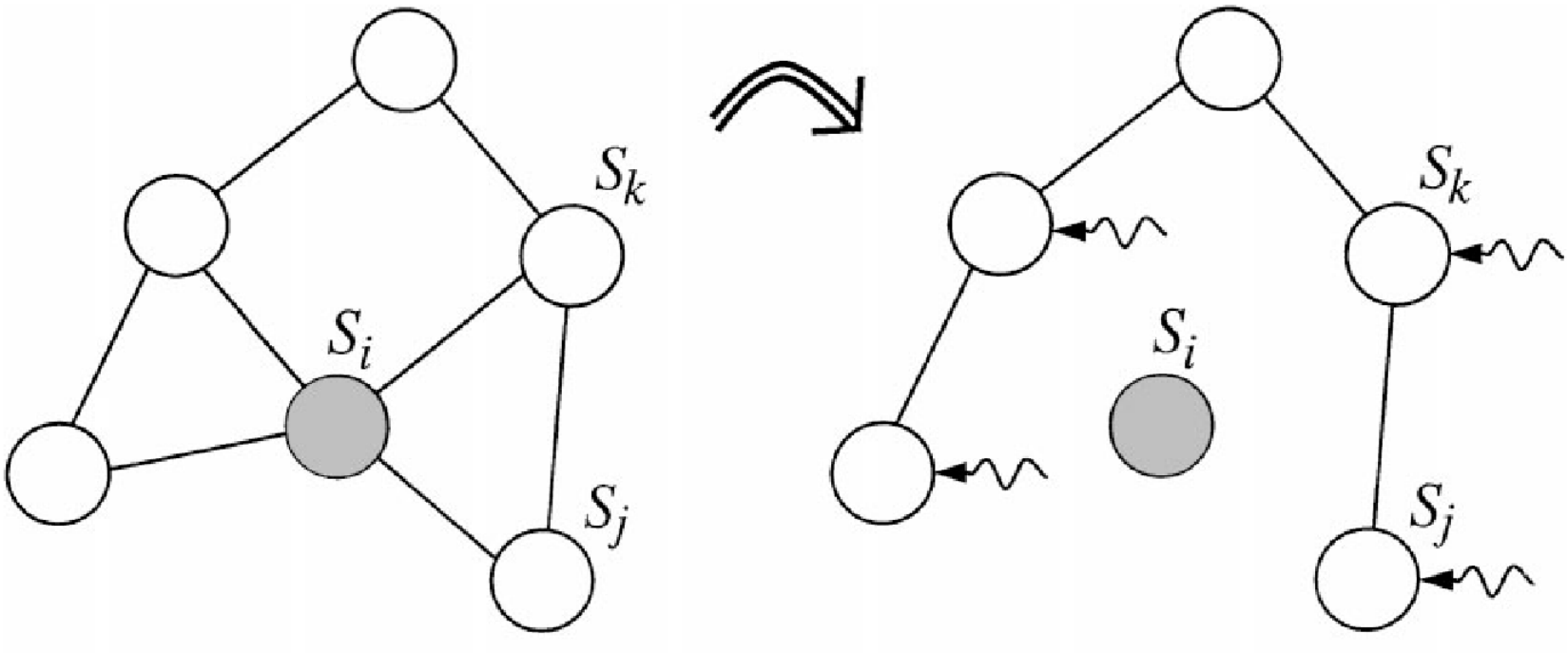
Нижняя оценка

- В графическом смысле мы отрезали S_i от соседей и добавили к соседям линейные члены:

$$\theta'_{jk} = \theta_{jk},$$
$$\theta'_{j0} = \theta_{j0} + \lambda_i^L \theta_{ij}.$$

- Но не соединили этих соседей, как получилось бы при точном выводе.
- Кроме того, добавился константный член $\lambda_i^L \theta_{i0} + H(\lambda_i^L)$.

Нижняя оценка



Верхняя оценка

- Можно аналогично найти и верхнюю оценку:

$$\ln(1+e^{-x}) = \ln(e^{\frac{x}{2}}+e^{-\frac{x}{2}}) + \frac{x}{2} \le \lambda x^2 + \frac{x}{2} - g^*(\lambda).$$

- Соответственно,

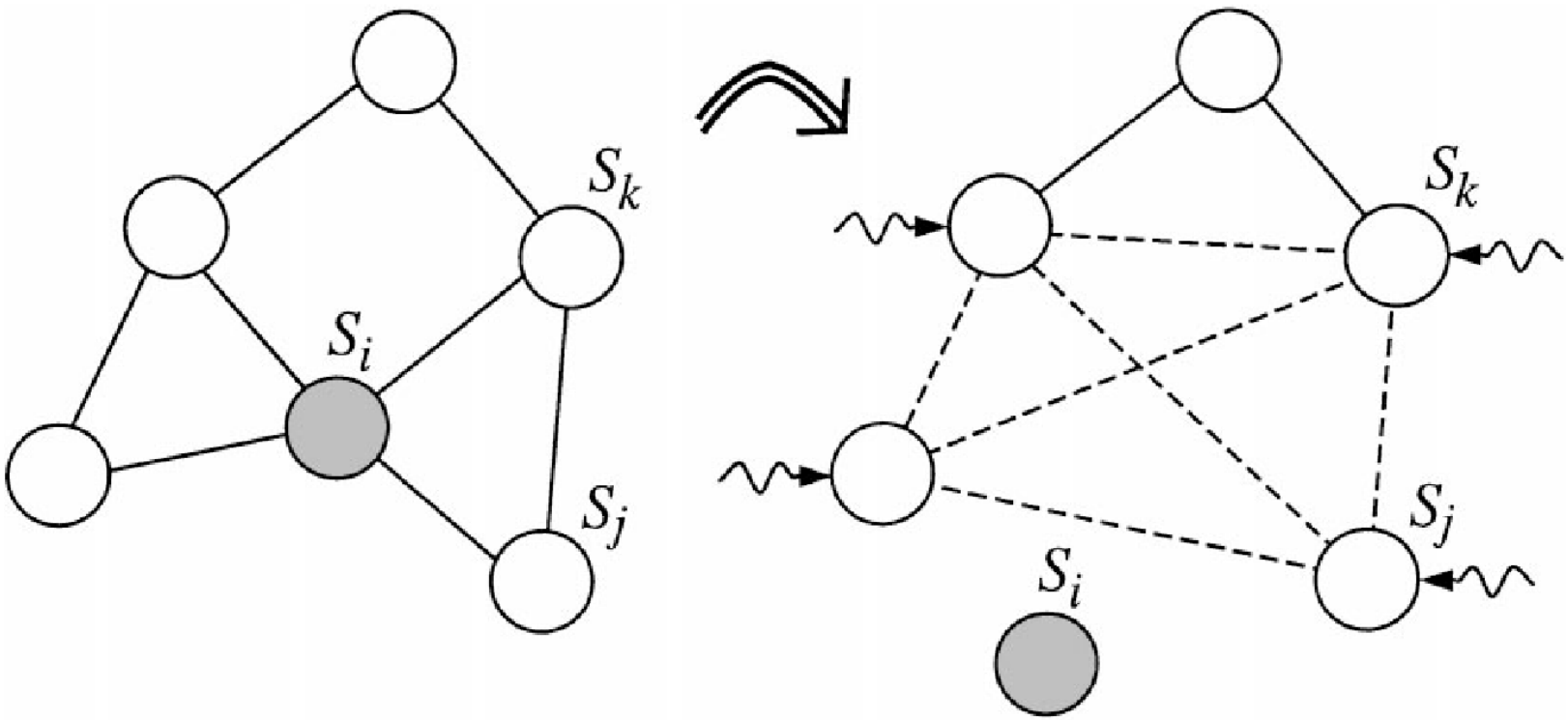
$$\ln\left[\sum_{S_i\in\{0,1\}} e^{\sum_{j<k}\theta_{ij}S_jS_k+\sum_j\theta_{j0}S_j}\right] \le \sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k+\sum_{j\neq i}\theta_{j0}S_j+$$
$$+\lambda_i^U\left(\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right)^2+\frac{1}{2}\left(\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right)-g^*(\lambda_i^U).$$

Верхняя оценка

- Графический смысл теперь немножко другой, т.к. добавились новые связи между соседями S_i:

$$\theta'_{jk} = \theta_{jk} + 2\lambda_i^U,$$
$$\theta'_{j0} = \theta_{j0} + \lambda_i^L \theta_{ij}.$$

Верхняя оценка



Итого

- В итоге мы получили верхнюю и нижнюю оценки.
- Нижняя оценка удобнее: вершины отщепляются целиком, можно в любом удобном порядке аппроксимировать, довести до структуры (например, дерева), на которой уже легко вести точный вывод.
- Зато верхняя оценка точнее (практические результаты это показывают).

Outline

Блочная аппроксимация: идея

- Мы до сих пор удаляли вершины по одной.
- Но, может быть, если сразу несколько вершин выбрать, можно найти более точную аппроксимацию?
- Идея:
  1. выбрать подструктуру в графе, для которой можно провести точный вывод (дерево, набор цепочек и т.п.);
  2. рассмотреть семейство вероятностных распределений на этой подструктуре с вариационными параметрами;
  3. выбрать одно распределение из этого семейства, желательно оптимально аппроксимирующее.

Блочная аппроксимация: идея

- Формально: есть $p(S)$, мы хотим оценить $p(H \mid E)$.
- Введём семейство приближений $q(H \mid E, \lambda)$, где λ – вариационные параметры.
- Выберем из них одно, минимизирующее расстояние Кульбака–Ляйблера:

$$\lambda^* = \arg\min_{\lambda} \mathrm{KL}(q(H \mid E, \lambda) \| p(H \mid E)), \text{ где}$$

$$\mathrm{KL}(q\|p) = \sum_{\{S\}} q(S) \ln \frac{q(S)}{p(S)}.$$

Расстояние Кульбака–Ляйблера

- Почему именно KL? Помимо прочего, это естественная нижняя оценка на правдоподобие $p(E)$.
- По неравенству Йенсена:

$$\ln p(E) = \ln \sum_{\{H\}} q(H \mid E) \frac{p(H \mid E)}{q(H \mid E)} \geq$$

$$\geq \sum_{\{H\}} q(H \mid E) \ln \frac{p(H \mid E)}{q(H \mid E)},$$

  и разница между левой и правой частями – это как раз $\mathrm{KL}(q \| p)$.
- Поэтому справа стоит нижняя оценка на $p(E)$, и для оптимального λ^* это оптимальная оценка.

Расстояние Кульбака–Ляйблера

- В итоге получается:

$$\ln p(E) \geq \sum_{\{H\}} q(H \mid E) \ln p(H \mid E) - \sum_{\{H\}} q(H \mid E) \ln q(H \mid E).$$

Упражнение. Это можно было бы и вариационным методом получить. Попробуйте получить эту оценку вариационным методом, используя вектор вероятностей $q(H \mid E, \lambda)$ как вектор вариационных параметров.

Обучение параметров

- Эту оценку можно использовать в рамках EM-алгоритма для обучения параметров модели.
- Добавим теперь в наши обозначения параметры θ: теперь $p(S \mid \theta)$.
- Введём функцию

$$\mathcal{L}(q, \theta) =$$
$$= \sum_{\{H\}} q(H \mid E) \ln p(H \mid E, \theta) - \sum_{\{H\}} q(H \mid E) \ln q(H \mid E) \leq \ln p(E).$$

Обучение параметров

- Если мы разрешили бы $q(H \mid E)$ быть любым, оптимальное значение было бы $q(H \mid E) = p(H \mid E, \theta)$.
- Давайте применим такую форму EM-алгоритма:

E-шаг $Q^{(k+1)} := \arg\max_Q \mathcal{L}\left(Q, \theta^{(k)}\right)$;
M-шаг $\theta^{(k+1)} := \arg\max_\theta \mathcal{L}\left(Q^{(k+1)}, \theta\right)$.

- Это просто покоординатный подъём для функции правдоподобия $\mathcal{L}(q, \theta)$.

Обучение параметров

- Если теперь $q(H \mid E)$ – это всё-таки аппроксимация, получается уже известный нам приём minorization–maximization.
- Мы на каждом шаге оптимизируем нижнюю оценку правдоподобия вместо него самого.
- Но в принципе алгоритм точно таким же остаётся.

Машины Больцмана

- Вернёмся к нашим машинам Больцмана:

$$p(S \mid \theta) = \frac{1}{Z} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0} S_i}.$$

- Что будет в $p(H \mid E, \theta)$?
  - для $S_i \in E$ и $S_j \in E$ $\theta_{ij} S_i S_j$ – это константа, и она исчезает при нормализации;
  - для $S_i \in H$, $S_j \in E$ квадратичный член становится линейным и вписывается в S_i;
  - линейные члены для $S_i \in E$ исчезают.

Машины Больцмана

- Итого:

$$p(H \mid E, \theta) = \frac{1}{Z_c} e^{\sum_{i<j} \theta_{ij} s_i s_j + \sum_i \theta^c_{i0} s_i},$$

где суммы только по узлам из H, и $\theta^c_{i0} = \theta_{i0} + \sum_{j \in E} \theta_{ij} S_j$.

- Теперь

$$Z_c = \sum_{\{H\}} e^{\sum_{i<j} \theta_{ij} s_i s_j + \sum_i \theta^c_{i0} s_i},$$

и мы получили машину Больцмана на подмножестве H.

Метод самосогласованного поля

- Ещё один термин из физики – метод самосогласованного поля (mean field theory).
- Смысл в том, чтобы искать приближение среди *полностью* факторизуемых распределений.
- Иначе говоря, мы ищем $q(H \mid E, \mu)$ в виде

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

Метод самосогласованного поля

- Теперь можно посчитать KL-расстояние:

$$\mathrm{KL}(q\|p) = \sum_{\{H\}} q(H \mid E, \mu) \ln \frac{q(H \mid E, \mu)}{p(H \mid E, \theta)} =$$

$$= \sum_i [\mu_i \ln \mu_i + (1-\mu_i)\ln(1-\mu_i)] - \sum_{i<j} \theta_{ij}\mu_i\mu_j - \sum_i \theta^c_{i0}\mu_i + \ln Z_c$$

(т.к. в распределении q S_i и S_j – независимые случайные величины со средними μ_i и μ_j).

- И мы хотим минимизировать это по q, т.е. по μ_i.

Метод самосогласованного поля

- Возьмём частные производные по μ_i и приравняем нулю; получим

$$\mu_i = \sigma(\sum_j \theta_{ij}\mu_j + \theta_{i0}),$$

  где $\sigma(z) = 1/(1 + e^{-z})$ – сигмоид.
- Эти уравнения называются «уравнениями самосогласованного поля»; их можно решить итеративно.

Обучение машин Больцмана

- Как мы уже говорили, можно это применить и для обучения параметров θ_{ij}.
- Выпишем нижнюю оценку для метода самосогласованного поля:

$$\ln p(E \mid \theta) \geq$$

$$\geq \sum_{\{H\}} q(H \mid E) \ln p(H \mid E) - \sum_{\{H\}} q(H \mid E) \ln q(H \mid E) =$$

$$= \sum_{i<j} \theta_{ij} \mu_i \mu_j + \sum_i \theta_{i0}^c \mu_i - \ln Z - \sum_i \left[\mu_i \ln \mu_i + (1 - \mu_i) \ln (1 - \mu_i) \right].$$

Обучение машин Больцмана

- Теперь можно взять производные по θ_{ij} и запустить градиентный подъём. Но для этого нужно знать $\frac{\partial \ln Z}{\partial \theta_{ij}}$.

Упражнение. Покажите, что $\frac{\partial \ln Z}{\partial \theta_{ij}} = \langle S_i S_j \rangle$, где $\langle \cdot \rangle$ – ожидание по распределению $p(S \mid \theta)$.

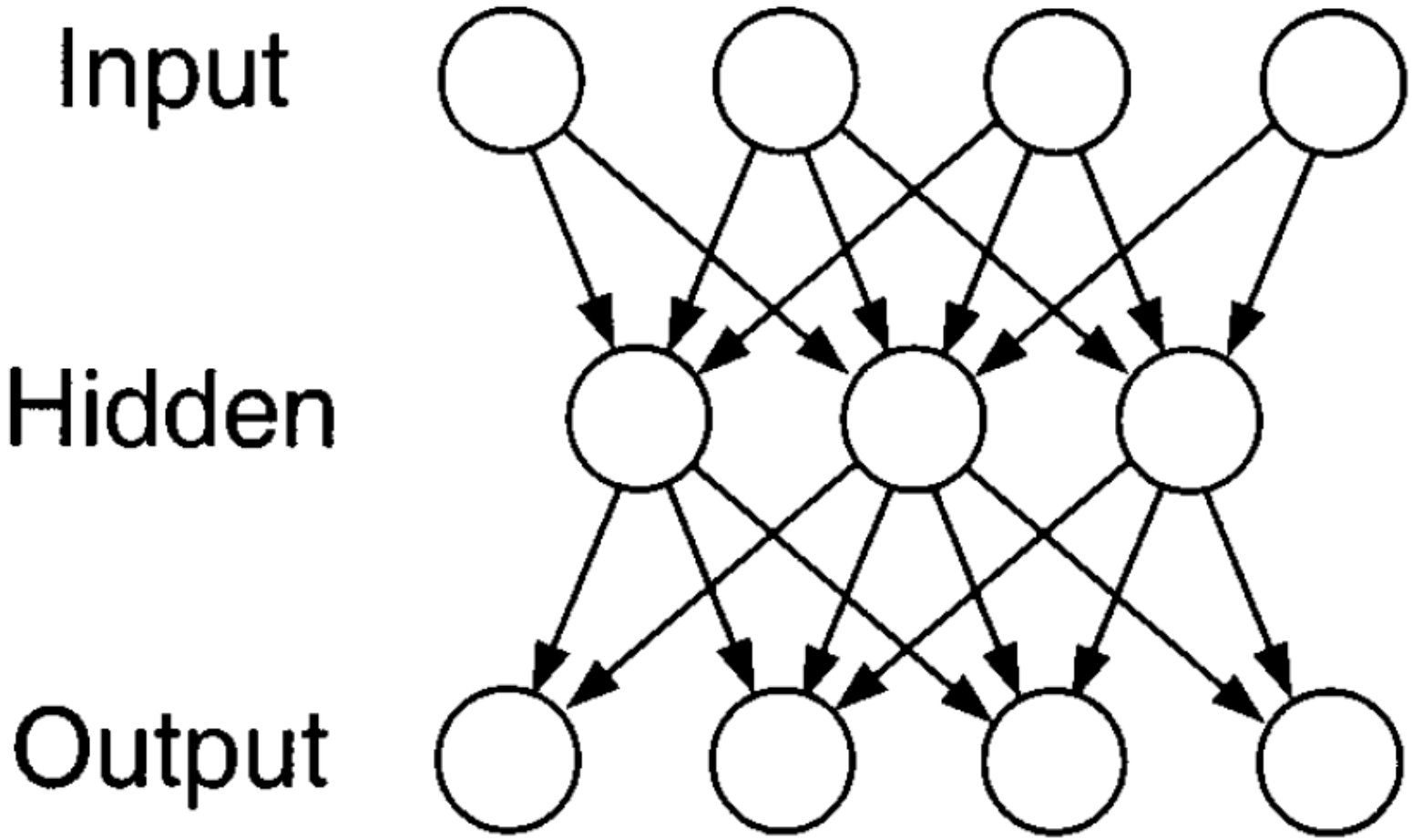
- В итоге получается правило обучения:

$$\Delta \theta_{ij} \propto \mu_i \mu_j - \langle S_i S_j \rangle.$$

Обучение машин Больцмана

- К сожалению, $\langle S_i S_j \rangle$ точно подсчитать не получится (если вообще точный вывод нельзя провести).
- Есть и более серьёзная проблема: мы же хотим несколько ассоциаций заложить в модель, т.е. получить мультимодальное распределение.
- А приближаем мы его унимодальным распределением (после факторизации).
- Один возможный подход – рассматривать мультимодальные распределения q, например, смеси распределений рассмотреть. Этого мы делать уже не будем.

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.

Нейронные сети

- Мы снова применяем метод самосогласованного поля, факторизуем полностью:

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

Нейронные сети

- 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.

Нейронные сети

- 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 $\sigma(z_i) = 1/(1 + e^{-z})$.

Нейронные сети

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

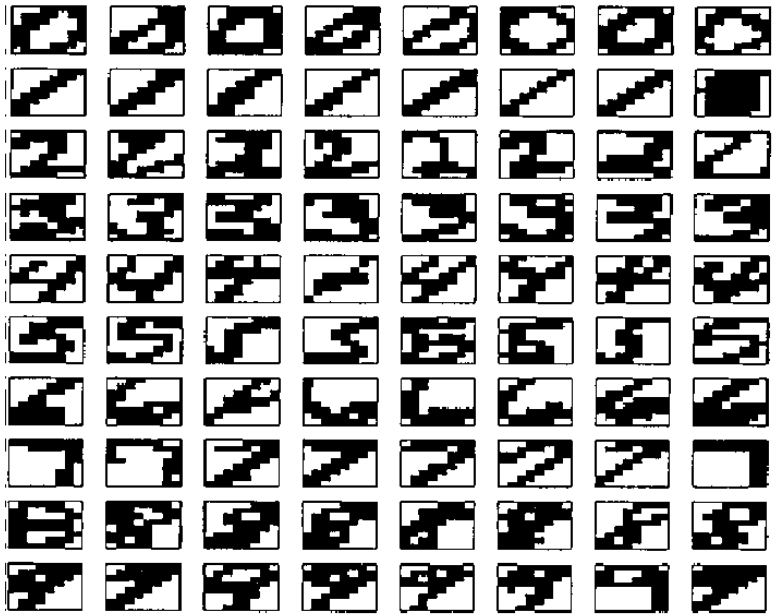
$$\mu_i = \sigma\left(\sum_j \theta_{ij}\mu_j + \theta_{i0} + \sum_j \theta_{ij}(\mu_j - \xi_j) + \sum_j K_{ji}\right),$$

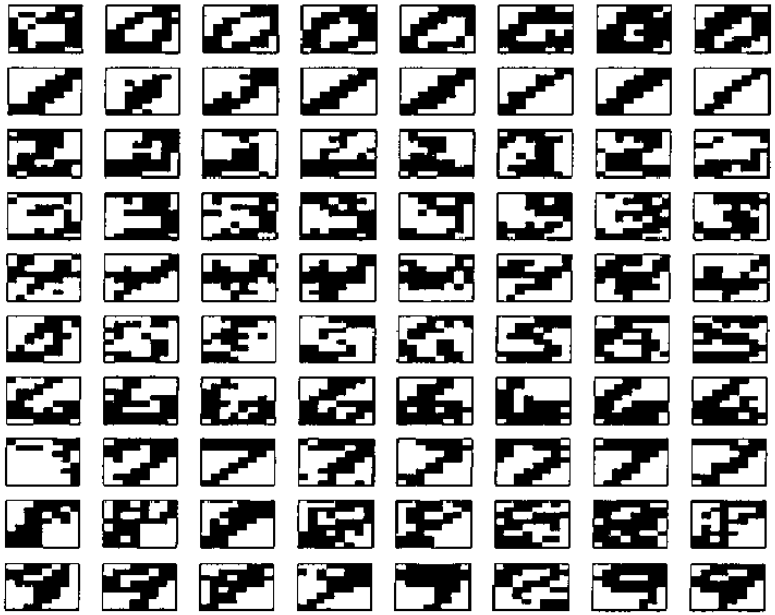
где $K_{ji} = \partial\left(-\ln\left\langle e^{-\xi_i z_i} + e^{(1-\xi_i)z_i}\right\rangle\right)/\partial\mu_i$.

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:

Пример: распознавание цифр





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.

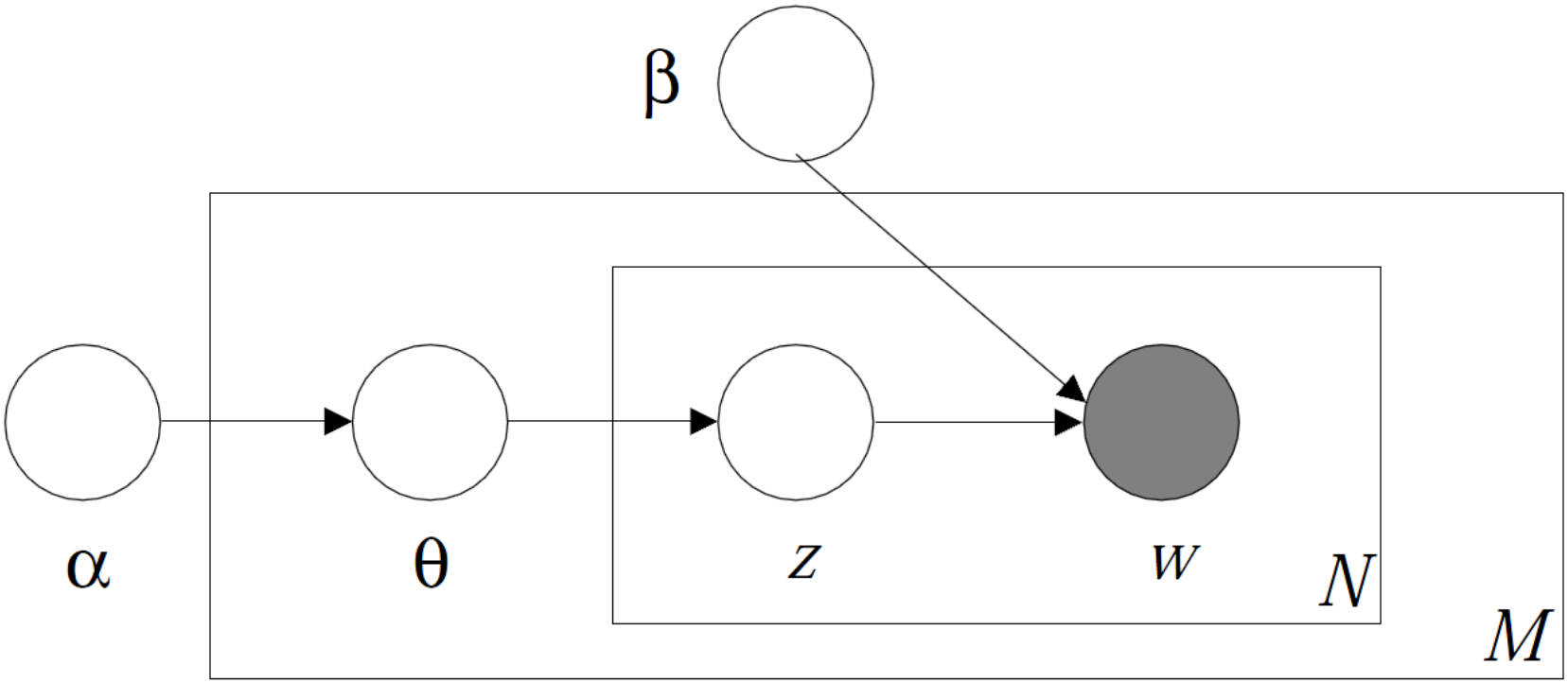
LDA

- 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.

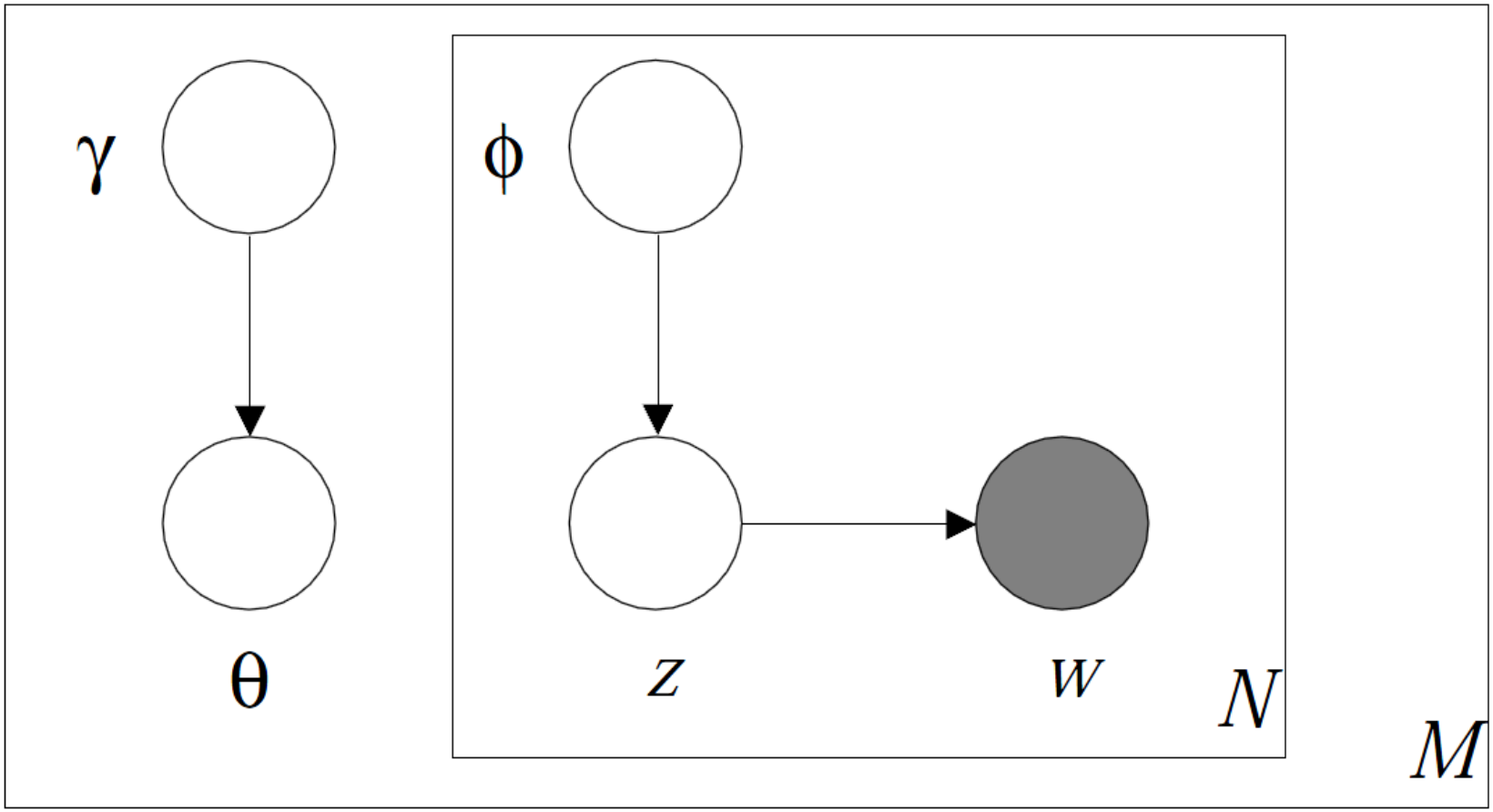
Вывод в LDA

- 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!