

ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург

1 апреля 2025 г.

Random facts:



- Ещё 1 апреля 1539 года один фламандский дворянин дал своим слугам шуточные поручения; в 1686 году Джон Обри упомянул «Fools Holiday»; 1 апреля 1696 г. несколько лондонцев зазывали людей на «мойку львов» в Тауэре; в Москве в 1703 году глашатаи приглашали всех прийти на «неслыханное представление» немецкой труппы, но зрители увидели лишь полотнище с надписью: «Первый апрель — никому не верь!»
- 1 апреля 1808 г. император Александр I провозгласил «навечное» присоединение Финляндии к России; примерно на век и получилось, до 1917 года
- 1 апреля 1939 г. радио Бургос передало: «Армия красных пленена и разоружена, национальные силы овладевают последними военными объектами. Война закончена. Бургос 1 апреля 1939 — года победы. Генералиссимус Франко»
- 1 апреля 1972 г. Intel представила микропроцессор 8008, а 1 апреля 1976 г. два Стива, Джобс и Возняк, основали Apple Computer Company
- 1 апреля 1979 г. по итогам всенародного референдума (99.3% за, 0.7% против) аятолла Хомейни провозгласил Иран исламской республикой

ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ

- Вариационный вывод: функционалы, производные по функциям... в общем, можно оптимизировать функционалы.
- Для нас это значит, что можно оптимизировать приближение q из какого-то класса к заданному p .
- Пусть есть $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ и $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$.
- Мы знаем $p(\mathbf{X}, \mathbf{Z})$ из модели, хотим найти приближение для $p(\mathbf{Z} \mid \mathbf{X})$ и $p(\mathbf{X})$.

- Как и в EM:

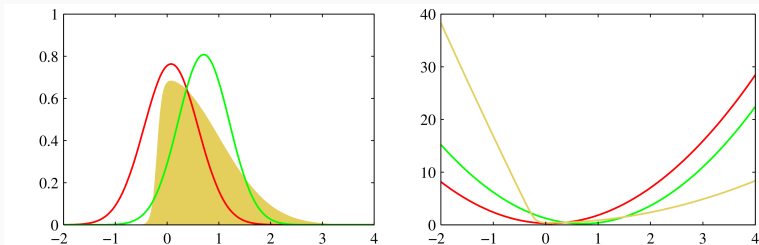
$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \text{KL}(q\|p), \text{ где}$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z},$$

$$\text{KL}(q\|p) = - \int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} | \mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

- $\mathcal{L}(q)$ — это вариационная нижняя оценка, её можно теперь максимизировать, и KL будет автоматически минимизироваться.

- Пример – сравним с лапласовским:



- Если $q(Z)$ произвольное, то мы просто получим $q(Z) = p(\mathbf{Z} | \mathbf{X})$; но это вряд ли получится.
- Будем ограничивать.

- Главный частный случай — пусть $\mathbf{Z} = \mathbf{Z}_1 \sqcup \dots \sqcup \mathbf{Z}_M$, и

$$q(\mathbf{Z}) = \prod_{i=1}^M q_i(\mathbf{Z}_i).$$

- Но больше никаких предположений! В этом прелесть — оптимизируем сразу функции!
- Это соответствует теории среднего поля в физике (mean field theory).

- Тогда:

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_i q_i \left(\ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right) d\mathbf{Z} \\ &= \int q_j \left[\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \right] d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + \text{const} \\ &= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + \text{const},\end{aligned}$$

где $\ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j} [\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$.

- Как максимизировать теперь $\mathcal{L}(q)$ по q_j ?

- Надо заметить, что мы получили там KL-дивергенцию между $q_j(\mathbf{Z}_j)$ и $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$.
- Т.е. оптимальное решение получится при

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E} [\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

- Константа здесь просто для нормализации.
- Оказывается, достаточно взять ожидание от логарифма совместного распределения.
- Но явных формул не получается, потому что ожидание считается по остальным $q_i^*, i \neq j$.
- И всё-таки обычно что-то можно сделать; давайте рассмотрим примеры.

- Первый пример — приблизим двумерный гауссиан одномерными:

$$p(\mathbf{z}) = N(\mathbf{z} \mid \mu, \Lambda^{-1}),$$
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

- И мы хотим приблизить $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$.
- Повычисляем...

- ...получится, что

$$\ln q_1^*(z_1) = -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_{11}\Lambda_{11} - z_1\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) + \text{const.}$$

- Чудесным образом получился гауссиан! Сам собой, без предположений.
- Найдём его среднее и дисперсию...

- ...получится

$$q_1^*(z_1) = N(z_1 \mid m_1, \Lambda_{11}^{-1}), \text{ где}$$

$$m_1 = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (E[z_2] - \mu_2).$$

- Аналогично,

$$q_2^*(z_2) = N(z_2 \mid m_2, \Lambda_{22}^{-1}), \text{ где}$$

$$m_2 = \mu_2 - \Lambda_{22}^{-1} \Lambda_{21} (E[z_1] - \mu_1).$$

- Какое решение у этой системы?

- Да просто $E[z_1] = m_1 = \mu_1$, $E[z_2] = m_2 = \mu_2$.
- А если бы мы минимизировали $KL(p\|q)$, получилось бы

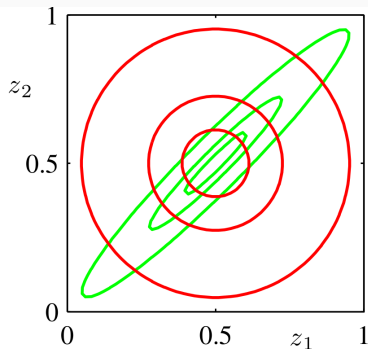
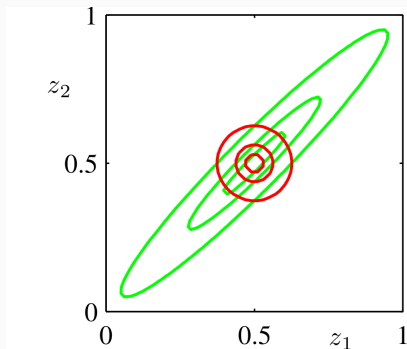
$$KL(p\|q) = - \int p(\mathbf{Z}) \left[\sum_i \ln q_i(\mathbf{Z}_i) \right] d\mathbf{Z} + \text{const},$$

и можно оптимизировать по отдельности:

$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j).$$

- Т.е. просто маргинализация.
- Почему бы так и не сделать? В чём разница?

- Разные дисперсии ответа:

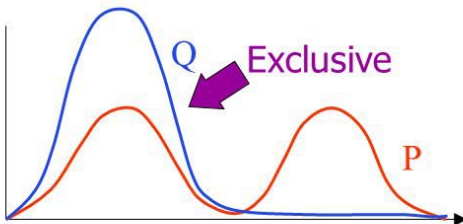


РАЗНЫЕ KL-ДИВЕРГЕНЦИИ

- Минимизация $KL(p||q)$ покрывает всю p , а $KL(q||p)$ строит всю q в регионе больших p :

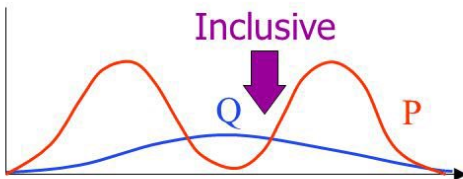
Minimising
 $KL(Q||P)$

$$= \sum_H Q(H) \ln \frac{Q(H)}{P(H|V)}$$

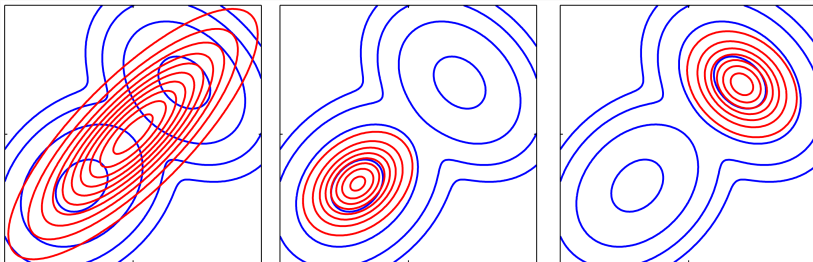


Minimising
 $KL(P||Q)$

$$= \sum_H P(H|V) \ln \frac{P(H|V)}{Q(H)}$$



- Например, для двумерного гауссиана:



- В машинном обучении гораздо интереснее, конечно, пик найти.

ВАРИАЦИОННОЕ ПРИБЛИЖЕНИЕ ДЛЯ ГАУССИАНА

- И ещё пример: давайте найдём параметры одномерного гауссиана по точкам $\mathbf{X} = \{x_1, \dots, x_N\}$. Правдоподобие:

$$p(\mathbf{X} \mid \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} e^{-\frac{\tau}{2} \sum_{n=1}^N (x_n - \mu)^2}.$$

- Вводим сопряжённые априорные распределения:

$$\begin{aligned} p(\mu \mid \tau) &= N(\mu \mid \mu_0, (\lambda_0 \tau)^{-1}), \\ p(\tau) &= \text{Gamma}(\tau \mid a_0, b_0). \end{aligned}$$

- Мы это только что подсчитали точно, но давайте приблизим теперь апостериорное распределение как

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau).$$

- На самом деле так не раскладывается!
- Это то, что мы делали для $q(\mathbf{Z}) = \prod_{i=1}^M q_i(\mathbf{Z}_i)$. Посчитаем...

- ... $q_\mu(\mu)$ – гауссиан с параметрами

$$\mu_N = \frac{\lambda_0 \mu_0 + N \bar{x}}{\lambda_0 + N}, \quad \lambda_N = (\lambda_0 + N) \mathbb{E}[\tau].$$

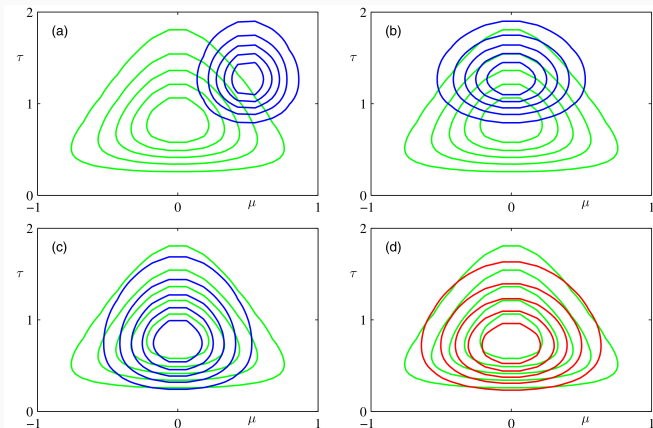
- А $q_\tau(\tau)$ – гамма-распределение с параметрами

$$a_N = a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{1}{2} \mathbb{E}_\mu \left[\sum_n (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right].$$

- Всё получилось как надо, но без предположений о форме q_τ и q_μ .

Одномерный ГАУССИАН

- Вот такой вывод в пространстве (μ, τ) :



- А для $\mu_0 = a_0 = b_0 = \lambda_0 = 0$ (non-informative priors) можно и точно посчитать...

- Получатся моменты для μ

$$E[\mu] = \bar{x}, \quad E[\mu^2] = \bar{x}^2 + \frac{1}{NE[\tau]}.$$

- Это можно подставить и найти $E[\tau]$:

$$\frac{1}{E[\tau]} = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2.$$

- Автоматически получили несмещённую оценку дисперсии!

ВАРИАЦИОННОЕ ПРИБЛИЖЕНИЕ ДЛЯ СМЕСИ ГАУССИАНОВ

- Смесь гауссианов: $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$,

$$p(\mathbf{Z} \mid \pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}},$$

$$p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda) = \prod_{n=1}^N \prod_{k=1}^K N(\mathbf{x}_n \mid \mu_k, \Lambda_k^{-1}).$$

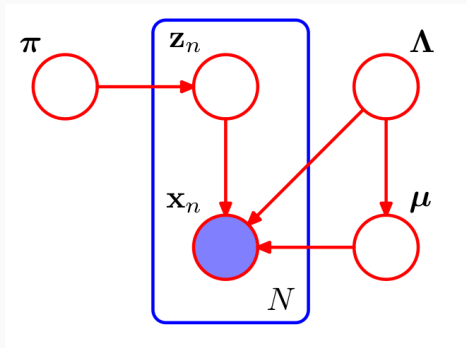
- Выберем сопряжённые априорные распределения:

$$p(\pi) = \text{Dir}(\pi \mid \alpha_0) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_0 - 1},$$

$$p(\mu, \Lambda) = p(\mu \mid \Lambda) p(\Lambda)$$

$$= \prod_{k=1}^K N(\mu_k \mid \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1}) W(\Lambda_k \mid \mathbf{W}_0, \nu_0).$$

- Вот такая графическая модель:



- Распределение Дирихле пусть будет симметричное для простоты; часто ещё $\mathbf{m}_0 = 0$.
- Заметьте разницу между латентными переменными и параметрами модели.

- Теперь вариационное приближение. Сначала сама факторизация:

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda)p(\mathbf{Z} \mid \pi)p(\pi)p(\mu \mid \Lambda)p(\Lambda).$$

- Мы наблюдаем только $\mathbf{X}$, остальное всё надо как-то оценить.
- Интересно, что единственное предположение про наше вариационное приближение выглядит так:

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda).$$

- И всё! Дальше всё само собой получится. Но не сразу...

- Сначала $q^*(\mathbf{Z})$:

$$\begin{aligned}\ln q^*(\mathbf{Z}) &= \mathbb{E}_{\pi, \mu, \Lambda} [\ln p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda)] + \text{const} \\ &= \mathbb{E}_{\pi, \mu, \Lambda} [\ln p(\mathbf{Z} \mid \pi)] + \mathbb{E}_{\mu, \Lambda} [\ln p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda)] + \text{const} \\ &= \dots = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \ln \rho_{nk} + \text{const},\end{aligned}$$

$$\begin{aligned}\text{где } \ln \rho_{nk} &= \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln(2\pi) - \\ &\quad - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k} [(\mathbf{x}_n - \mu_k)^\top \Lambda_k (\mathbf{x}_n - \mu_k)].\end{aligned}$$

- Нормируем:

$$q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}}, \quad \text{где } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}}.$$

- Теперь $E[z_{nk}] = r_{nk}$, т.е. r_{nk} – то, насколько точка $\mathbf{x}_n$ принадлежит кластеру k .
- Можно определить статистики с их учётом, как обычно:

$$N_k = \sum_{n=1}^N r_{nk},$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} \mathbf{x}_n,$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top.$$

- То же самое происходило и в EM-алгоритме.

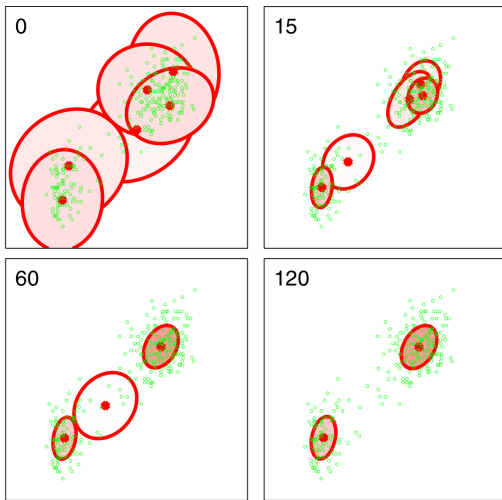
- Теперь $q^*(\pi, \mu, \Lambda)$:

$$\begin{aligned}\ln q^*(\pi, \mu, \Lambda) &= \ln p(\pi) + \sum_{k=1}^K \ln p(\mu_k, \Lambda_k) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} \mid \pi)] \\ &\quad + \sum_{k=1}^K \sum_{n=1}^N \mathbb{E}[z_{nk}] \ln N(\mathbf{x}_n \mid \mu_k \Lambda_k^{-1}) + \text{const.}\end{aligned}$$

- Вот уже получилось, что $q^*(\pi, \mu, \Lambda)$ раскладывается в $q^*(\pi)q^*(\mu, \Lambda)$, опять же без предположений.
- Более того, $q^*(\mu, \Lambda) = \prod_{k=1}^K q(\mu_k, \Lambda_k)$.
- И теперь можно по отдельности посчитать (упражнение), получится типичный М-шаг.
- Причём распределения останутся той же формы (т.к. были сопряжённые).

ВАРИАЦИОННОЕ ПРИБЛИЖЕНИЕ

- Теперь даже model selection автоматически получается, просто у некоторых компонент $N_k \approx 0$:



- Никакого оверфиттинга или коллапса компонент.

- Есть другие примеры вариационных приближений.
- Обращение матриц; например, для линейной регрессии надо посчитать $\beta^* = C^{-1}\mathbf{b}$:

$$J(\beta) = \frac{1}{2}(\beta^* - \beta)^\top C(\beta^* - \beta) = \dots = \text{Const} - \beta^\top \mathbf{b} + \frac{1}{2}\beta^\top C\beta,$$

и теперь можно решать такую задачу выпуклой оптимизации.

- Метод конечных элементов – для уравнения Пуассона $-u''(x) = f(x)$, $x \in (a, b)$:

$$J(u) = \frac{1}{2} \int_a^b (u'(x) - u^{*'}(x))^2 dx = \dots = \text{Const} - \int_a^b u(x)f(x)dx + \frac{1}{2} \int_a^b u'(x)^2 dx,$$

и если ищем в подпространстве $\tilde{u}(x) = \sum_{i=1}^k \alpha_i \phi_i(x)$, то опять

$$\tilde{J}(\alpha) = \alpha^\top \mathbf{b} + \frac{1}{2}\alpha^\top C\alpha.$$

- В графических моделях – теория среднего поля (mean field theory). Пусть дано $p(\mathbf{x})$, $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_h)$, и надо найти

$$\log p(\mathbf{x}_v) = \log \sum_{\mathbf{x}_h} p(\mathbf{x}_v, \mathbf{x}_h), \quad p(\mathbf{x}_h | \mathbf{x}_v) = p(\mathbf{x}_h, \mathbf{x}_v) / p(\mathbf{x}_v).$$

- Опять делаем тот же трюк:

$$J(q) = \log p(\mathbf{x}_v) - \text{KL}(q_{\mathbf{x}_h} \| p_{\mathbf{x}_h | \mathbf{x}_v}) = \log p(\mathbf{x}_v) - \sum_{\mathbf{x}_h} q(\mathbf{x}_h) \log \frac{q(\mathbf{x}_h)}{p(\mathbf{x}_h | \mathbf{x}_v)}$$

$$\dots = H(q) + \mathbb{E}_q [\log p(\mathbf{x}_h, \mathbf{x}_v)] = H(q) + \sum_{C \in \mathcal{C}} \sum_{\mathbf{x}_{C \cap h}} q(\mathbf{x}_{C \cap h}) \log \Psi_C(\mathbf{x}_C),$$

где $q(\mathbf{x}_{C \cap h})$ – маргинальная вероятность по скрытым переменным из клики C .

- Теория среднего поля – это когда $q(\mathbf{x}_h) = \prod_{i \in h} q_i(x_i)$.

Спасибо за внимание!

