Генеративные модели

Лекция 3: Модели со скрытыми переменными

Идея нормализующих потоков

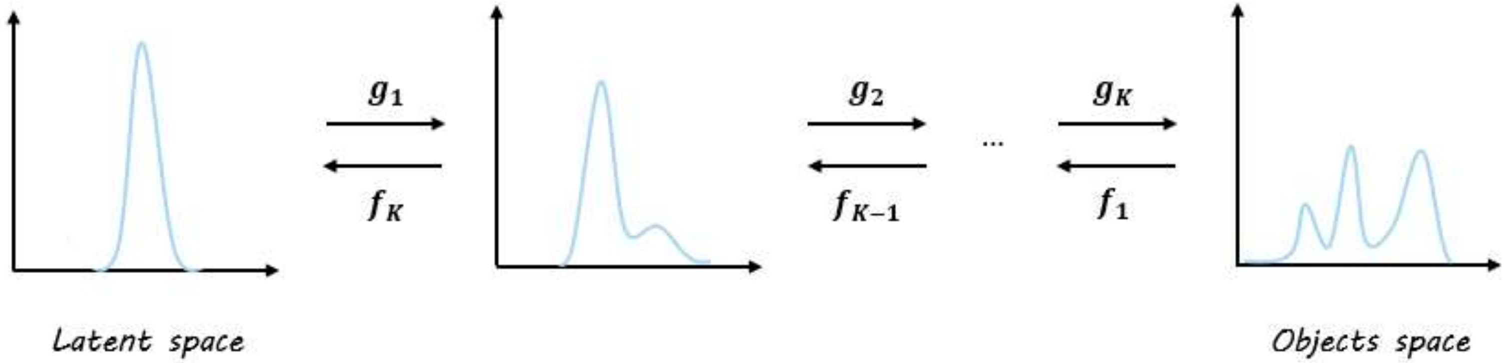
Особенность потоков – наличие двух направлений:

- ***Нормализующее направление*** (***для обучения***) $\mathbf{x} \to \boldsymbol{z}$:

  Берём реальный объект $\mathbf{x}$ и с помощью нормализующей функции $\boldsymbol{f}$ превращаем его в простой $\boldsymbol{z}$

- **Генеративное направление** $\mathbf{z} \to \mathbf{x}$:

  Берём простой объект $\boldsymbol{z}$ и с помощью генеративной функции $\boldsymbol{g} = \boldsymbol{f}^{-1}$ создаём объект $\mathbf{x}$



Теорема о замене переменных

Наша цель – найти распределение $p(\mathbf{x})$

Связать неизвестную плотность $p(\mathbf{x})$ с известной плотностью простого распределения $p(\mathbf{z})$ нам позволяет ***формула замены переменных***:

$$p(\mathbf{x}) = p(\mathbf{z})\left|\det\left(\boldsymbol{J_f}\right)\right| = p(\mathbf{z})\left|\det\left(\frac{\partial\mathbf{z}}{\partial\mathbf{x}}\right)\right| = p(\boldsymbol{f}(\mathbf{x}))\left|\det\left(\frac{\partial\boldsymbol{f}(\mathbf{x})}{\partial\mathbf{x}}\right)\right|$$

Здесь $\boldsymbol{J_f}(\mathbf{x})$ – матрица Якоби:

$$\boldsymbol{J_f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_D}{\partial x_1} & \cdots & \frac{\partial f_D}{\partial x_D} \end{pmatrix}$$

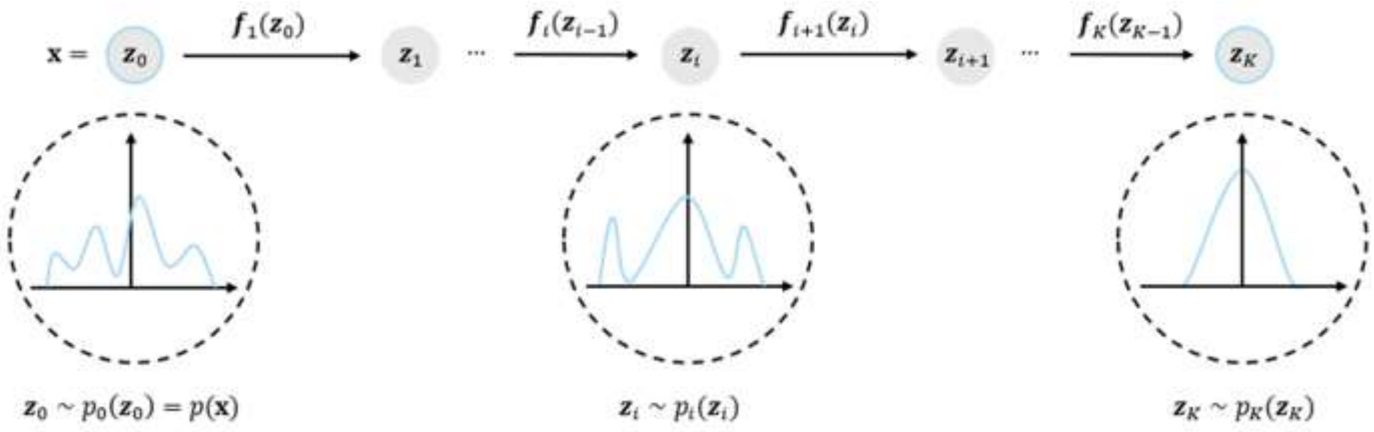
Композиция преобразований

- Одной функции $\boldsymbol{f}$ недостаточно, чтобы из $\mathcal{N}(0,1)$ получить сложное распределение наших данных

- Потоки используют последовательность из K простых и обратимых преобразований:

$$\mathbf{x} = \boldsymbol{z}_0 \xrightarrow{f_1} \boldsymbol{z}_1 \xrightarrow{f_2} \dots \xrightarrow{f_K} \boldsymbol{z}_K$$

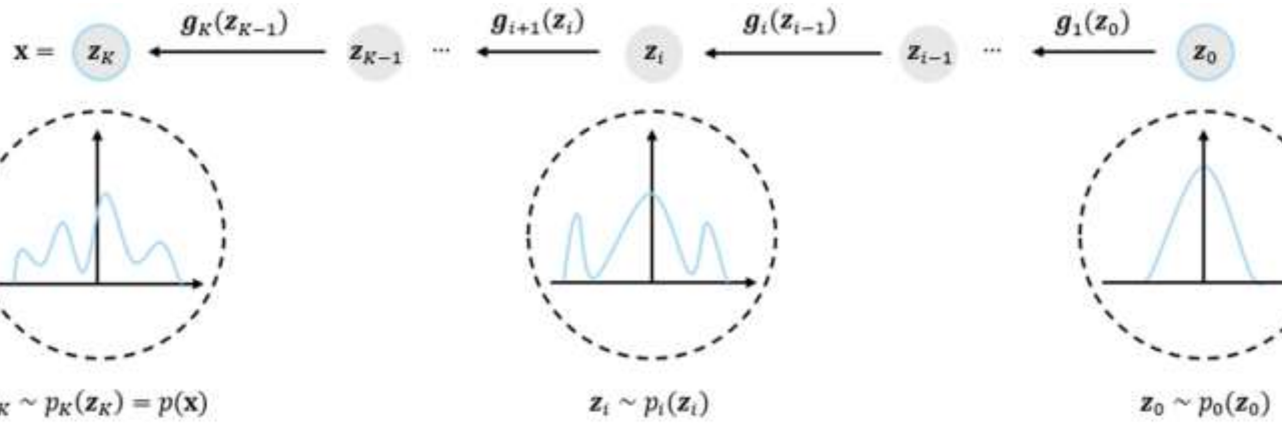
- Во время обучения мы используем нормализующее направление $\mathbf{x} \to \mathbf{z}$:



Композиция преобразований

- Генеративное направление $\boldsymbol{z} \to \mathbf{x}$ используется для создания новых объектов:

$$\boldsymbol{z}_0 \xrightarrow{g_1} \boldsymbol{z}_1 \xrightarrow{g_2} \dots \xrightarrow{g_K} \boldsymbol{z}_K = \mathbf{x}$$



Композиция преобразований

Нужно найти определитель якобиана всей композиции

Два математических факта:

- Матрица Якоби для композиций функций равна их произведению:

$$\frac{\partial \boldsymbol{f}}{\partial \mathbf{x}} = \frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{z}_{K-1}} \cdot \frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{z}_{K-2}} \cdot \ldots \cdot \frac{\partial \boldsymbol{f}_1}{\partial \boldsymbol{z}_0}$$

- Определитель произведения матриц равен произведению определителей:

$$\det\left(\boldsymbol{J}_{\boldsymbol{f}}\right) = \det\left(\boldsymbol{J}_{\boldsymbol{f}_K}\right) \cdot \det\left(\boldsymbol{J}_{\boldsymbol{f}_{K-1}}\right) \cdot \ldots \cdot \det\left(\boldsymbol{J}_{\boldsymbol{f}_1}\right)$$

Будем считать определители для каждого шага $\boldsymbol{f}_k$ и перемножать их:

$$\det\left(\frac{\partial \boldsymbol{f}}{\partial \mathbf{x}}\right) = \prod_{k=1}^{K} \det\left(\frac{\partial \boldsymbol{f}_k}{\partial \boldsymbol{z}_{k-1}}\right)$$

Обучение нормализующих потоков

Чтобы сделать эту композицию обучаемой, представим $\boldsymbol{f}_k$ как нейросеть с параметрами $\boldsymbol{\theta}$

Плотность, которую моделирует наша модель:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}))\left|\det\left(\frac{\partial \boldsymbol{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}}\right)\right| = p(\boldsymbol{z}_K)\left|\prod_{k=1}^{K}\det\left(\frac{\partial \boldsymbol{f}_{k,\boldsymbol{\theta}}}{\partial \boldsymbol{z}_{k-1}}\right)\right|$$

$\boldsymbol{z}_K$ – финальный вектор в латентном пространстве

Максимизируем правдоподобие:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log p(\boldsymbol{z}_K) + \sum_{k=1}^{K}\log\left|\det\left(\frac{\partial \boldsymbol{f}_{k,\boldsymbol{\theta}}}{\partial \boldsymbol{z}_{k-1}}\right)\right|$$

- $\log p(\boldsymbol{z}_K)$ отвечает за то, чтобы выход потока соответствовал базовому распределению
- $\sum_{k=1}^{K}\log\left|\det\left(\frac{\partial \boldsymbol{f}_{k,\boldsymbol{\theta}}}{\partial \boldsymbol{z}_{k-1}}\right)\right|$ отслеживает суммарное изменение объема после всех преобразований

План

- Модели со скрытыми переменными

- EM алгоритм

- Вариационный вывод

Модели со скрытыми переменными

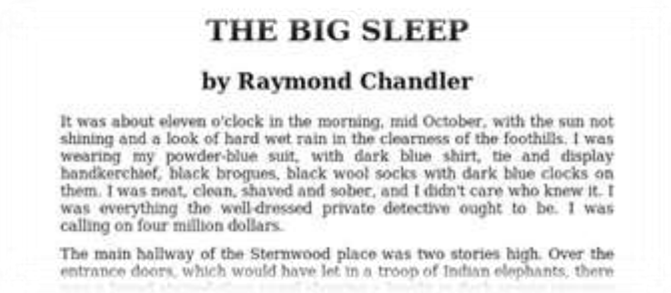
Проблема высокой размерности

Ищем $\boldsymbol{\theta}$, которые максимизируют правдоподобие:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$$

Объекты в реальном мире – многомерные случайные величины

Прямое моделирование $p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{D})$ – сложная задача из-за проклятия размерности

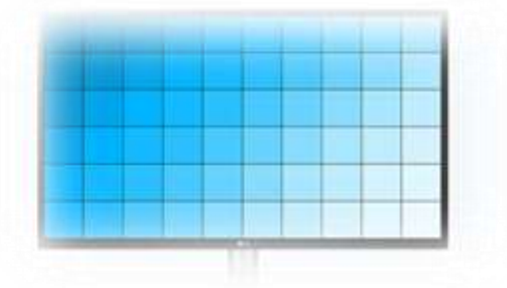


THE BIG SLEEP

by Raymond Chandler

It was about eleven o'clock in the morning, mid October, with the sun not shining and a look of hard wet rain in the clearness of the foothills. I was wearing my powder-blue suit, with dark blue shirt, tie and display handkerchief, black brogues, black wool socks with dark blue clocks on them. I was neat, clean, shaved and sober, and I didn't care who knew it. I was everything the well-dressed private detective ought to be. I was calling on four million dollars.

The main hallway of the Sternwood place was two stories high. Over the entrance doors, which would have let in a troop of Indian elephants, there



Модели со скрытыми переменными

Предположим, что помимо наблюдаемой переменной $\mathbf{x}$ в нашей задаче есть некоторая скрытая переменная $\boldsymbol{z}$, которая влияет на $\mathbf{x}$

Мы не можем измерить ее напрямую, но знаем, что она есть

Мы можем записать совместную плотность и разложить ее на условную и априорную:

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta}) \cdot p(\boldsymbol{z}|\boldsymbol{\theta})$$

$p(\boldsymbol{z}|\boldsymbol{\theta})$ – априорное распределение на скрытых переменных
$p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta})$ – условное распределение наблюдаемых данных

Проинтегрируем по всем $\mathbf{z}$, чтобы найти правдоподобие:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\, d\boldsymbol{z} = \int p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta}) \cdot p(\boldsymbol{z}|\boldsymbol{\theta})\, d\boldsymbol{z}$$

Модели со скрытыми переменными

Наша цель – по-прежнему максимизация правдоподобия:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \qquad \rightarrow \qquad \boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \int p(\mathbf{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \cdot p(\boldsymbol{z}_i|\boldsymbol{\theta})\, d\boldsymbol{z}_i$$

Мы заменили одно сложное распределение $p(\mathbf{x}|\boldsymbol{\theta})$ на интеграл от двух более простых:

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \int p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta}) \cdot p(\boldsymbol{z}|\boldsymbol{\theta})\, d\boldsymbol{z} \quad \rightarrow \quad \max_{\boldsymbol{\theta}}$$

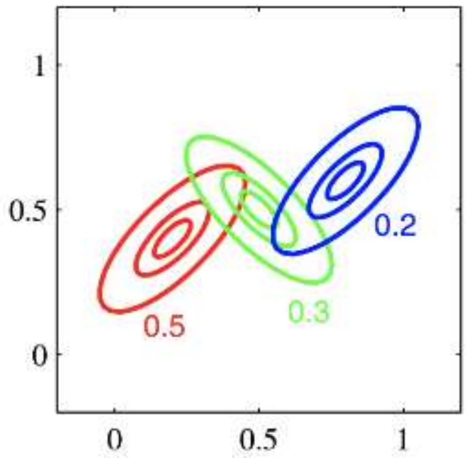
Выбирая подходящие распределения для $p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta})$ и $p(\boldsymbol{z}|\boldsymbol{\theta})$, мы можем либо получить аналитическое решение для $\boldsymbol{\theta}^*$, либо использовать методы оптимизации

Gaussian Mixture Models

- Реальные данные редко описываются одним простым распределением
- Мы можем представить сложное распределение как взвешенную сумму (смесь) более простых распределений

Самый популярный представитель – ***модели гауссовских смесей*** (***Gaussian Mixture Models***)

- Сначала случайно берём одну из K компонент
- Затем семплируем вектор $\mathbf{x}$ из выбранного распределения $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Номер компоненты, из которой пришел вектор, является скрытой переменной $\boldsymbol{z}$ (мы её не наблюдаем)



Gaussian Mixture Models

Скрытая переменная $\boldsymbol{z}$ – это бинарный вектор $\boldsymbol{z} = \{0,1\}^K$, где $\boldsymbol{z}_k$ означает принадлежность к k-ой компоненте

Априорная вероятность $p(\boldsymbol{z}|\boldsymbol{\theta})$ задается весами смеси π_k, при этом:

$$\sum_{k=1}^{K} \pi_k = 1$$

Если точка пришла из k-ой компоненты, то ее вероятность описывается гауссианой:

$$p(\mathbf{x}|\boldsymbol{z}_k = 1, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

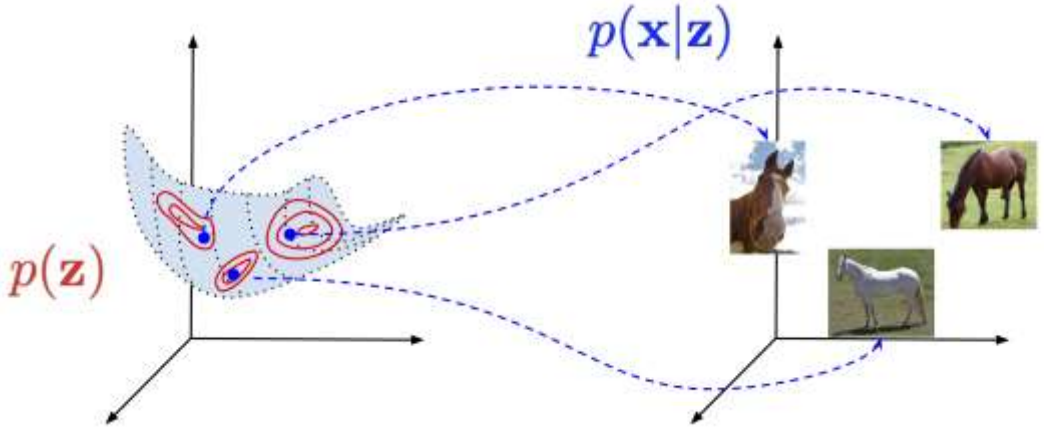
Итоговое правдоподобие:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\boldsymbol{z}_k = 1|\boldsymbol{\theta}) \cdot p(\mathbf{x}|\boldsymbol{z}_k = 1, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

MLE для моделей со скрытыми переменными

Нужно вычислить сложный интеграл:

$$\sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) \cdot p(\mathbf{z}_i)\, d\mathbf{z}_i$$

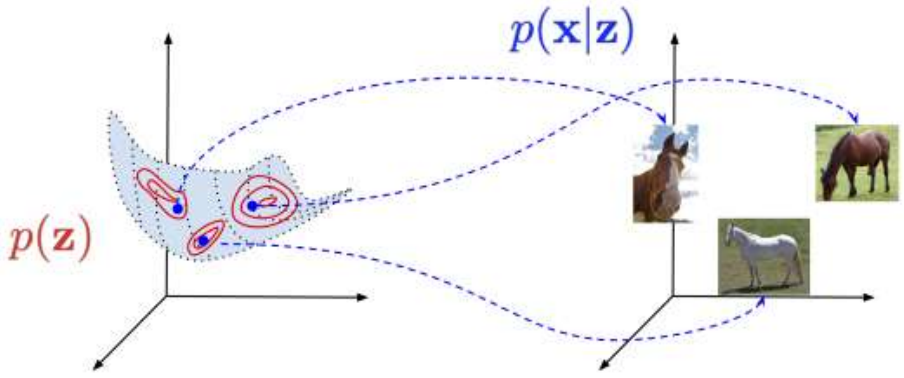


Попробуем оценить этот интеграл методом Монте-Карло:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) \cdot p(\mathbf{z})\, d\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})] \approx \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{x}|\mathbf{z}_k, \boldsymbol{\theta})$$

MLE для моделей со скрытыми переменными

- Большинство сэмплов $\boldsymbol{z}_k$ попадут в «бесполезные» области и декодер сгенерирует шум



Наше априорное распределение $p(\boldsymbol{z})$ ничего не знает об изображениях $\mathbf{x}$

В многомерном пространстве объем «пустого» пространства огромен по сравнению с объемом полезной области

Чтобы получить надежную оценку интеграла, нам понадобиться экспоненциально большое число сэмплов K

EM алгоритм

Модели со скрытыми переменными

В моделях со скрытыми переменными наша цель – максимизировать логарифм правдоподобия $\log p\,(\boldsymbol{X}|\boldsymbol{\theta})$

Из-за наличия скрытых переменных правдоподобие выражается через интеграл (сумму)

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \int p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\theta})\, d\boldsymbol{Z}$$

Итоговая функция, которую мы максимизируем, содержит интеграл внутри логарифма:

$$\log p\,(\boldsymbol{X}|\boldsymbol{\theta}) = \log \int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \quad \rightarrow \quad \max_{\boldsymbol{\theta}}$$

Наивная оценка не работает:

Попытка оценить эти интегралы методом Монте-Карло проваливается из-за проклятия размерности

Формула Байеса

Введем понятия:

$\boldsymbol{X}$ – наши реальные, неполные данные
$\{\boldsymbol{X}, \boldsymbol{Z}\}$ – полные данные

Мы не знаем точные значения $\boldsymbol{Z}$, но можем вычислить апостериорное распределение:

$$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}) \cdot p(\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{X}|\boldsymbol{\theta})}$$

- $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ – ***апостериорное распределение***, вероятность того, что скрытые переменные приняли значение $\boldsymbol{Z}$, после того как мы увидели данные $\boldsymbol{X}$
- $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta})$ – ***правдоподобие***, вероятность увидеть данные $\boldsymbol{X}$, если скрытые переменные равны $\boldsymbol{Z}$
- $p(\boldsymbol{Z}|\boldsymbol{\theta})$ – ***априорное распределение***, насколько вероятны значения $\boldsymbol{Z}$ до наблюдений
- $p(\boldsymbol{X}|\boldsymbol{\theta})$ – ***обоснованность*** (***evidence***) вероятность увидеть данные $\boldsymbol{X}$

ЕМ алгоритм

- Пусть у нас есть уже есть некоторая оценка параметров $\boldsymbol{\theta}_{old}$
- Будем итеративно выполнять 2 шага

Е-шаг (***Expectation***):

Используя текущие значения $\boldsymbol{\theta}_{old}$, находим апостериорное распределение $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old})$ и вычисляем матожидание логарифма правдоподобия для полных данных $\{\boldsymbol{X}, \boldsymbol{Z}\}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \int p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old}) \cdot \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})\, d\boldsymbol{Z}$$

М-шаг (***Maximization***):

Обновляем оценку параметров $\boldsymbol{\theta}$, новые значения выбираются таким образом, чтобы максимизировать $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$:

$$\boldsymbol{\theta}_{new} = \arg \underset{\boldsymbol{\theta}}{max}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$$

ЕМ алгоритм

Чтобы вычислить знаменатель $p(\boldsymbol{X}|\boldsymbol{\theta})$ в формуле для апостериорного распределения, нам нужно взять интеграл:

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z}$$

В случае $\boldsymbol{GMM}$ мы можем получить аналитическое решение:

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\boldsymbol{Z} = k|\boldsymbol{\theta}) \cdot p(\boldsymbol{X}|\boldsymbol{Z} = k, \boldsymbol{\theta})$$

ЕМ алгоритм

ЕМ алгоритм прекрасен, но для более сложных моделей вычислить истинное апостериорное распределение $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}_{old})$ на ***Е-шаге*** аналитически невозможно

- Будем аппроксимировать распределение $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}_{old})$
- В этом и состоит идея ***вариационного вывода*** ($\boldsymbol{Variational\ Inference}$)

Вариационный вывод

Вариационный вывод

Наша цель – оценить $\log p(\mathbf{x}|\boldsymbol{\theta})$:

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \int p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z}$$

Введем вспомогательное распределение $q(\boldsymbol{z})$ – произвольную плотность над скрытыми переменными $\boldsymbol{z}$

Идея:

Мы не знаем $\log p(\mathbf{x}|\boldsymbol{\theta})$, но можем попытаться построить для него некую ***нижнюю границу*** (***lower bound***), которую будем оптимизировать

Вывод $ELBO$ (1)

Умножим и разделим выражение для $\log p\,(\mathbf{x}|\boldsymbol{\theta})$ на $q(\boldsymbol{z})$:

$$\log p\,(\mathbf{x}|\boldsymbol{\theta}) = \log \int p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z} = \log \int \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z} =$$

$$\log \mathbb{E}_q \left[\frac{p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta})}{q(\boldsymbol{z})}\right] \geq \mathbb{E}_q \log \frac{p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta})}{q(\boldsymbol{z})} = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

Неравенство Йенсена:

$$\log \mathbb{E}[X] \geq \mathbb{E}[\log X]$$

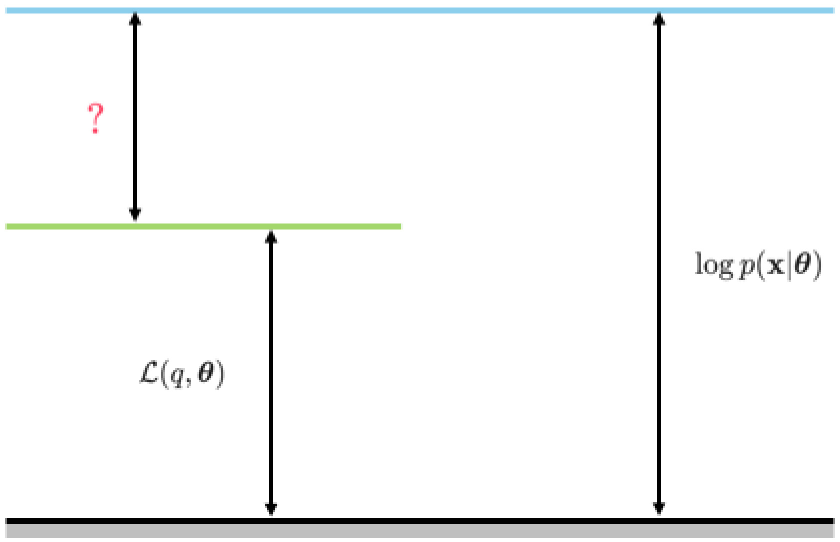
Мы получили ***вариационную нижнюю оценку*** (***Evidence Lower Bound***, ***ELBO***), которая утверждает, что для любого $q(\boldsymbol{z})$:

$$\log p\,(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

Вывод $ELBO$ (1)

$q(\boldsymbol{z})$ будем называть ***вариационным распределением***

Насколько далеко наша нижняя граница $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ находится от истинного значения $\log p\,(\mathbf{x}|\boldsymbol{\theta})$?



Вывод $ELBO$ (2)

Перепишем выражение для $\boldsymbol{ELBO}$:

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\boldsymbol{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\boldsymbol{z})} d\boldsymbol{z} = \int q(\boldsymbol{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) \cdot p(\mathbf{x}|\boldsymbol{\theta})}{q(\boldsymbol{z})} d\boldsymbol{z} =$$

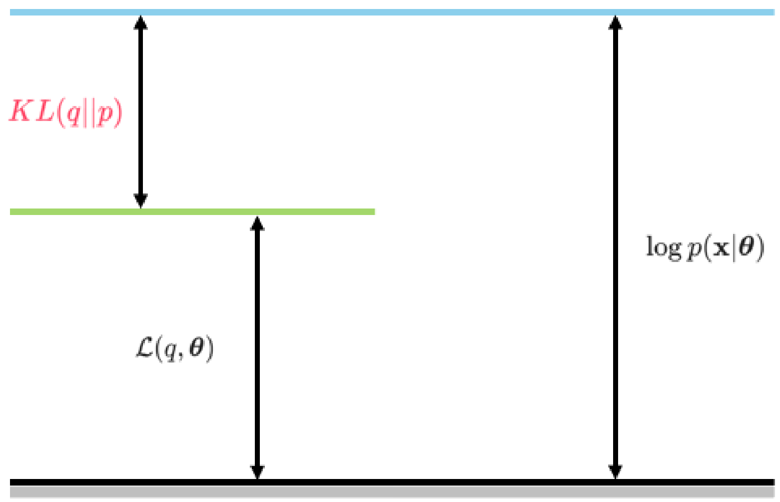
$$\int q(\boldsymbol{z}) \log p(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{z} + \int q(\boldsymbol{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{z})} d\boldsymbol{z} = \log p(\mathbf{x}|\boldsymbol{\theta}) \int q(\boldsymbol{z}) d\boldsymbol{z} + \int q(\boldsymbol{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{z})} d\boldsymbol{z} =$$

$$\log p(\mathbf{x}|\boldsymbol{\theta}) + \int q(\boldsymbol{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{z})} d\boldsymbol{z} = \log p(\mathbf{x}|\boldsymbol{\theta}) + KL(q(\boldsymbol{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))$$

Вывод $ELBO$ (2)

Зазор между $\log p\,(\mathbf{x}|\boldsymbol{\theta})$ и $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ равен KL-дивергенции между вариационным распределением $q(\boldsymbol{z})$ и истинным апостериорным распределением $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})$:

$$\log p\,(\mathbf{x}|\boldsymbol{\theta}) - \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}))$$



Вывод $ELBO$ (2)

Получим ещё одно выражение для $\boldsymbol{ELBO}$, используя разложение $p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta}) \cdot p(\boldsymbol{z})$

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\boldsymbol{z}) \log \frac{p(\mathbf{x}, \boldsymbol{z}|\boldsymbol{\theta})}{q(\boldsymbol{z})} d\boldsymbol{z} = \int q(\boldsymbol{z}) \log p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta}) d\boldsymbol{z} + \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z}$$

$$\underbrace{\mathbb{E}_q[\log p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta})]}_{\textit{Reconstruction loss}} - \underbrace{KL(q(\boldsymbol{z})||p(\boldsymbol{z}))}_{\textit{Regularization loss}}$$

Итоговое правдоподобие:

$$\log p\,(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + KL\big(q(\boldsymbol{z})\big|\big|p(\boldsymbol{z}|\mathbf{x}, \boldsymbol{\theta})\big) =$$

$$\mathbb{E}_q[\log p(\mathbf{x}|\boldsymbol{z}, \boldsymbol{\theta})] - KL(q(\boldsymbol{z})||p(\boldsymbol{z})) + KL\big(q(\boldsymbol{z})\big|\big|p(\boldsymbol{z}|\mathbf{x}, \boldsymbol{\theta})\big)$$

Выводы

Вместо прямой максимизации правдоподобия $\log p(\mathbf{x}|\boldsymbol{\theta})$ будем максимизировать его нижнюю границу $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ по параметрам вариационного распределения q и параметрам модели $\boldsymbol{\theta}$:

$$\max_{q,\boldsymbol{\theta}} \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

Максимизируя $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$, мы одновременно стараемся сделать 2 вещи:

1. Увеличить правдоподобие наблюдаемых данных

2. Уменьшить зазор $KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\mathbf{x},\boldsymbol{\theta}))$, аппроксимируя $p(\boldsymbol{z}|\mathbf{x},\boldsymbol{\theta})$ с помощью $q(\boldsymbol{z})$

EM алгоритм

Мы получили следующее выражение:

$$\log p\,(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) + KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\mathbf{x},\boldsymbol{\theta}))$$

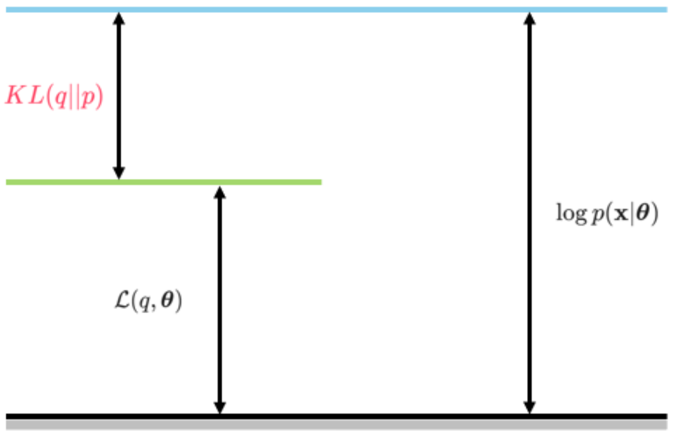
- EM алгоритм можно интерпретировать как метод координатного подъёма (***coordinate ascent***) по ***ELBO***

- Он итеративно оптимизирует $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$, поочерёдно фиксируя либо параметры вариационного распределения $q(\boldsymbol{z})$, либо параметры модели $\boldsymbol{\theta}$

E – шаг

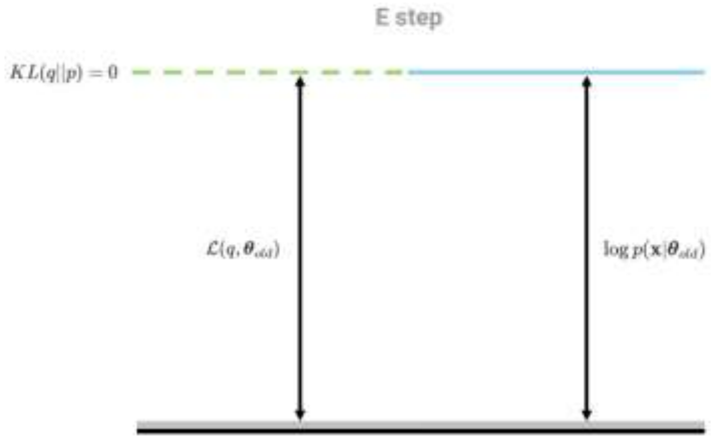
E-шаг:

- Фиксируем $\boldsymbol{\theta}_{old}$ и ищем оптимальное $q(\boldsymbol{z})$, которое максимизирует $\mathcal{L}_{q,\boldsymbol{\theta}_{old}}(\mathbf{x})$



- $\log p\,(\mathbf{x}|\boldsymbol{\theta})$ не зависит от $q(\boldsymbol{z})$, поэтому $\mathcal{L}_{q,\boldsymbol{\theta}_{old}}(\mathbf{x})$ максимальна, когда $KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})) = 0$

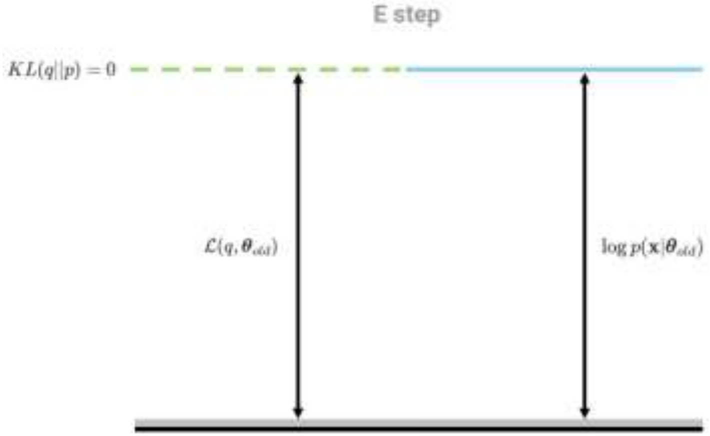
- Наилучший выбор для $q(\boldsymbol{z})$ – это апостериорное распределение $p(\boldsymbol{z}|\mathbf{x},\boldsymbol{\theta})$



M – шаг

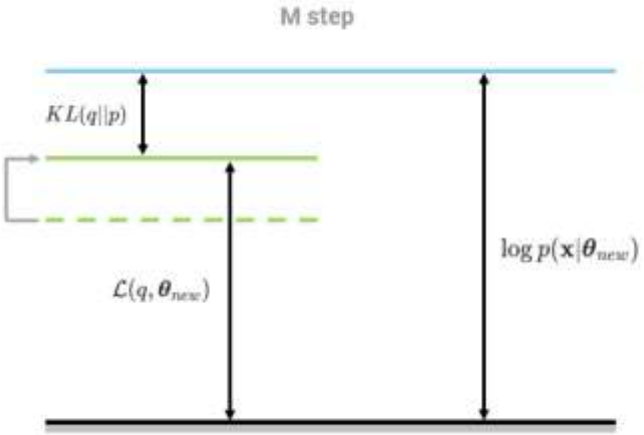
М-шаг:

- Фиксируем найденное распределение $q(\boldsymbol{z})$ и пытаемся увеличить $\mathcal{L}_{q,\boldsymbol{\theta}_{old}}(\mathbf{x})$



- Увеличение нижней границы гарантирует $\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$ увеличение $\log p\,(\mathbf{x}|\boldsymbol{\theta})$ (если мы не в максимуме)

- Распределение $q(\boldsymbol{z})$ было определено на основе старых параметров $\boldsymbol{\theta}_{old}$, поэтому оно не равно апостериорному распределению $p(\boldsymbol{z}|\mathbf{x}, \boldsymbol{\theta}_{new})$

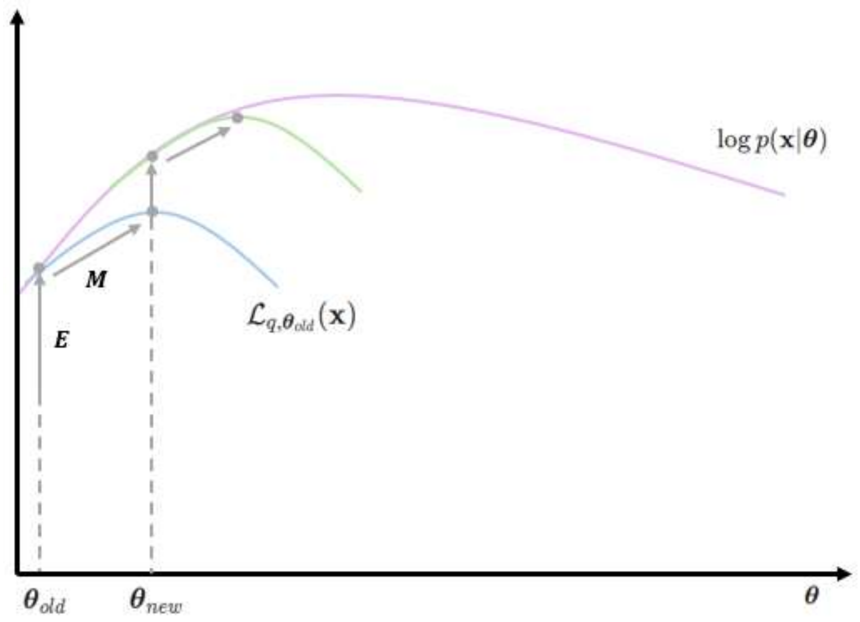


EM – алгоритм

- Работу EM-алгоритма можно рассмотреть в пространстве параметров

- На E-шаге мы строим нижнюю границу $\mathcal{L}_{q,\boldsymbol{\theta}_{old}}(\mathbf{x})$, которая касается $\log p\,(\mathbf{x}|\boldsymbol{\theta})$ в точке $\boldsymbol{\theta}_{old}$

- На M-шаге ищем максимум этой нижней границы (точка $\boldsymbol{\theta}_{new}$), что приводит к большему значению $\log p\,(\mathbf{x}|\boldsymbol{\theta}_{new})$



Спасибо за внимание!