Скрытые марковские модели

Цепь Маркова

Определение: Цепь Маркова

Набор состояний $Q = \{ 1, ..., K \}$

Вероятности переходов a_{st} между любыми двумя состояниями *s* и *t*

$$a_{st} = P(x_i=t \mid x_{i-1}=s)$$
$$a_{i1} + \ldots + a_{iK} = 1, \text{ для всех состояний } i = 1 \ldots K$$

Основное свойство: Вероятность текущего состояния x_i зависит только от предыдущего состояния x_{i-1}:

$$P(x_i|x_{i-1},\ldots,x_1) = P(x_i|x_{i-1})$$

Вероятность последовательности x:

$$P(x) = P(x_L, x_{L-1}, \ldots, x_1) = P(x_L \mid x_{L-1}, \ldots, x_1) P(x_{L-1} \mid x_{L-2}, \ldots, x_1) \ldots P(x_1)$$

используя формулу условной вероятности $P(X,Y) = P(X \mid Y) P(Y)$

$$P(x) = P(x_L \mid x_{L-1}) P(x_{L-1} \mid x_{L-2}) \ldots P(x_2 \mid x_1) P(x_1) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

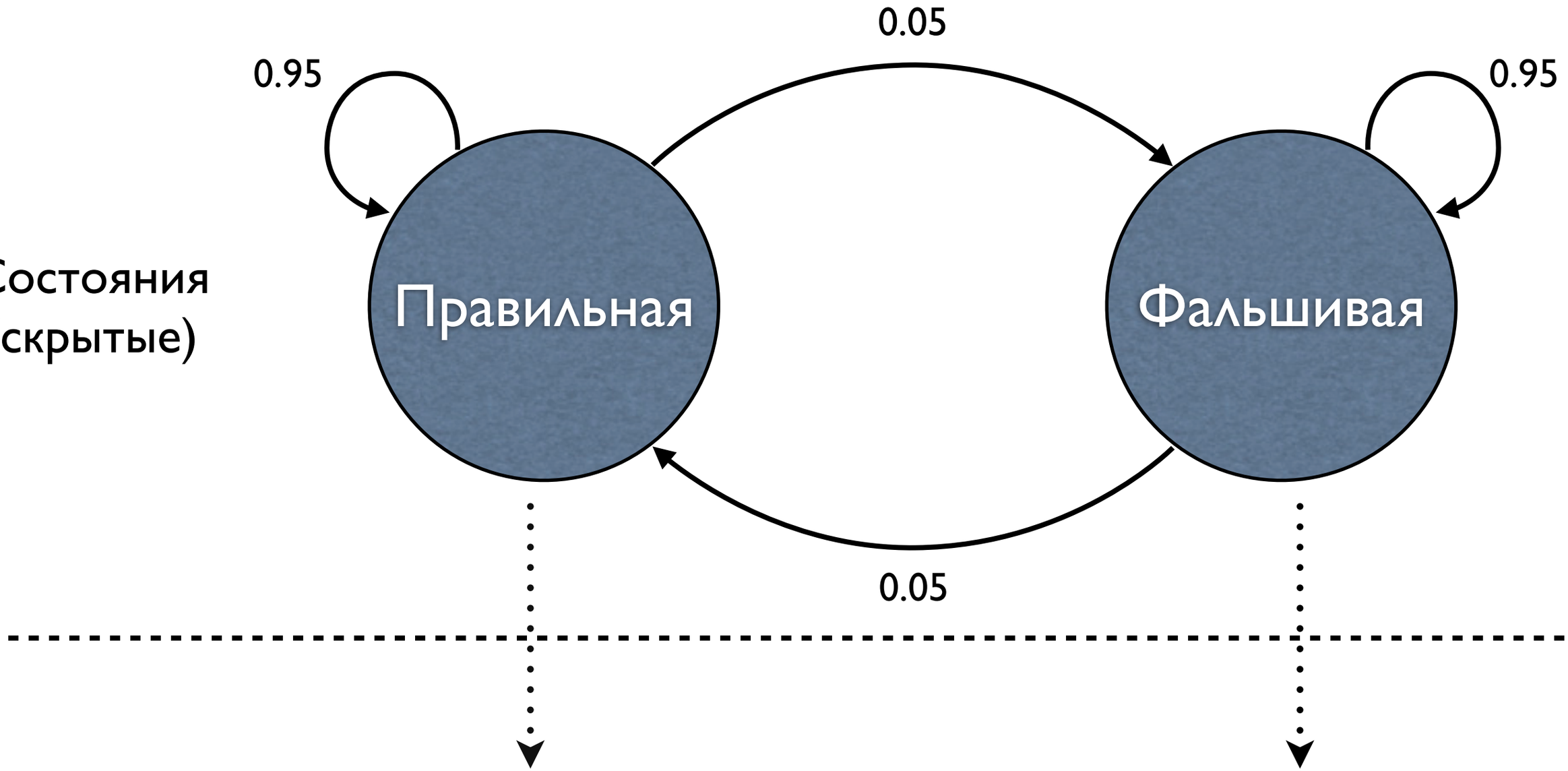
Скрытые марковские модели. Пример: обман в казино

- Обычная кость:
  P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=1/6

- Фальшивая кость:
  P(1)=P(2)=P(3)=P(4)=P(5)=1/10
  P(6)=1/2

- Крупье подменяет кость примерно каждые 20 бросков

Пример: обман в казино. Модель.



Состояния (скрытые)

Символы (наблюдаемые)

P(1|П)=1/6
P(2|П)=1/6
P(3|П)=1/6
P(4|П)=1/6
P(5|П)=1/6
P(6|П)=1/6

P(1|Ф)=1/10
P(2|Ф)=1/10
P(3|Ф)=1/10
P(4|Ф)=1/10
P(5|Ф)=1/10
P(6|Ф)=1/2

Скрытая марковская модель (Hidden Markov Model, HMM)

Определение: Скрытая марковская модель

Набор состояний $Q = \{ 1, ..., K \}$

Вероятности переходов a_{st} между любыми двумя состояниями *s* и *t*

$$a_{st} = P(x_i = t \mid x_{i-1} = s)$$
$$a_{i1} + \ldots + a_{iK} = 1, \quad \text{for all states } i = 1 \ldots K$$

Начальные вероятности a_{0i}

$$a_{01} + \ldots + a_{0K} = 1$$

Набор наблюдаемых символов $= \{ b_1, b_2, \ldots, b_M \}$

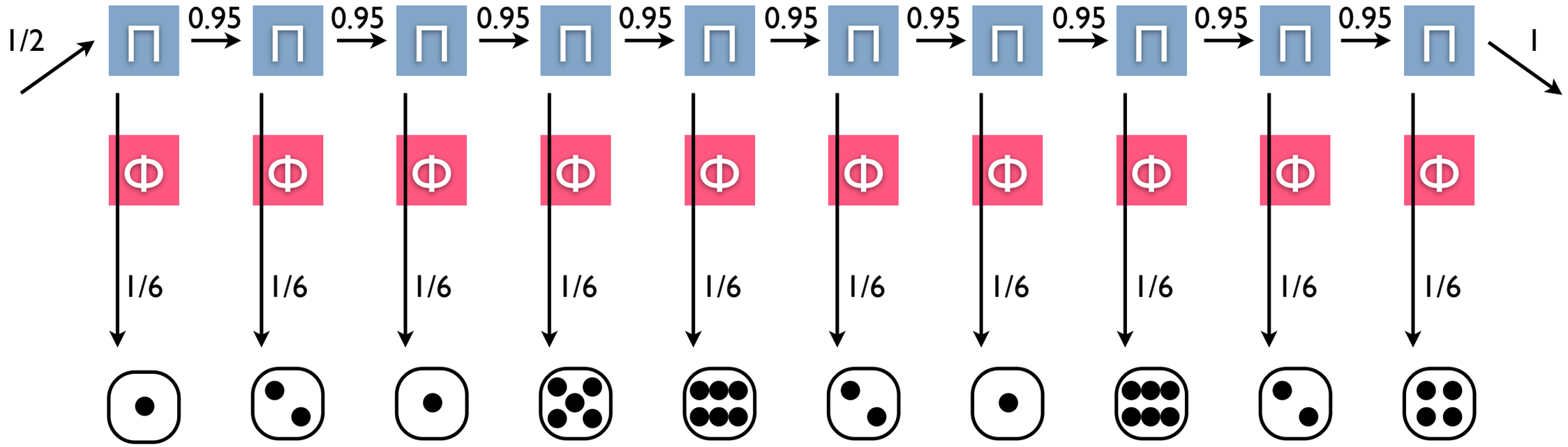
Матрица эмиссионных вероятностей $E = \{e_k(b)\}$

$$e_k(b) = P(x_i = b \mid a_i = s)$$

Совместная вероятность последовательностей символов x и состояний π:

$$P(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

Вероятность последовательностей наблюдений и состояний



Какова совместная вероятность последовательности состояний

π = Правильная, Правильная, Правильная, Правильная, Правильная, Правильная, Правильная, Правильная, Правильная, Правильная

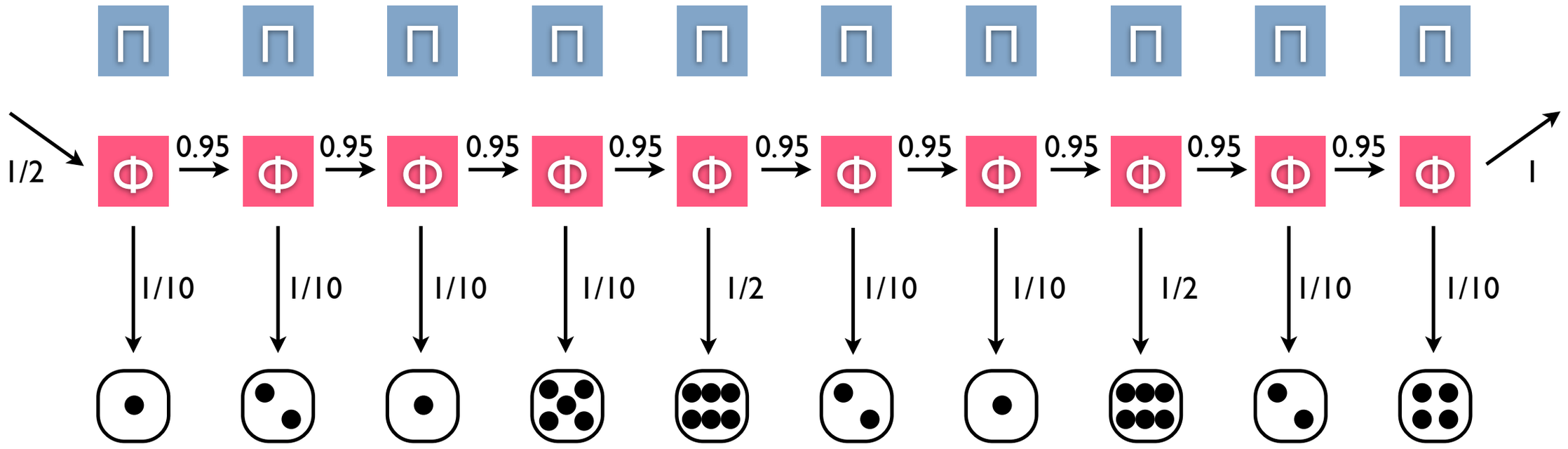
и последовательности символов

x = 1,2,1,5,6,2,1,6,2,4

$P(x,\pi) = 1/2 \times P(1|\text{Правильная}) \times P(\text{Правильная}_2|\text{Правильная}_1) \times P(2|\text{Правильная}) \times P(\text{Правильная}_3|\text{Правильная}_2) \times \ldots \times P(4|\text{Правильная}) =$

$= 1/2 \times (1/6)^{10} \times (0.95)^9 = 0.5 \times 10^{-9}$

Вероятность последовательностей наблюдений и состояний



Какова совместная вероятность последовательности состояний

π = Фальшивая, Фальшивая, Фальшивая, Фальшивая, Фальшивая, Фальшивая, Фальшивая, Фальшивая, Фальшивая, Фальшивая

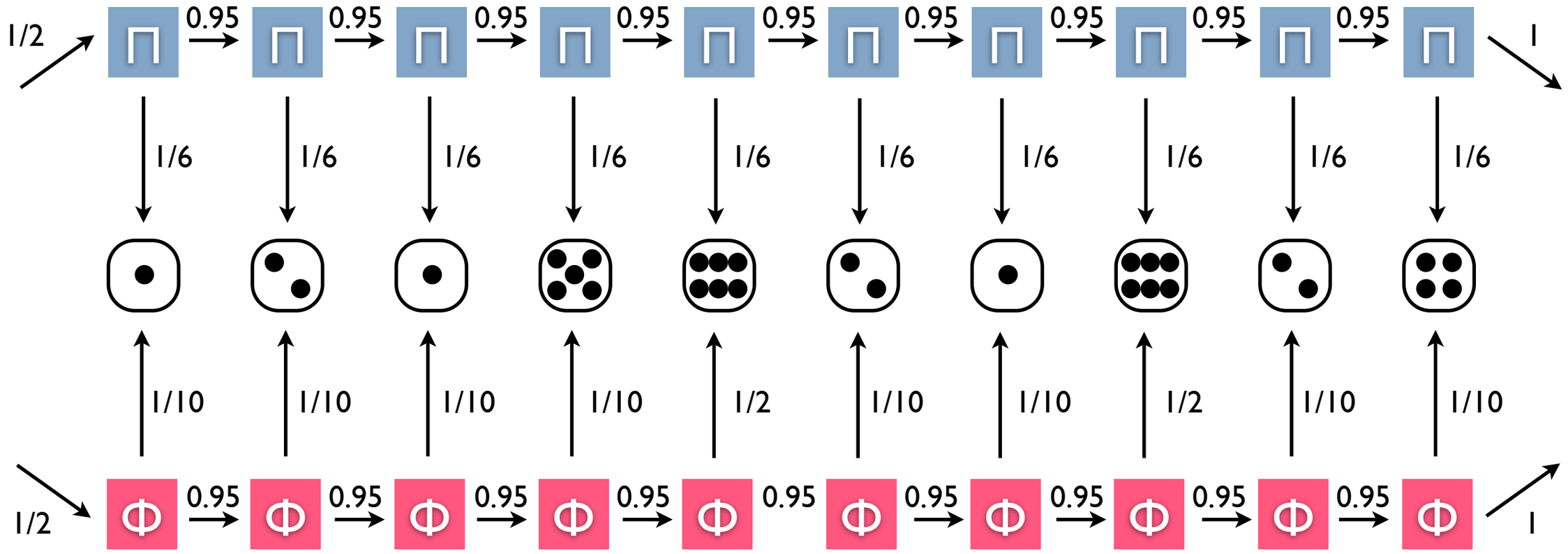
и последовательности символов

x = 1,2,1,5,6,2,1,6,2,4

$P(x,\pi) = 1/2 \times P(1|\text{Фальшивая}) \times P(\text{Фальшивая}_2|\ \text{Фальшивая}_1) \times P(2|\text{Фальшивая}) \times P(\text{Фальшивая}_3|\text{Фальшивая}_2) \times \ldots \times P(4|\text{Фальшивая}) =$

$= 1/2 \times (1/10)^8 \times (1/2)^2 \times (0.95)^9 = 7.9 \times 10^{-10}$

Сравнение двух моделей



Две последовательности состояний:

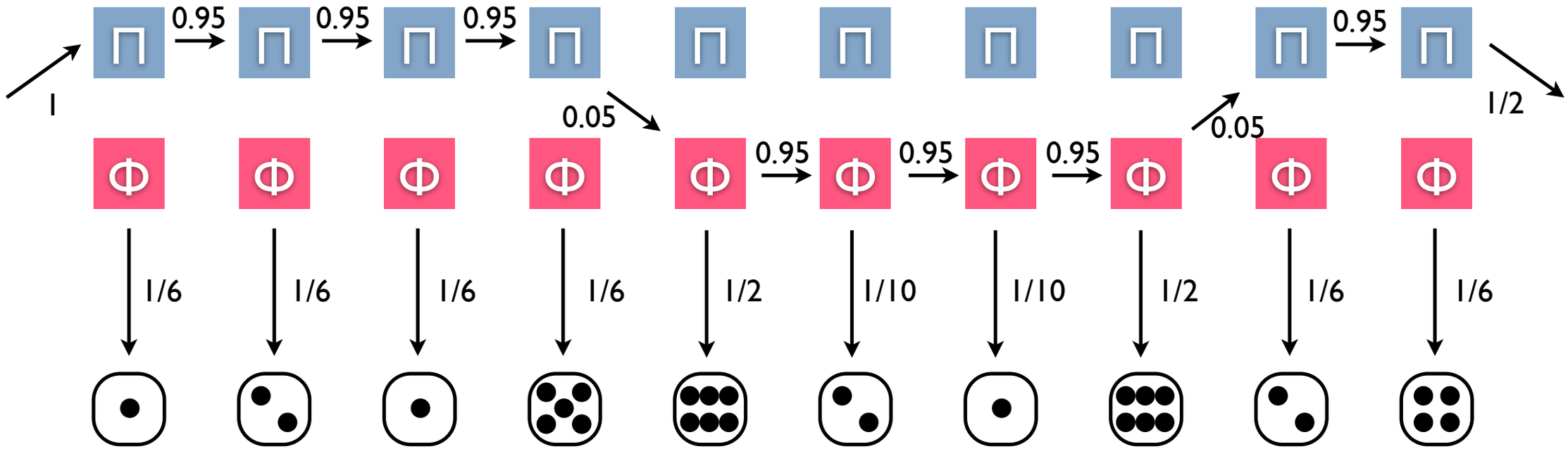
P(x, все-П) = 0.5×10^{-9}

P(x, все-Ф) = 7.9×10^{-10}

Отношение правдоподобия:

P(x, все-П) в 6.59 раз вероятнее, чем P(x, все-Ф)

Вероятность последовательностей наблюдений и состояний для случая подмены кости



Какова совместная вероятность последовательности состояний

π = Правильная, Правильная, Правильная, Правильная, Фальшивая, Фальшивая, Фальшивая, Фальшивая, Правильная, Правильная

и последовательности символов

x = 1,2,1,5,6,2,1,6,2,4

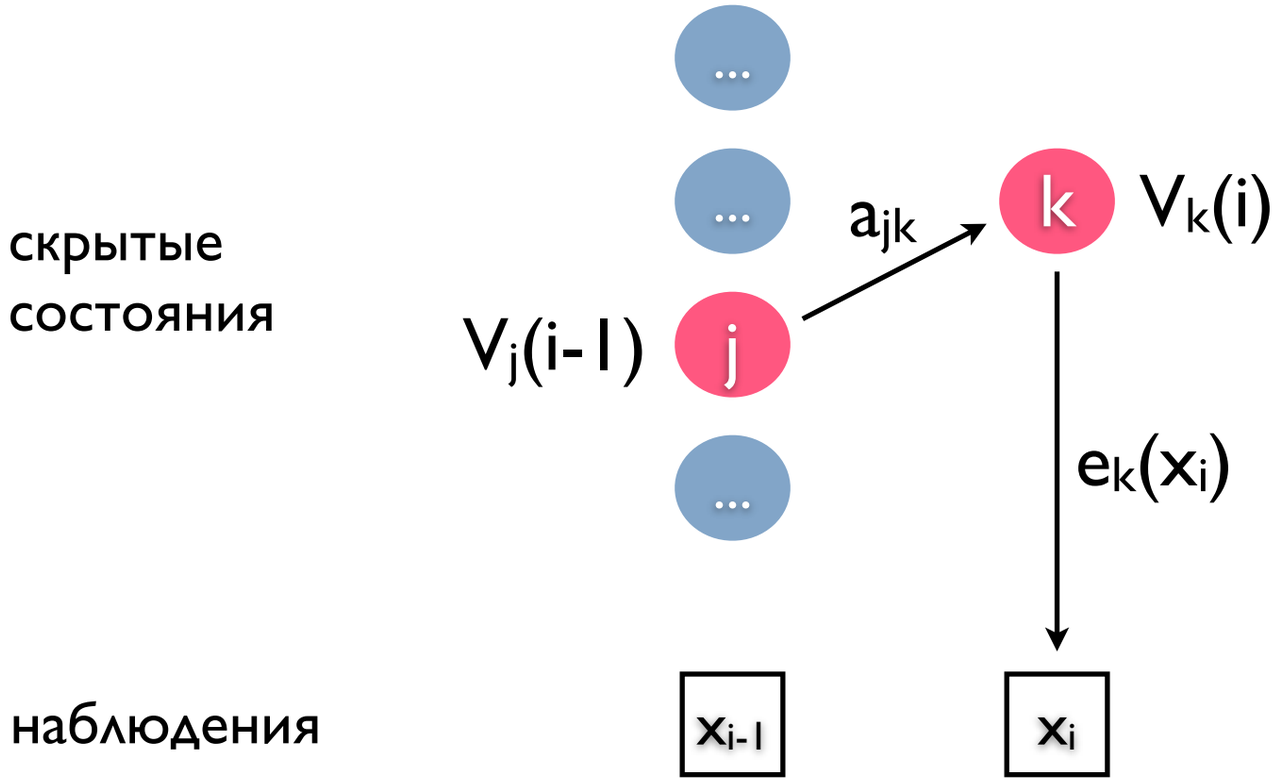
$P(x,\pi) = 1/2 \times P(1|\text{Правильная}) \times P(\text{Правильная}_2|\text{Правильная}_1) \times P(2|\text{Правильная}) \times P(\text{Правильная}_3|\text{Правильная}_2) \times \ldots \times P(4|\text{Правильная}) =$

$= 1/2 \times (1/2)^2 \times (1/10)^2 \times (1/6)^6 \times (0.95)^7 \times (0.05)^2 = 1.87 \times 10^{-9}$

Поиск оптимального пути

- Имея последовательность символов (наблюдений), мы умеем вычислять вероятность любого пути - последовательности скрытых состояний
- Как найти оптимальный - наиболее вероятный путь?
- Оптимальный путь можно определить рекурсивно (алгоритм динамического программирования Витерби):
  - пусть $V_k(i-1)$ - вероятность оптимального пути, проходящего через состояние k в момент времени i-1
  - мы можем вычислить вероятности для состояний в момент времени i, как функцию $max_k\{V_k(i)\}$

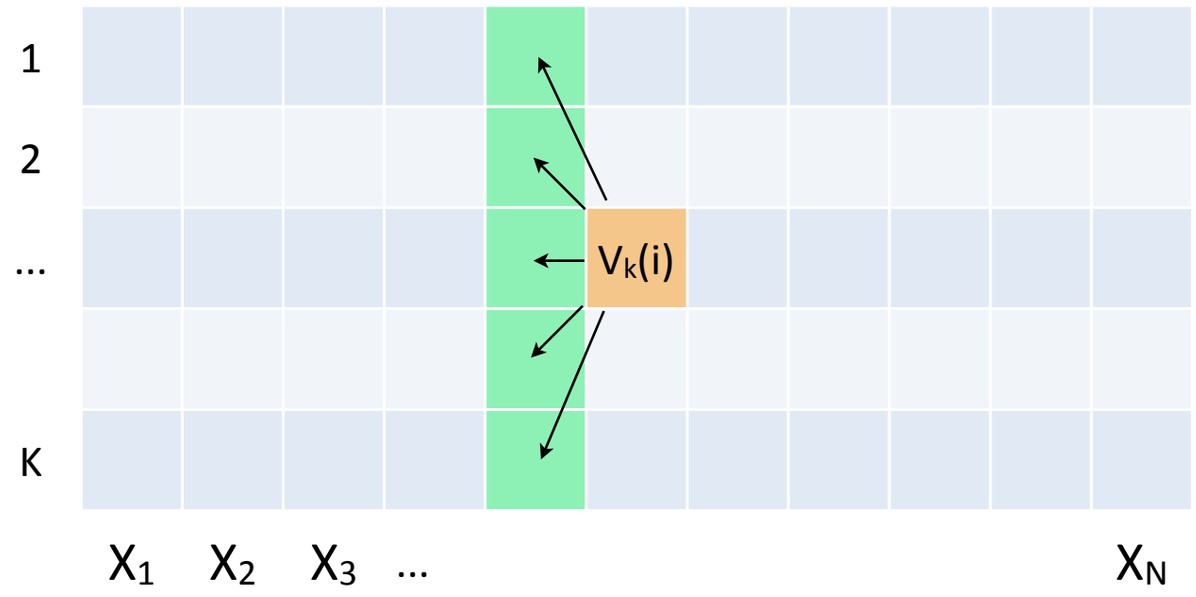
Рекурсивное вычисление оптимального пути



- Пусть мы знаем V_j для всех состояний момента времени i-1
- Тогда,

$$V_k(i) = e_k(x_i) \times \max_j (V_j(i-1) \times a_{jk})$$

Алгоритм Витерби



Входные аргументы: $x = x_1 ... x_N$

Инициализация:

$V_0(0) = 1$, $V_k(0) = 0$, для всех $k>0$

Рекурсия:

$V_k(i) = e_k(x_i) \times \max_j (a_{jk} \cdot V_k(i-1))$

Завершение:

$P(x,\pi^*) = \max_j V_j(N)$

Процедура обратного прохода:

Оптимальный путь находим проходя по ссылкам в обратном направлении (аналогично алгоримам выравнивания)

На практике:

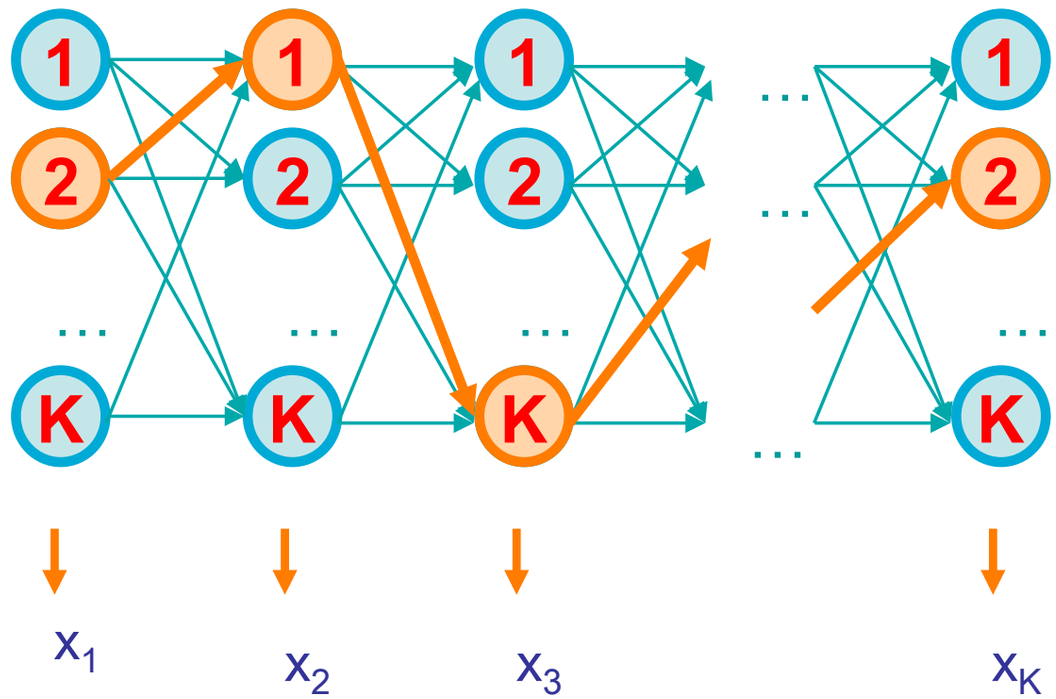
Для избежания потерь значимости при перемножении маленьких чисел выполняется в логорифмическом пространстве

Вычислительная сложность:

Время: $O(K^2N)$

Пространство: $O(KN)$

Алгоритм просмотра вперед



Задача:

Дана последовательность наблюдений x. Определить вероятность, того что данная последовательность сгенерирована заданной HMM.

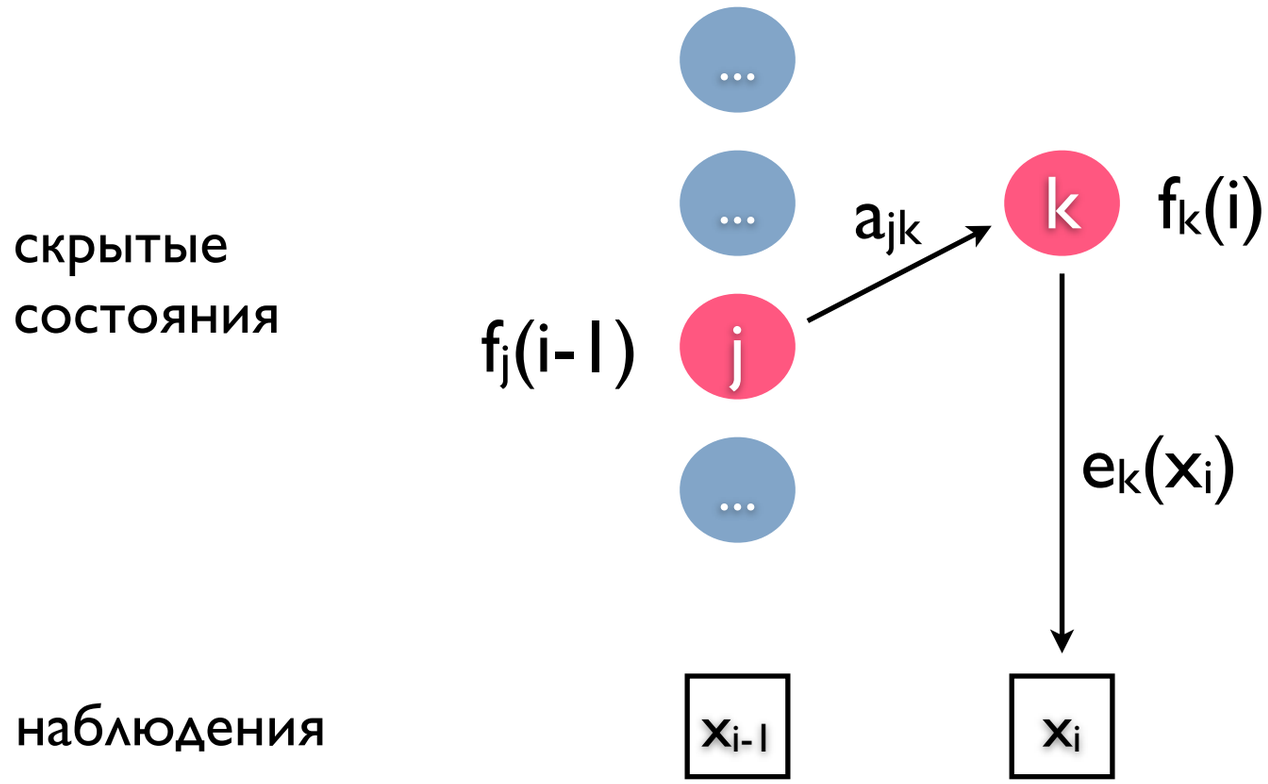
Приблеженное решение:

Вычислить вероятность последовательности наблюдений P(x) для наиболее вероятного пути π*.

Точное решение:

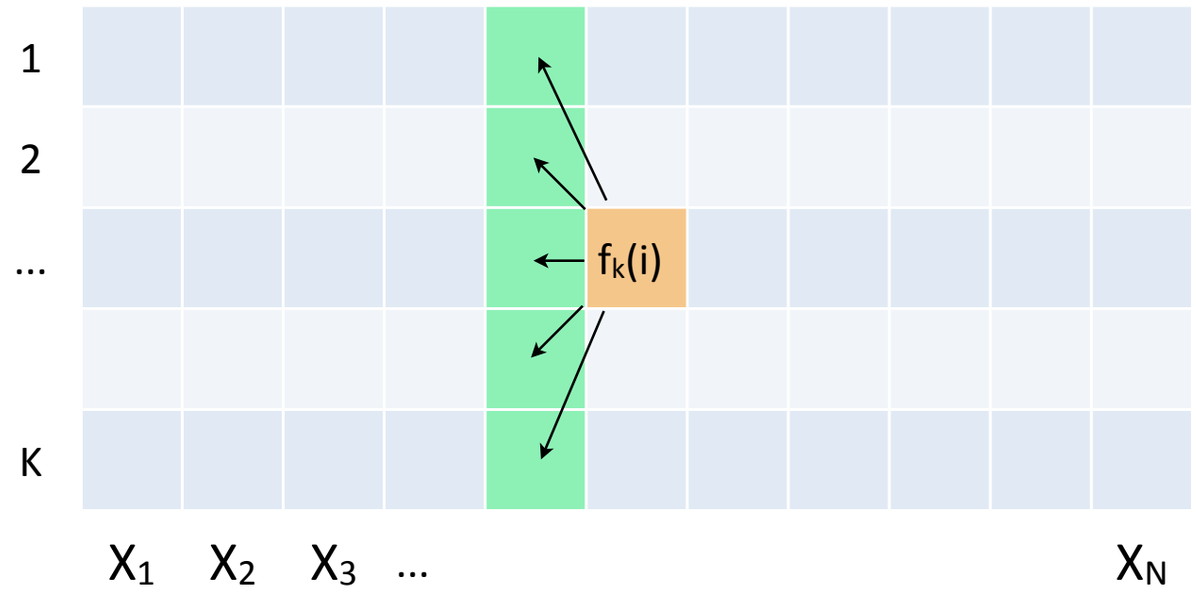
$$P(x) = \sum_{\pi} P(x,\pi)$$

Алгоритм просмотра вперед



$$f_k(i) = e_k(x_i) \times \mathrm{sum}_j \left(f_j(i-1) \times a_{jk} \right)$$

Алгоритм просмотра вперед



Входные аргументы: $x = x_1 ... x_N$

Инициализация:

$f_0(0) = 1,\ f_k(0) = 0$, для всех $k>0$

Рекурсия:

$f_k(i) = e_k(x_i) \times sum_j (a_{jk} \cdot f_k(i-1))$

Завершение:

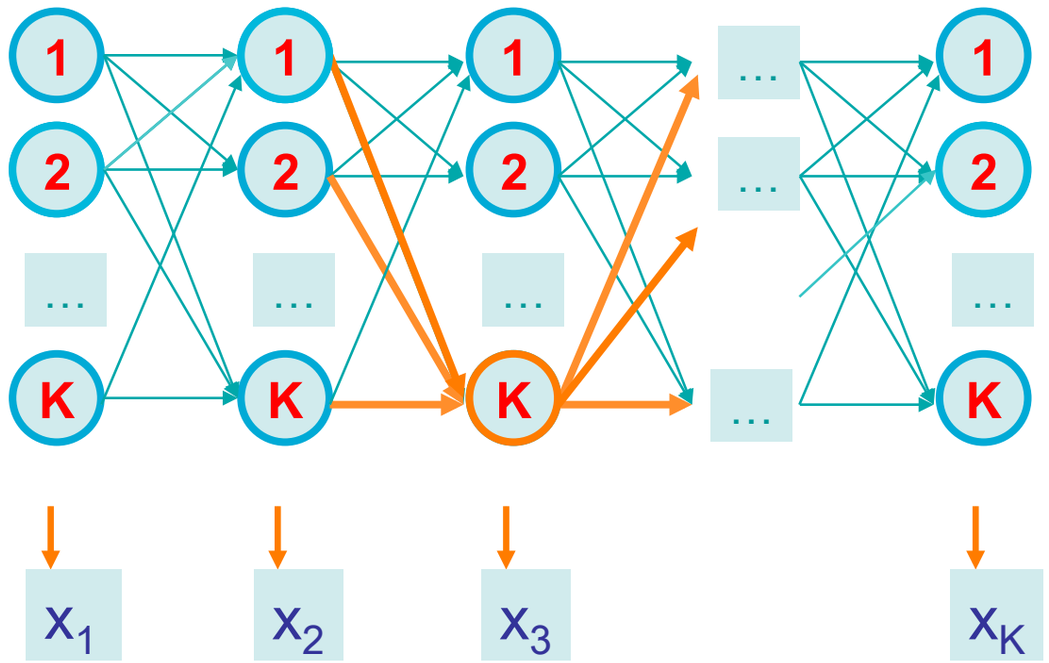
$P(x,\pi^*) = sum_j f_j(N)$

Вычислительная сложность:

Время: $O(K^2N)$

Пространство: $O(KN)$

Вероятность выбранного состояния



- Каково наиболее вероятное состояния в момент времени i для заданной последовательности x?

- Еще один способ определения оптимального пути:

$$\hat{\pi}_i = \arg\max_k P(\pi_i = k \mid x)$$

Алгоритм просмотра назад

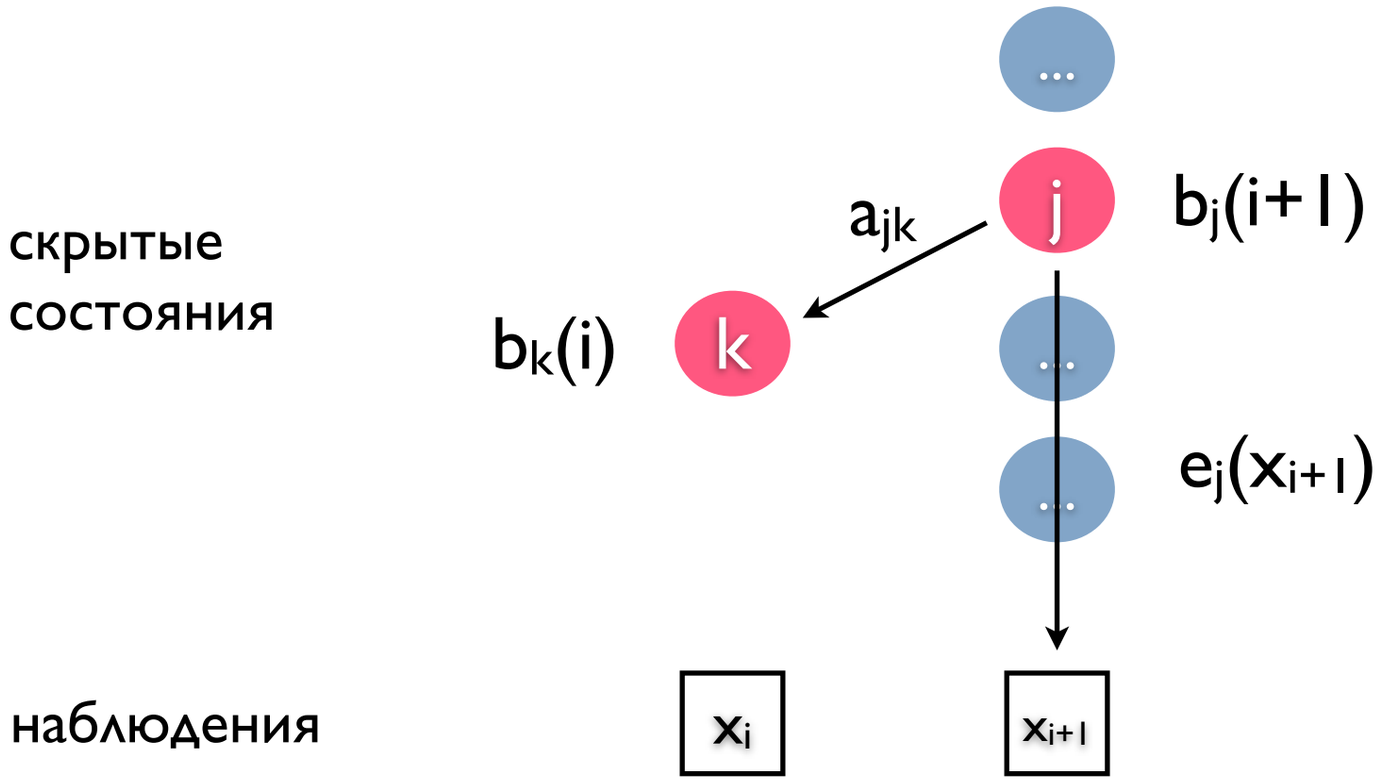
Наша цель определить $P(\pi_i=k \mid x)$ - вероятность состояния k при i-м наблюдении и заданной последовательности x.

$$P(\pi_i = k \mid x) = \frac{P(\pi_i = k, x)}{P(x)}$$

$$P(\pi_i = k, x) = P(x_1 \ldots x_i, \pi_i = k, x_{i+1} \ldots x_N) =$$
$$= P(x_1 \ldots x_i, \pi_i = k) P(x_{i+1} \ldots x_N \mid x_1 \ldots x_i, \pi_i = k) =$$
$$= \underbrace{P(x_1 \ldots x_i, \pi_i = k)}_{\text{просмотр вперед } f_k(i)} \underbrace{P(x_{i+1} \ldots x_N \mid \pi_i = k)}_{\text{просмотр назад } b_k(i)}$$

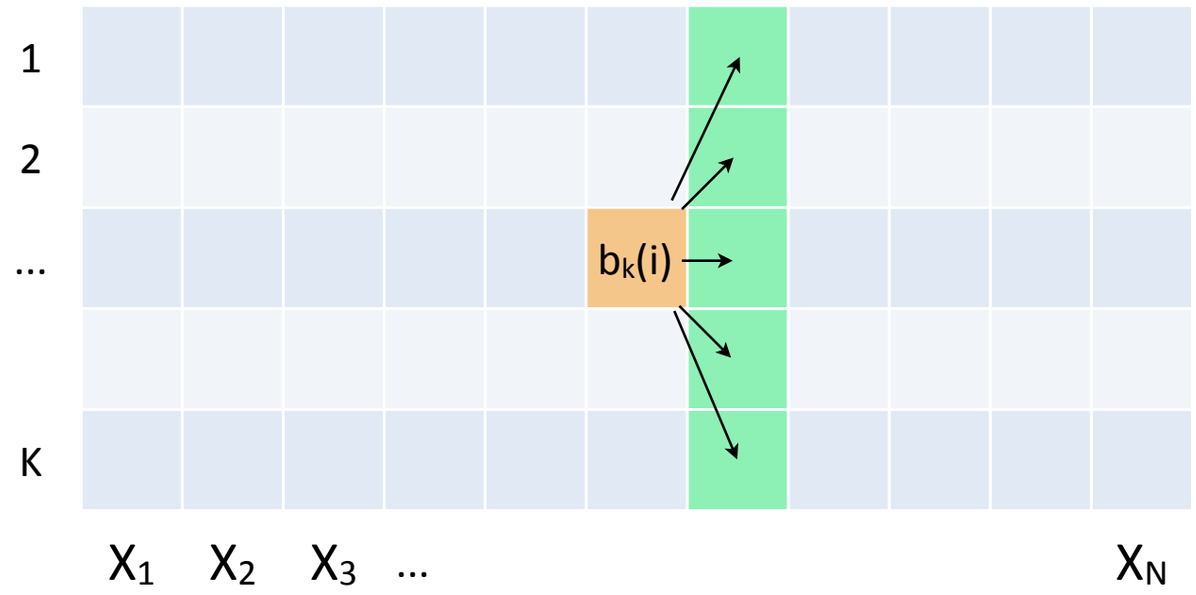
$$P(\pi_i = k \mid x) = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$

Алгоритм просмотра назад



$$b_k(i) = \text{sum}_j \left(e_j(x_{i+1}) \times b_j(i+1) \times a_{jk} \right)$$

Алгоритм просмотра назад



Входные аргументы: $x = x_1 ... x_N$

Инициализация:

$b_K(N) = a_{k0}$, для всех k

Рекурсия:

$b_k(i) = sum_j\ (e_j(x_{i+1}) \cdot a_{jk} \cdot b_j(i+1))$

Завершение:

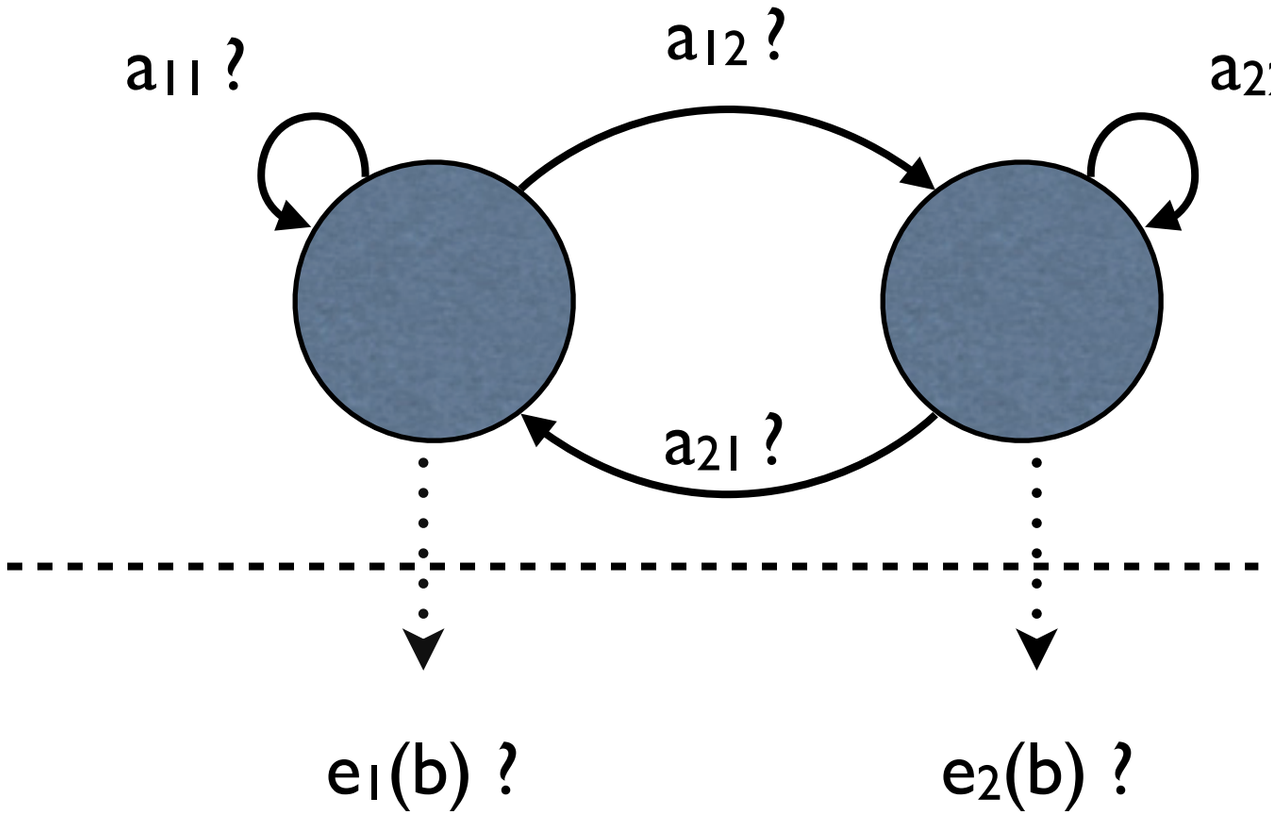
$P(x) = sum_j\ (e_j(x_1) \cdot a_{0j} \cdot b_j(1))$

Вычислительная сложность:

Время: $O(K^2N)$

Пространство: $O(KN)$

Оценка параметров HMM



- Рассмотрим два случая имеющихся данных:

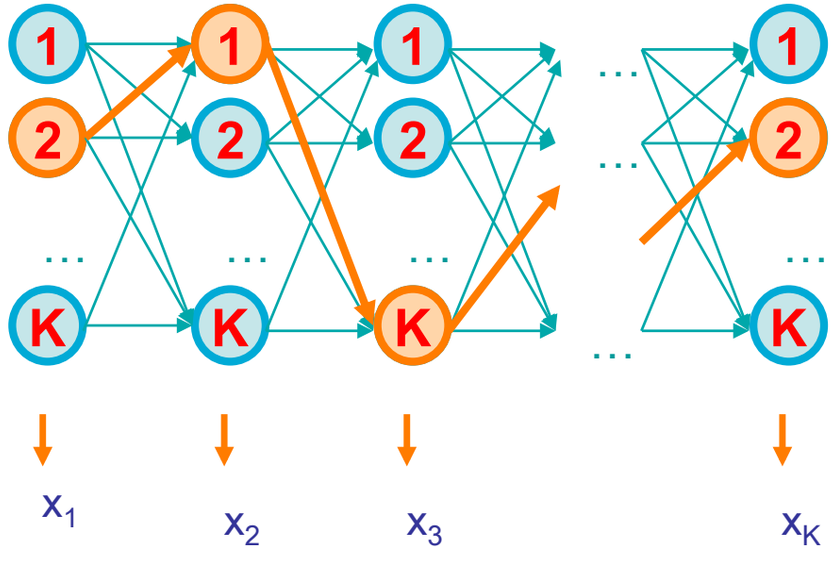
  1. Имеются множества наблюдений и известна последовательность смены состояний

  2. Имеются множества наблюдений. Последовательность смены состояний неизвестна.

Случай 1: путь известен

<u>Известны</u>:

- последовательность наблюдений $x = x_1 ... x_N$
- последовательность состояний (путь) $\pi = \pi_1 ... \pi_N$



<u>Определим</u>:

- A_{kl} - количество переходов из состояния k в состояние l вдоль пути π
- $E_k(b)$ - количество наблюдаемых символов b, сгенерированных в состоянии k

<u>Оценки параметров модели:</u>

Оценки параметров модели θ, вычисленные методом максимального правдаподобия

$$a_{kl} = \frac{A_{kl}}{\sum_i A_{kl}} \qquad e_k(b) = \frac{E_k(b)}{\sum_c E_k(c)}$$

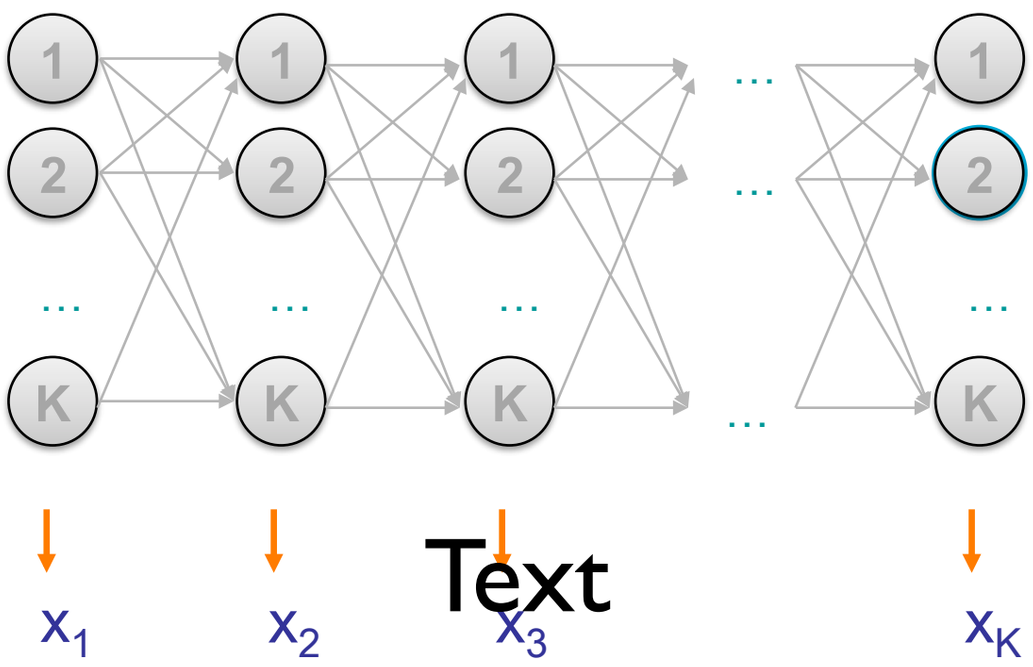
Псевдокаунты

При малом объеме обучающей выборке возникает проблема переобучения - оценки a_{kl} и $e_k(b)$ в некоторых случаях могут быть равны нулю

Определим:

- A_{kl} - количество переходов из k в l + r_{kl}
- $E_k(b)$ - количество генераций b в состоянии k + $r_k(b)$

Случай 2: путь неизвестен



Два метода:

- экономный - обучение Витерби (используется оптимальный путь)
- правильный - метод Баума-Уэлча, максимизация ожидания

Обучение Витерби

<u>Инициализация:</u>
Выбираем параметры модели случайным образом или на основе априорных знаний

<u>Итерации (выполняем до сходимости):</u>

1. Находим оптимальный путь π^* алгоритмом Витерби
2. Вычисляем A_{kl} и $E_k(b)$ вдоль пути π^*, добавляем псевдоотчеты
3. Вычисляем новые параметры модели a_{kl} и $e_k(b)$

Примечания:

- процедура не максимизирует $P(x^1,x^2,...,x^N|\theta)$, а максимизирует $P(x^1,x^2,...,x^N|\theta, \pi^*(x^1), \pi^*(x^2),..., \pi^*(x^{2N}))$
- в целом метод работает хуже, чем максимизация ожидания (метод Баума-Уэлча)

Максимизация ожидания (Expectation Maximization)

- Случайный выбор параметров модели
- Применение модели для оценки отсутствующих данных (E-шаг)
- Использование полученных данных для обновления параметров модели (M-шаг)

Метод Баума-Уэлча

Считая известными параметры модели (на первом шаге выбираются случайно, на последующих - оцениваются), вычислим вероятность $P_{kl}(x|\theta)$ перехода их k-го состояния шага i в l-е состояние шага i+1

$$P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{P(\pi_i = k, \pi_{i+1} = l, x_1 \ldots x_N)}{P(x \mid \theta)}$$

$$P(\pi_i = k, \pi_{i+1} = l, x_1 \ldots x_N) = P(x_1, \ldots, x_i, \pi_i = k, \pi_{i+1} = l, x_{i+1}, \ldots, x_N) =$$

$$= P(\pi_{i+1} = l, x_{i+1} \ldots x_N \mid \pi_i = k) P(x_1 \ldots x_i, \pi_i = k) =$$

$$= P(\pi_{i+1} = l, x_{i+1} \ldots x_N \mid \pi_i = k) f_k(i) =$$

$$= P(x_{i+2} \ldots x_N \mid \pi_{i+1} = l) P(x_{i+1} \mid \pi_{i+1} = l) P(\pi_{i+1} = l \mid \pi_i = k) f_k(i) =$$

$$= b_i(i+1) e_i(x_{i+1}) a_{kl} f_k(i)$$

$$P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{f_k(i) a_{kl} e_i(x_{i+1}) b_i(i+1)}{P(x \mid \theta)}$$

Метод Баума-Уэлча: оценка параметров модели

Суммируем вероятности переходов из состояния *k* в состояние *l* по всем моментам времени *i* и по всем обучающим последовательностям *x*

$$A_{kl} = \sum_x \sum_i P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \sum_x \sum_i \frac{f_k(i) a_{kl} e_i(x_{i+1}) b_i(i+1)}{P(x \mid \theta)}$$

Аналогично,

$$E_k(b) = \sum_x \sum_{\{i \mid x_i = b\}} \frac{f_k(i) b_k(i)}{P(x \mid \theta)}$$

Алгоритм Баума-Уэлча

<u>Инициализация:</u>
Выбираем параметры модели случайным образом или на основе априорных знаний

<u>Для каждой последовательности:</u>

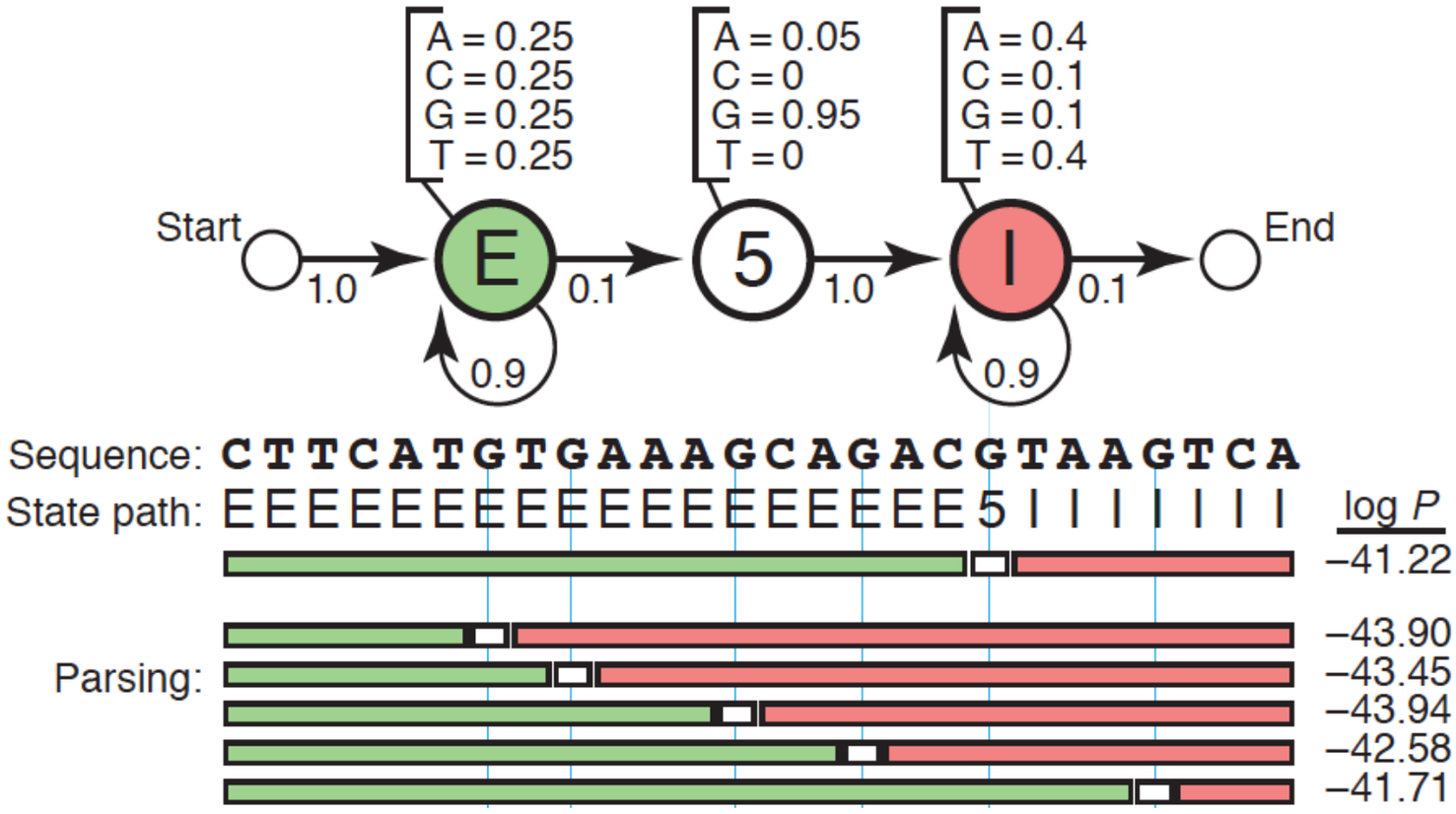
1. Вычисляем $f_k(i)$ алгоритмом просмотра вперед
2. Вычисляем $b_k(i)$ алгоритмом просмотра назад
3. Добавляем вклад последовательности в A_{kl} и $E_k(b)$

Вычисляем новые параметры модели a_{kl} и $e_k(b)$ и повторяем итерации
Вычислим значение правдоподобия модели

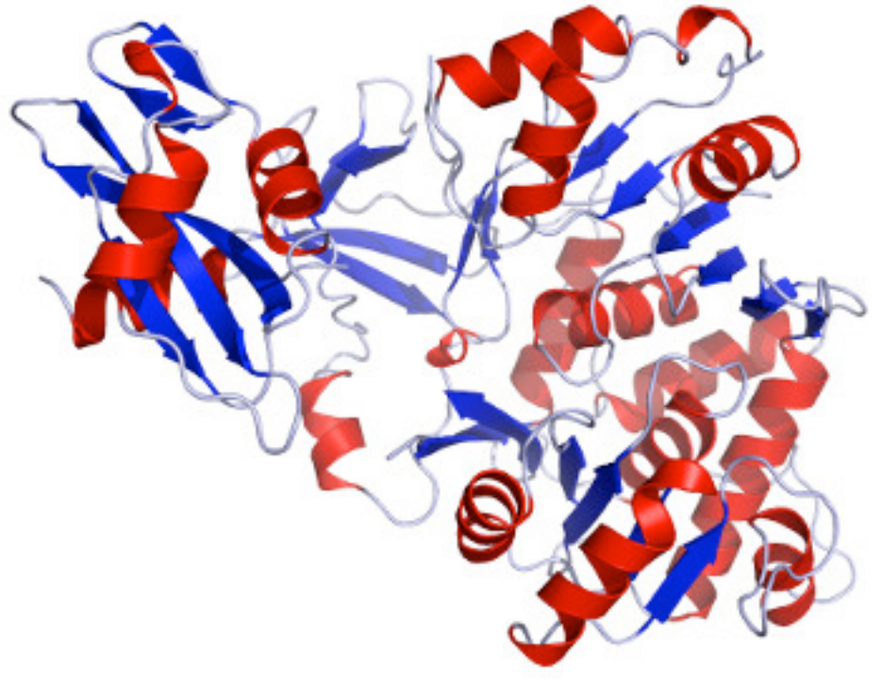
Останавливаемся, когда улучшение правдоподобия достигает максимума или ее изменение меньше порога

Применение HMM в биоинформатике

Пример:
распознавание 5’ сайта сплайсинга.



- распознавание генов
- предсказание типов вторичной структуры белков
- распознавание CpG островков
- участки белков - сигнальные пептиды, эпитопы и т.д.



Благодарности

- При подготовке слайдов использовались материалы лекций:
  - Михаила Гельфанда (ИППИ)
  - Андрея Миронова (МГУ)
  - Serafim Batzoglou (Stanford)
  - Manolis Kellis (MIT)
  - Pavel Pevzner (UCSD)