УДК 519.872 DOI:10.31854/1813-324X-2021-7-4-10-17

Применение EM-алгоритма для аппроксимации гиперэкспонентами плотностей вероятностей коррелированного трафика

М.А. Буранова[1]*, **И.В. Карташевский**[1]

[1]Поволжский государственный университет телекоммуникаций и информатики,
Самара, 443010, Российская Федерация
*Адрес для переписки: buranova-ma@psuti.ru

Информация о статье
Поступила в редакцию 18.09.2021
Поступила после рецензирования 11.10.2021
Принята к публикации 18.10.2021

Ссылка для цитирования: Буранова М.А., Карташевский И.В. Применение EM-алгоритма для аппроксимации гиперэкспонентами плотностей вероятностей коррелированного трафика // Труды учебных заведений связи. 2021. Т. 7. № 4. С. 10–17. DOI:10.31854/1813-324X-2021-7-4-10-17

Аннотация: *Точная оценка параметров качества обслуживания в современных инфокоммуникационных сетях является весьма важной задачей. В работе предложено использование гиперэкспоненциальных распределений для решения задачи аппроксимации произвольной плотности вероятностей в системе G/G/1 для случая, когда предполагается аппроксимация системой типа H2/H2/1. Для определения параметров плотности вероятностей гиперэкспоненциального распределения предлагается использовать EM-алгоритм, который дает достаточно простые варианты использования для некоррелированных потоков. В данной работе предложен вариант реализации алгоритма EM для определения параметров гиперэкспоненциального распределения при наличии корреляционных свойств анализируемого потока.*

Ключевые слова: *EM-алгоритм, распределения случайных величин, корреляция, аппроксимация, QoS.*

Введение

Традиционно для оценки параметров качества обслуживания используют методы теории массового обслуживания [1], при этом в простейшем случае такие системы представляются математической моделью системы обработки потоков на основе $M/M/1$, где интервалы времени между пакетами и интервалы длительности пакетов представляют собой последовательности независимых случайных величин с показательными плотностями вероятностей. Учитывая, что трафик современных приложений, обрабатываемый в IP-сети, отличается от простейшего потока и обладает фрактальными свойствами [2, 3], системы обработки таких потоков следует описывать моделью $G/G/1$. Следует учесть, что для потоков, генерируемых современными приложениями, часто характерно наличие корреляций внутри структур трафика [2, 3].

Использование известных математических подходов для оценки параметров качества обработки трафика современных приложений может привести к ошибочным результатам, потому что большинство подходов по оценке QoS основано на предположении отсутствия корреляций внутри структур потоков [1, 4–6]. При этом есть многочисленные доказательства того, что для современного трафика характерно наличие корреляционных свойств [7–10]. Это связано с характером формирования потоков, поведением пользователей, агрегированием потоков, сетевым управлением и т. д.

В то же время, работ по оценке параметров функционирования сетей с учетом обработки коррелированных потоков весьма мало, что осложняет возможность получения точных прогнозов функционирования инфокоммуникационных сетей в современных условиях. Существуют некоторые подходы к оценке параметров функционирования с учетом автокорреляции потоков, например, в работах [7–9, 11]. Но представленные подходы не всегда дают требуемую точность и связаны с необходимостью выполнения сложных вычислений.

Одним из возможных вариантов использования методов классической теории массового обслуживания при обработке коррелированного трафика является вариант аппроксимации системы $G/G/1$ системой $H_2/H_2/1$ [12], имеющей гиперэкспоненциально распределенное время поступления и об-

служивания. В [13, 14] было установлено, что применение смеси экспоненциальных распределений являются теоретически обоснованным подходом к моделированию IP-трафика. Получение аппроксимации такого типа возможно при использовании подходов кластерного анализа, например, алгоритма EM (*аббр. от англ.* Expectation-Maximization), который позволяет определить параметры распределений и число экспонент аппроксимации реальной системы.

Кроме алгоритма EM, существуют и другие подходы по определению параметров экспоненциальных распределений, такие как определение параметров по двум моментам (среднее, дисперсия) исходного распределения для независимых случайных величин [15]. При использовании данного подхода возможно получение аналитических выражений начальных моментов гиперэкспоненциальных распределений до второго порядка.

EM-алгоритм часто применяют в разнообразных задачах оптимизации, он является удобным инструментом реализации итерационной процедуры поиска экстремума целевой функции с использованием численных методов. Такие подходы востребованы при поиске оценок максимального правдоподобия параметров вероятностных моделей в случае, если есть предположение о присутствии в модели скрытых переменных.

EM-алгоритм является весьма эффективным для нахождения аппроксимации наблюдаемых реализаций как одномерных, так и многомерных распределений, но применение EM-алгоритма, как правило, используется для разделения смеси нормальных распределений. Однако в теории очередей чаще всего используется показательное распределение, так как параметры, описывающие поведение трафика в современной IP-сети, являются непрерывными случайными величинами, принимающими неотрицательные значения.

Ранее в работах [6, 15] были представлены подходы по определению параметров гиперэкспоненциальных распределений с использованием EM-алгоритма для систем с некоррелированными реализациями мгновенных значений случайных величин. Применение алгоритма EM в случае наличия корреляционных свойств потока случайных величин вызывает безусловный интерес.

Реализация EM–алгоритма для некоррелированных величин

Основными составляющими при оценке параметров QoS с использованием теории очередей являются следующие случайные величины: интервалы времени между пакетами и длины пакетов обрабатываемого в системе потока. Реализации потоков данных случайных интервалов времени обычно описываются различными распределениями для положительных случайных величин.

Смеси распределений являются весьма эффективным механизмом для описания плотностей распределения вероятностей любой реализации случайной величины в случае, если описание конкретным законом распределения с достаточной точностью вызывает затруднения.

Плотность вероятностей показательного распределения имеет вид:

$$f(x) = \lambda e^{-\lambda x},$$

где λ – параметр распределения, имеющий смысл интенсивности потока.

Если произвольную плотность вероятностей в системе $G/G/1$ на любой позиции G (в данном рассмотрении) обозначить как $w(x)$, то задача аппроксимации $w(x)$ с использованием гиперэкспоненциальных распределений может быть представлена в виде:

$$w(x) \approx H_K(x) = \sum_{i=1}^{K} p_i\, f_i(x), \qquad (1)$$

где $f_i(x)$ – функция правдоподобия i-ой компоненты смеси, $f_i(x) = \lambda_i\, e^{-\lambda_i x}$, $p_i \geq 0$ – априорная вероятность:

$$\sum_{i=1}^{K} p_i = 1.$$

Для определения параметров гиперэкспоненциальных распределений можно воспользоваться подходом, основанным на использовании EM-алгоритма [15–17], который может реализовывать получение достоверных оценок максимального правдоподобия для многих приложений, включая оценку параметров плотности смеси.

Будем предполагать, что в выражении (1) вид функций $f_i(x)$ известен, а параметры p_i и λ_i неизвестны, совокупность которых обозначим вектором $\theta = (p_1, \cdots p_K;\ \lambda_1, \cdots \lambda_K)$.

Теперь при разделении смеси (1) будем решать задачу статистического оценивания параметров θ по известной реализации случайной величины X – $(x_1, \cdots, x_N)$. Предполагая независимость значений x_i, в рамках модели (1) логарифм функции правдоподобия $L(\theta;\ x)$ параметра θ имеет вид [18]:

$$\log L(\theta;\ x) = \log \prod_{j=1}^{N} \phi_\theta^X(x_j) = \\ = \sum_{j=1}^{N} \log\left(\sum_{i=1}^{K} p_i f_i(x_j;\ \theta_i)\right). \qquad (2)$$

Далее, следуя [18], введем в рассмотрение ненаблюдаемую случайную величину Y, которая играет роль индикатора выбора компоненты плотности $f_{\theta,y}^X(x_j)$, где y принимает значения $y_j \in (1, 2, \ldots, K)$ ненаблюдаемой случайной величины Y. Другими

словами, предположим, что в процессе поступления x_j, $j = 1, \dots, N$, реализуется также ненаблюдаемое значение $y_j \in (1, 2, \dots, K)$.

Рассмотрим теперь совместную плотность случайных величин X и Y, обозначаемую как $\phi_\theta(x, y)$. С учетом того, что маргинальная плотность величины Y равна $\phi_\theta^Y(i) = p_i$, $i = 1, \dots, K$, для условной плотности случайной величины X при фиксированном $Y = i$ можно записать $\phi_\theta(x|i) = f_i(x;\ \theta_i)$.

Для полной функции правдоподобия, учитывающей присутствие в эксперименте случайной величины Y, следует:

$$\begin{aligned} \log L(\theta;\ x, y) &= \log \prod_{j=1}^{N} \phi_\theta(x_j, y_j) = \\ = \sum_{j=1}^{N} \log \phi_\theta(x_j, y_j) &= \sum_{j=1}^{N} \log\left[\phi_\theta(x_j|y_j)\phi_\theta^Y(y_j)\right] = \\ &= \sum_{j=1}^{N} \log\left[p_{y_j} f_{y_j}(x_{y_j};\ \theta_{y_j})\right] = \\ &= \sum_{j=1}^{N} \log p_{y_j} + \sum_{j=1}^{N} \log f_{y_j}(x_{y_j};\ \theta_{y_j}). \end{aligned} \quad (3)$$

С учетом вышеприведенных выражений для $\phi_\theta^Y(i)$ и $\phi_\theta(x|i) = f_i(x;\ \theta_i)$ совместная плотность $\phi_\theta(x, y)$ записывается как:

$$\phi_\theta(x, y) = p_y f_y(x;\ \theta_y) = \phi_\theta(y|x)\phi_\theta^X(x),$$

и на основе формулы Байеса для апостериорной вероятности того, что наблюдение x получено в соответствии с распределением, задаваемым компонентой f_i, можно получить:

$$\phi_\theta(y|x) = \frac{\phi_\theta(x, y)}{\phi_\theta^X(x)} = \frac{p_y f_y(x;\ \theta_y)}{\sum_{i=1}^{K} p_i f_i(x;\ \theta_i)}.$$

Далее, учитывая независимость реализаций X и Y, последняя формула записывается в виде:

$$\phi_\theta(y|x) = \prod_{j=1}^{N} \phi_\theta(y_j|x_j) = \prod_{j=1}^{N} \frac{p_{y_j} f_{y_j}(x_j;\ \theta_{y_j})}{\sum_{i=1}^{K} p_i f_i(x_j;\ \theta_i)}. \quad (4)$$

Как показано в [18], если определить условное среднее (по возможным итерациям) логарифма полной функции правдоподобия при известном значении наблюдаемой компоненты X, то максимизация этого условного среднего методом неопределенных множителей Лагранжа дает следующие формулы для формирования оценок максимального правдоподобия на каждой (m)-ой итерации алгоритма:

$$p_l^{(m+1)} = \frac{1}{N} \sum_{j=1}^{N} \frac{p_l^{(m)} f_l(x_j;\ \theta_l^{(m)})}{\sum_{i=1}^{K} p_i^{(m)} f_i(x_j;\ \theta_i^{(m)})}, \quad (5)$$

$$\lambda_l^{(m+1)} = \frac{\sum_{j=1}^{N} \frac{f_l^{(m)}(x_j) \cdot x_j}{\sum_{l=1}^{K} f_l^{(m)}(x_j)}}{\sum_{j=1}^{N} \frac{f_l^{(m)}(x_j)}{\sum_{l=1}^{K} f_l^{(m)}(x_j)}}, \quad l = 1, \dots, K. \quad (6)$$

В формуле (6) учитывается, что при гиперэкспоненциальной аппроксимации $f_l(x) = \lambda_l\, e^{-\lambda_l x}$ и для модели $H_2(x)$ оцениванию подлежат только три параметра: p, λ_1, λ_2, т. к. модель $H_2(x)$ может быть представлена в виде:

$$g(x; p,\ \lambda_1,\ \lambda_2) = p\lambda_1 e^{-\lambda_1 x} + (1 - p)\lambda_2 e^{-\lambda_2 x}.$$

С введением обозначения $g(x;\ p,\ \lambda_1,\ \lambda_2)$ формулы (5) и (6) могут быть записаны проще:

$$p^{(m+1)} = \frac{1}{N} \sum_{j=1}^{N} \frac{p^{(m)} f_1(x_j;\ \lambda_1^{(m)})}{g^{(m)}(x_j)}, \quad (7)$$

$$\lambda_l^{(m+1)} = \frac{\sum_{j=1}^{N} \frac{f_l^{(m)}(x_j) \cdot x_j}{g^{(m)}(x_j)}}{\sum_{j=1}^{N} \frac{f_l^{(m)}(x_j)}{g^{(m)}(x_j)}}, \quad l = 1, 2. \quad (8)$$

Рассмотренный подход применим для последовательностей с независимыми (в крайнем случае некоррелированными) случайными величинами. В случае наличия корреляции внутри последовательностей необходимо применять другие подходы.

Реализация EM–алгоритма для коррелированных величин

Известно, что при вычислении оценок максимального правдоподобия для ненаблюдаемых параметров априорная информация играет очень важную роль. Для последовательности $(x_1, x_2, \dots, x_N)$, которая обладает корреляционными связями соседних элементов, коэффициент корреляции $R = R(x_{k-1}, x_k)$ является необходимой информацией для построения оценок плотностей вероятностей смеси.

Поэтому на этапе определения априорных сведений о последовательности $(x_1, x_2, \dots, x_N)$ целесообразно получить оценку одномерной плотности вероятностей $\hat{w}(x)$ и ее первых двух моментов m_x, σ_x^2, а также коэффициента корреляции соседних элементов R. Кроме того, для инициализации вычислений при нулевой итерации $(m = 0)$ следует в последовательности $(x_1, x_2, \dots, x_N)$ задать элемент x_0 (возможно $x_0 = 0$), т. е. для $g(x; p,\ \lambda_1,\ \lambda_2)$ задать значения λ_1, λ_2 и p. Значение отсчета x_0, в принципе, может выбираться произвольно, главное, чтобы это значение было сопоставимо с m_x.

Если $(x_1, x_2, \dots, x_N)$ – заданная последовательность временных интервалов, то, очевидно, что $\lambda_x = m_x^{-1}$ имеет смысл интенсивности наступления некоторых событий. Тогда при нулевой итерации $(m = 0)$ при обработке отсчета x_0 можно ограничиться заданием одного параметра интенсивности

(например, λ_1), т. к. λ_1 и λ_2 связаны очевидным соотношением:

$$\lambda_x = \frac{\lambda_1\lambda_2}{p\lambda_2 + (1-p)\lambda_1}, \quad (9)$$

из которого при заданном λ_1 и априорно известном λ_x можно легко определить λ_1.

Выбор значения для λ_1 достаточно произволен, но всегда надо следить за тем, чтобы полученное из (9) значение λ_1 было положительным. В противном случае необходимо изменить выбор для λ_1. Выбор значения p из интервала $(0,1)$ также достаточно произволен, но обычно выбирается $p = 0{,}5$.

Корреляционные свойства последовательности $(x_1, x_2, \dots, x_N)$ благодаря рекуррентности формул (7) и (8) могут быть явно учтены в этих формулах, если установить аналитическую связь соседних элементов последовательности.

Это можно сделать при использовании уравнения регрессии, записанного в виде [11]:

$$x_k = Rx_{k-1} + m_x(1-R) + \left(\sigma_x\sqrt{1-R^2}\right)S, \quad (10)$$

где S – случайная нормально распределенная величина с нулевым средним и единичной дисперсией, не коррелированная с x_k.

Учтем теперь выражение (10) в рекуррентном соотношении (8). Выпишем несколько первых слагаемых из числителя выражения при $l = 1$, включив в него слагаемое для x_0, служащее для инициализации процесса вычислений:

$$\frac{f_1(x_0)\cdot x_0}{g(x_0)} + \frac{f_1(x_1)\cdot x_1}{g(x_1)} + \frac{f_1(x_2)\cdot x_2}{g(x_2)} + \dots. \quad (11)$$

Рассмотрим формирование значения $f_1(x_i)x_i$ в формуле (8) с учетом соотношения (10). Например, при обработке x_1:

$$f_1(x_1)\cdot x_1 = f_1(x_1)\left[\frac{1}{R}\left(x_2 - m_x(1-R) - S^{(2)}\right)\right] = \\ = \frac{1}{R}f_1(x_1)\left[x_2 - G^{(2)}\right]. \quad (12)$$

где $G^{(2)} = m_x(1-R) + \left(\sigma_x\sqrt{1-R^2}\right)S^{(2)}$ – случайная величина, которая должна генерироваться в данном случае на втором шаге вычислений, когда проводятся вычисления, связанные с элементом последовательности x_2.

Очевидно, для произвольного x_i:

$$f_1(x_i)x_i = \frac{1}{R}f_1(x_i)\left[x_{i+1} - G^{(i+1)}\right]. \quad (13)$$

Стоящее в правой части выражения (13) значение $f_1(x_i)$ вычисляется по информации о λ_1, полученной на предыдущей итерации вычислений:

$$f_1^{(m)}(x_i) = \lambda_1^{(m-1)}e^{-\lambda_1^{(m-1)x_i}}. \quad (14)$$

Соответственно:

$$g^{(m)}(x_i) = p^{(m-1)}\lambda_1^{(m-1)}e^{-\lambda_1^{(m-1)x_i}} + \\ + (1-p^{(m-1)})\lambda_2^{(m-1)}e^{-\lambda_2^{(m-1)x_i}}.$$

Теперь можно определить порядок действий при реализации соотношений (7) и (8).

1) Методами статистического анализа, реализованного соответствующими программными продуктами, определяется плотность вероятностей $\widehat{w}(x)$, аппроксимирующая истинную плотность анализируемой выборки. Определяются также моменты m_x, σ_x^2 и коэффициент корреляции соседних элементов R. При этом предполагается, что следующим шагом аппроксимации $\widehat{w}(x)$, является выбор функции:

$$g(x) = pf_1(x) + (1-p)f_2(x).$$

2) Осуществляется выбор исходных данных p, $\lambda_1(0), \lambda_2(0)$ для x_0.

3) Текущие вычисления производятся по формулам (7) и (8) с использованием (10) и (12). Особенностью вычислений является то, что на каждом шаге необходимо генерировать нормальную случайную величину со средним значением $m_x(1-R)$ и дисперсией $\sigma_x^2(1-R^2)$.

4) Выбор момента останова вычислений можно выбирать следующим образом.

Если ошибку аппроксимации на каждом (ν)-шаге представить в виде:

$$\varepsilon_{(\nu)} = \frac{\sum_{i=1}^{(\nu)}\log\left[g^{(\nu)}(x_i)\right] - \sum_{i=1}^{(\nu)}\log\left[g^{(\nu-1)}(x_i)\right]}{\sum_{i=1}^{(\nu)}\log\left[g^{(\nu-1)}(x_i)\right]},$$

где при фиксированном N должно выполняться $(\nu)_{\max} < N$, то при выполнении условия $\varepsilon_{(\nu)} \le \varepsilon_{\text{пороговое}}$, можно считать алгоритм сходящимся и фиксировать полученные значения $p^{(\nu)}, \lambda_1^{(\nu)}, \lambda_2^{(\nu)}$. В противном случае следует искать другую «траекторию» вычислений, задав другое начальное значение x_0 [18]. При данном вычислении и значение $\varepsilon_{\text{пороговое}}$, изменив которое, можно получить сходящийся алгоритм, оказывает влияние на сходимость алгоритма.

Анализ ошибки аппроксимации при выполнении условия $\varepsilon_{(\nu)} \le \varepsilon_{\text{пороговое}}$ можно осуществить, используя для оценивания ошибки аппроксимации меру Кульбака – Лейблера [19]:

$$\Delta = \sum_{i=1}^{N} w(x_i)\ln\frac{w(x_i)}{g(x_i)}.$$

Анализ точности аппроксимации через параметр Δ целесообразно проводить для выбора среди сходящихся «траекторий» наилучшей, т. е. обладающей наименьшим значением Δ.

Моделирование работы ЕМ–алгоритма при коррелированной реализации случайных величин

Проверка работоспособности рассмотренного алгоритма расчета параметров аппроксимирующей плотности (1) при коррелированной выборке была проведена на основе эксперимента, реализованного методом статистического моделирования. Для простоты моделирования была выбрана последовательность случайных событий с интервалами времени между событиями, имеющими логнормальное распределение с заданным коэффициентом корреляции соседних интервалов.

Плотность вероятностей логнормального распределения имеет вид:

$$w(x) = \frac{1}{\sigma_0 x\sqrt{2\pi}} \exp\left(-\frac{\ln^2 x}{2\sigma_0^2}\right),$$

$$m_x = \sigma_0\sqrt{e}, \quad \sigma_x^2 = e(e-1)\sigma_0^2.$$

Моделирование *стационарного* логарифмически-нормального случайного процесса с заданным коэффициентом корреляции R можно осуществить следующим образом [20]:

– моделируется нормальный стационарный случайный процесс $\{y_i\}$, $i = 1, 2, \ldots$ с коэффициентом корреляции:

$$\rho = \ln(1 + (e-1)R),$$

– каждый элемент последовательности y_i подвергается преобразованию $x = e^y$ для получения последовательности $\{x_i\}$, $i = 1, 2, \ldots$.

При этом нормальный процесс с заданным коэффициентом корреляции моделируется разностным уравнением первого порядка:

$$y(n) = a_0\xi(n) + b_1 y(n-1),$$

где $a_0 = \sigma_0\sqrt{1-\rho^2}$; $b_1 = \rho$; $\xi(n)$ – последовательность нормальных случайных чисел с нулевым средним и единичной дисперсией – $N(0,1)$.

Итак, если положить, что $R = 0{,}7$, то при $\sigma_0 = 1$ получается $a_0 = 0{,}613$, $b_1 = 0{,}79$.

Сгенерированная трасса из 5000 логнормальных интервалов с коэффициентом корреляции соседних интервалов $R = 0{,}7$, с начальными условиями, согласованными с (9), в виде:

$$\lambda_x = 0{,}606,\ p = 0{,}5,\ \lambda_1 = 0{,}5,\ \lambda_2 = 0{,}77$$

при обработке алгоритмом ЕМ по формулам (7) и (8) с $\varepsilon_{\text{пороговое}} = 0{,}01$ дала при $(\nu) = 499$ следующие результаты:

$$\hat{p} = 0{,}45, \quad \hat{\lambda}_1 = 2{,}72, \quad \hat{\lambda}_2 = 2{,}78.$$

Это означает, что для трассы с коэффициентом корреляции соседних элементов $R = 0{,}7$ вместо модели плотности в виде гиперэкспоненциального распределения $g(x;\ p,\ \lambda_1,\ \lambda_2)$ можно использовать (в рамках допустимой погрешности) просто экспоненциальное распределение с показателем $\lambda \approx 2{,}75$. Такой вывод основан на более подробном анализе корреляционных свойств сгенерированной трассы, представленном на рисунке 1 и на сравнении данных результатов с результатами, полученными в работах [9, 21].

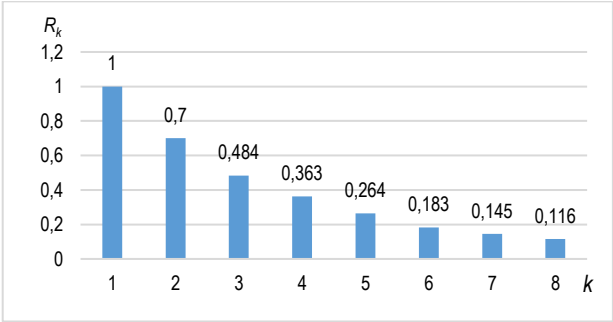

Рис. 1. Коэффициенты корреляции последовательности

Fig. 1. Sequence Correlation Coefficients

В этих работах рассмотрены методы синтеза параметров гиперэкспоненциального распределения $H_2(x)$ при полном учете корреляционных свойств последовательности (таких, как на рисунке 1), который основан на использовании индекса дисперсии интервалов (или отсчетов) последовательности. Практически все результаты в данных работах указывают на то, что одна из экспонент в распределении $H_2(x)$ в данной ситуации присутствует с весом p из интервала (0,95 ... 0,98). Поэтому для достаточно простой модели плотности вероятностей мгновенных значений коррелированной последовательности случайных интервалов времени может использоваться экспоненциальная плотность с показателем, учитывающим корреляционные свойства последовательности. Справедливость данного утверждения демонстрируют результаты, показанные на рисунке 2, где приведены плотности распределений вероятностей при полученных значениях параметров гиперэкспоненциальных распределений для коррелированного трафика ($f_k(t)$), для некоррелированного трафика ($f_n(t)$), а также в случае использования экспоненциальной плотности распределений коррелированного трафика ($f_{1\exp}(t)$) и некоррелированного трафика ($f_{2\exp}(t)$).

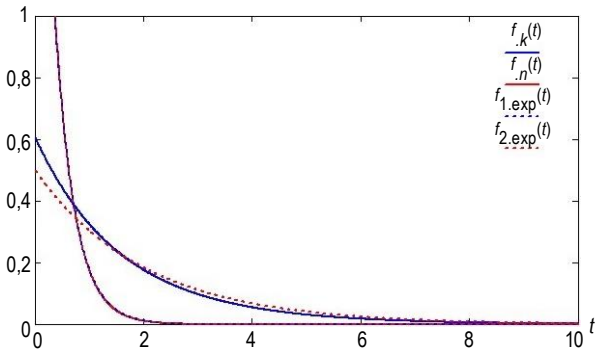

Рис. 2. Плотности распределений вероятностей

Fig. 2. Densities of Probability Distributions

Анализ полученных результатов (см. рисунок 2) показывает, что график плотности экспоненциального распределения практически совпал с графиком плотности гиперэкспоненциального распределения для некоррелированной последовательности, и весьма близок для коррелированной последовательности.

Использование EM-алгоритма подтвердило вывод о том, что при рекуррентном формировании оценок максимального правдоподобия для параметров гиперэкспоненциального распределения при априорной неопределенности относительно значений этих параметров, оценивания параметров гиперэкспоненциального распределения при наличии корреляционных свойств у рассматриваемых последовательностей.

Заключение

В работе представлена методика аппроксимации распределения смеси гиперэкспонентами для систем обработки коррелированного потока с использованием EM-алгоритма. В результате были получены выражения для программной реализации EM-алгоритма для случаев некоррелированных и коррелированных потоков.

Проверка на основе использования методов статистического моделирования работоспособности представленного EM-алгоритма для коррелированных последовательностей была проведена для случайных величин, имеющих логнормальное распределение с коэффициентом корреляции соседних интервалов $R = 0{,}7$. Результаты проверки показали точность в пределах порогового значения $\varepsilon_{\text{пороговое}} = 0{,}01$. Были получены следующие параметры гиперэкспоненциального распределения: $\hat{p} = 0{,}45$, $\hat{\lambda}_1 = 2{,}72$, $\hat{\lambda}_2 = 2{,}78$. Данные результаты согласуются с решениями, полученными на основе других подходов оценивания параметров гиперэкспоненциальных распределений.

Установлено, что для достаточно простой модели плотности вероятностей мгновенных значений коррелированной последовательности случайных величин может использоваться экспоненциальная плотность с параметром, учитывающим корреляционные свойства последовательности.

Список используемых источников

1. Kleinrock L. Queueing Systems. Vol. I. Theory. John Wiley & Sons, 1975. 432 p.
2. Шелухин О.И., Тенякшев А.М., Осин А.В. Фрактальные процессы в телекоммуникациях. М.: Радиотехника, 2003. 480 с.
3. Буранова М.А. Исследование статистических характеристик самоподобного телекоммуникационного трафика // Инфокоммуникационные технологии. 2012. Т. 10. № 4. С. 35–41.
4. Kartashevskiy V., Kireeva N., Buranova M., Chupakhina L. Study of queuing system G/G/1 with an arbitrary distribution of time parameter system // Proceedings of the 2nd International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T, Kharkiv, Ukraine, 13–15 October 2015). IEEE, 2015. DOI:10.1109/INFOCOMMST.2015.7357297
5. Kartashevskiy V., Buranova M. Analysis of Packet Jitter in Multiservice Network // Proceedings of the 5th International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T, Kharkiv, Ukraine, 9–12 October 2018). IEEE, 2015. PP. 797–802. DOI:10.1109/INFOCOMMST.2018.8632085.
6. Буранова М.А., Карташевский В.Г., Латыпов Р.Т. Оценка джиттера в системе G/M/1 на основе использования гиперэкспоненциальных распределений // Инфокоммуникационные технологии. 2020. Т. 18. № 1. С. 13–20. DOI:10.18469/ikt.2020.18.1.02
7. Kartashevskiy I., Buranova M. Calculation of Packet Jitter for Correlated Traffic // Proceedings of the 19th International Conference on Next Generation Wired/Wireless Networking (NEW2AN 2019) and Proceedings of the 12th Conference on Internet of Things and Smart Spaces (ruSMART, 2019, St. Petersburg, Russia, 26–28 August 2019). Lecture Notes in Computer Science. Computer Communication Networks and Telecommunications. Vol. 11660. PP. 610–620. Cham: Springer, 2019. DOI:10.1007/978-3-030-30859-9_53.
8. Vishnevskii V.M., Dudin A.N. Queueing Systems with Correlated Arrival Flows and Their Applications to Modeling Telecommunication Networks // Automation and Remote Control. 2017. Vol. 78(8). PP. 1361–1403. DOI:10.1134/S0005117917080001X
9. Balcioğlu B., Jagerman D.L., Altiok T. Approximate mean waiting time in a $GI/D/1$ queue with autocorrelated times to failures // IIE Transactions. 2007. Vol. 39. Iss. 10. PP. 985–996. DOI:10.1080/07408170701275343
10. Бузов А.Л., Букашкин С.А. Специальная радиосвязь. Развитие и модернизация оборудования и объектов. М.: Радиотехника, 2017. 448 с.
11. Balcıoglu B., Jagerman D.L., Altıok T. Merging and splitting autocorrelated arrival processes and impact on queueing performance // Performance Evaluation. 2008. Vol. 65. Iss. 9. PP. 653–669. DOI:10.1016/j.peva.2008.02.003
12. Keilson J., Machihara F. Hyperexponential waiting time structure in hyperexponential H_N /H_K/1 system // Journal of the Operation Society of Japan. 1985. Vol. 28(3). PP. 242–250.
13. Feldmann A., Whitt W. Fitting Mixtures of Exponentials to Long-Tail Distributions to Analyze Network Performance Models // Performance Evaluation. 1998. Vol. 31. Iss. 3-4. PP. 245–279. DOI:10.1016/S0166-5316(97)00003-5
14. Buranova M., Ergasheva D., Kartashevskiy V. Using the EM-algorithm to approximate the Distribution of a Mixture by Hyperexponents // Proceedings of the International Conference on Engineering and Telecommunication (EnT, Dolgoprudny, Russia, 20–21 November 2019). IEEE, 2019. DOI:10.1109/EnT47717.2019.9030551
15. Baird S.R. Estimating mixtures of exponential distributions using maximum likelihood and the EM algorithm to improve

simulation of telecommunication networks. University of British Columbia, 2002. URL: https://open.library.ubc.ca/collections/ubctheses/831/items/1.0090805 (дата обращения 18.10.2021)

16. Королев В.Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. М.: ИПИ РАН, 2007. 94 с.

17. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения 18.10.2021)

18. Ip E.H. A Stochastic EM Estimator in the Presence of Missing Data: Theory and Practice. PhD Dissertation. Stanford University, 1994. 127 p.

19. Kullback S., Leibler R.A. On information and sufficiency // The Annals of Mathematical Statistics. 1951. Vol. 22. PP. 79–86.

20. Быков В.В. Цифровое моделирование в статистической радиотехнике. М.: Советское радио, 1971. 328 с.

21. Карташевский И.В., Сапрыкин А.В. Анализ времени ожидания заявки в очереди для системы массового обслуживания общего вида // T-Comm: Телекоммуникации и транспорт. 2018. Т. 12. № 2. С. 4–10. DOI:10.24411/2072-8735-2018-10024

* * *

Application of EM-Algorithm for Approximation of Correlated Traffic Probabilities Density by Hyperexponents

M. Buranova[1]*, **I. Kartashevskiy**[1]

[1]Povolzhskiy State University of Telecommunications and Informatics,
Samara, 443010, Russian Federation

Article info
DOI:10.31854/1813-324X-2021-7-4-10-17
Received 18th September 2021
Revised 11th October 2021
Accepted 18th October 2021

For citation: Buranova M., Kartashevskiy I. Application of EM-Algorithm for Approximation of Correlated Traffic Probabilities Density by Hyperexponents. *Proc. of Telecom. Universities.* 2021;7(4):10–17. (in Russ.) DOI:10.31854/1813-324X-2021-7-4-10-17

Abstract: *An accurate assessment of the quality of service parameters in modern information communication networks is a very important task. This paper proposes the use of hyperexponential distributions to solve the problem of approximating an arbitrary probability density in the G/G/1 system for the case when the approximation by a system of the type $H_2/H_2/1$ is assumed. To determine the parameters of the probability density of the hyperexponential distribution, it is proposed to use EM- algorithm that provides fairly simple use cases for uncorrelated flows. In this paper, we propose a variant of the EM algorithm implementation for determining the parameters of the hyperexponential distribution in the presence of correlation properties of the analyzed flow.*

Keywords: *EM-algorithm, distributions of random variables, correlation, approximation, QoS.*

References

1. Kleinrock L. *Queueing Systems. Vol. I. Theory.* John Wiley & Sons; 1975. 432 p.

2. Sheluhin O.I., Tenyakshev A.M., Osin A.V. *Fractal Processes in Telecommunications*. Moscow: Radiotehnika Publ.; 2003. 480 p. (in Russ.)

3. Buranova M.A. Research of statistical characteristics of the self-similar telecommunication traffic. *Infokommunikacionnie tehnologii*. 2012;10(4):35–40. (in Russ.)

4. Kartashevskiy V., Kireeva N., Buranova M., Chupakhina L. Study of queuing system G/G/1 with an arbitrary distribution of time parameter system. *Proceedings of the 2nd International Scientific-Practical Conference Problems of Infocommunications Science and Technology, PIC S&T, 13–15 October 2015, Kharkiv, Ukraine*. IEEE; 2015. DOI:10.1109/INFOCOMMST.2015.7357297

5. Kartashevskiy V., Buranova M. Analysis of Packet Jitter in Multiservice Network // *Proceedings of the 5th International Scientific-Practical Conference Problems of Infocommunications Science and Technology, PIC S&T, 9–12 October 2018, Kharkiv, Ukraine*. IEEE; 2018. p.797–802. DOI:10.1109/INFOCOMMST.2018.8632085

6. Buranova M.A., Kartashevskiy V.G., Latypov R.T. Estimation of jitter in the G/M/1 system based on the use of hyperexponential distributions. *Infokommunikacionnie tehnologii*. 2020;18(1):13–20. (in Russ.) DOI:10.18469/ikt.2020.18.1.02

7. Kartashevskiy I., Buranova M. Calculation of Packet Jitter for Correlated Traffic. *Proceedings of the 19th International Conference on Next Generation Wired/Wireless Networking, NEW2AN 2019, and Proceedings of the 12th Conference on Internet of Things and Smart Spaces, ruSMART 2019, 26–28 August 2019, St. Petersburg, Russia. Lecture Notes in Computer Science. Computer Communication Networks and Telecommunications.* vol.11660. p.610–620. Cham: Springer; 2019. DOI:10.1007/978-3-030-30859-9_53

8. Vishnevskii V.M., Dudin A.N. Queueing Systems with Correlated Arrival Flows and Their Applications to Modeling Telecommunication Networks. *Automation and Remote Control*. 2017;78(8):1361–1403. DOI:10.1134/S000511791708001X

9. Balcioğlu B., Jagerman D.L., Altiok T. Approximate mean waiting time in a *GI*/*D*/1 queue with autocorrelated times to failures. *IIE Transactions*. 2007;39(10):985–996. DOI:10.1080/07408170701275343

10. Buzov A.L., Bukashkin S.A. *Special Radio Communication. Development and Modernization of Equipment and Facilities.* Moscow: Radiotekhnika Publ.; 2017. 448 p. (in Russ.)

11. Balcıoglu B., Jagerman D.L., Altıok T. Merging and splitting autocorrelated arrival processes and impact on queueing performance. *Performance Evaluation*. 2008;65(9):653–669. DOI:10.1016/j.peva.2008.02.003

12. Keilson J., Machihara F. Hyperexponential waiting time structure in hyperexponential H_N /H_K/1 system. *Journal of the Operation Society of Japan*. 1985;28(3):242–250.

13. Feldmann A., Whitt W. Fitting Mixtures of Exponentials to Long-Tail Distributions to Analyze Network Performance Models. *Performance Evaluation*. 1998;31(3-4):245–279. DOI:10.1016/S0166-5316(97)00003-5

14. Buranova M., Ergasheva D., Kartashevskiy V. Using the EM-algorithm to approximate the Distribution of a Mixture by Hyperexponents. *Proceedings of the International Conference on Engineering and Telecommunication, EnT, 20–21 November 2019, Dolgoprudny, Russia*. IEEE; 2019. DOI:10.1109/EnT47717.2019.9030551

15. Baird S.R. *Estimating mixtures of exponential distributions using maximum likelihood and the EM algorithm to improve simulation of telecommunication networks*. University of British Columbia; 2002. Available from: https://open.library.ubc.ca/collections/ubctheses/831/items/1.0090805 [Accessed 18th October 2021]

16. Korolyov V.Yu. The EM-Algorithm, its Modifications, and Their Application to the Problem of Separating Mixtures of Probability Distributions. Theoretical Review. Moscow: Institute of Informatics Problems of the Russian Academy of Sciences Publ.; 2007. 94 p. (in Russ.)

17. Voroncov K.V. Mathematical Teaching Methods on Precedents. (in Russ.) Available from: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf [Accessed 18th October 2021]

18. Ip E.H. *A Stochastic EM Estimator in the Presence of Missing Data: Theory and Practice*. PhD Dissertation. Stanford University; 1994. 127 p.

19. Kullback S., Leibler R.A. On information and sufficiency. *The Annals of Mathematical Statistics*. 1951;22:79–86.

20. Bykov V.V. *Digital Modeling in Statistical Radio Engineering*. Moscow: Soviet radio Publ.; 1971. 328 p. (in Russ.)

21. Kartashevskiy I.V., Saprykin A.V. Waiting time analysis for the request in general queuing system. *T-Comm*. 2018;12(2): 4–10. (in Russ.) DOI:10.24411/2072-8735-2018-10024

Сведения об авторах:

БУРАНОВА Марина Анатольевна — кандидат технических наук, доцент, доцент кафедры информационной безопасности Поволжского государственного университета телекоммуникаций и информатики, buranova-ma@psuti.ru
https://orcid.org/0000-0003-2986-8252

КАРТАШЕВСКИЙ Игорь Вячеславович — доктор технических наук, профессор кафедры программного обеспечения и управления в технических системах Поволжского государственного университета телекоммуникаций и информатики, ivk@psuti.ru
https://orcid.org/0000-0002-1388-4867