САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
02.03.02 «Фундаментальная информатика и информационные технологии»
ООП: Программирование и информационные технологии

ОТЧЕТ О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Тема задания: **Оценка трудоемкости алгоритма на основе эмпирического анализа**

Выполнил: **Васильев** В. В. 16.Б13-пу
Фамилия И. О. номер группы

Руководитель научно-исследовательской работы: Никифоров К. А.,
ФИО, должность, ученая степень
кандидат физ.–мат. наук, доцент

Санкт-Петербург, 2019

Содержание

1 Введение

Одним из ключевых свойств алгоритма является его эффективность. Это свойство связано с вычислительными ресурсами, используемыми алгоритмом. В настоящее время наиболее широко используемой оценкой является функция сложности в среднем, на основе которой с достаточно хорошей точностью могут быть прогнозированы (в статистическом смысле) временные оценки программной реализации алгоритма.

Однако проблема заключается в том, что средняя оценка, будучи статистической и точечной, не дает никакой информации о поведении алгоритма на конкретных входах, что важно как для задач большой размерности, так и для систем реального времени. Использование оценки в худшем случае приводит к существенному завышению временного прогноза из-за малой вероятности входных данных, обеспечивающих максимум функции трудоемкости алгоритма при фиксированной размерности задачи. Таким образом, интерес представляет задача построения практически значимой интервальной оценки трудоемкости алгоритма.

Возможным решением задачи повышения точности результатов эмпирического анализа алгоритма является рассмотрение функции трудоемкости при фиксированной входной длине как дискретной ограниченной случайной величины, имеющей некоторое неизвестное распределение. Подход заключается в построении доверительного интервала трудоемкости на основе аппроксимации неизвестного дискретного распределения значений трудоемкости непрерывным распределением с ограниченной вариацией, для чего предлагается использовать бета-распределение. Полученное решение позволяет задать более реальную правую границу сложности алгоритма при фиксированной входной длине с заданной доверительной вероятностью.

Данный метод включает в себя два этапа. Первый этап — предварительное исследование, целью которого является проверка гипотезы о виде закона распределения значений трудоемкости алгоритма как дискретной ограниченной случайной величины [1]. Второй этап — основное исследование,в ходе которого значения доверительной трудоемкости $f_\gamma(n)$ определяются в зависимости от входной длины алгоритма [2].

2 Постановка задачи

Целью данной работы является исследование применимости вероятностного анализа функции трудоемкости алгоритма и сравнение полученных результатов с классическим эмпирическим подходом. Основными задачами являются:

- выбор алгоритма для анализа;
- реализация выбранного алгоритма и генератора входных данных на одном из языков программирования;
- проведение этапа предварительного исследования и рассмотрение гипотезы о выбранном распределении;
- в случае принятия гипотезы о распределении проведение этапа основного исследования и выявление зависимости доверительной трудоемкости от входной длины алгоритма;
- анализ полученных результатов.

В качестве алгоритма исследования возьмем один из алгоритмов нахождения кратчайших путей в графе — алгоритм Паллоттино.

3 Описание выбранного алгоритма

Нахождение кратчайшего пути от одной вершины до всех остальных является одной из основных задач оптимизации сети. Существует несколько распространенных и хорошо известных алгоритмов, которые решают эту проблему. Одна категория состоит из алгоритма Дейкстры [3] и его модификаций, другая категория содержит алгоритм Беллмана – Форда – Мура [4–6] и его модификации. Есть также ряд других алгоритмов, которые не входят в эти категории. Однако они не будут рассмотрены в силу невозможности охватить всевозможные алгоритмы нахождения кратчайших путей в одной работе.

Алгоритм Паллоттино — алгоритм, который находит кратчайшее расстояние от одной из вершин до всех остальных на графах без петель, является модификацией алгоритма Беллмана – Форда – Мура. Этот алгоритм также работает для графов с ребрами отрицательного веса при определенных условиях. Он широко используется для решения задач оптимального распределения грузопотоков по транспортной сети и выбора наиболее выгодных путей ее развития.

В литературе есть небольшая путаница с наименованием этого алгоритма. Пейп развил предложение Д'Эсопо [7] и предложил улучшенный алгоритм [8]. В то же время Левит и Лившиц разработали свою версию с той же идеей [9]. Таким образом, у нас есть алгоритм Д'Эсопо – Пейпа – Левита, который использует двухстороннюю очередь для поддержания помеченных вершин в графе. Позже Паллоттино предложил использовать две очереди вместо двухсторонней очереди, чтобы избежать экспоненциальной сложности алгоритма [10]. Именно последний алгоритм является объектом исследований в данной работе.

Входными данными для алгоритма является граф G, подаваемый в виде упорядоченной совокупности множеств вершин V и ребер E: $G = (V, E)$. Оценка размера входных данных производится по размеру множеств вершин и ребер. Пусть n — количество вершин в множестве V, m — количество ребер в множестве E: $n = |V|, m = |E|$. В качестве основной операции алгоритма выберем операцию релаксации ребра.

Для анализа алгоритма Паллоттино был использован генератор полных графов без мультиребер. Стоит отметить, что генератор входных данных должен обеспечивать репрезентативность выборки, т. е. генерировать такие входные данные, которые по вероятности соответствуют особенностям применения данного алгоритма. Поскольку алгоритм Паллоттино используется для решения транспортных задач, в реализацию генератора было добавлено условие: для любых трех узлов графа должно выполняться неравенство треугольника. Пусть $G = (V, E)$ — полный неориентированный граф с n вершинами, $W : E \to R_+$ — функция весов ребер. Тогда неравенство треугольника имеет вид:

$$W((u, w)) \leq W((v, u)) + W((u, w)), \forall u, v, w \in V. \tag{1}$$

Отметим, что алгоритм Паллотино имеет оценки

$$O(n, m) = n, \tag{2}$$

$$O(n, m) = n \cdot m, \tag{3}$$

$$O(n, m) = n^2 \cdot m \tag{4}$$

в лучшем, среднем и худшем случаях соответственно [10].

4 Этап предварительного исследования

4.1 Основные этапы

Организация экспериментального исследования сложности алгоритма привела к необходимости модификации исходного кода реализации алгоритма Паллоттино, связанной с размещением счетчика для определения значения числа выполняемых основных операций на этом входе.

Рассмотрим более подробно этап планирования предварительного исследования [1]. Основной задачей этого этапа является определение рационального размера выборки при фиксированной длине входа для проведения экспериментальных исследований трудоемкости алгоритма в среднем. Все формулы для расчетов взяты из методики [1].

Выдвигается гипотеза, что функция трудоемкости алгоритма Паллоттино имеет бета-распределение. Основные этапы предварительного исследования:

1. Фиксация некоторого значения длины входа n из реального сегмента длин в области применения алгоритма Паллоттино. В рассматриваемом случае $n = 80$.

2. Определение необходимого числа экспериментов m с программной реализацией для получения гистограммы относительных частот значений трудоемкости, например, по методике, изложенной в [1].

3. Проведение экспериментального исследования и получение значений трудоемкости f_i для сгенерированных случайных допустимых входов D_i: $f_i = f_A(D_i), i = \overline{1, m}$, где $f_A(D_i)$ — произвольное значение функции трудоемкости.

4. Получение теоретических функций трудоемкости алгоритма для лучшего и худшего случаев, как функций длины входа. Для алгоритма Паллоттино эти функции имеют вид:

$$f_A^{\vee}(n) = n, \tag{5}$$

$$f_A^{\wedge}(n) = \frac{n^3 \cdot (n-1)}{2}. \tag{6}$$

Последняя формула следует из того, что алгоритм Паллоттино в худшем случае имеет оценку (4), а поскольку для полных графов $m = \frac{n \cdot (n-1)}{2}$, получаем (6).

5. Выбор числа s полусегментов для гистограммы частот значений трудоемкости.

6. Нормирование значений экспериментальной трудоемкости и построение на основе полученных данных гистограммы относительных частот в полусегментах.

7. Вычисление нормированного выборочного среднего и нормированной исправленной выборочной дисперсии по формулам:

$$\bar{t} = \frac{\overline{f_t}(n) - f^{\wedge}}{f^{\wedge} - f^{\vee}}, \tag{7}$$

$$s^2 = \frac{1}{m-1} \sum_{i=1}^{m} \frac{(f_i - \overline{f_t}(n))^2}{(f^{\wedge} - f^{\vee})^2}, \tag{8}$$

где $f^{\wedge}$ и $f^{\vee}$ — соответственно максимальное и минимальное значение теоретических функций трудоемкости, $\overline{f_t}(n)$ — выборочное среднее.

8. Формулировка гипотезы и расчет параметров аппроксимирующего закона распределения. В данном случае выдвигается гипотеза о бета-распределении. Параметры бета-распределения рассчитываются по формулам:

$$\alpha = \frac{\bar{t}}{s^2}(\bar{t} - (\bar{t})^2 - s^2), \tag{9}$$

$$\beta = \frac{(1-\bar{t})}{s^2}(\bar{t} - (\bar{t})^2 - s^2). \tag{10}$$

9. Расчет теоретических частот по функции плотности по формуле:

$$p_i = \int_{x_i}^{x_i+\Delta x_i} b(x, \alpha, \beta)dx, \tag{11}$$

где b — функция бета-распределения.

10. Расчет наблюдаемого значения критерия Пирсона по формуле:

$$\chi^2_{\text{набл}} = m \sum_{i=1}^{s} \frac{(w_i - p_i)^2}{p_i}, \tag{12}$$

где w_i — относительные частоты.

11. Проверка гипотезы о законе распределения. Если нет оснований отвергнуть нулевую гипотезу, то переход к основному этапу исследования. В противном случае – выбор другого закона распределения и повторная проверка гипотезы.

4.2 Результаты предварительного исследования

Для оценки необходимого числа экспериментов с программной реализацией алгоритма Паллоттино для фиксированной длины входа ($n = 80$) в соответствие с изложенной методикой был проведен этап предварительного исследования с коэффициентом доверия $\gamma = 0.95$.

Сначала была извлечена выборка объемом 200, вычислен коэффициент вариации $V_f = \frac{\sigma^2}{\bar{t}}$, вычислено значение $m^*_{(1)}$, результаты приведены в таблице 1.

Таблица 1

Предварительный объем выборки	200
Выборочное среднее	9372.87
Выборочная дисперсия	2774677.39005
Выборочное отклонение	1665.736291
Коэффициент вариации	0.177718915
Рассчитанный объем выборки	121334

Далее была извлечена выборка объемом 121334, результаты ее обработки приведены в таблице 2. Поскольку рассчитанный объем выборки оказался больше, чем объем выборки текущего эксперимента, то $m^*_{(1)} = 148178$.

Таблица 2

Предварительный объем выборки	121334
Выборочное среднее	9524.10
Выборочная дисперсия	3498806.02004
Выборочное отклонение	1870.509562
Коэффициент вариации	0.1963974169
Рассчитанный объем выборки	148178

Далее была извлечена выборка объемом 148178, результаты ее обработки приведены в таблице 3. Поскольку рассчитанный объем выборки оказался больше, чем объем выборки текущего эксперимента, то $m^*_{(1)} = 149628$.

Таблица 3

Предварительный объем выборки	148178
Выборочное среднее	9531.38
Выборочная дисперсия	3538446.02776
Выборочное отклонение	1881.075763
Коэффициент вариации	0.197356002
Рассчитанный объем выборки	149628

Далее была извлечена выборка объемом 149628, результаты ее обработки приведены в таблице 4. Поскольку рассчитанный объем выборки оказался меньше, чем объем выборки текущего эксперимента, то $m = 149192$.

Таблица 4

Предварительный объем выборки	149628
Выборочное среднее	9526.48
Выборочная дисперсия	3524490.40991
Выборочное отклонение	1877.362621
Коэффициент вариации	0.197067772
Рассчитанный объем выборки	149192

Далее была извлечена итоговая выборка объемом 149192 и построена гистограмма относительных частот на 387 полусегментах. Данное значе-

ние для количества полусегментов было получено с помощью функции создания гистограммы относительных частот из стандартного пакета для анализа данных среды Microsoft Excel. Данный пакет использует формулу Скотта для вычисления количества интервалов, делит исходную выборку на равные сегменты, вычисляет частоты и строит гистограмму на основу полученных данных. Нормирование значений экспериментальной трудоемкости и построение на основе полученных данных гистограммы в полусегментах представлены на 1. Значения параметров бета-распределения для итоговой выборки: $\alpha = 25.24821, \beta = 53988.69414$.

Результаты расчетов теоретических частот по функции плотности приведены на рис. 1. Наблюдаемое значения критерия Пирсона в данном случае $\chi^2_{\text{набл}} = 282.24386$. Поскольку $\chi^2_{\text{кр}}(0.05, 384) = 430.69192$ (вычислено стандартной функцией пакета Microsoft Excel), получаем $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(0.05, 384)$. Следовательно, нет оснований отвергать нулевую гипотезу и можно перейти к основному этапу исследования.

5 Этап основного исследования

5.1 Основные этапы

1. Определение сегмента значений длин входа, соответствующего особенностям применения данного алгоритма в разрабатываемой программной системе. В данном случае алгоритма Паллоттино будет применяться для массивов длиной от 80 до 2560;

2. Определение сегмента значений длин входа, для которого будут проводиться экспериментальные исследования. В данном примере таким сегментом является сегмент от 80 до 320;

3. Выбор шага изменения длины входа в экспериментальном исследовании. В данном случае значение шага равно 10;

4. Выбор необходимого числа m экспериментов с программной реализацией алгоритма для фиксированной длины входа для определения выборочной средней и дисперсии. В данном случае по результатам предварительного исследования $m = 149192$;

5. Расчет на основе экспериментальных данных значений выборочной средней и дисперсии для каждого значения n. В данном случае n изменяется от 80 до 320 с шагом 10;

6. Анализ экспериментальных данных — построение уравнения регрессии для выборочной средней и выборочной дисперсии;

7. Расчет на основе полученных результатов параметров аппроксимирующего бета-распределения по формулам (9), (10) как функций длины входа $\alpha(n), \beta(n)$;

8. Выбор значения доверительной вероятности и вычисление значений левого γ-квантиля бета-распределения [2]: $x_\gamma(n) = B^{-1}(\gamma, \alpha(n), \beta(n))$;

9. Вычисление значений функции доверительной трудоемкости для исследуемого сегмента длин входа по формуле [2]:

$$f_\gamma(n) = f^\vee(n) + x_\gamma(n)(f^\wedge(n) - f^\vee(n)). \tag{13}$$

5.2 Результаты основного исследования

Было проведено основное исследование с выборками объемом 149192 для входных данных в сегменте $n = [80, 2560]$. Результаты представлены на рисунках 2–7.

Результаты построения уравнения регрессии для нормированной выборочной средней и выборочной дисперсии показаны на рис. 2 и рис. **??** соответственно. В первом случае уравнение регрессии имеет вид $\bar{t} = 0.9449 \cdot n^{-1.1735}$, во втором — $s^2 = y = 0.1165 \cdot n^{-3.741}$. В данном случае $y = ax^{-b}$ — наилучший в смысле максимума значения коэффициента детерминации R^2 вид функции, подбор коэффициентов осуществлен с помощью метода наименьших квадратов (расчеты выполнены в Microsoft Excel).

График функции $\alpha(n)$ показан на рис. 4, график функции $\beta(n)$ показан на рис. 5.

В рассматриваемом примере $\gamma = 0.95, \alpha < \beta$, график значений $x_\gamma(n)$ показан на 6. На **??** показан график значений доверительной трудоемкости и трудоемкости в худшем случае для алгоритма Паллоттино на сегменте $n = [80, 2560]$. Отдельно отметим, что доверительная трудоемкость получена для значения доверительной вероятности $\gamma = 0.95$, т. е. в 95% случаев наблюдаемая в единичном эксперименте трудоемкость алгоритма не будет превышать значение доверительной трудоемкости. Для рассматриваемого примера эти значения в разы меньше трудоемкости в худшем случае на исследуемом сегменте длин входа.

6 Заключение

Оценка функции сложности алгоритма Паллоттино в худшем случае приводит к существенному завышению временного прогноза из-за малой вероятности входных данных, обеспечивающих максимум функции трудоемкости для рассматриваемой задачи нахождения кратчайших путей в графе.

Полученные результаты подтверждают возможность повышения достоверности прогнозирования временной эффективности компьютерных алгоритмов и более эффективного решения задачи выбора рациональных алгоритмов на основе сравнительного анализа функций доверительной трудоемкости вместо традиционного сравнения трудоемкости в среднем случае.

В качестве дальнейшей работы планируется провести ряд экспериментов для графов с «географическим» происхождением (это графы, построенные на основе транспортных сетей и реальных расстояний). Необходимость в дополнительном исследовании обусловлена тем, что требуется удостовериться в практической актуальности полученных результатов, т. к. в реальном мире графы практически никогда не являются полными.

Отдельный интерес представляет проведение аналогичных экспериментов для алгоритмов Беллмана – Форда – Мура и Дейкстры, а также сравнение полученных результатов с результатами анализа алгоритма Паллоттино, поскольку эти алгоритмы являются ближайшими конкурентами алгоритма Паллоттино для решения задачи нахождения кратчайших путей в графе.

Наконец, в исследованиях [1, 2] рассматриваются алгоритмы с функцией трудоемкости от одного параметра входа. В данной работе в качестве объекта исследований был взят алгоритм Паллоттино с функцией трудоемкости от двух параметров входа (количество вершин и количество ребер). Чтобы использовать методологию, изложенную в [2], потребовалось зафиксировать второй параметр. Именно поэтому в качестве входных данных рассмотрены только полные графы, для которых можно легко рассчитать количество ребер при заданном количестве вершин. Однако не всегда можно выразить один параметр функции трудоемкости через другой. Таким

образом, еще одной возможностью для улучшения работы является исследование функции трудоемкости при вариации нескольких параметров, что позволит применять данную методологию к более широкому классу алгоритмов.

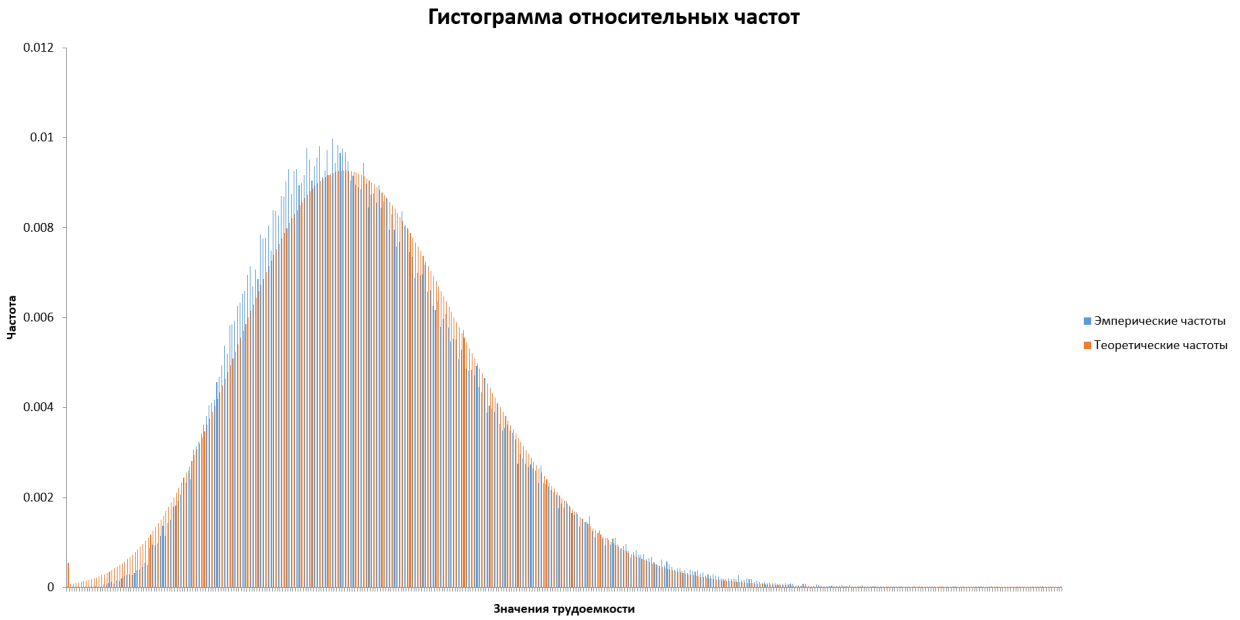

Рис. 1: теоретические и эмпирические частоты для алгоритма Паллоттино при $n = 80$ с разбиением нормированного сегмента [0, 1] на 387 полусегментов

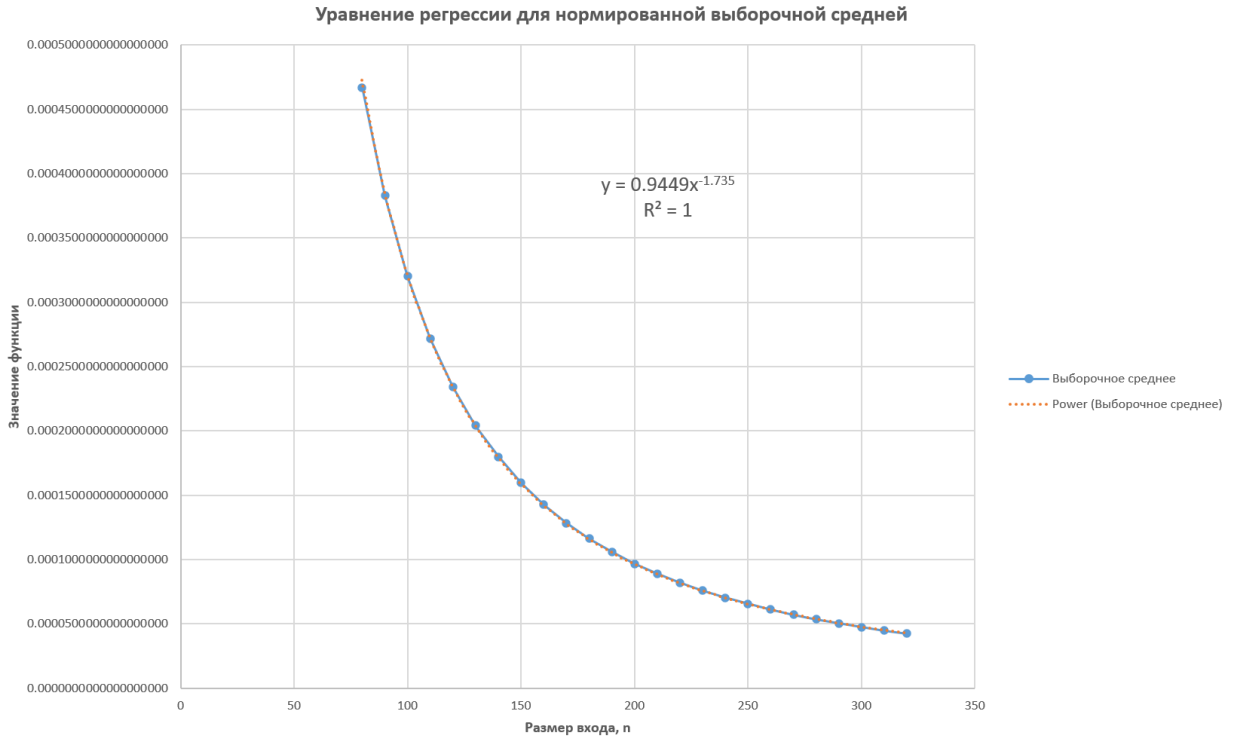

Рис. 2: экспериментальные данные и уравнение регрессии для нормированной выборочной средней значений трудоемкости алгоритма Паллоттино

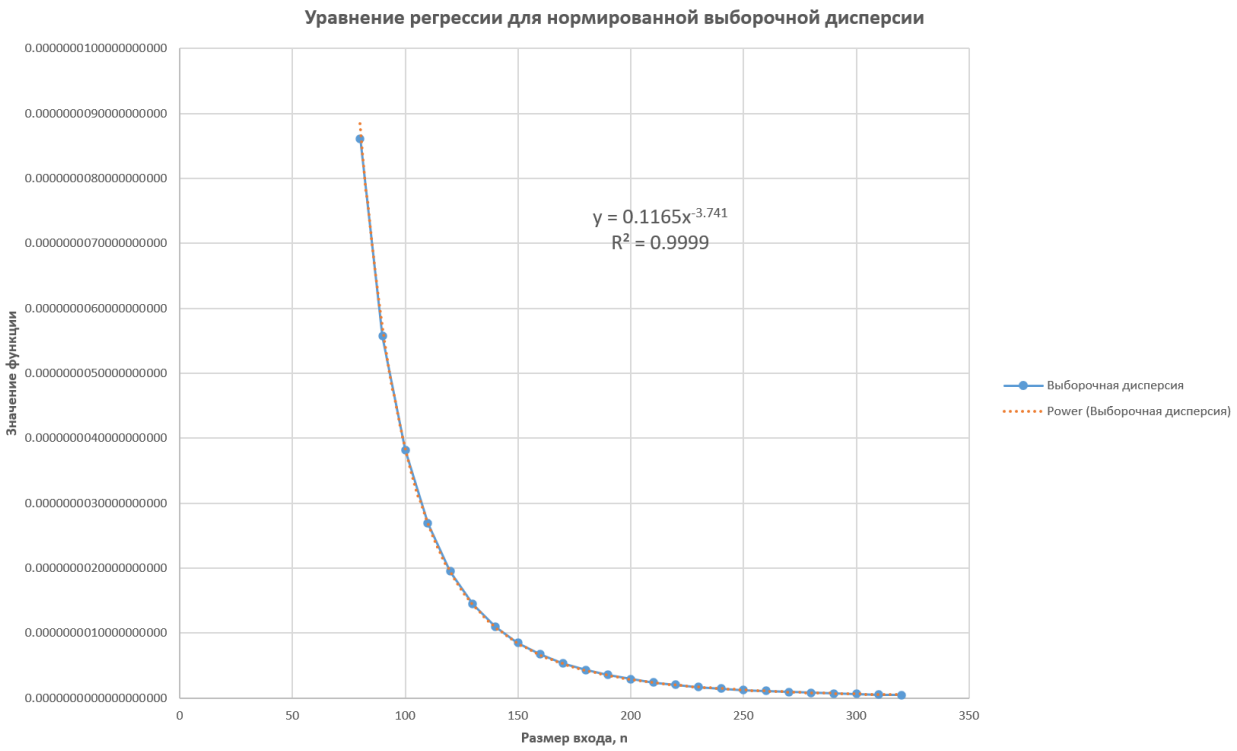

Рис. 3: экспериментальные данные и уравнение регрессии для нормированной выборочной дисперсии значений трудоемкости алгоритма Паллоттино

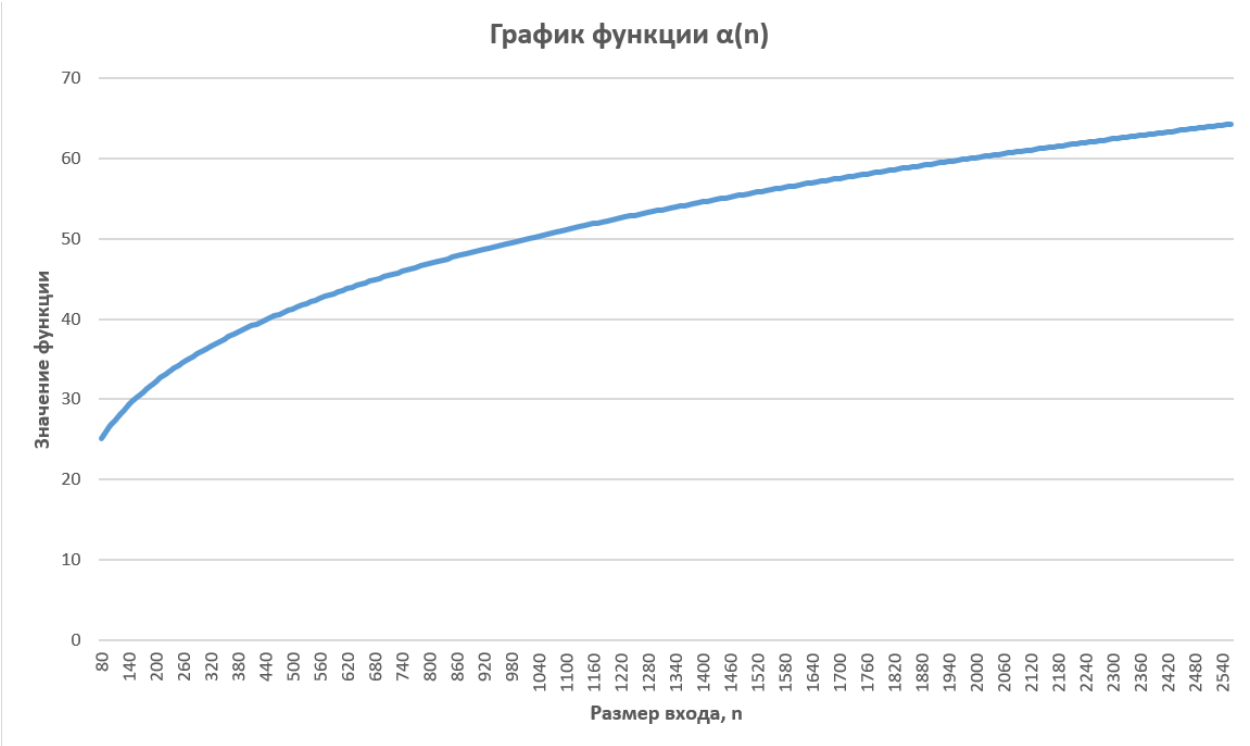

Рис. 4: график функции $\alpha(n)$ — параметра α аппроксимирующего бета-распределения для алгоритма Паллоттино

Рис. 5: график функции $\beta(n)$ — параметра β аппроксимирующего бета-распределения для алгоритма Паллоттино

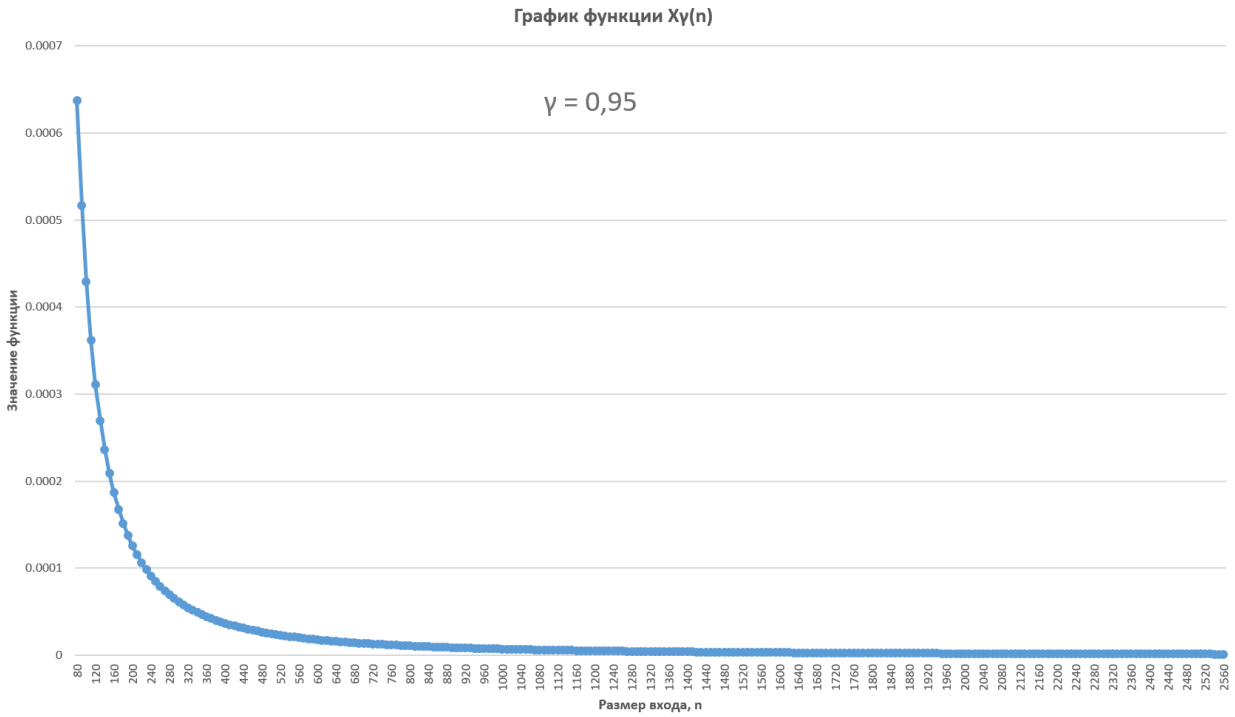

Рис. 6: график зависимости левого γ-квантиля бета-распределения $x_\gamma(n)$ от длины входа для алгоритма Паллоттино

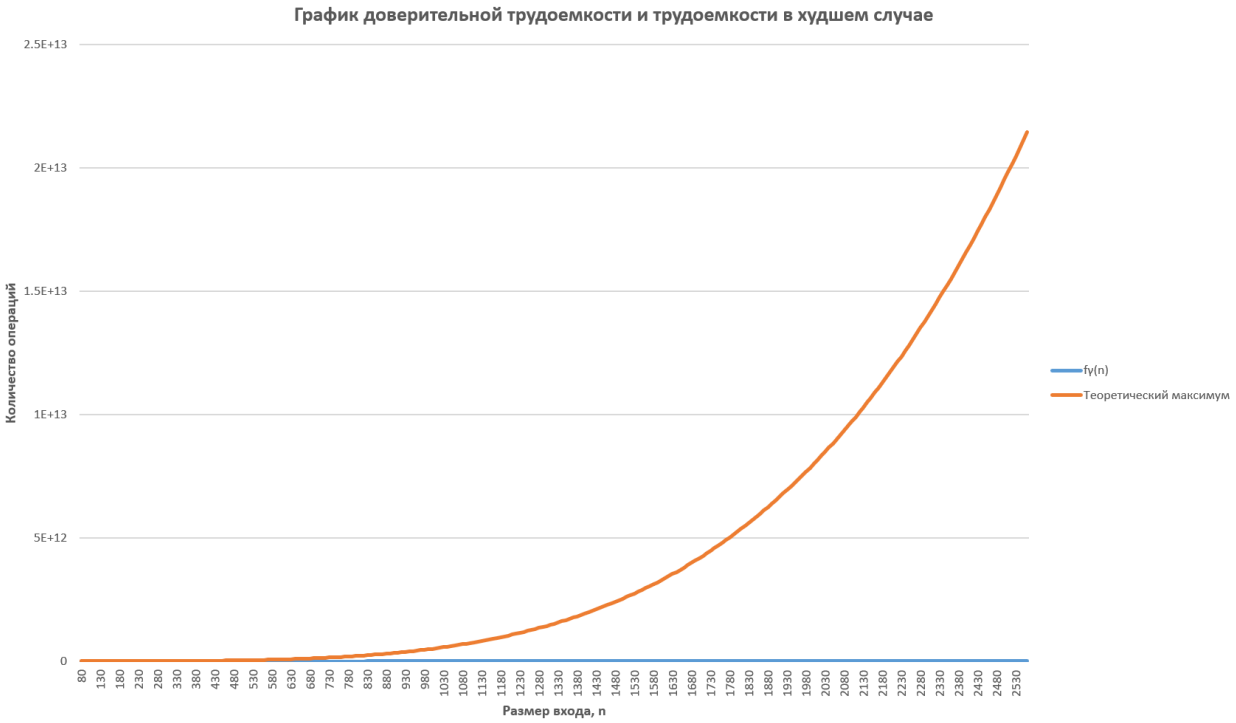

Рис. 7: график доверительной трудоемкости и трудоемкости в худшем случае для алгоритма Паллоттино

Список литературы

[1] Петрушин В. Н., Ульянов М. В. Планирование экспериментального исследования трудоемкости алгоритмов на основе бета-распределения // Информационные технологии и вычислительные системы. 2008. № 2. С. 81–91.

[2] Петрушин В. Н., Ульянов М. В., Кривенцов А. С. Доверительная трудоемкость — новая оценка качества алгоритмов // Информационные технологии и вычислительные системы. 2009. № 2. С. 23–37.

[3] Dijkstra E. W. A note on two problems in connexion with graphs // Numer. Math. 1959. Vol. 1. P. 269–271.

[4] Bellman R. On a routing problem // Quart. Appl. Math. 1958. Vol. 16. P. 87–90.

[5] Ford L. R. and Fulkerson D. R. Flows in Networks // Princeton: Princeton University Press. 1962.

[6] Moore E. F. The shortest path through a maze // Bell Telephone System. Technical publications. 1959. Vol. 3523.

[7] Pollack M. and Wiebenson W. Solutions of the shortest-route problem – A review // Oper. Res. 8. 1960. P. 224–230.

[8] Pape U. Implementation and efficiency of Moore-algorithms for the shortest route problem // Math. Program. 1974. Vol. 7. P. 212–222.

[9] Левит Б. Ю., Лившиц В. Н. Нелинейные сетевые транспортные задачи // Институт комплексных транспортных проблем. М., Изд-во «Транспорт». 1972. С. 1–144.

[10] Pallottino S. Shortest-path methods: Complexity, interrelations and new propositions // Networks. 1984. Vol. 14. P. 257–267.