ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том –, с. –

МАТЕМАТИКА

УДК –

ОПТИМАЛЬНОЕ РАЗДЕЛЕНИЕ ДАННЫХ В РАСПРЕДЕЛЕННОЙ ОПТИМИЗАЦИИ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

© **2023 г.** **Медяков Даниил[1], Молодцов Глеб[1], Александр Безносиков[1], Александр Гасников[1]**

Задача распределенной оптимизации в последнее время становится все более актуальной. Эта постановка имеет множество преимуществ, например, обработка большого объема данных за меньшее время по сравнению с нераспределенными методами. Однако, большинство распределенных подходов страдают от существенного недостатка – большой стоимости коммуникаций. Поэтому в последнее время большое количество исследований было направлено на решение этой проблемы. Один из таких подходов использует локальное сходство данных. В частности, существует алгоритм, доказательно оптимально использующий свойство подобия.. Однако этот результат, а также результаты других работ устраняют проблему коммуникаций, фокусируясь только на том факте, что они значительно дороже локальных вычислений и не учитывает различные мощности устройств в сети и различное соотношение между временем коммуникаций и затртами на локальные вычисления. Такая проблема и рассматривается в данном исследовании, целью которого является достижение оптимального распределения данных между сервером и локальными машинами при любой стоимости коммуникаций и локальных вычислений. Время работы сети сравнивается при равномерном и оптимальном распределениях данных. Ускорение, которое получается за счет последнего, подтверждено экспериментально.

1. ВСТУПЛЕНИЕ

1.1. РАСПРЕДЕЛЕННАЯ ОПТИМИЗАЦИЯ

Рассматривается задача оптимизации следующего вида:

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

где $x \in \mathbb{R}^d$ содержит параметры статистической модели для обучения, n – количество устройств/узлов в сети, а f_i – эмпирическая функция потерь на i-ом устройстве, то есть $f_i(x) = \frac{1}{b_i}\sum_{j=1}^{b_i} l(x, z_i^j)$, с $z_i^1, \ldots, z_i^{b_i}$ - набором выборок b_i, принадлежащих i-му устройству, и $l(x, z_i^j)$ измеряет несоответствие между параметром x и образцом z_i^j. Это формулировка задачи распределенной оптимизации. В настоящее время есть несколько причин сделать такую постановку задачи.

Чтобы достичь наилучших результатов в современных оптимизационных задачах в машинном обучении, исследователи и практики сталкиваются с различными вызовами. Иметь дело с современными моделями машинного обучения остается чрезвычайно сложной задачей, в первую очередь потому, что модели обучаются на все более объемных выборках данных. Наличие обучающей выборки большого объёма в наборе данных повышает устойчивость и обобщаемость полученной модели. В этом случае данные обычно обрабатываются с использованием сети устройств, т.е. собираются распределенным образом и хранятся в крайних узлах сети, например, как при классическом кластеризационном [1] и федеративном [2]—[4] обучении.

Для решения задачи (1) было предложено несколько методов решения. Стандартный подход предполагает чередование локальных вычислений на краевых устройствах (узлах $i = 2, \ldots, n$) с коммуникациями с сервером ($i = 1$), который поддерживает и обновляет актуальную копию оптимизационных переменных, в конечном итоге формируя окончательную оценку решения. При распределенном обучении сложных моделей проблемным местом часто становятся коммуникационные расходы между устройствами в сети. Такая проблема обусловливает необходимость разработки более эффективных методов распределенного обучения, некоторые из которых были описаны в [2], [5]—[7].

1.2. РАСПРЕДЕЛЁННАЯ ОПТИМИЗАЦИЯ В УСЛОВИЯХ СХОЖЕСТИ ДАННЫХ

Сегодня в машинном обучении модно использовать методы, основанные на идее инерционного моментума. Одним из путей решения распределенной оптимизационной задачи является применение ускорения Нестерова [8], которое представляет собой оптимальный метод для гладких нераспределенных оптимизационных задач. Для распределенных сетей он может быть применен следующим образом. На каждой итерации локально вычисляется градиент и результаты отправляются на сервер. Он, в свою очередь, усредняет полученные градиенты и делает шаг метода. Тогда количество коммуникаций будет равно количеству итераций. В этом случае получаются оптимальные оценки для локальных вычислений - $\sqrt{\kappa}, \kappa = {}^{L}/_{\mu}$, где L и μ - константы гладкости и сильной выпуклости целевой функции f. В случае, когда κ мала, такой подход вполне приемлем. Однако для плохо обусловленных функций с большой κ полиномиальная зависимость от κ может оказаться неудовлетворительной из-за высокой стоимости коммуникаций. Это часто имеет место для многих задач Минимизации Эмпирического Риска (МЭР), когда оптимальный параметр регуляризации для тестового прогнозирования очень мал.

Для дальнейшего решения пробелемы стоимости коммуникаций можно использовать дополнительную структуру, обычно встречающуюся в задачах МЭР, известную как схожесть данных [9]—[11]. Ее можно определить как разность градиентов функций, т.е. $\|\nabla f_i(x) - \nabla f_j(x)\| < \delta \ \ \forall x$. Но такой подход не является "естественным так как если задача не ограниченна, то такая δ не может существовать. Рассмотрим, например, квадратичную постановку: $\nexists\, \delta : \|(A_i - A_j)x\| < \delta$, если $x \to \infty$. Поэтому, сфокусируемся на другом подходе, а именно, на подобии Гессианнов. В частности, для всех x из рассматриваемой области и всех $i \neq j;\ i, j \in \{1, \ldots, n\}$, разность между матрицами Гессиана локальных потерь, обозначаемая $\|\nabla^2 f_i(x) - \nabla^2 f_j(x)\|$, ограничена δ, где $\delta > 0$ измеряет степень сходства. При таком предположении можно оценить $\delta \sim \mathcal{O}({}^{1}/_{\sqrt{N}})$, где N - размер выборки на одно устройство [9]. Впервые такой подход был исследован в [10]. После этого в [9] были доказаны нижние оценки для этой задачи, где затраты на коммуникации пропорциональны $\sqrt{{}^{\delta}/_{\mu}}$. Затем, долгое время исследователи пытались найти методы, позволяющие достичь этих оценок. В частности, были получены такие алгоритмы, как [12]—[16]. В 2022 году удалось получить оптимальный метод, который описан в [17].

1.3. РАЗЛИЧНЫЕ ЗАТРАТЫ НА КОММУНИКАЦИИ И ЛОКАЛЬНЫЕ ВЫЧИСЛЕНИЯ

Во всех вышеперечисленных работах авторы выдвигали предположение о том, что коммуникации обходятся значительно дороже локальных вычислений. Более того, в целом в работах по распределенной оптимизации, не только в области схожести Гессианов, делалось такое предположение. Данный вопрос будет рассмотрен с другой стороны, уходя от фиксированных больших коммуникаций и сделав 2 **предположения**:

1. Устройства в сети имеют разную мощность, то есть выполняют локальные вычисления одного и того же объема данных за разное время.
2. Отношение коммуникационных затрат к локальному времени вычислений – величина переменная, которая может быть либо $\ll 1$, либо $\gg 1$, либо даже ~ 1.

При таких предположениях необходим новый подход к задаче распределенной оптимизации, основанный на уже полученных оптимальных алгоритмах. Это приводит к исследуемому в данной статье вопросу.

Можно ли найти такое распределение данных между устройствами в сети, чтобы сократить реальное время работы оптимального алгоритма [17] *при любых коммуникационных затратах и времени локальных вычислений?*

На практике сети могут работать в течение длительного времени, и, как следствие, в них могут возникать шумы. Другими словами, стоимость связи и мощность устройств не являются постоянными величинами в течении работы сети. Таким образом, делается еще одно **предположение**:

3. Стоимости связи и мощности устройств задаются как случайные величины, затем на длительное время запускается работа сети и измеряются их математическое ожидание и дисперсия. Поскольку распределение данных по устройствам зависит от постоянных времени связи и мощности устройств, то в реальности оптимальное распределение будет отличаться из-за шума.

Поэтому в связи с этим предположением возникает вопрос об измерении погрешности времени работы программы при оптимальном распределении данных.

1.4. ВКЛАД

В целом наш вклад заключается в следующем:

- **Обобщение модели вычислений.** Строится общая модель времени вычислений в сетях при распределенной оптимизации. Модель основана на оптимальном алгоритме [17] и учитывает разницу в мощностях устройств при различных затратах на связь.
- **Всеобъемлющий анализ.** Особое внимание уделяеется частным случаям и полученным в них результатам. Рассматривается случай, когда коммуникации слишком дорогие, а также случай дешёвых коммуникаций (не настолько дорогих, чтобы связь занимала больше времени, чем обработка всех данных одним устройством). Более того, результаты получаются не только с учетом разницы во временных затратах, но и при рассмотрении различных оценок δ.
- **Различные техники получения решения.** Получаются результаты в различных случаях, в том числе для различных оценок δ. Также используются следующие методы: формула Кардано, верхние оценки в предельных случаях и нахождение нуля функции простейшими численными методами.
- **Погрешность решения из-за шума.** При третьем предположении приводена теоретическую погрешность времени работы программы при шуме времени коммуникаций и локальных вычислений.
- **Эксперименты.** На основе проведенных экспериментов было получено подтверждение того, что с выведенным распределением решение выбранной задачи занимает меньшее время. Кроме того, соответствующие эксперименты были дополнены шумом в сети.

2. ПОСТАНОВКА ЗАДАЧИ

На данном этапе остановимся на первых двух предположениях из Секции 1.3. Для достижения меньшей коммуникационной и локально-градиентной сложности обратимся к алгоритму 1 из [17]. Для этого необходимо представить функцию в виде суммы гладкой выпуклой функции f_1 и гладкой потенциально невыпуклой функции $f - f_1$. Тогда алгоритм будет переписан в следующем более общем виде:

Алгоритм 1 Accelerated Extragradient

1: **Input:** $x^0 = x_f^0 \in \mathbb{R}^d$
2: **Parameters:** $\tau \in (0,1), \eta, \theta, \alpha > 0, K \in \{1, 2, \ldots\}$
3: **for** $k = 0, 1, 2, \ldots, K-1$ **do**
4: $\quad x_g^k = \tau x^k + (1-\tau) x_f^k$
5: $\quad x_f^{k+1} \approx \arg\min\limits_{x \in \mathbb{R}^d} [A_\theta^k(x) := \langle \nabla(f - f_1)(x_g^k), x - x_g^k \rangle + \frac{1}{2\theta}\|x - x_g^k\|^2 + f_1(x)]$
6: $\quad x^{k+1} = x^k + \eta\alpha(x_f^{k+1} - x^k) - \eta \nabla f(x_f^{k+1})$
7: **end for**
8: **Output:** x^K

Нам необходимо проанализировать работу этого алгоритма, а именно, выяснить, сколько операций выполняет этот алгоритм за итерацию. В строке 5, при решении подзадачи arg min, на устройствах выполняется одно локальное вычисление для вычисления $f_i(x_g^k)$, затем одна коммуникация для передачи этих результатов и дополнительные вычисления на сервере для нахождения решения x_f^{k+1}. Затем, в строке 6, выполняется одно локальное вычисление и одна коммуникация. Получается выражение для общего времени работы алгоритма. Введем следующие обозначения: τ_i – время одного локального вычисления на i-ом устройстве, K – количество итераций, τ_{comm} – время одной коммуникации, k_{some} – дополнительные вычисления центрального узла, n – количество узлов в сети. С учетом этого общее время работы алгоритма можно записать в виде:

$$T_{sum} = 2 \cdot \max(\tau_1, \tau_2, \ldots, \tau_n) \cdot K + 2 \cdot K \cdot \tau_{comm} + \tau_1 \cdot k_{some}.$$

Наша задача - минимизировать время T_{sum}. С учетом утверждения (1) и вида функций f_i представим время τ_i в виде $\tau_i = \tau_i^{loc} \cdot b_i$, где τ_i^{loc} - мощность, т.е. время, затрачиваемое i-м устройством на обработку единицы информации, поступающей на его вход, а b_i - размер набора данных, поступающего на i-е устройство. b_i должно удовлетворять следующим ограничениям: $\sum\limits_{i=1}^{n} b_i = N$, где N - размер всего набора данных, $\delta = \frac{L}{\sqrt{b_i}}$ или $\delta = \frac{L}{b_i}$ [15].

В итоге была получена следующая задача минимизации:

$$\min_{\sum\limits_{i=1}^{n} b_i=N;\delta=\frac{L}{b_1\gamma}} [2\cdot\max(\tau_1^{loc}\cdot b_1,\tau_2^{loc}\cdot b_2,\ldots,\tau_n^{loc}\cdot b_n)\cdot K+2\cdot K\cdot\tau_{comm}+\tau_1\cdot k_{some}],\ \gamma\in\{\frac{1}{2},1\}. \quad (2)$$

3. КАК РЕШИТЬ (2)

3.1. ПЕРВИЧНАЯ ЗАДАЧА МИНИМИЗАЦИИ

В [17] представлены оценки K и k_{some}, а именно: $2\cdot K=\mathcal{O}(\max\{1,\sqrt{\frac{\delta}{\mu}}\}\log(\frac{1}{\varepsilon}))$, $k_{some}=\mathcal{O}(\max\{1,\sqrt{\frac{L}{\delta}},\sqrt{\frac{\delta}{\mu}},\sqrt{\frac{L}{\mu}}\}\log(\frac{1}{\varepsilon}))$.

В таком случае, (2) принимает вид:

$$\begin{aligned}\min_{\sum\limits_{i=1}^{n} b_i=N;\delta=\frac{L}{b_1\gamma}} &[(\max(\tau_1^{loc}\cdot b_1,\tau_2^{loc}\cdot b_2,\ldots,\tau_n^{loc}\cdot b_n)+\tau_{comm})\cdot\mathcal{O}(\max\{1,\sqrt{\frac{\delta}{\mu}}\log(\frac{1}{\varepsilon})\})\\ &+\tau_1^{loc}\cdot b_1\cdot\mathcal{O}(\max\{1,\sqrt{\frac{L}{\delta}},\sqrt{\frac{\delta}{\mu}},\sqrt{\frac{L}{\mu}}\}\log(\frac{1}{\varepsilon}))],\ \gamma\in\{\frac{1}{2},1\}.\end{aligned} \quad (3)$$

3.2. ВСПОМОГАТЕЛЬНАЯ ЗАДАЧА

Рассматриваем следующую вспомогательную задачу:

$$\min_{\sum\limits_{i=2}^{n} b_i=N} [\max(\tau_2^{loc}\cdot b_2,\tau_3^{loc}\cdot b_3,\ldots,\tau_n^{loc}\cdot b_n)]. \quad (4)$$

Лемма 1. *Решением задачи* (4) *является* $\overrightarrow{b}=(b_2,b_3,\ldots,b_n)^T$, *такой что* $\tau_2^{loc}\cdot b_2=\tau_3^{loc}\cdot b_3=\ldots=\tau_n^{loc}\cdot b_n$.

Доказательство. Не умаляя общности, примем фиксированные значения $\tau_2^{loc}\leqslant\tau_3^{loc}\leqslant\ldots\leqslant\tau_n^{loc}$. Затем произвольно выберем $b_2\geqslant b_3\geqslant\ldots\geqslant b_n$. Это действительно так, иначе имела бы место ситуация, когда $\exists\ i\neq j:\ i,j\in\{2,\ldots,n\}:\max(\tau_i^{loc}\cdot b_i,\tau_j^{loc}\cdot b_j)>\max(\tau_i^{loc}\cdot b_j,\tau_j^{loc}\cdot b_i)$, и, следовательно, распределение было бы неоптимальным.

Главная цель – минимизировать функцию $g(\overrightarrow{b})=\max(\tau_2^{loc}\cdot b_2,\tau_3^{loc}\cdot b_3,\ldots,\tau_n^{loc}\cdot b_n)$. Предположим, что существует такое распределение, что $\exists i\in\{2,\ldots,n\}:g(\overrightarrow{b}^0)=\tau_i^{loc}\cdot b_i^0$ является минимумом, и $\forall j:j\geqslant 2,j\neq i\hookrightarrow\tau_i^{loc}\cdot b_i^0>\tau_j^{loc}\cdot b_j^0$. Из этого следует, что $b_i^0>\frac{\tau_j^{loc}}{\tau_i^{loc}}b_j^0>\frac{\tau_{j_1}^{loc}}{\tau_i^{loc}}b_{j_1}^0>\ldots>\frac{\tau_{j_k}^{loc}}{\tau_i^{loc}}b_{j_k}^0$.

Далее, учитывая $\sum\limits_{i=2}^{n} b_i=N\hookrightarrow b_i^0+\frac{\tau_j^{loc}}{\tau_i^{loc}}b_j^0+\frac{\tau_{j_1}^{loc}}{\tau_i^{loc}}b_{j_1}^0+\ldots+\frac{\tau_{j_k}^{loc}}{\tau_i^{loc}}b_{j_k}^0>N$, получаем $b_i^0>N(1+\tau_i^{loc}\sum\limits_{\substack{j=2\\j\neq i}}^{n}\frac{1}{\tau_j^{loc}})^{-1}$. Затем рассмотрим $b_i=N(1+\tau_i^{loc}\sum\limits_{\substack{j=2\\j\neq i}}^{n}\frac{1}{\tau_j^{loc}})^{-1},\quad b_j=\frac{\tau_i^{loc}}{\tau_j^{loc}}\cdot b_i\ \ \forall j\in\{2,\ldots,n\}$. Такое распределение дает минимум $g(\overrightarrow{b})=\tau_i^{loc}\cdot b_i=\tau_j^{loc}\cdot b_j\ \ \forall j\in\{2,\ldots,n\}$, и $g(\overrightarrow{b})<g(\overrightarrow{b}^0)$. Это противоречит предположению о минимальности. Таким образом, для распределения, минимизирующего функцию $g(\overrightarrow{b})=\max(\tau_2^{loc}\cdot b_2,\tau_3^{loc}\cdot b_3,\ldots,\tau_n^{loc}\cdot b_n)$, справедливо $\tau_2^{loc}\cdot b_2=\tau_3^{loc}\cdot b_3=\ldots=\tau_n^{loc}\cdot b_n$. □

Вернемся к задаче (3). Помимо минимального выражения, уже изученного в (4), в задаче (3) имеются дополнительные члены. δ в (3) зависит от величины b_1, но не зависит от $b_i,i=\overline{2,n}$. Отсюда и из леммы 1 следует, что в исходной задаче (3) обмен данными между 2-м, 3-м и последующими устройствами должен быть пропорциональным. Таким образом, задача (3) сводится к новой задаче с дополнительными ограничениями:

$$\min_{\substack{\sum_{i=1}^{n} b_i=N;\delta=\frac{L}{b_1\gamma};\\ \tau_2^{loc}\cdot b_2=\ldots=\tau_n^{loc}\cdot b_n}} [(\max(\tau_1^{loc}\cdot b_1,\tau_2^{loc}\cdot b_2,\ldots,\tau_n^{loc}\cdot b_n)+\tau_{comm})\cdot\mathcal{O}(\max\{1,\sqrt{\frac{\delta}{\mu}}\log(\frac{1}{\varepsilon})\}) + \tau_1^{loc}\cdot b_1\cdot\mathcal{O}(\max\{1,\sqrt{\frac{L}{\delta}},\sqrt{\frac{\delta}{\mu}},\sqrt{\frac{L}{\mu}}\}\log(\frac{1}{\varepsilon}))],\ \gamma\in\{\frac{1}{2},1\}. \tag{5}$$

3.3. ОПРЕДЕЛИМ ОКОНЧАТЕЛЬНУЮ ЗАДАЧУ МИНИМИЗАЦИИ

Из леммы 1 следует, что $b_i\cdot\tau_i^{loc}=const\ \forall i\in\overline{2,n}$. Таким образом,

$$N-b_1=\sum_{i=2}^{n}b_i=\sum_{i=2}^{n}\frac{\tau_2^{loc}\cdot b_2}{\tau_i^{loc}}=\tau_2^{loc}\cdot b_2\cdot\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}}\Rightarrow b_2=\frac{N-b_1}{\tau_2^{loc}}(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}.$$

Как уже говорилось ранее, рассматривается случай $\delta=\frac{L}{b_1}$ и case of $\delta=\frac{L}{\sqrt{b_1}}$.

3.3.1. СЛУЧАЙ $\delta=\frac{L}{b_1}$

Здесь выполняются следующие соотношения:

$$\gamma=1,\ \mu\leqslant\delta\leqslant L\Rightarrow\begin{cases}2\cdot K=\mathcal{O}(\sqrt{\frac{L}{\mu b_1}}\log(\frac{1}{\varepsilon}))\\ k_{some}=\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}))\end{cases}.$$

Подставив эти оценки в (5), задача примет следующий вид:

$$\min_{\sum_{i=1}^{n} b_i=N}[(\max\{\tau_1^{loc}\cdot b_1,\tau_2^{loc}\cdot b_2\}+\tau_{comm})\cdot\mathcal{O}(\sqrt{\frac{L}{\mu b_1}}\log(\frac{1}{\varepsilon}))+\tau_1^{loc}\cdot b_1\cdot\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}))].$$

В результате, оставив в функции единственную переменную b_1, осуществляется переход к окончательному виду задачи минимизации:

$$\min_{0<b_1\leqslant N}[(\max\{\tau_1^{loc}\cdot b_1;\ (N-b_1)\cdot(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\}+\tau_{comm})\cdot\mathcal{O}(\sqrt{\frac{L}{\mu b_1}}\log(\frac{1}{\varepsilon}))+ + \tau_1^{loc}\cdot b_1\cdot\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}))]. \tag{6}$$

Исследуем эту задачу далее. Для этого найдем точку, в которой выражения под максимумом совпадают.

$$b_1^0\cdot(\tau_1^{loc}+(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1})=N(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\Rightarrow b_1^0=\frac{N(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}}{\tau_1^{loc}+(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}}.$$

Таким образом, получились два полуинтервала, на каждом из которых можно сформулировать свою задачу минимизации:

$$\begin{cases}(a)\ \ 0<b_1\leqslant b_1^0\Rightarrow\max\{\tau_1^{loc}\cdot b_1;\ (N-b_1)\cdot(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\}=(N-b_1)\cdot(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\\ (b)\ \ b_1^0<b_1\leqslant N\Rightarrow\max\{\tau_1^{loc}\cdot b_1;\ (N-b_1)\cdot(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\}=\tau_1^{loc}\cdot b_1\end{cases}.$$

Строим функции одной переменной $\mathcal{F}_1(b_1),\mathcal{F}_2(b_1)$ на соответствующих полуинтервалах, которые необходимо минимизировать в соответствии с задачей (6).

$$\begin{cases}(a):\ \mathcal{F}_1(b_1)=[N(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}+\tau_{comm}]\cdot c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1^{-\frac{1}{2}}-c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}b_1^{\frac{1}{2}}+\tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1\\ (b):\ \mathcal{F}_2(b_1)=\tau_{comm}\cdot c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1^{-\frac{1}{2}}+c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})\tau_1^{loc}b_1^{\frac{1}{2}}+\tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1\end{cases}.$$

Кроме того, сразу же найдем их производные для дальнейшего анализа.

$$\begin{cases} (a): \ \mathcal{F}'_1(b_1) = -\frac{1}{2}c_1 b_1^{-\frac{3}{2}}[N(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_{comm}] \cdot \sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) - \frac{1}{2}c_1 b_1^{-\frac{1}{2}}\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) \\ (b): \ \mathcal{F}'_2(b_1) = -\frac{1}{2}c_1 b_1^{-\frac{3}{2}}\tau_{comm}\cdot\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) + \frac{1}{2}c_1 b_1^{-\frac{1}{2}}\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})\tau_1^{loc} + \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) \end{cases}.$$

3.3.2. СЛУЧАЙ $\delta = \frac{L}{\sqrt{b_1}}$

Здесь будем действовать аналогично предыдущему пункту. Сначала представим необходимые в данном случае соотношения.

$$\gamma = \frac{1}{2},\ \mu \leqslant \delta \leqslant L \Rightarrow \begin{cases} 2\cdot K = \mathcal{O}(\sqrt{\frac{L}{\mu\sqrt{b_1}}}\log(\frac{1}{\varepsilon})) \\ k_{some} = \mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})) \end{cases}.$$

Подставляя эти соотношения в (5), получаем:

$$\min_{\sum\limits_{i=1}^{n} b_i = N}\ [(\max\{\tau_1^{loc}\cdot b_1, \tau_2^{loc}\cdot b_2\} + \tau_{comm})\cdot\mathcal{O}(\sqrt{\frac{L}{\mu\sqrt{b_1}}}\log(\frac{1}{\varepsilon})) + \tau_1^{loc}\cdot b_1\cdot\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}))].$$

Снова избавившись от всех переменных, кроме b_1, запишем окончательную задачу минимизации в этом случае:

$$\begin{aligned}\min_{0<b_1\leqslant N}[(\max\{\tau_1^{loc}\cdot b_1;\ (N-b_1)\cdot(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\} + \tau_{comm})\cdot\mathcal{O}(\sqrt{\frac{L}{\mu\sqrt{b_1}}}\log(\frac{1}{\varepsilon})) \\ +\ \tau_1^{loc}\cdot b_1\cdot\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}))].\end{aligned} \tag{7}$$

Аналогично выбираем точку b_1^0, она оказывается такой же, как и в предыдущем пункте. После этого можно получить два полуинтервала, на каждом из которых можно сформулировать свою задачу минимизации:

$$\begin{cases} (a)\ \ 0 < b_1 \leqslant b_1^0 \Rightarrow \max\{\tau_1^{loc}\cdot b_1;\ (N-b_1)\cdot(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\} = (N-b_1)\cdot(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} \\ (b)\ \ b_1^0 < b_1 \leqslant N \Rightarrow \max\{\tau_1^{loc}\cdot b_1;\ (N-b_1)\cdot(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\} = \tau_1^{loc}\cdot b_1 \end{cases}.$$

Мы строим функции одной переменной $\mathcal{F}_1(b_1), \mathcal{F}_2(b_1)$ на соответствующих полуинтервалах, которые необходимо минимизировать в соответствии с задачей (7).

$$\begin{cases} (a): \ \mathcal{F}_1(b_1) = [N(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_{comm}]\cdot c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1^{-\frac{1}{4}} - c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}b_1^{\frac{3}{4}} + \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1 \\ (b): \ \mathcal{F}_2(b_1) = \tau_{comm}\cdot c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1^{-\frac{1}{4}} + c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})\tau_1^{loc}b_1^{\frac{3}{4}} + \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})b_1 \end{cases}.$$

Кроме того, сразу же найдем их производные для дальнейшего анализа.

$$\begin{cases} (a): \ \mathcal{F}'_1(b_1) = -\frac{1}{4}c_1 b_1^{-\frac{5}{4}}[N(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_{comm}]\cdot\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) - \frac{3}{4}c_1 b_1^{-\frac{1}{4}}\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) \\ (b): \ \mathcal{F}'_2(b_1) = -\frac{1}{4}c_1 b_1^{-\frac{5}{4}}\tau_{comm}\cdot\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) + \frac{3}{4}c_1 b_1^{-\frac{1}{4}}\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})\tau_1^{loc} + \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) \end{cases}.$$

3.4. ФИНАЛЬНОЕ РЕШЕНИЕ

3.4.1. СЛУЧАЙ $\delta = \frac{L}{b_1}$

Наша цель - найти минимум уже полученных функций $\mathcal{F}_1(b_1), \mathcal{F}_2(b_1)$. Для этого будем искать нули $\mathcal{F}'_1(b_1), \mathcal{F}'_2(b_1)$. Здесь получаем кубическое уравнение. Для его решения можно воспользоваться формулой Кардано.

Рассмотрим уравнение $ax^{-\frac{1}{2}} + bx^{-\frac{3}{2}} + c = 0$,
где в случае $(a):\ \ 0 < b_1 \leqslant b_1^0$ и $(b):\ \ b_1^0 < b_1 \leqslant N$ получаем:

$$\begin{cases} (a):\ \ a = \frac{1}{2}c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1};\ b = -\frac{1}{2}c_1[N(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_{comm}]\cdot\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon});\ c = \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) \\ (b):\ \ a = \frac{1}{2}c_1\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon})\tau_1^{loc};\ b = -\frac{1}{2}c_1\tau_{comm}\cdot\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon});\ c = \tau_1^{loc}\cdot c_2\sqrt{\frac{L}{\mu}}\log(\frac{1}{\varepsilon}) \end{cases}.$$

Тогда при условии, что

$$N \geqslant \frac{a^2}{3c^2} + \frac{\sqrt[3]{2a^6 + 3\sqrt{3}\sqrt{4a^3b^3c^6 + 27b^4c^8} + 18a^3bc^2 + 27b^2c^4}}{3\sqrt[3]{2}c^2} - \frac{\sqrt[3]{2}\left(-a^4 - 6abc^2\right)}{3c^2\sqrt[3]{2a^6 + 3\sqrt{3}\sqrt{4a^3b^3c^6 + 27b^4c^8} + 18a^3bc^2 + 27b^2c^4}},$$

получаем решение:

$$x = \frac{a^2}{3c^2} + \frac{\sqrt[3]{2a^6 + 3\sqrt{3}\sqrt{4a^3b^3c^6 + 27b^4c^8} + 18a^3bc^2 + 27b^2c^4}}{3\sqrt[3]{2}c^2} - \frac{\sqrt[3]{2}\left(-a^4 - 6abc^2\right)}{3c^2\sqrt[3]{2a^6 + 3\sqrt{3}\sqrt{4a^3b^3c^6 + 27b^4c^8} + 18a^3bc^2 + 27b^2c^4}}.$$

Отсюда тривиально следует искомое решение. Поскольку на каждом из полуинтервалов получили одно значение b_1, которое является минимумом функции на нем, то, выбрав из них то, на котором функция меньше, получим оптимальное значение b_1.

3.4.2. СЛУЧАЙ $\delta = \frac{L}{\sqrt{b_1}}$

Поступить аналогично предыдущему параграфу не получается, так как выписать решение этих уравнений в аналитическом виде не представляется возможным из-за их степеней. Поэтому рассмотрим следующие предельные реализации:

1. $\forall i \hookrightarrow \tau_{comm} \ll \tau_i^{loc}$;
2. $\forall i \hookrightarrow \tau_{comm} \gg \tau_i^{loc}, \forall i \neq j \hookrightarrow \tau_i^{loc} = \tau_j^{loc}$.

Для упрощения записи, введём обозначения: $\alpha = c_1 \cdot \sqrt{\frac{L}{\mu}} \cdot \log(\frac{1}{\varepsilon}), \beta = c_2 \cdot \sqrt{\frac{L}{\mu}} \cdot \log(\frac{1}{\varepsilon})$. Теперь всё готово к рассмотрению двух случаев по-отдельности.

Реализация 1:

(a): $0 < b_1 \leqslant b_1^0$ и $\mathcal{F}_1(b_1) = [N(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} + \tau_{comm}]\cdot\alpha b_1^{-\frac{1}{4}} - \alpha(\sum\limits_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}b_1^{\frac{3}{4}} + \tau_1^{loc}\cdot\beta b_1$. Предположим, что $\tau_1^{loc} \leqslant \tau_2^{loc} \leqslant \ldots \leqslant \tau_n^{loc}$. Используя данное предположение, а так же что $\tau_{comm} \ll \tau_i^{loc}$, можно получить следующие оценки:

$$\begin{aligned} (\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1} &= \frac{1}{\frac{1}{\tau_1^{loc}} + \ldots + \frac{1}{\tau_n^{loc}}} \\ &= \frac{\tau_2^{loc}\cdot\ldots\cdot\tau_n^{loc}}{\tau_3^{loc}\cdot\ldots\cdot\tau_n^{loc} + \tau_2^{loc}\cdot\tau_4^{loc}\cdot\ldots\cdot\tau_n^{loc} + \ldots + \tau_2^{loc}\cdot\ldots\cdot\tau_{n-1}^{loc}} \\ &\geqslant \frac{\tau_2^{loc}}{n-1} \gg \tau_{comm}. \end{aligned} \tag{8}$$

Получим оценку (8), функции $\mathcal{F}_1(b_1)$ и соответственно $\mathcal{F'}_1(b_1)'$ можно приблизительно упростить следующим образом:

$$\begin{aligned} \mathcal{F}_1(b_1) &= \alpha(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\cdot b_1^{-\frac{1}{4}}(N - b_1) + \tau_1^{loc}\beta\cdot b_1, \\ \mathcal{F'}_1(b_1) &= \alpha(\sum_{i=2}^{n}\frac{1}{\tau_i^{loc}})^{-1}\cdot(-\frac{1}{4}b_1^{-\frac{5}{4}}N - \frac{3}{4}b_1^{-\frac{1}{4}}) + \tau_1^{loc}\beta. \end{aligned}$$

Получаем уравнение в тех же степенях, и поэтому снова не можем выписать аналитическое решение, однако для этой задачи проще найти численное решение.

(b): $b_1^0 \leqslant b_1 \leqslant N$ and $\mathcal{F}_2(b_1) = \tau_{comm} \cdot \alpha b_1^{-\frac{1}{4}} + \alpha \tau_1^{loc} b_1^{\frac{3}{4}} + \tau_1^{loc} \cdot \beta b_1 = \alpha \cdot b_1^{-\frac{1}{4}}(\tau_{comm} + \tau_1^{loc} b_1) + \beta \cdot \tau_1^{loc} \cdot b_1$. С предположением $\tau_1^{loc} \leqslant \tau_2^{loc} \leqslant \ldots \leqslant \tau_n^{loc}$, мы получаем

$$\tau_1^{loc} b_1 \geqslant \frac{\tau_1^{loc} N \frac{\tau_2^{loc}}{n-1}}{\tau_1^{loc} + \frac{\tau_n^{loc}}{n-1}} \geqslant \frac{\tau_1^{loc}\tau_2^{loc} N}{(n-1)(\tau_1^{loc} + \tau_n^{loc})} \geqslant \frac{\tau_1^{loc}\tau_2^{loc} N}{2(n-1)\tau_n^{loc}} \gg \tau_{comm} \frac{N}{2(n-1)} \gg \tau_{comm}. \tag{9}$$

Здесь, используя (9), можно также упростить $F_2(b_1)$ и затем $\mathcal{F}'_2(b_1)$:

$$\mathcal{F}_2(b_1) = \alpha \cdot \tau_1^{loc} \cdot b_1^{\frac{3}{4}} + \beta \tau_1^{loc} \cdot b_1, \quad \mathcal{F}'_2(b_1) = \frac{3}{4}\alpha \cdot \tau_1^{loc} \cdot b_1^{-\frac{1}{4}} + \beta \cdot \tau_1^{loc} > 0.$$

Так как производная функции положительна, то функция возрастает, а значит, минимум будет находиться в точке $b_1 = b_1^0 = \frac{N(\sum_{i=2}^{n} \frac{1}{\tau_i^{loc}})^{-1}}{\tau_1^{loc} + (\sum_{i=2}^{n} \frac{1}{\tau_i^{loc}})^{-1}}$. Таким образом, в случае малых τ_{comm} получаем следующий результат: $b_{1_{\min}} \leqslant b_1^0 = \frac{N(\sum_{i=2}^{n} \frac{1}{\tau_i^{loc}})^{-1}}{\tau_1^{loc} + (\sum_{i=2}^{n} \frac{1}{\tau_i^{loc}})^{-1}}$.

Реализация 2:

Здесь тоже можно определить: $\tau := \tau_i^{loc}\ \forall i \in 1, \ldots, n$. Затем мы можем переписать целевую функцию из (6) следующим способом:

$$\mathcal{F}(b_1) = (\max\{\tau b_1; (N - b_1)\frac{\tau}{n-1}\} + \tau_{comm}) \cdot \frac{\alpha}{\sqrt[4]{b_1}} + \tau\beta b_1.$$

Рассмотрим случай $\tau_{comm} = N^2\tau$. Можно также считать, что размер данных N большой, следовательно $\tau_{comm} \gg N\tau$. И тогда:

$$\max\{\tau b_1; (N - b_1)\frac{\tau}{n-1}\} < \tau N \ll \tau_{comm} \Rightarrow \mathcal{F} \approx \frac{\alpha\tau_{comm}}{\sqrt[4]{b_1}} + \beta\tau b_1$$

$$\mathcal{F}'(b_1) = -\frac{\alpha\tau_{comm}}{4b_1\sqrt[4]{b_1}} + \beta\tau = 0 \Rightarrow b_{1_{\min}}^{\frac{5}{4}} = \frac{\tau_{comm}\alpha}{4\beta\tau} \Rightarrow b_{1_{\min}} = (\frac{\tau_{comm}\alpha}{4\beta\tau})^{\frac{4}{5}}.$$

Предполагая, что найденное значение b_1 лежит на интервале $(0, N)$, то есть при $0 < (\frac{\tau_{comm}\alpha}{4\beta\tau})^{\frac{4}{5}} < N$, она будет точкой минимума функции $\mathcal{F}$. Тогда:

$$\mathcal{F}(b_{1_{\min}}) = (\alpha\tau_{comm})^{\frac{4}{5}} \cdot (4\beta\tau)^{\frac{1}{5}} + (\beta\tau)^{\frac{1}{5}} \cdot (\frac{\alpha\tau_{comm}}{4})^{\frac{4}{5}} = (\alpha\tau_{comm})^{\frac{4}{5}} \cdot (\beta\tau)^{\frac{1}{5}}(4^{\frac{1}{5}} + 4^{-\frac{4}{5}}).$$

В противном случае минимум будет достигнут на правой границе, так как можно утверждать, что функция возрастает, начиная с нулевого значения. Обобщая все вышесказанное на этот случай, отметим, что для очень больших значений N вторая реализация сводится к следующему условию:

$$\begin{gathered} \forall i \hookrightarrow \tau_{comm} = \mathcal{O}(N^k \tau_i^{loc}) \text{ with } k > 1, \text{ and } \forall i \neq j \hookrightarrow \tau_i^{loc} = \tau_j^{loc} = \tau, \\ \min \mathcal{F}(b_1) = \begin{cases} (\alpha\tau_{comm})^{\frac{4}{5}} \cdot (\beta\tau)^{\frac{1}{5}}(4^{\frac{1}{5}} + 4^{-\frac{4}{5}}), & 0 < (\frac{\tau_{comm}\alpha}{4\beta\tau})^{\frac{4}{5}} < N \\ \frac{\alpha\tau_{comm}}{N} + \beta\tau N, & (\frac{\tau_{comm}\alpha}{4\beta\tau})^{\frac{4}{5}} \geqslant N \end{cases}. \end{gathered} \tag{10}$$

3.5. ЧИСЛЕННОЕ РЕШЕНИЕ

Поскольку аналитическое решение найдено не для всех случаев, приведем общее численное решение нашей задачи. Для определения минимума этих функций на соответствующих полуинтервалах будем рассматривать точки, в которых производные $\mathcal{F}'_1(b_1)$ и $\mathcal{F}'_2(b_1)$ приближаются к нулю. Следует отметить, что, учитывая характер этих функций, их производные могут быть равны нулю лишь один раз на нужном полуинтервале. Поэтому, применив метод Ньютона [18] для $\mathcal{F}'_1(b_1)$ и $\mathcal{F}'_2(b_1)$, можно найти их нули. Далее необходимо сравнить значения соответствующей функции в этих точках со значением в крайней точке интервала. Одна из этих точек даст минимальное решение и, тем самым, послужит окончательным решением задачи (6) и (7).

4. ШУМ В СЕТЯХ

Перейдем к третьему предположению из Секции 1.3. Как уже говорилось выше, пусть τ_{comm} и τ_i^{loc} – случайные величины: $\exists\ \mathbb{E}[\tau_{comm}], \mathbb{E}[\tau_i^{loc}], \mathbb{D}[\tau_{comm}] < \infty, \mathbb{D}[\tau_i^{loc}] < \infty$. Здесь рассматривается случай $\delta = \frac{L}{\sqrt{b_1}}$. Аналитическое решение в этой постановке было получено в предельных случаях для малых и больших коммуникаций с учетом локального времени вычислений. Рассмотрим их по отдельности.

4.1. СЛУЧАЙ БОЛЬШОГО ВРЕМЕНИ КОММУНИКАЦИЙ

Наложим дополнительные ограничения на случайные величины τ_{comm} и τ_i^{loc}: $\forall \{\tau_{comm}^k\}_{k=1}^m, \{\tau_i^{loc,\ k}\}_{k=1}^m\ \forall k, i \hookrightarrow \tau_{comm}^k \gg \tau_i^{loc,\ k}$. Рассмотрим функцию времени работы задачи (7) в точке минимума. Ранее в 3 было получено:

$$\mathcal{F}(b_{1_{\min}}) = (\alpha \cdot \tau_{comm})^{4/5} \cdot (\beta \cdot \tau_1^{loc})^{1/5} \cdot (4^{1/5} + 4^{-4/5}), \tag{11}$$

где $\alpha = c_1 \cdot \sqrt{\frac{L}{\mu}} \cdot \log(\frac{1}{\varepsilon}), \beta = c_2 \cdot \sqrt{\frac{L}{\mu}} \cdot \log(\frac{1}{\varepsilon})$.

Для получения результата произведем следующее равенство: $X, Y-$ независимые случайные величины $\Rightarrow \mathbb{D}[XY] = \mathbb{E}[(XY - \mathbb{E}[XY])^2] = \mathbb{E}[(XY)^2] - 2\mathbb{E}^2[XY] + \mathbb{E}^2[XY] = \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}^2[X]\mathbb{E}^2[Y] = (\mathbb{D}[X]+\mathbb{E}^2[X]) \cdot (\mathbb{D}[Y]+\mathbb{E}^2[Y]) - \mathbb{E}^2[X]\mathbb{E}^2[Y] = \mathbb{D}[X]\mathbb{D}[Y] + \mathbb{D}[X]\mathbb{E}^2[Y] + \mathbb{D}[Y]\mathbb{E}^2[X]$. Применив это равенство к (11), получим искомую ошибку:

$$\begin{aligned}\mathbb{D}[\mathcal{F}(b_{1,\min})] =& [\alpha^{4/5} \cdot \beta^{1/5} \cdot (4^{1/5} + 4^{-4/5})] \cdot \{\mathbb{D}[(\tau_{comm})^{4/5}]\mathbb{D}[(\tau_1^{loc})^{1/5}] \\ &+ \mathbb{D}[(\tau_{comm})^{4/5}]\mathbb{E}^2[(\tau_1^{loc})^{1/5}] + \mathbb{D}[(\tau_1^{loc})^{1/5}]\mathbb{E}^2[(\tau_{comm})^{4/5}]\}.\end{aligned}$$

4.2. СЛУЧАЙ МАЛОГО ВРЕМЕНИ КОММУНИКАЦИЙ

Здесь будет рассматриваться шум только на коммуникациях, т.е. время коммуникаций будем представлять как случайную величину с математическим ожиданием и конечной дисперсией, а время локальных вычислений для каждого устройства - как постоянную величину. Аналогично предыдущему пункту наложим дополнительные ограничения: $\forall \{\tau_{comm}^k\}_{k=1}^m\ \forall k, i \hookrightarrow \tau_{comm}^k \ll \tau_i^{loc}$. Рассмотрим функцию времени работы задачи (7) в точке минимума (здесь возьмем точку b_1^0):

$$\begin{aligned}\mathcal{F}(b_{1,\min}) &= [(N - b_1^0) \cdot (\sum_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1} + \tau_{comm}] \cdot \alpha \cdot \frac{1}{(b_1^0)^{1/4}} + \tau_1^{loc} \cdot b_1^0 \cdot \beta \\ &= \left[\frac{\tau_1^{loc} \cdot N \cdot (\sum\limits_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1}}{\tau_1^{loc} + (\sum\limits_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1}} + \tau_{comm}\right] \cdot \alpha \cdot \frac{(\tau_1^{loc} + (\sum\limits_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1})^{1/4}}{(N \cdot (\sum\limits_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1})^{1/4}} + \tau_1^{loc} \cdot b_1^0 \cdot \beta,\end{aligned}$$

где $\alpha = c_1 \cdot \sqrt{\frac{L}{\mu}} \cdot \log(\frac{1}{\varepsilon}), \beta = c_2 \cdot \sqrt{\frac{L}{\mu}} \cdot \log(\frac{1}{\varepsilon})$.

Тогда искомая ошибка будет следующей:

$$\mathbb{D}[\mathcal{F}(b_{1,\min})] = \left[\alpha \cdot \frac{(\tau_1^{loc} + (\sum\limits_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1})^{1/4}}{(N \cdot (\sum\limits_{i=2}^n \frac{1}{\tau_i^{loc}})^{-1})^{1/4}}\right]^2 \cdot \mathbb{D}[\tau_{comm}].$$

5. ЭКСПЕРИМЕНТЫ

5.1. ЭКСПЕРИМЕНТЫ С РАСПРЕДЕЛЕНИЕМ ДАННЫХ

Для экспериментальной проверки теоретических результатов рассмотрена задача гребневой регрессии:

$$\min_\omega [\frac{1}{2N}\|X\omega - y\|_2^2 + \frac{\lambda}{2}\|\omega\|_2^2], \tag{12}$$

где ω - вектор весов модели, $\{x_i, y_i\}_{i=1}^N$ - обучающий набор данных, $\lambda > 0$ - параметр регуляризации. Рассматриваем сеть с 21 устройством, смоделированную на однопроцессорной машине. Кроме того, используется набор данных из библиотеки LIBSVM [19]. Значение $\tau_1^{loc} = 1$, значения для остальных

i $\tau_i^{loc}, i \neq 1$, взяты относительно и равномерно генерируются от 3 до 7. τ_{comm} выбраны так, что $\frac{\tau_{comm}}{\tau_1^{loc}} = 10^l, l = \overline{-6, 12}$.

Был реализован алгоритм 1 на Python 3.9.6, используя итерационный метод OGM-G из статьи [20] для нахождения $\arg\min$ в 1 (именно это рекомендуется в оригинальной статье [17]). Подсчитав необходимое количество итераций для достижения определенной точности, находим значения констант c_1, c_2 и, соответственно, α, β. С их помощью стало возможным распределение данных по устройствам в соответствии с приведенными выше формулами.

Далее был запущен запускаем алгоритм и измерено время работы на полученном распределении данных по устройствам и равномерном распределении. Наша цель - найти ускорение полученного нами распределения данных относительно равномерного разбиения.

Рассматриваются два случая различных δ: $\delta = \frac{L}{\sqrt{b_1}}$ и $\delta = \frac{L}{b_1}$. Для случая $\delta = \frac{L}{\sqrt{b_1}}$ для нахождения $b_{1,\min}$ используем следующие подходы: 1) для всех случаев времени связи используем метод Ньютона для численного нахождения решения; 2) для малых и больших связей используем также результаты раздела 3.4.2. Для случая $\delta = \frac{L}{b_1}$ для нахождения $b_{1,\min}$ мы также используем метод Ньютона и дополнительно формулу Кардано из раздела 3.4.1. Результаты приведены на рисунке 1.

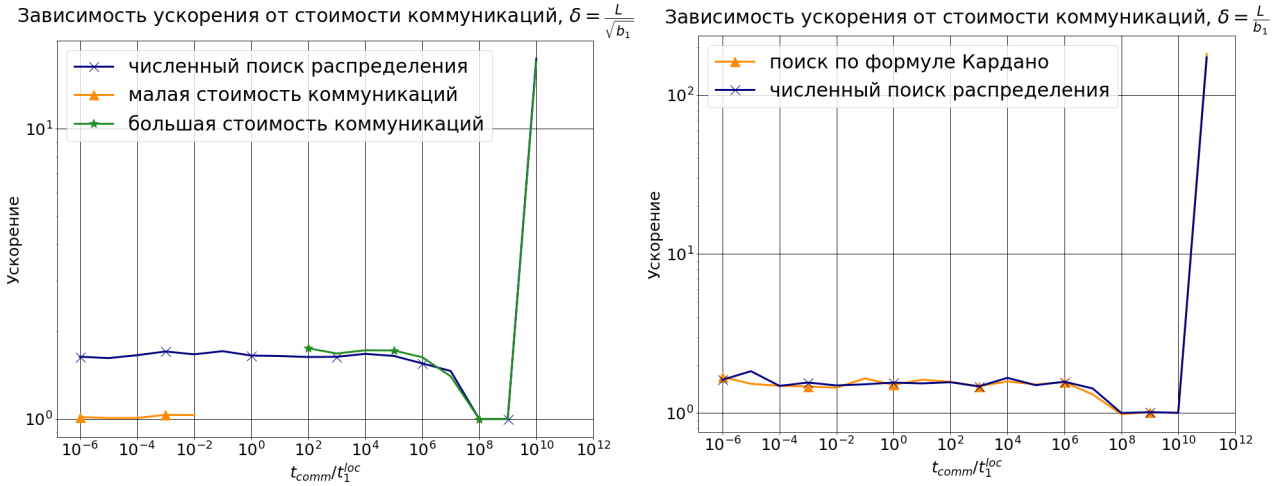

Рис. 1: Эксперименты с распределением данных

5.2. ЭКСПЕРИМЕНТЫ С ШУМОМ

Мы модифицировали моделирование алгоритма 1, добавив шум в коммуникации и мощности устройств. Шум генерировался из равномерного распределения, и его величина составляла 10, 20, 30, 50 и 100 % соответственно от абсолютного значения времени коммуникаций и мощности устройств. В новой модели шума были проведены измерения времени выполнения задачи Гребневой регрессии (Ридж-регрессии) с полученным распределением данных и с равномерным распределением, получив ускорение, дающее правильное распределение данных. При этом проводились измерения математического ожидания затрат на коммуникации и мощности устройств, а впоследствии было получено ускорение при этих ожидаемых значениях. Эксперименты проводились только в случае больших затрат на связь 4.1. На рисунке 2 приведен график отношения этих ускорений и доверительные интервалы.

Проанализируем полученные результаты. Из графика видно, что все прямые попадают в доверительные интервалы, а значит для любого значения шума, приведенные в Секции 4.1 теоретические расчеты подтверждаются экспериментом. Также отметим, что вблизи значения $t_{comm}/t_1^{loc} = 10^{10}$ шум практически перестает влиять на результаты.

6. ЗАКЛЮЧЕНИЕ

В данной работе был представлен новый метод разбиения данных для задачи распределенной оптимизации. Новое решение основано на построении функции времени работы алгоритма 1 и нахождении ее минимума. Такой метод хорошо работает в сетях с различной стоимостью связи между сервером и локальными устройствами и различной мощностью устройств. Теоретические результаты были подтверждены экспериментально. Это показывает, что данный метод дает ускорение при решении такого

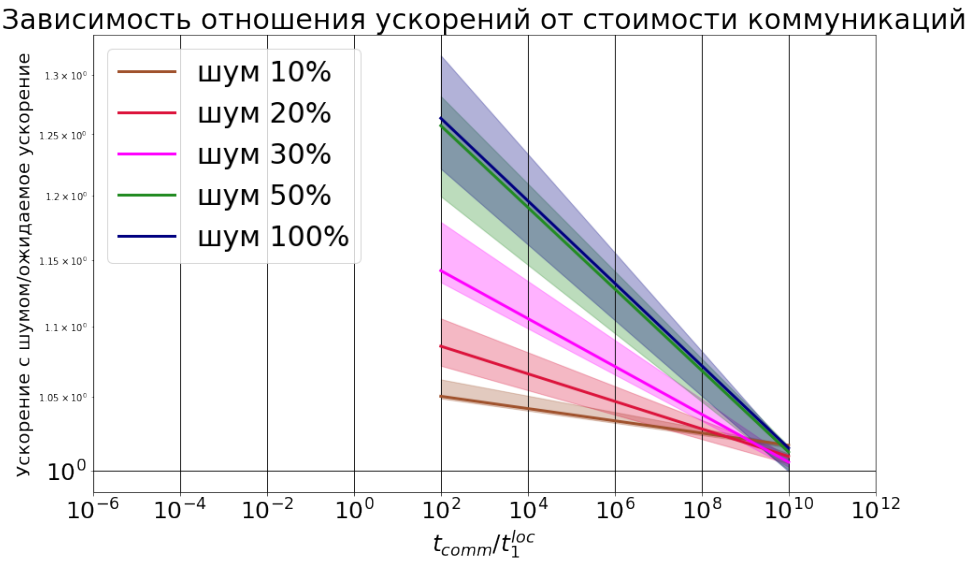

Рис. 2: Эксперименты с шумом в сети

рода задач. Кроме того, предполагая наличие шума в сетях, была найдена погрешность оптимального решения и проведены соответствующие эксперименты.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Исследования А. Безносикова были поддержаны Российским научным фондом (проект № 23-11-00229).

ИСТОЧНИКИ

[1] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen и J. S. Rellermeyer, "A survey on distributed machine learning," *Acm computing surveys (csur)*, т. 53, № 2, с. 1—33, 2020.

[2] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh и D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[3] T. Li, A. K. Sahu, A. Talwalkar и V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, т. 37, № 3, с. 50—60, 2020.

[4] P. Kairouz, H. B. McMahan, B. Avent и др., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, т. 14, № 1–2, с. 1—210, 2021.

[5] A. Ghosh, R. K. Maity, A. Mazumdar и K. Ramchandran, "Communication efficient distributed approximate Newton method," в *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, с. 2539—2544.

[6] V. Smith, S. Forte, M. Chenxin, M. Takáč, M. I. Jordan и M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *Journal of Machine Learning Research*, т. 18, с. 230, 2018.

[7] E. Gorbunov, K. P. Burlachenko, Z. Li и P. Richtárik, "MARINA: Faster non-convex distributed learning with compression," в *International Conference on Machine Learning*, PMLR, 2021, с. 3788—3798.

[8] Y. Nesterov и др., *Lectures on convex optimization*. Springer, 2018, т. 137.

[9] Y. Arjevani и O. Shamir, "Communication complexity of distributed convex learning and optimization," *Advances in neural information processing systems*, т. 28, 2015.

[10] O. Shamir, N. Srebro и T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," в *International conference on machine learning*, PMLR, 2014, с. 1000—1008.

[11] S. Matsushima, H. Yun, X. Zhang и S. Vishwanathan, "Distributed stochastic optimization of the regularized risk," *arXiv preprint arXiv:1406.4363*, 2014.

[12] Y. Tian, G. Scutari, T. Cao и A. Gasnikov, "Acceleration in distributed optimization under similarity," в *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, с. 5721—5756.

[13] Y. Sun, G. Scutari и A. Daneshmand, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM Journal on Optimization*, т. 32, № 2, с. 354—385, 2022.

[14] S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós и A. Smola, "Aide: Fast and communication efficient distributed optimization," *arXiv preprint arXiv:1608.06879*, 2016.

[15] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach и L. Massoulie, "Statistically preconditioned accelerated gradient method for distributed optimization," в *International conference on machine learning*, PMLR, 2020, с. 4203—4227.

[16] A. Beznosikov, G. Scutari, A. Rogozin и A. Gasnikov, "Distributed saddle-point problems under data similarity," *Advances in Neural Information Processing Systems*, т. 34, с. 8172—8184, 2021.

[17] D. Kovalev, A. Beznosikov, E. Borodich, A. Gasnikov и G. Scutari, "Optimal gradient sliding and its application to optimal distributed optimization under similarity," *Advances in Neural Information Processing Systems*, т. 35, с. 33 494—33 507, 2022.

[18] B. T. Polyak, "Newton's method and its use in optimization," *European Journal of Operational Research*, т. 181, № 3, с. 1086—1096, 2007.

[19] C.-C. Chang и C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, т. 2, № 3, с. 1—27, 2011.

[20] D. Kim и J. A. Fessler, "Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions," *Journal of optimization theory and applications*, т. 188, № 1, с. 192—219, 2021.