Распределенная оптимизация в условиях Поляка-Лоясиевича*

***И. О. Автор*[1], *И. О. Соавтор*[2], *И. О. Фамилия*[1,2]**
author@site.ru; co-author@site.ru; co-author@site.ru
[1]Организация, адрес; [2]Организация, адрес

В статье рассматривается новый метод децентрализованного распределенного решения больших систем нелинейных уравнений в условиях Поляка-Лоясиевича. Суть метода состоит в постановке эквивалентной задачи распределенной оптимизации и последующем ее сведении сперва к задаче ограниченной оптимизации, а затем к задаче композитной оптимизации, но уже без ограничений. Предложенный метод сравнивается с градиентным спуском, ускоренным градиентным спуском, а также с последовательным и параллельным алгоритмом обратного распространения ошибки при обучении многослойной нейронной сети с нелинейной функцией активации нейрона.

Ключевые слова: *большие нелинейные системы; распределенная оптимизация; условия Поляка-Лоясиевича; многослойные нейронные сети*

DOI: 10.21469/22233792

1 Введение

За последние несколько лет наблюдается скачок популярности задач, связанных с анализом больших данных. Это вызванно увеличением количества параметров моделей как в машинном обучении, так и в других областях прикладной математики. Примерами таких задач могут послужить задачи обработки непрерывно поступающих данных с измерительных устройств, аудио и видеоматериалов; изучения потоков сообщений в социальных сетях или метеорологических данных; анализа данных о местонахождении абонентов сетей и оптимальное распределение мощности между вышками сотовой связи. В связи с этим возникла потребность решать эти задачи распределенно. Суть многих из этих задач заключается в решении огрромных сислем нелинейных уравнений.

В этой статье мы раасматриваем распределенный способ решения системы нелинейных уравнений:

$$\begin{cases} f_1(x_1, \dots, x_n) = 0 \\ f_2(x_1, \dots, x_n) = 0 \\ \dots \\ f_m(x_1, \dots, x_n) = 0 \end{cases} \qquad f_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \dots, m \tag{1}$$

Эту систему можно переписать в виде эквивалентной задачи оптимизации:

$$g(x) := \sum_{i=1}^{m} f_i^2(x) \to \min_{x \in \mathbb{R}^n} \tag{2}$$

В [1] эта задача представляется в децентрализованном виде:

$$\begin{aligned} \min_{x_i \in \mathbb{R}^n} \quad & \sum_{i=1}^{m} f_i^2(x_i) \\ \text{s.t.} \quad & x_1 = x_2 = \ldots = x_m \end{aligned} \tag{3}$$

*Работа выполнена при финансовой поддержке РФФИ, проекты № №00-00-00000 и 00-00-00001.

Таким образом, возникает задача ограниченной оптимизации, которую авторы статьи решают методом прямодвойственного градиентного спуска. Причем, если функция g удовлетворяет условиям Поляка-Лоясиевича с константой $\nu > 0$, то есть:

$$\frac{1}{2}||\nabla g(x)||^2 \geqslant \nu(g(x) - g^*),\ \forall x \in \mathbb{R}^n;\ g^* = \min_{x\in\mathbb{R}^n} g(x) \tag{4}$$

то метод будет иметь линейную скорость сходимости.

Мы, в свою очередь, сводим задачу ограниченной оптимизации к задаче композитной оптимизации, смягчив жесткие условия на совпадение $x_{i i=1}^{m}$ в задаче (3). И предлагаем решать полученную задачу аналогами метода подобных треугольников или слайдинга [4].

Наш метод сравнивается с методами градиентного спуска и ускоренного градиентного спуска, описанными в [2] и [5]. Также мы сравнили наш метод с последовательным и параллельным вариантами самого распространенного на данный момент алгоритма обратного распространения ошибки для обучения нейронных сетей с нелинейной функцией активации нейрона, предложенными в [3]. Сравнение производится в ходе вычислительного эксперимента при обучении нейронных сетей с различным количеством слоев и функцией активации нейрона – сигмоидой. Обучение производится на классических данных (CIFAR, MNIST, IMAGNET).

2 Постановка задачи

Итак, перед нами стоит задача решения большой системы нелинейных уравнений (1), которую мы переписали в виде задачи децентрализованной оптимизации (3). Предполагается, что каждая из функций $f_i(x_i)$ будет минимизироваться на отдельном процессоре. А связи между этими процессорами будут обеспечивать равенство решений $\{x_i\}_{i=1}^m$. Связи между процессорами можно представить в виде взвешенного неориентированного графа G. И пусть L – это матрица Кирхгофа этого графа.

Теперь введем обозначения:

1) $X = (x_1, \ldots, x_m)$ – матрица, стобцы которой есть вектора аргументов функций $\{f_i\}_{i=1}^m$

2) $F(X) = \sum\limits_{i=1}^{m} f_i^2(x_i)$ – целевая функция задачи (3)

3) $W = L \otimes I_n$, где $\otimes$ означает произведение Кронекера

В [1] так же показано, что задача (3) эквивалентна следующей задаче условной оптимизации:

$$\begin{aligned} \min_{X\in\mathbb{R}^{m\times n}} \quad & F(X) \\ \text{s.t.} \quad & W^{1/2}X = 0 \end{aligned} \tag{5}$$

Мы же предлагаем убрать жесткие условия и свести задачу (5) к задаче композитной оптимизации:

$$\min_{X\in\mathbb{R}^{m\times n}} \qquad F(X) + R||W^{1/2}X|| \tag{6}$$

Здесь R – это некоторая правильно подобранная положительная константа.

Алгоритм решения поставленной задачи основан на методе подобных треугольников, описанном в [4].

3 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

[1] Hamed Karimi, Julie Nutini, and Mark W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *CoRR*, abs/1608.04636, 2016.

[2] Hamed Karimi, Julie Nutini, and Mark W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *CoRR*, abs/1608.04636, 2016.

[3] G. Sandhya Prafulla. Speaker independent vowel recognition using backpropagation neural network on master-slave architecture jv.s. srinivas,, October 02 2013.

[4] А. В. Гасников. Универсальный метод для задач стохастической композитной оптимизации. 2016.

[5] А. В. Гасников. Современные численные методы оптимизации, метод универсального градиентного спуска. 2018.

Поступила в редакцию 01.01.2017