

Распределенные методы второго порядка с быстрой скоростью сходимости и компрессией

Исламов Рустем Ильфакович

Московский физико-технический институт
Кафедра Интеллектуальных систем

Научный руководитель: д.ф.-м.н. Стрижов В.В.
Консультант: Ph.D. (к.ф.-м.н) П. Рихтарик

Июнь, 2021

Постановка задачи

Оптимизационная задача

Определить оптимальные параметры модели машинного обучения путем решения оптимизационной задачи:

$$\min_{x \in \mathbb{R}^d} \left\{ P(x) := f(x) + \frac{\lambda}{2} \|x\|^2 \right\}, \quad (1)$$

где x — параметры модели, а f — функция потерь.

Предполагается, что данные для обучения распределены между n клиентами, каждый клиент $i \in \{1, \dots, n\}$ имеет доступ к m векторам признаков объектов $a_{ij} \in \mathbb{R}^d$, $j \in \{1, \dots, m\}$. Функция f имеет вид

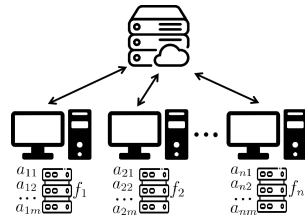
$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x), \quad f_i(x) = \frac{1}{m} \sum_{j=1}^m f_{ij}(x), \quad f_{ij}(x) = \varphi_{ij}(a_{ij}^\top x). \quad (2)$$

- [1] Konstantin Mishchenko, Eduard Gorbunov, Martin Takac, and Peter Richtarik.
Distributed learning with compressed gradient differences.
arXiv preprint arXiv:1901.09269, 2019.
- [2] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtarik.
Acceleration for compressed gradient descent in distributed and federated optimization.
In International Conference on Machine Learning, 2020.
- [3] Rixon Crane and Fred Roosta.
DINGO: Distributed Newton-type method for gradient-norm optimization.
Advances in Neural Information Processing Systems, volume 32, pages 9498, 2019.
- [4] Rustem Islamov, Xun Qian, and Peter Richtarik.
Distributed Second Order Methods with Fast Rates and Compressed Communication.
Proceedings of the 38th International Conference on Machine Learning, 2021.

Модель распределенной оптимизации

Достоинства и недостатки модели

- + Возможно обучать модели на больших объемах данных, распределенных между устройствами;
- + Возможно параллелизовать вычисления на устройствах;
- Скорость обмена данными между клиентом и сервером намного медленнее, чем скорость вычислений на самих устройствах и сервере.



Архитектура модели
«клиент-сервер».

Существующие подходы и их недостатки

- Скорость сходимости методов первого порядка зависит от числа обусловленности поставленной оптимизационной задачи;
- Скорость сходимости методов второго порядка зависит от числа обусловленности поставленной оптимизационной задачи;
- Стоимость коммуникации между сервером и клиентом для методов второго порядка очень дорогая.

Цель

Предложить эффективный с точки зрения коммуникации метод второго порядка, чья скорость сходимости не зависит от числа обусловленности.

Предположения и структура Гессианов

Предположения

Поставленная оптимизационная задача имеет хотя бы одно решение x^* . Для всех i, j функция потерь $\varphi_{ij} : \mathbb{R} \rightarrow \mathbb{R}$ является дважды непрерывно дифференцируемой функцией с ν -липшецевой второй производной.

Гессианы функций

Гессианы функций f_{ij}, f_i, f соответственно имеют вид

$$\mathbf{H}_{ij}(x) = \varphi''(a_{ij}^\top x) a_{ij} a_{ij}^\top, \quad \mathbf{H}_i(x) = \frac{1}{m} \sum_{j=1}^m \mathbf{H}_{ij}(x), \quad \mathbf{H}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i(x). \quad (3)$$

Основная идея: NEWTON-STAR

NEWTON-STAR

Предположим, что Серверу известен Гессиан $\mathbf{H}(x^*)$ функции f в оптимуме. Шаг метода NEWTON-STAR имеет вид:

$$x^{k+1} = x^k - (\nabla^2 P(x^*))^{-1} \nabla P(x^k) = x^k - (\mathbf{H}(x^*) + \lambda \mathbf{I})^{-1} \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k) + \lambda x^k \right). \quad (4)$$

Теорема 1 (Исламов, 2021)

Предположим, что $\mathbf{H}(x^*) \succeq \mu^* \mathbf{I}$, $\mu^* \geq 0$, причем $\mu^* + \lambda > 0$. Тогда NEWTON-STAR сходится локально квадратично

$$\|x^{k+1} - x^*\| \leq \frac{\nu}{2(\mu^* + \lambda)} \left(\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|a_{ij}\|^3 \right) \|x^k - x^*\|^2. \quad (5)$$

Достоинства и недостатки NEWTON-STAR

- Локальная квадратичная сходимость, наследованная от стандартного метода Ньютона. Скорость сходимости не зависит от числа обусловленности;
- Стоимость коммуникаций между сервером и клиентом $\mathcal{O}(d)$ — такая же, как и у градиентных методов. Каждый клиент пересылает серверу только градиент $\nabla f_i(x^k)$;
- Метод имеет только теоретическую значимость, Гессиан в оптимуме не известен.

Дополнительные предположения

Каждая функция φ_{ij} является выпуклой, параметр регуляризации λ положительный.

Основная идея метода

Аппроксимируем матрицу $\mathbf{H}(x^*)$ на шаге k матрицей \mathbf{H}^k вида

$$\mathbf{H}^k = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{m} \sum_{j=1}^m h_{ij}^k a_{ij} a_{ij}^\top \right), \quad x^{k+1} = x^k - \left(\mathbf{H}^k + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k) + \lambda x^k \right). \quad (6)$$

Требования:

- $h_{ij}^k \rightarrow \varphi_{ij}''(a_{ij}^\top x^*)$ при $k \rightarrow \infty$;
- обновление элементов вектора $h_i^k := (h_{i1}^k, \dots, h_{im}^k)^\top$ должно быть слабым, т.е. вектор $h_i^{k+1} - h_i^k$ разрежен.

Оператор несмещенной компрессии

Определение

Рандомизированное отображение $\mathcal{C} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ называется **оператором несмещенной компрессии**, если оно удовлетворяет условиям:

$$\mathbb{E}[\mathcal{C}(h)] = h, \quad \mathbb{E}[\|\mathcal{C}(h)\|^2] \leq (\omega + 1) \|h\|^2, \quad \forall h \in \mathbb{R}^m. \quad (7)$$

Оператор Rand- r

Оператор Rand- r , определенный по формуле $\mathcal{C}(h) := \frac{m}{r} \cdot \xi \circ h$, где ξ — случайный вектор из равномерного распределения над множеством $\{h \in \{0, 1\}^m : \|h\|_0 = r\}$, а \circ определяет поэлементное умножение. Данный оператор имеет параметр $\omega = \frac{m}{r} - 1$.

$$h = \begin{pmatrix} 1 \\ -15 \\ -7 \\ 10 \end{pmatrix}$$

$$\mathcal{C}(h) = \frac{4}{2} \begin{pmatrix} 1 \\ 0 \\ -7 \\ 0 \end{pmatrix}$$

Пример: Rand-2.

Механизм обновления коэффициентов

Введем вектор вторых производных функций потерь:

$$h_i(x) := \left(\varphi''_{i1}(a_{i1}^\top x), \dots, \varphi''_{im}(a_{im}^\top x) \right)^\top.$$

Механизм обновления (DIANA-trick [1])

Вектор $h_i^k = (h_{i1}^k, \dots, h_{im}^k)^\top$ обновляется согласно формуле:

$$h_i^{k+1} = \left[h_i^k + \eta \mathcal{C}_i^k(h_i(x^k) - h_i^k) \right]_+, \quad (8)$$

где \mathcal{C}_i^k — оператор несмещенной компрессии, а η — числовой коэффициент.

Algorithm 1 Шаг метода NEWTON-LEARN

- 1: **каждый клиент** $i \in \{1, \dots, n\}$:
 - 2: вычисляет $\nabla f_i(x^k)$
 - 3: обновляет $h_i^{k+1} = [h_i^k + \eta \mathcal{C}_i^k(h_i(x^k) - h_i^k)]_+$
 - 4: отправляет $\nabla f_i(x^k)$, $h_i^{k+1} - h_i^k$ и $\{a_{ij} : h_{ij}^{k+1} - h_{ij}^k \neq 0\}$ на сервер
 - 5: **сервер:**
 - 6: $x^{k+1} = x^k - (\mathbf{H}^k + \mathbf{I})^{-1} \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k) + \lambda x^k \right)$
 - 7: $\mathbf{H}^{k+1} = \mathbf{H}^k + \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m (h_{ij}^{k+1} - h_{ij}^k) a_{ij} a_{ij}^\top$
-

Сходимость NEWTON-LEARN

Введем функцию Ляпунова $\Phi_1^k := \|x^k - x^*\|^2 + \frac{1}{3mn\eta\nu^2 R^2} \sum_{i=1}^n \|h_i^k - h_i(x^*)\|^2$, где $R = \max_{i,j} \|a_{ij}\|$.

Теорема 2 (Исламов, 2021)

Пусть $\eta \leq \frac{1}{\omega+1}$ и $\|x^k - x^*\|^2 \leq \frac{\lambda^2}{12\nu^2 R^6}$ для всех $k \geq 0$. Тогда выполнено

$$\mathbb{E} [\Phi_1^k] \leq \theta_1^k \Phi_1^0, \quad \mathbb{E} \left[\frac{\|x^{k+1} - x^*\|^2}{\|x^k - x^*\|^2} \right] \leq \theta_1^k \left(6\eta + \frac{1}{2} \right) \frac{\nu^2 R^6}{\lambda^2} \Phi_1^0,$$

где $\theta_1 = 1 - \min \left\{ \frac{\eta}{2}, \frac{5}{8} \right\}$.

Лемма: при использовании оператора Rand- r достаточно предположить, что $\|x^0 - x^*\|^2 \leq \frac{\lambda^2}{12\nu^2 R^6}$, для следующих x^k неравенство будет выполнено автоматически.

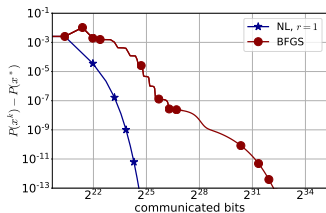
Достоинства и недостатки NEWTON-LEARN

- Локальная линейная и сверхлинейная сходимость, независимые от числа обусловленности;
- Стоимость коммуникаций между сервером и клиентом есть $\mathcal{O}(d)$, если $r = \|h_i^{k+1} - h_i^k\|_0 = \mathcal{O}(1)$ — такая же, как и у градиентных методов;
- Из сходимости функции Ляпунова следует, что алгоритм итеративно приближает Гессиан в оптимуме.

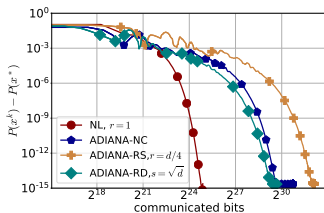
Эксперименты

Эксперименты проведены для логистической регрессии на различных наборах данных библиотеке LibSVM.

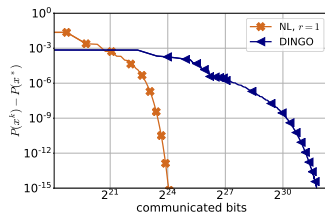
$$P(x) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \log \left(1 + \exp(-b_{ij} a_{ij}^\top x) \right) + \frac{\lambda}{2} \|x\|^2, \quad a_{ij} \in \mathbb{R}^d, b_{ij} \in \{-1, 1\}. \quad (9)$$



w8a, $\lambda = 10^{-3}$



a9a, $\lambda = 10^{-4}$



phishing, $\lambda = 10^{-5}$

Сравнение NL с BFGS, ADIANA и DINGO в терминах сложности коммуникаций.

Результаты, выносимые на защиту

Полученные результаты

- Экспериментальное и теоретическое подтверждение сходимости предложенного метода;
- Экспериментальные данные показывают превосходство предложенного метода над существующими SOTA методами в терминах сложности коммуникаций;
- Придуман первый метод второго порядка в дистрибутивной оптимизации, скорость сходимости которого не зависит от числа обусловленности функции.

Дальнейшие исследования

- Обобщение теории для произвольных дважды непрерывно дифференцируемых функций;
- Разработка алгоритма, которому не требуется раскрывать локальные данные клиента; покрывающего случаи частичного участия клиентов, двухсторонней компрессии.

- [1] Rustem Islamov, Xun Qian, and Peter Richtarik
Distributed Second Order Methods with Fast Rates and Compressed Communication.
Proceedings of the 38th International Conference on Machine Learning, 2021.
- [2] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtarik
FedNL: Making Newton-Type Methods Applicable to Federated Learning
arXiv preprint arXiv: 2106.02969, 2021.