

Дистрибутивные методы второго порядка с компрессией и агрегацией Бернулли

Исламов Рустем Ильфакович

Московский физико-технический институт
Кафедра Интеллектуальных систем

Научный руководитель: д.ф.-м.н. Стрижов В.В.

21.06.2023, Москва

Оптимизационная задача

Определить оптимальные параметры модели путем решения оптимизационной задачи:

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

где x — параметры модели, n — число участвующих клиентов, f_i — локальная функция потерь i -го клиента.

Вводятся предположения на функции

- 1 функция f является μ -сильно выпуклой:

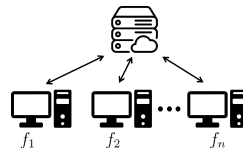
$$\nabla^2 f(x) \succeq \mu I.$$

- 2 каждая функция f_i имеет Липшицев гессиан:

$$\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq L\|x - y\|.$$

Достоинства и недостатки модели

- + Возможно обучать модели на бóльших 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.

Для уменьшения количества передаваемой информации на практике прибегают к ее компрессии.

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.
- Для разных видов компрессии необходимо доказывать сходимость отдельно.

Поставленные цели:

- Предложить и охарактеризовать более широкий класс операторов компрессии \mathcal{S} , покрывающий существующие;
- Предложить эффективный с точки зрения коммуникации метод второго порядка, чья локальная скорость сходимости не зависит от числа обусловленности, а также работающий для любого компрессора из \mathcal{S} ;
- Предложить способы улучшения вычислительной сложности метода.

Пусть серверу известен гессиан $\nabla^2 f(x^*)$ функции f в оптимуме. Шаг метода Newton Star имеет вид:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^*) \right)^{-1} \nabla f(x^k).$$

Теорема о сходимости Newton Star

Newton Star сходится локально квадратично

$$\|x^{k+1} - x^*\| \leq \frac{L}{2\mu} \|x^k - x^*\|^2. \quad (2)$$

Данный метод имеет следующие характеристики:

- + Стоимость коммуникаций одного шага метода $\mathcal{O}(d)$;
- + Локально квадратичная сходимость (такая же, как у классического метода Ньютона);
- Не может быть использован на практике, т.к. гессиан в оптимуме неизвестен.

Основная идея предложенного подхода

Аппроксимируем гессиан $\nabla^2 f_i(x^*)$ на шаге k матрицей \mathbf{H}_i^k и выполняем шаг типа Ньютон

$$x^{k+1} = x^k - \left(\frac{1}{n} \sum_{i=1}^n \mathbf{H}_i^k \right)^{-1} \nabla f(x^k).$$

Требования:

- $\mathbf{H}_i^k \rightarrow \nabla^2 f_i(x^*)$ при $k \rightarrow \infty$;
- обновление матрицы \mathbf{H}_i^k должно быть эффективным с точки зрения коммуникаций: $\mathbf{H}_i^{k+1} - \mathbf{H}_i^k$ должно быть сжатым.

Определение (класса сжимающих компрессоров)

Рандомизированное отображение $\mathcal{C} : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d}$, удовлетворяющее условию

$$\mathbb{E} \left[\|\mathcal{C}(\mathbf{X}) - \mathbf{X}\|^2 \right] \leq (1 - \alpha) \|\mathbf{X}\|^2, \quad \alpha \in (0, 1], \quad (3)$$

называется **оператором сжатия**.

Определение (класса 3PC-компрессоров)

Рандомизированное отображение $\mathcal{C}_{\mathbf{H}, \mathbf{Y}} : \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d}$, удовлетворяющее условию

$$\mathbb{E} \left[\|\mathcal{C}_{\mathbf{H}, \mathbf{Y}}(\mathbf{X}) - \mathbf{X}\|^2 \right] \leq (1 - A) \|\mathbf{H} - \mathbf{Y}\|^2 + B \|\mathbf{X} - \mathbf{Y}\|^2, \quad (4)$$

называется **оператором 3PC-компрессии**.

Классическим примером сжимающего оператора является Top-K:

$$\begin{pmatrix} 1.9 & -2 \\ -0.2 & 1 \end{pmatrix} \xrightarrow{\text{Top-1}} \begin{pmatrix} 0 & -2 \\ 0 & 0 \end{pmatrix},$$

который удовлетворяет (3) с $\alpha = \frac{K}{d^2} - 1$.

На основе произвольного оператора сжатия можно построить виды ЗРС-компрессоров. Например, шаг алгоритма Error Feedback (EF21)² является ЗРС-компрессором.

Лемма (Исламов, 2023)

ЗРС-компрессор EF21, заданный формулой $C_H(\mathbf{X}) := \mathbf{H} + C(\mathbf{X} - \mathbf{H})$, где C — произвольный оператор сжатия, является ЗРС-компрессором с параметрами $\mathbf{Y} = \mathbf{X}$, $A = \alpha$, $B = 0$.

¹Safaryan et. al, FedNL: Making Newton-Type Methods Applicable to Federated Learning, ICML 2022

²Richtárik et. al, EF21: A new, simpler, theoretically better, and practically faster error feedback, NeurIPS 2021

Лемма (Исламов, 2023)

ЗРС-компрессор СВAG, заданный формулой

$$C_H(\mathbf{X}) := \begin{cases} \mathbf{H} + C(\mathbf{X} - \mathbf{H}) & \text{with prob. } p \\ \mathbf{H} & \text{otherwise.} \end{cases},$$

где C — произвольный оператор сжатия, является ЗРС-компрессором с параметрами $A = 1 - (1 - p\alpha)(1 + s)$, $B = (1 - p\alpha)(1 + s^{-1})$, где $s = \frac{p\alpha}{2(1-p\alpha)}$ и \mathbf{Y} — произвольная матрица.

Лемма (Исламов, 2023)

ЗРС-компрессор CLAG, заданный формулой

$$C_{H,Y}(\mathbf{X}) := \begin{cases} \mathbf{H} + C(\mathbf{X} - \mathbf{H}) & \text{if } \|\mathbf{X} - \mathbf{H}\|^2 > \zeta \|\mathbf{X} - \mathbf{Y}\|^2 \\ \mathbf{H} & \text{otherwise.} \end{cases},$$

где C — произвольный оператор сжатия, является ЗРС-компрессором с параметрами $A = (1 - \alpha)(1 + s)$, $B = \max\{(1 - \alpha)(1 + s^{-1}), \zeta\}$, где $s = \frac{\alpha}{2(1-\alpha)}$.

Algorithm Newton-3PC

- 1: **Input:** $x^0 \in \mathbb{R}^d, \mathbf{H}_1^0, \dots, \mathbf{H}_n^0 \in \mathbb{R}^{d \times d}, \mathbf{H}^0 = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i^0, l^0 = \frac{1}{n} \sum_{i=1}^n \|\mathbf{H}_i^0 - \nabla^2 f_i(x^0)\|_F^2$
- 2: **on server**
- 3: *Option 1:* $x^{k+1} = x^k - [\mathbf{H}^k]_{\mu}^{-1} \nabla f(x^k)$
- 4: *Option 2:* $x^{k+1} = x^k - [\mathbf{H}^k + l^k \mathbf{I}]^{-1} \nabla f(x^k)$
- 5: **for** each device $i = 1, \dots, n$ **in parallel do**
- 6: compute local gradient $\nabla f_i(x^{k+1})$ and local Hessian $\nabla^2 f_i(x^{k+1})$
- 7: apply 3PC compressor and update $\mathbf{H}_i^{k+1} = \mathcal{C}_{\mathbf{H}_i^k, \nabla^2 f_i(x^k)}(\nabla^2 f_i(x^{k+1}))$
- 8: send $\nabla f_i(x^{k+1}), \mathbf{H}_i^{k+1}$, and $l_i^{k+1} = \|\mathbf{H}_i^{k+1} - \nabla^2 f_i(x^{k+1})\|_F$ to the server
- 9: **end for**
- 10: **on server**
- 11: Aggregate $\nabla f(x^{k+1}) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^{k+1}), \mathbf{H}^k = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i^{k+1}$
- 12: $l^{k+1} = \frac{1}{n} \sum_{i=1}^n l_i^{k+1}$

Примечание: Метод аналогичен стандартному методу Ньютона, где настоящие гессианы $\nabla^2 f_i(x^k)$ заменены на их аппроксимацию \mathbf{H}_i^k , которые обновляются с использованием компрессии.

Введем функцию Ляпунова вида

$$\Phi^k := \frac{1}{n} \sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_F^2 + 6(A^{-1} + 3AB)L^2 \|x^k - x^*\|^2.$$

Теорема о сходимости Newton-3PC, Исламов 2023

Пусть f является μ -сильно выпуклой, а каждая локальная функция f_i имеет L -Липшицев гессиан. Пусть $\|x^0 - x^*\| \leq \frac{\mu}{\sqrt{2}L}$. Тогда Newton-3PC сходится со следующими скоростями:

$$\mathbb{E} [\Phi^k] \leq \left(1 - \min \left\{ \frac{A}{2}, \frac{1}{3} \right\}\right)^k \Phi^0,$$
$$\mathbb{E} \left[\frac{\|x^{k+1} - x^*\|^2}{\|x^k - x^*\|^2} \right] \leq \left(1 - \min \left\{ \frac{A}{2}, \frac{1}{3} \right\}\right)^k \left(2 + \frac{AL}{12(1 + 3AB)L^2}\right) \frac{\Phi^0}{\mu^2}.$$

Используя результаты данной теоремы, можно сделать вывод:

- Скорости сходимости не зависят от числа обусловленности функции;
- Из вида функции Ляпунова и ее сходимости следует, что $\mathbf{H}_i^k \rightarrow \nabla^2 f_i(x^*)$.

3PC-компрессор CBAG

3PC-компрессор CBAG, заданный формулой

$$C_H(\mathbf{X}) := \begin{cases} \mathbf{H} + C(\mathbf{X} - \mathbf{H}) & \text{with prob. } p \\ \mathbf{H} & \text{otherwise.} \end{cases},$$

позволяет пропускать вычисления локальных гессианов с вероятностью $1 - p$.

Sketch&Project

Оператор Sketch&Project, заданный формулой

$$C(\mathbf{X}) := \mathbf{S}(\mathbf{S}^\top \mathbf{S})^\dagger \mathbf{S}^\top \mathbf{X},$$

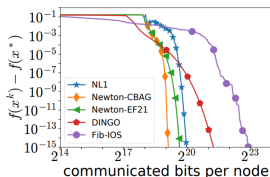
где $\mathbf{S} \in \mathbb{R}^{d \times \tau} \sim \mathcal{D}$, $\tau \ll d$, является оператором сжатия с $\alpha = \lambda_{\min}(\mathbb{E} [\mathbf{S}(\mathbf{S}^\top \mathbf{S})^\dagger \mathbf{S}^\top])$.
Данный оператор может вычислен при помощи только векторно-матричных вычислений, что позволяет снизить стоимость вычислений.

Эксперименты проводятся на логистической регрессии с ℓ_2 регуляризацией.

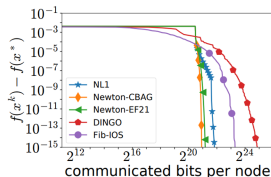
$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) + \frac{\lambda}{2} \|x\|^2 \right\} \quad f_i(x) = \frac{1}{m} \sum_{j=1}^m \log \left(1 + \exp(-b_{ij} a_{ij}^\top x) \right),$$

где $\{a_{ij}, b_{ij}\}_{j=1}^m$ — локальный датасэт i -го клиента. В экспериментах были использованы датасэты из библиотеки LibSVM.

Сравнение с другими методами второго порядка



a1a, $\lambda = 10^{-3}$



w2a, $\lambda = 10^{-4}$

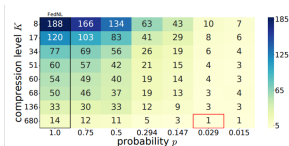
Рис.: Сравнение Newton-CBAG и Newton-EF21 с другими методами второго порядка: NL1³, DINGO⁴, Fib-IOS⁵ с точки зрения стоимости коммуникаций.

Исходя из результатов эксперимента, можно сделать вывод, что Newton-3PC превосходит другие методы второго порядка, в некоторых случаях на несколько порядков.

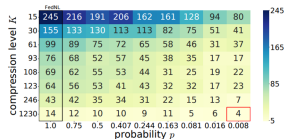
³Islamov et. al, Distributed Second Order Methods with Fast Rates and Compressed Communication, ICMML 2021

⁴Crane & Roosta. Dingo: Distributed newton-type method for gradient-norm optimization, NeurIPS 2019

⁵Fabbro et. al, A newton-type algorithm for federated learning based on incremental hessian eigenvector sharing, arXiv preprint arXiv: 2202.05800, 2022



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



a1a, $\lambda = 10^{-4}$

Рис.: Стоимость коммуникаций Newton-СВАG, основанного на Top- K . В экспериментах меняются параметры K и p . Результаты представлены в Mb.

Исходя из результатов эксперимента, можно сделать вывод, что Newton-СВАG более эффективен с точки зрения коммуникаций, если вероятность $p < 1$, что означает, что нет необходимости обновлять матрицы \mathbf{H}_i^{k+1} на каждом шаге.

Полученные результаты

- Экспериментальное и теоретическое подтверждение сходимости предложенного метода;
- Экспериментальные данные показывают превосходство предложенного метода над существующими методами в терминах сложности коммуникаций;
- Предложенный класс ЗРС-компрессоров обобщает существующие подходы, тем самым унифицирует теорию методов второго порядка;
- Предложены способы улучшения метода с точки зрения вычислительных затрат. Показана экспериментальная эффективность предложенного подхода.

Дальнейшие исследования

- Переход от предложенного метода к методам типа квази-Ньютона.
- Исследование метода с локальными шагами.

- [1] **Rustem Islamov**, Xun Qian, Slavomir Hanzely, Mher Safaryan, Peter Richtárik. *Distributed Newton-Type Methods with Communication Compression and Bernoulli Aggregation*
arXiv preprint arXiv: 2206.03588, NeurIPS workshop 2022.
- [2] Maksim Makarenko, Elnur Gasanov, **Rustem Islamov**, Abdurakhmon Sadiev and Peter Richtárik. *Adaptive Compression for Communication-Efficient Distributed Training*
arXiv preprint arXiv: 2211.00188, 2022. To appear in Transactions of Machine Learning Research.
- [3] Konstantin Mishchenko, **Rustem Islamov**, Eduard Gorbunov, Samuel Horváth. *Partially Personalized Federated Learning: Breaking the Curse of Data Heterogeneity*
arXiv preprint arXiv 2305.18285, 2023.
- [4] Sarit Khirirat, Eduard Gorbunov, Samuel Horváth, **Rustem Islamov**, Fakhri Karray, Peter Richtárik. *Clip21: Error Feedback for Gradient Clipping*
arXiv preprint: arXiv 2305.18929