Распределенные методы оптимизации и федеративное обучение

Малиновский Григорий Станиславович

"Московский Физико-Технический Институт (НИУ)"
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

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

2021г

Задача поиска стационарной точки оператора

Цель: предложить унифицированный анализ для класса локальных методов оптимизации

Задачи

- Описать класс локальных методов оптимизации в операторном виде
- Получить верхние оценки скорости сходимости для данного класса
- Исследовать поведение методов при разном числе локальных шагов

Исследуемая проблема

- Уменьшение стоимости коммуникаций в алгоритмах распределенной оптимизации и федеративного обучения

Методы решения

- Предлагается использование методов с локальными итерациями для уменьшения числа коммуникаций

Мотивация

- Передача данных (коммуникация) является преобладающей проблемой в распределенной оптимизации и федеративном обучении, а не недостаток вычислительных ресурсов локальных машин. Соотвественно, главной задачей является уменьшение числа коммуникаций и их стоимости.

Список литературы

- Chraibi S. et al. Distributed fixed point methods with compressed iterates //arXiv preprint arXiv:1912.09925. – 2019.
- Khaled A., Mishchenko K., Richtárik P. First analysis of local gd on heterogeneous data //arXiv preprint arXiv:1909.04715. – 2019.
- Khaled A., Mishchenko K., Richtárik P. Tighter theory for local SGD on identical and heterogeneous data //International Conference on Artificial Intelligence and Statistics. – PMLR, 2020. – С. 4519-4529.
- Stich S. U. Local SGD converges fast and communicates little //arXiv preprint arXiv:1805.09767. – 2018.

Задача поиска стационарной точки оператора

Пусть $\mathcal{T}_i, i = 1, \ldots, M$ операторы в $\mathbb{R}^d$. Определим усредненный оператор

$$\mathcal{T} : x \in \mathbb{R}^d \mapsto \frac{1}{M} \sum_{i=1}^{M} \mathcal{T}_i(x).$$

Задача заключается в поиске стационарной точки усредненного оператора:

$$x^\star \in \mathbb{R}^d : \mathcal{T}\left(x^\star\right) = x^\star \ .$$

Предположения об операторах:

- Оператор называется устойчиво нерасширяющимся, если для любых $x, y \in \mathbb{R}^d$ выполнено

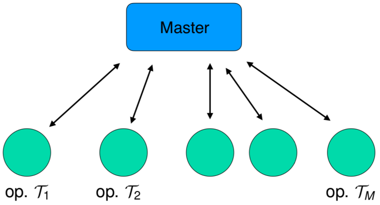
$$\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 \leq \|x - y\|^2 - \|\mathcal{T}(x) - x - \mathcal{T}(y) + y\|^2.$$

- Оператор называется ко-коэрсивным, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$(1 + \rho)\left\|\mathcal{T}(x) - \mathcal{T}(y)\right\|^2 \leq \|x - y\|^2 - \|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2$$

для $\rho > 0$.

Описание алгоритма



Локальные вычисления:

$$h_i^{k+1} := (1-\lambda)x_i^k + \lambda \mathcal{T}_i(x_i^k)$$

Усреднение на сервере:

$$\hat{x}^{k+1} := \frac{1}{M}\sum_{i=1}^{M} h_i^{k+1}$$

В детерминированном случае коммуникации происходят раз в несколько итераций. При этом количество итераций во внутреннем цикле ограничено.

$$1 \leq t_n - t_{n-1} \leq H, \text{ для всех } n \geq 1$$

В рандомизированном случае мы на каждой итерации генерируем Бернуллиевскую случайную величину, так что с вероятностью p происходит коммуникация, а с вероятностью $1-p$ продолжаются локальные вычесления.

Скорость сходимости детерминированного алгоритма

Теорема (Малиновский 2020)

Предположим, что каждый оператор $\mathcal{T}_i$ является устойчиво нерасширяющимся. Пусть $\lambda \leq \frac{1}{8\max(1,H-1)}$. Тогда $\forall T \in \mathbb{N}$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k - \mathcal{T}\left(\hat{x}^k\right)\right\|^2 \leq \frac{3\left\|\hat{x}^0 - x^\star\right\|^2}{\lambda T} + \frac{36\lambda^2(H-1)^2}{M}\sum_{i=1}^{M}\left\|x^\star - \mathcal{T}_i\left(x^\star\right)\right\|^2$$

Следствие 1 (Малиновский 2020)

Предположим, что $H \geq 2$ и $\lambda \leq \frac{1}{8}$. Тогда достаточным условием на число коммуникаций, чтобы получить ε точность является

$$\frac{T}{H-1} \geq \frac{24\left\|\hat{x}^0 - x^\star\right\|^2}{\varepsilon}\max\left\{2, \frac{3\sigma}{\sqrt{\varepsilon}}\right\},$$

где $\sigma^2 := \frac{1}{M}\sum_{i=1}^{M}\left\|x^\star - \mathcal{T}_i\left(x^\star\right)\right\|^2$

Следствие 2 (Малиновский 2020)

Пусть $T \in \mathbb{N}$ и $H \geq 1$, такие что $H \leq \frac{\sqrt{T}}{\sqrt{M}}$; положим $\lambda = \frac{1}{8}\frac{\sqrt{M}}{\sqrt{T}}$. Тогда

$$\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k - \mathcal{T}\left(\hat{x}^k\right)\right\|^2 \leq \frac{24\left\|\hat{x}^0 - x^\star\right\|^2}{\sqrt{MT}} + \frac{3M(H-1)^2\sigma^2}{8T}$$

Скорость сходимости рандомизированного алгоритма

Теорема 2 (Малиновский 2020)

Определим функцию Ляпунова для $k \in \mathbb{N}$:

$$\Psi^k := \left\|\hat{x}^k - x^\star\right\|^2 + \frac{5\lambda}{p}\frac{1}{M}\sum_{i=1}^{M}\left\|x_i^k - \hat{x}^k\right\|^2.$$

Пусть каждый оператор является ко-коэрсивным, и если $\lambda \leq \frac{p}{15}$ мы получаем для $k \in \mathbb{N}$, что

$$\mathbb{E}\Psi^k \leq \left(1 - \min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right)\right)^k \Psi^0 + \frac{150}{\min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right)p^2}\lambda^3\sigma^2$$

Следствие 3 (Малиновский 2020)

Достаточным условием на число коммуникаций, чтобы получить ε точность является

$$pT \geq p\max\left\{\frac{15(1+\rho)}{\rho p}, \frac{18\sigma(1+\rho)^{\frac{1}{3}}}{p\rho^{\frac{3}{2}}\varepsilon^{\frac{1}{2}}}, \frac{40\sigma^{\frac{2}{3}}(1+\rho)}{p\rho\varepsilon^{\frac{1}{3}}}\right\} \times \log\frac{2\Psi_0}{\varepsilon}$$

Описание эксперимента

Рассматривается функция ошибки логистической регрессии:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i a_i^\top x\right)\right) + \frac{\kappa}{2}\|x\|^2.$$

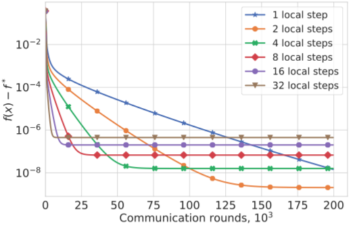
Эксперименты проведены на выборках 'a9a' и 'a4a' библиотеки LIBSVM. Все методы иплементированы на языке Python с использованием пакета MPI4PY. Эксперементы проведены на Intel(R) Xeon(R) Gold 6146 CPU с частотой 3.20GHz, с 24 ядрами.

Локальный градиентный спуск

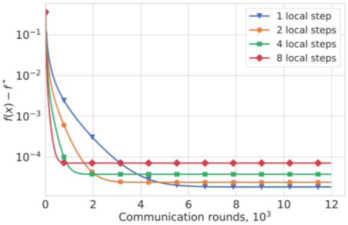
Рассматривается задача минимизации конечной суммы $f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x)$, где каждая функция f_i является выпуклой и L-гладкой. В качестве операторов мы используем $\mathcal{T}_i\left(x_i^k\right) := x_i^k - \frac{1}{L}\nabla f_i\left(x_i^k\right)$.

Локальный циклический градиентный спуск

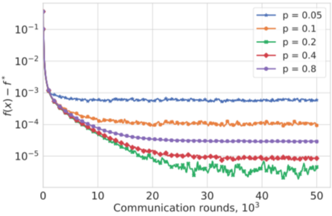
Рассматривается задача минимизации двойной конечной суммы $f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x)$, где каждая функция f_i также является конечной суммой $f_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij}$. Вместо применения полных шагов градиента, мы применяем N поэлементных шагов градиента в последовательном порядке. В качестве операторов мы используем $\mathcal{T}_i\left(x_i^k\right) := S_{i1}\left(S_{i2}\left(\dots S_{in}\left(x_i^k\right)\right)\right)$, где $S_{ij} : y \mapsto y - \frac{1}{NL}\nabla f_{ij}$.



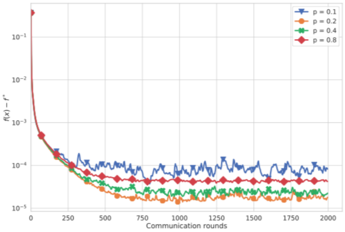
(a) Локальный градиентный спуск



(b) Локальный циклический градиентный спуск



(c) Локальный рандомизированный градиентный спуск



(d) Локальный рандомизированный циклический градиентный спуск

Выносится на защиту

- Описан класс локальных методов оптимизации в операторном виде
- Были предложены два локальных метода для поиска стационарной точки
- Были изучены теоретические свойства и получены оценки сходимости
- Эффективность методов была проверена в различных практических постановках

Публикации

- **Malinovsky G. et al.** From Local SGD to Local Fixed-Point Methods for Federated Learning //International Conference on Machine Learning. – PMLR, 2020. – C. 6692-6701.
- Condat L., **Malinovsky G.**, Richtárik P. Distributed proximal splitting algorithms with rates and acceleration //arXiv preprint arXiv:2010.00952. – 2020.
- **Malinovsky G.**, Sailanbayev A., Richtárik P. Random Reshuffling with Variance Reduction: New Analysis and Better Rates //arXiv preprint arXiv:2104.09342. – 2021.

Доклады:

- Метод усредненного тяжелого мяча, доклад, 62-я научная конференция в МФТИ, Секция анализа данных, распознавания и прогнозирования, Москва, Россия
- Определение сложности выборки с помощью универсальной аппроксимирующей модели, доклад, Математические методы распознавания образов, 19-я Всероссийская конференция с международным участием
- Random Reshuffling with Variance Reduction: New Analysis and Better Rates, онлайн выступление на KAUST Conference on Artificial Intelligence