Применение операторов сжатия в задачах распределенной оптимизации

Касюк Вадим KASIUK.VA@PHYSTECH.EDU

Editor:

Abstract

В работе предлагаются 3 новых смещенных оператора сжатия. Рассматривается их применения в задачах распределенной оптимизации. Получены теоретические гарантии сходимости методов с этими операторами в случаях одного и нескольких устройств, оценена асимптотика сходимости. Проведены численные эксперименты, показывающие превосходство предложенных операторов над стандартными: Top-k, Rand-K, для случая 1 устройства на примере классических алгоритмов машинного обучения и нейросетей.

Keywords: Compression operators, biased compressors, distributed learning, linear convergence

1 Введение

Применение методов распределенной оптимизации сейчас активно используется в задачах, которые невозможно эффективно решить на одном вычислительном устройстве из-за ограничений памяти или вычислительных мощностей. Кроме того, потребность в более совершенных способах вычислений только увеличивается с каждым годом.

Важную роль в распределенной оптимизации занимают методы, которые пытаются сократить расходы на общение между устройствами в процессе вычислений. Особый интерес представляют операторы сжатия, которые уменьшают количество передаваемой информации.

Операторы сжатия делятся на 2 группы: смещенные и несмещенные. Несмещенные давно и хорошо изученные операторы, которые на практике демонстрируют меньшую производительность, в сравнении с смешенными. Несмотря на это, лишь сравнительно недавно появилась математическая теория, описывающая вторых.

1.1 Распределенная оптимизация

Решается задача оптимизации :

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

где $x \in R^d$ - пространство признаков размерности d, n - количество устройст/узлов, $f_i(x) : R^d \to R$ - потери, которые получает модель на устройстве под номером i. Часто функция потерь имеем вид :

$$f_i(x) = E_{\xi \sim P_i}[f_\xi(x)],$$

где P_i - это распределение тренировочных данных на устройстве i. То, как данные распределяются между работниками оказывает большое влияние на процесс обучения.

В данной работе рассматривается случай централизованного типа :

1. Имеется n работников, каждый производит вычисления градиента функции $f_i(x)$(или стохастического градиента) парралельно с другими.

2. Устройства посылают посчитанный градиент на главное устройство - сервер

3. Сервер обрабатывает посчитанный градиент, делает шаг градиентного спуска и передает новую информацию каждому устройству, для выполнения новый вычислений. И затем процесс повторяется.

Большие временные потери приходятся на общение между сервером и устройствами. Именно для их уменьшения используется операторы сжатия.

1.2 Базовое решение

Базовым решением задачи (1) является обычный градиентный спуск (GD), который имеет вид :

$$x^{k+1} = x^k - \frac{\eta^k}{n} \sum_{i=1}^{n} \nabla f_i(x^k) \tag{2}$$

где, $\eta^k > 0$ - темп обучения или размер шага. Базовое решение уже подвергалось улучшению: использование ускорение(Нестеров), использование моментума(Метод тяжелого шарика)(Тут вообще общие ускоренные методы), уменьшение количества обмена информации с сервером, посредством выполения нескольких шагов локально(Надо найти источник), а также уменьшения размеров передаваемой информации за счет использования операторов сжатия(Тут можно вставить Безноса и его предшественника)(Здесь уже другие подходы к распределенным вычислениям).

1.3 Связанные работы

Ранее была проведена обширная работа, связанная со сжатием сообщений, в основном сосредоточенная на несмещенном сжатии Alistarh et al. (2017), поскольку его гораздо легче анализировать. В частности, было показано Gorbunov et al. (2019), что как классический метод с несмещенным сжатием Alistarh et al. (2017),

так и более продвинутые модификации Mishchenko et al. (2023) могут рассматриваться как специальные версии SGD. Впоследствии результаты Gorbunov et al. (2019) для сильно выпуклых задач были перенесены на выпуклые Li and Richtárik (2020) и невыпуклые целевые функции Vogels et al. (2020). В то же время работы, касающиеся смещенных сжатий, показывают более сильные резульаты.

Получены эмпирические результаты, но с ограниченным анализом или без него Wu et al. (2018). Было предпринято несколько попыток решить эту проблему, например, "проведен анализ для квадратичных чисел в распределенной среде".Был проведен анализ для momentum SGD с определенным смещенным сжатием, но с необоснованными предположениями, т.е. с ограниченной нормой градиента и памятью. Первый результат, который позволил получить линейную скорость сходимости при смещенном сжатии, был получен с помощью Shi et al. (2024), но только для одного узла и в предположении об ограниченной норме градиента, которое позже было преодолено с помощью Karimireddy et al. (2019).

Использование смещенного сжатия для распределенного обучения улучшило теоретические гарантии. Недавно был разработан новый вариант механизма обратной связи по ошибкам.Введено в Fatkhullin et al. (2021), показывающее улучшенные показатели для распределенных невыпуклых задач.

Более всего автор опирается на работу Beznosikov et al. (2024), где автор вводит 3 семейства смещенных операторов, анализирует их свойства, получает гарантии сходимости для них и в общем развивает идею : "Смещенные операторы вычислительно эффективнее несмещенных".

1.4 Нотация и определения

Мы используем $\langle x, y\rangle \coloneqq \sum_{i=1}^{d} x_i y_i$ для обозначения стандартного скалярного произведения 2-ух векторов $x, y \in \mathbb{R}^d$, где x_i обозначает i-ую компонетну x в стандартном базисе $\mathbb{R}^d$. Скалярное произведение индуцирует ℓ_2-норму в $\mathbb{R}^d$ следующим образом : $\|x\|_2 \coloneqq \sqrt{\langle x, x\rangle}$. Также ℓ_p-норма определяется как $\|x\|_p \coloneqq (\sum_{i=1}^{d} |x_i|^p)^{1/p}$ для $p \in (1, \infty)$. С помощью $\mathrm{E}[\cdot]$, мы обозначамаем математическое ожидание. Дифференцируемая функция $f : \mathbb{R}^d \to \mathbb{R}$ называется L-гладкой, если :

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2, \qquad \forall x, y \in \mathbb{R}^d.$$

Функция $f : \mathbb{R}^d \to \mathbb{R}$ μ-сильно выпуклая, если :

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2, \qquad \forall x, y \in \mathbb{R}^d.$$

2 Смещенные операторы

Под операторами сжатия мы подразумеваем (возможно рандомизированное) отображение $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ с некоторыми требованиями. Как правило, в литературе рассматриваются *несмещенные* операторы сжатия $\mathcal{C}$ с ограниченным вторым моментом, т.е. :

Definition 1 *Пусть $\zeta \geq 1$. Тогда $\mathcal{C} \in \mathbb{U}(\zeta)$ если $\mathcal{C}$ является несмещенным (то есть, $\mathrm{E}[\mathcal{C}(x)] = x$ для любых x) и втором момент ограничен* [*]

$$\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \zeta \|x\|_2^2, \qquad \forall x \in \mathbb{R}^d. \tag{3}$$

2.1 3 класса смещенных операторов

Definition 2 *Будем говорить, что $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$ для некторых $\alpha, \beta > 0$ если*

$$\alpha \|x\|_2^2 \leq \mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \beta \langle \mathrm{E}[\mathcal{C}(x)], x\rangle, \qquad \forall x \in \mathbb{R}^d. \tag{4}$$

И в статье(Безносикова вставить ссылку) доказывается, что (4) влечет за собой $\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right] \leq \beta^2 \|x\|_2^2$.

Во втором классе мы требуем, чтобы скалярное произведение между несжатым x и сжатым $\mathcal{C}(x)$ вектором мажорировало квадраты норм этих векторов Берется математическое ожидание от сжатого вектора, поскольку оператор может быть рандомизированным.

Definition 3 *Будем говорить, что $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$ для некторых $\gamma, \beta > 0$ если*

$$\max\left\{\gamma \|x\|_2^2, \tfrac{1}{\beta}\mathrm{E}\left[\|\mathcal{C}(x)\|_2^2\right]\right\} \leq \langle \mathrm{E}[\mathcal{C}(x)], x\rangle \qquad \forall x \in \mathbb{R}^d. \tag{5}$$

Наконец, в третьем классе мы требуем, чтобы ошибка сжатия $\|\mathcal{C}(x) - x\|_2^2$ была строго меньше чем 2-норма $\|x\|_2^2$ входного вектора x в среднем.

Definition 4 *Будем говорить, что $\mathcal{C} \in \mathbb{B}^3(\delta)$ для некоторого $\delta > 0$ если*

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|_2^2\right] \leq \left(1 - \frac{1}{\delta}\right)\|x\|_2^2, \qquad \forall x \in \mathbb{R}^d. \tag{6}$$

2.2 Примеры смещенных и несмещенных операторов

(a) Для $k \in [d] \coloneqq \{1, \ldots, d\}$, **несмещенный случайный (aka Rand-k) сжимающий** определяется как :

$$\mathcal{C}(x) \coloneqq \frac{d}{k}\sum_{i \in S} x_i e_i, \tag{7}$$

где $S \subseteq [d]$ подмножество индексов размера k выбранных равновероятно, и $e_1, \ldots, e_d$ стандартные базисные вектора $\mathbb{R}^d$.

Lemma 5 *Rand-k оператор (7) принадлежит $U(\frac{d}{k})$.*

[*](3) Что также может быть переписанное в виде $\mathrm{E}\left[\|\mathcal{C}(x) - x\|_2^2\right] \leq (\zeta - 1)\|x\|_2^2$.

(b) Пусть $S \subseteq [d]$ случайное множество индексов, с вектором вероятности получения элементов $p \coloneqq (p_1, \ldots, p_d)$, где $p_i \coloneqq \mathsf{P}(i \in S) > 0$ для любых i . Определим **случайный смещенный сжимающий** оператор как :

$$\mathcal{C}(x) \coloneqq \sum_{i \in S} x_i e_i. \tag{8}$$

Lemma 6 *Пусть $q \coloneqq \min_i p_i$, тогда случайный смещенный сжимающий оператор (8) принадлежит $\mathbb{B}^1(q, 1)$, $\mathbb{B}^2(q, 1)$, $\mathbb{B}^3(1/q)$.*

(c) **Адаптивный случайный смещенный оператор** определяется, как :

$$\mathcal{C}(x) \coloneqq x_i e_i \quad \text{с вероятностью} \quad \frac{|x_i|}{\|x\|_1}. \tag{9}$$

Lemma 7 *Адаптивный случайный смещенный оператор (9) принадлежит $\mathbb{B}^1(\frac{1}{d}, 1)$, $\mathbb{B}^2(\frac{1}{d}, 1)$, $\mathbb{B}^3(d)$.*

(d) **Жадный (aka Top-k) сжимающий** оператор определяется, как :

$$\mathcal{C}(x) \coloneqq \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)}, \tag{10}$$

где координаты упорядочены по их абсолютным значениям $|x_{(1)}| \leq |x_{(2)}| \leq \cdots \leq |x_{(d)}|$.

Lemma 8 *Top-k сжимающий оператор (10) принадлежит $\mathbb{B}^1(\frac{k}{d}, 1)$, $\mathbb{B}^2(\frac{k}{d}, 1)$, и $\mathbb{B}^3(\frac{d}{k})$.*

Доказательство лемм представлено в Beznosikov et al. (2024)

3 Градиентный спуск с компрессией

Теперь поставим задачу оптимизаци, которую будем решать с помощью градиентного спуска с сжатием(CGD - compressed gredient descent). Пускай решается задача:

$$\min_{x \in R^d} f(x) \tag{11}$$

, где $f : R^d \to R$, L - гладкая и μ-сильно-выпуклая, методом :

$$x^{k+1} = x^k - \eta C(\nabla f(x)) \tag{12}$$

, где $C : R^d \to R^d$ - оператор сжатия и $\eta > 0$ - темп обучения.

3.1 Новые операторы сжатия

Предлагается рассмотреть 3 новый оператора сжатия вида :

$$C(\nabla f(x)) = Top_k(\nabla f(x)|w) \tag{13}$$

, где w имеет смысл вектора "важности", а $Top_k(\circ|w)$ - выбор самых важных координат(обладающих самыми большими весами "важности" - координатами w_i). В зависимости от того, как определеяется вектор w, меняется и сам оператор.

(a) $w_i := f\left(x^k\right) - f\left(x^k - \eta\left[\nabla f(x^k)\right]_i\right)$.

В таком способе определения важности, мы смотрим на сколько мы можем уменьшить значение функции спустившить только по одной координате. Наибольшей важностью будет обладать координата, которая более других минимизирует функцию.

(b) $w := \arg\min\limits_{w\in\Delta} f\left(x^k - \eta\sum_{i=1}^d w_i\left[\nabla f(x^k)\right]_i\right)$.

Решается задача оптимизации на симплексе, где итоговый вектор w интерпретируется, как вектор направления наибольшего уменьшения функциона в среднем при семплировании из мультиномиального распределения индексов с весами w_i.

(c) $w := \arg\min\limits_{w\in[0,1]^d} f\left(x^k - \eta\sum_{i=1}^d w_i\left[\nabla f(x^k)\right]_i\right)$

Здесь вектор w "подбирает" размеры шагов относительно спуска, будучи ограниченным лишь k координатами.

3.2 Гарантии сходимости

Lemma 9 *Метод (c) лучше (b), метод (b) лучше (a)*

Пусть решается задача оптимизации (11) методом градиентного спуска с сжатием (12), где оператор сжатия имеет вид (13). Тогда если имеет место сходимость метода (b), то сходится и метод (b). Если сходится метод (b), то сходится метод (a).

Proof

С очевидностью, решение задачи (a) явлсяется решением задачи (b), которая в свою очередь есть частный случай решения задачи (c), так как имеется вложение множест, на которых решается оптимизационная подзадача. То есть имеем неравество:

$$f\left(x^k - \eta W_c\left[\nabla f(x^k)\right]\right) \le f\left(x^k - \eta W_b\left[\nabla f(x^k)\right]\right) \le f\left(x^k - \eta W_a\left[\nabla f(x^k)\right]\right)$$

Где W_a, W_b, W_c соответствующие диагональные матрицы, которые имеют вид : $W = diag(1, 1, 0, 1...)$, где на $i = j$ месте стоит 1, если данная координата вошла в $Top - k$ по важности и 0 иначе. ■

Получается, что достаточно доказать сходимость только для оператора (a). Для других операторов сходимость будет следовать автоматически

Theorem 10

Proof

■

4 Численные эксперименты

Рассмотрим работу новых операторов на примере задачи логистической регрессии классическим методом и с использованием нейросетей. Следует обратить внимание, что на каждом шаге алгоритмов (b) и (c) необходимо решать свою задачи оптимизации. Конкретно для случая (b) воспользуемся методом зеркального спуска.

На каждой итерации спуска нам нужно решать задачу:

$$\min_{w \in \Delta^d} g_k(w),$$

где

$$g_k(w) := f\left(x^k - \eta\, w^T \cdot \mathcal{P}\right),$$

$$\mathcal{P} := \mathcal{I} \odot \nabla f\left(x^k\right) = \left(\left[\nabla f\left(x^k\right)\right]_1, \left[\nabla f\left(x^k\right)\right]_2, ..., \left[\nabla f\left(x^k\right)\right]_d\right)$$

Решать ее будем методом зеркального спуска с KL-дивергенцией.
Решение:

$$w_i^{k+1} := \frac{w_i^k e^{-\eta\left[\nabla g(w^k)\right]_i}}{\sum\limits_{j=1}^{d} w_j^k e^{-\eta\left[\nabla g(w^k)\right]_j}}$$

Здесь :

$$\nabla g(w) = -\eta\, \mathcal{P} \cdot \nabla f\left(x^k - \eta\, w^T \cdot \mathcal{P}\right)$$

4.1 Случай 1 устройства

Классическая логистическая регрессия :

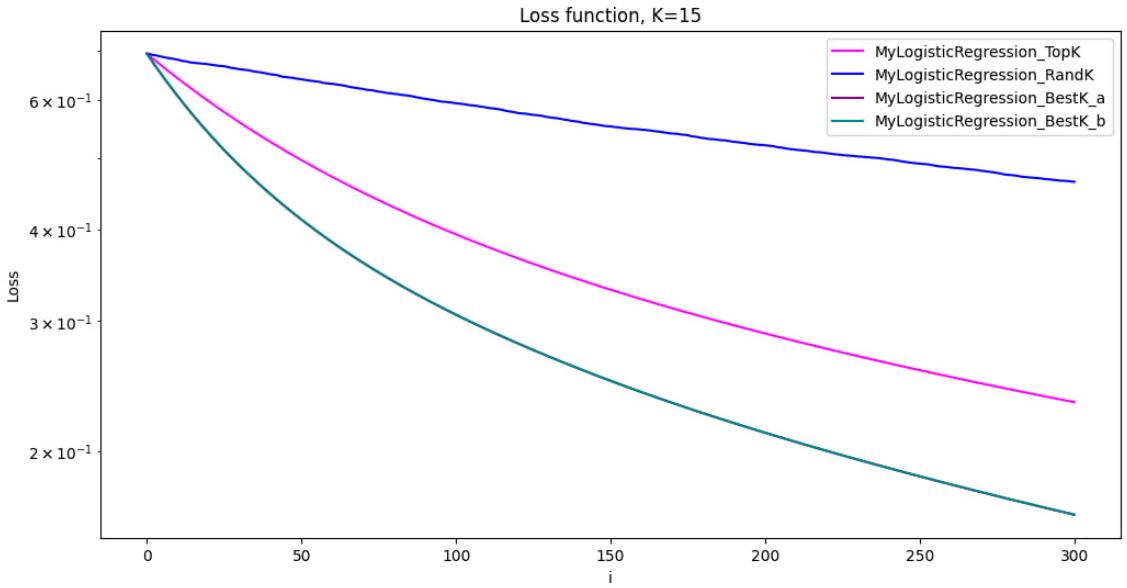

Рис. 1: Зависимость функции потерь от итерации

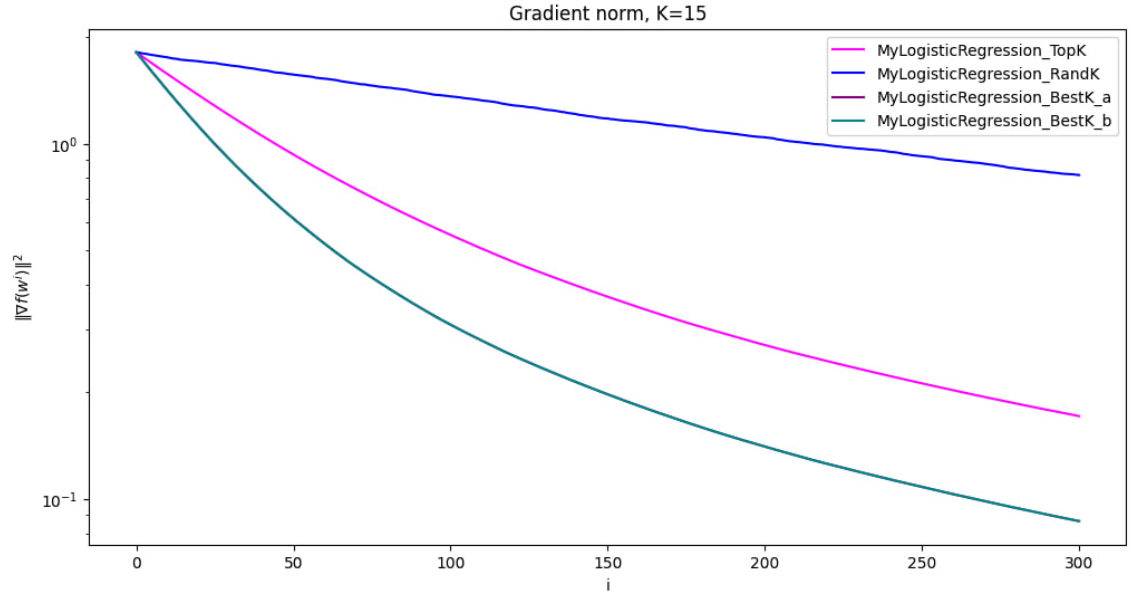

Рис. 2: Зависимость нормы градиента от итерации

Теперь будем обучать нейросеть для датасета mushrooms. Получаем следующие результаты :

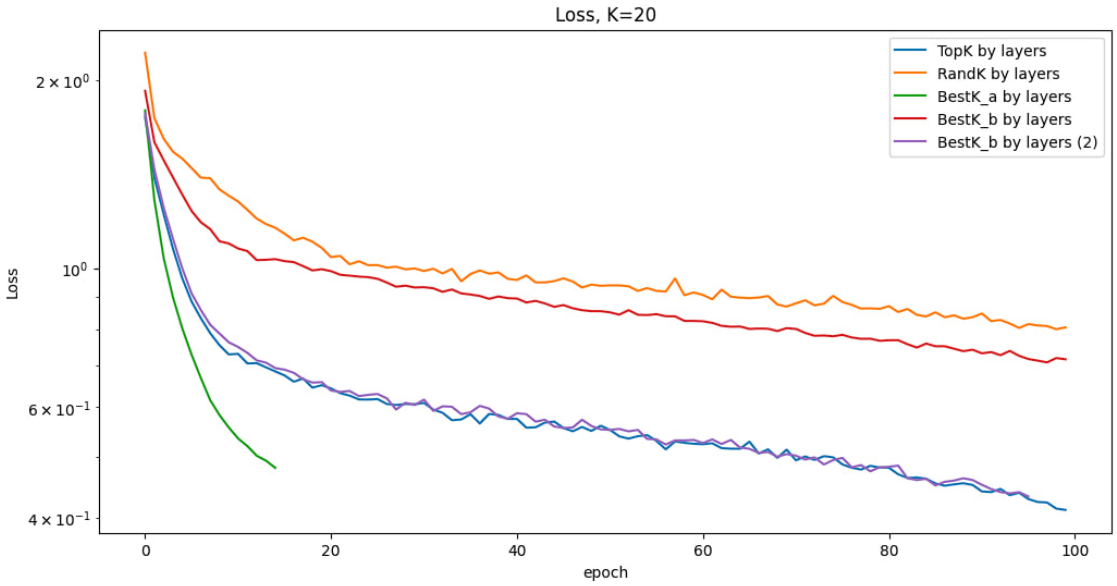

Рис. 3: Зависимость функции потерь от эпохи

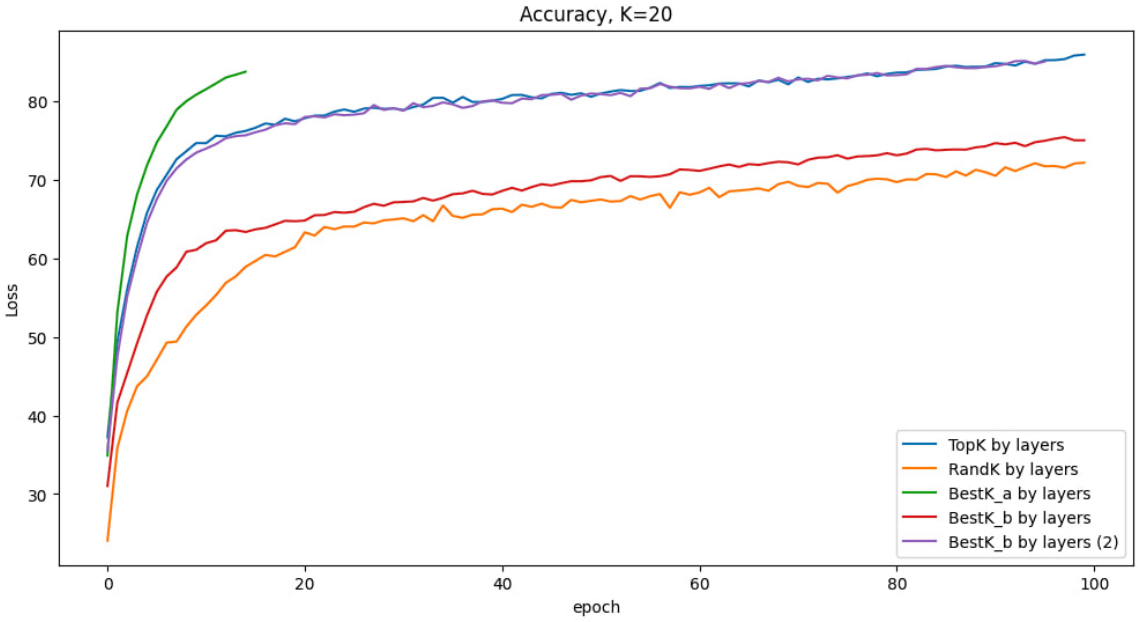

Рис. 4: Зависимость точности от эпохи

Хочу обратить внимание на реальное время для обучения :

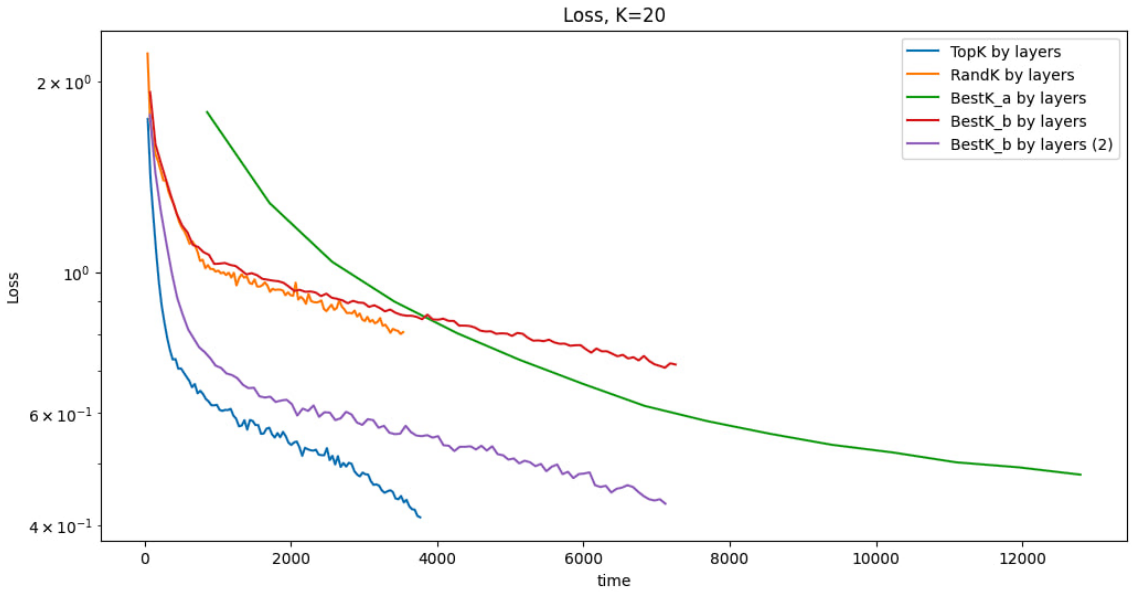

Рис. 5: Зависимость функции потерь от времени обучения

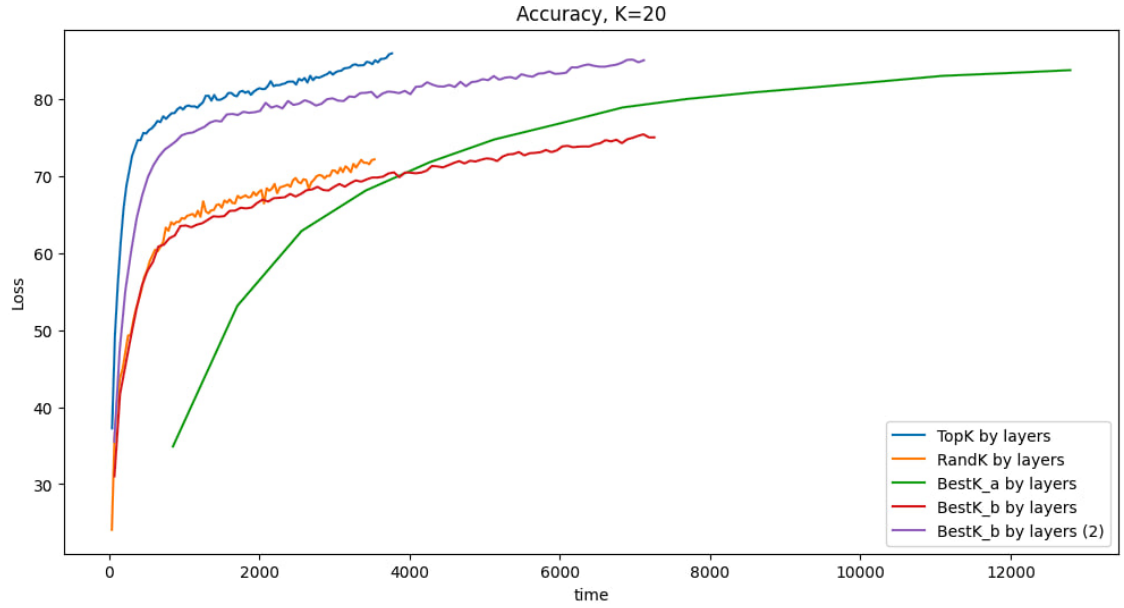

Рис. 6: Зависимость точности от времени

4.2 Анализ полученных результатов

Можно сделать следующие выводы относительно результатов эксперимента

1. Методы действительно сходятся.
2. Методы показывают скорость сходимости более высокую, чем $Rand_k$, что согласуется с теорией.

3. Имеет место более высокая скорость сходимости методов (a), (b) и (c) в некоторых задачах. Однако они обходят вычислительно дороже, чем стандартные.

5 Итог

1. Предложены и опробованы 3 новых оператора сжатия
2. Получены гарантии сходимости этих методов
3. Показано превосходство новых методов над стандартными.

Список литературы

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning, 2024. URL `https://arxiv.org/abs/2002.12410`.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. Ef21 with bells whistles: Practical algorithmic extensions of modern error feedback, 2021. URL `https://arxiv.org/abs/2110.03294`.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent, 2019. URL `https://arxiv.org/abs/1905.11261`.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes, 2019. URL `https://arxiv.org/abs/1901.09847`.

Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization, 2020. URL `https://arxiv.org/abs/2006.07013`.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences, 2023. URL `https://arxiv.org/abs/1901.09269`.

Chang-Wei Shi, Shen-Yi Zhao, Yin-Peng Xie, Hao Gao, and Wu-Jun Li. Global momentum compression for sparse communication in distributed learning, 2024. URL `https://arxiv.org/abs/1905.12948`.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization, 2020. URL https://arxiv.org/abs/1905.13727.

Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5325–5333. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/wu18d.html.