DOI: 10.15514/ISPRAS-2024-36(5)-7

CC BY

Эффективный метод с компрессией для распределенных и федеративных кокоэрсивных вариационных неравенств

[1,2] *Д.О. Медяков, ORCID 0009-0003-2007-8446 <mediakov.do@phystech.edu>*
[1,2] *Г.Л. Молодцов, ORCID 0009-0004-3516-1848 <molodtsov.gl@phystech.edu>*
[1,2,3] *А.Н. Безносиков, ORCID 0000-0002-3217-3614 <anbeznosikov@gmail.com>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*
[2] *Московский физико-технический институт (НИУ), Россия, 117303, г. Москва, ул. Керченская, д. 1А, корп. 1.*
[3] *Университет Иннополис, Россия, 420500, г. Иннополис, ул. Университетская, д. 1.*

Аннотация. Вариационные неравенства как эффективный инструмент для решения прикладных задач, в том числе задач машинного обучения, в последние годы привлекают всё больше внимания исследователей. Области применения вариационных неравенств охватывают широкий спектр направлений – от обучения с подкреплением и генеративных моделей до традиционных приложений в экономике и теории игр. В то же время, невозможно представить современный мир машинного обучения без подходов распределенной оптимизации, которые позволяют значительно ускорить процесс обучения на огромных объемах данных. Однако, сталкиваясь с большими затратами на коммуникации между устройствами в вычислительной сети, научное сообщество стремится к разработке подходов, делающих вычисления дешевыми и стабильными. В этой работе исследуется техника сжатия передаваемой информации применительно к задаче распределенных вариационных неравенств. В частности, предлагается метод на основе продвинутых техник, исконно разработанных для задач минимизации. Для нового метода приводится исчерпывающий теоретический анализ сходимости для кокоэрсивных сильно монотонных вариационных неравенств. Проведенные эксперименты подчеркивают высокую производительность представленной техники и подтверждают практическую применимость.

Ключевые слова: Вариационные неравенства; распределенная оптимизация; федеративное обучение; техники сжатия информации.

Для цитирования: Медяков Д.О., Молодцов Г.Л., Безносиков А.Н. Эффективный метод с компрессией для распределенных и федеративных кокоэрсивных вариационных неравенств. Труды ИСП РАН, 2024, том 36, вып. 5, с. 93-108. Doi: 10.15514/ISPRAS-2024-36(5)-7.

Благодарности: Работа выполнена в Лаборатории проблем федеративного обучения ИСП РАН при поддержке Минобрнауки России (доп. соглашение № 2 от «19» апреля 2024 г. к Соглашению № 075-03-2024-214 от «18» января 2024 года).

Effective Method with Compression for Distributed and Federated Cocoercive Variational Inequalities

[1,2] *D.O. Medyakov, ORCID 0009-0003-2007-8446 <mediakov.do@phystech.edu>*
[1,2] *G.L. Molodtsov, ORCID 0009-0004-3516-1848 <molodtsov.gl@phystech.edu>*
[1,2,3] *A.N. Beznosikov, ORCID 0000-0002-3217-3614 <anbeznosikov@gmail.com>*

[1] *Ivannikov Institute for System Programming of the RAS,*
25, Alexander Solzhenitsyn str., Moscow, 109004, Russia.
[2] *Moscow Institute of Physics and Technology,*
1A, Kerchenskaya Str., Moscow, 117303, Russia.
[3] *Innopolis University,*
1, Universitetskaya str., Innopolis, 420500, Russia.

Abstract. Variational inequalities as an effective tool for solving applied problems, including machine learning tasks, have been attracting more and more attention from researchers in recent years. The use of variational inequalities covers a wide range of areas – from reinforcement learning and generative models to traditional applications in economics and game theory. At the same time, it is impossible to imagine the modern world of machine learning without distributed optimization approaches that can significantly speed up the training process on large amounts of data. However, faced with the high costs of communication between devices in a computing network, the scientific community is striving to develop approaches that make computations cheap and stable. In this paper, we investigate the compression technique of transmitted information and its application to the distributed variational inequalities problem. In particular, we present a method based on advanced techniques originally developed for minimization problems. For the new method, we provide an exhaustive theoretical convergence analysis for cocoersive strongly monotone variational inequalities. We conduct experiments that emphasize the high performance of the presented technique and confirm its practical applicability.

Keywords: Variational inequalities; distributed optimization; federated learning; compression techniques.

For citation: Medyakov D.O., Molodtsov G.L., Beznosikov A.N. Effective Method with Compression for Distributed and Federated Cocoercive Variational Inequalities. *Trudy ISP RAN/Proc. ISP RAS*, 2024, vol. 36, issue 5, pp. 93-108. Doi: 10.15514/ISPRAS-2024-36(5)-7.

Acknowledgements. The work was carried out in the Laboratory for Problems of Federal Education of the ISP RAS with the support of the Ministry of Education and Science of Russia (additional agreement No 2 dated April 19, 2024 to Agreement No 075-03-2024-214 dated January 18, 2024).

1. Введение

Вариационные неравенства (ВН) привлекают внимание исследователей в различных областях уже более полувека [1]. В данной работе рассматривается общая постановка задачи вариационных неравенств на неограниченном множестве:

$$\text{Найти } z^* \in \mathbb{R}^d \text{ такое, что } F(z^*) = 0, \tag{1}$$

где $F: R^d \to R^d$ – заданный оператор. Такая общая постановка покрывает множество известных задач. Приведем несколько примеров.

Пример 1. *Классическая задача минимизации:*

$$\min_{z \in R^d} f(z),$$

где f(z) – некоторая целевая функция. Для того, чтобы свести задачу 1 к задаче минимизации, достаточно рассмотреть оператор вида $F(z) = \nabla f(z)$. Более того, для выпуклых функций f(z) точка z^* будет решением задачи 1 тогда и только тогда, когда она будет решением этой задачи.

Пример 2. *Задача поиска седловой точки (минимаксная задача):*

$$\min_{x\in R^d}\max_{y\in R^d} g(x,y),$$

где $g(z)=g(x,y)$ *– некоторая целевая функция. Для того, чтобы свести задачу 1 к минимаксной задаче, достаточно рассмотреть оператор вида* $F(z)=F(x,y)=\left[\nabla_x g(x,y), -\nabla_y g(x,y)\right]$. *Более того, для выпуклых-вогнутых функций* $g(x,y)$ *точка* $z^*=(x^*,y^*)$ *будет решением задачи 1 тогда и только тогда, когда она будет решением минимаксной задачи.*

Тем не менее, несмотря на общность постановки, она практически неприменима в современных прикладных задачах, в первую очередь в задачах обучения. Дело в том, что считать полное значение оператора на одном вычислительном устройстве слишком затратно по времени из-за огромных объемов тренировочных данных. Поэтому на практике используют распределенные подходы [2-4]. В такой парадигме используется сеть из нескольких устройств, каждое из которых содержит часть тренировочной выборки. Эти устройства обычно объединены в звездную топологию, где крайние узлы вычисляют значение оператора, исходя из хранящейся локально информации, а сервер агрегирует полученные данные и производит обновление оптимизационных переменных. Отдельно выделяют постановку федеративного обучения [5-7]. Она подразумевает наличие на каждом устройстве уникальных тренировочных выборок, причем выборки на разных устройствах необязательно одинаково распределены и, как правило, содержат приватную информацию.

Таким образом, чтобы формально перейти к общей распределенной постановке, представим оператор F в виде конечной суммы:

$$F(z)=\frac{1}{n}\sum_{i=1}^{n} F_i(z), \tag{2}$$

где n – количество устройств в вычислительной сети, а $F_i(\cdot)$ – локальный оператор, вычисляемый на основе данных на i-ом устройстве. Комбинация (1) и (2) отражает всевозможный спектр задач распределенного машинного обучения от простых регрессий до нейронных сетей [8]. Несмотря на то, что такой подход применим к классической задаче минимизации (Пример 1), больший интерес он вызывает для поиска седловых точек (Пример 2) [9]. На практике это нашло особенно яркое применение в так называемом состязательном подходе, будь то обучение генеративных сетей (GAN) [10], или робастное обучение моделей [11-12]. Кроме того, вариационные неравенства также находят широкое применение в различных классических задачах, включая эффективное матричное разложение [13], устранение зернения в изображениях [14-15], робастную оптимизацию [16], постановки из экономики, теории игр [17] и оптимального управления [18].

Для решения такой задачи ВН (1) + (2) необходима адаптация классических методов оптимизации, например, таких как градиентный спуск. Ее можно осуществить по аналогии с Примером 1, рассмотрев $\nabla f(\cdot)\to F(\cdot)$. Однако, из-за распределенной постановки 2, чтобы собрать полное значение оператора $F(\cdot)$, вычислительным устройствам необходимо "общаться" друг с другом, как правило, посредством пересылки данных на сервер. Подобные затраты на коммуникацию представляют собой серьезное ограничение распределенных и федеративных подходов. Передача информации часто занимает много времени и нарушает процесс обучения. В некоторых случаях временные затраты могут значительно превышать сложность локальных вычислений, что делает архитектурные решения неэффективными для практического применения.

Для снижения издержек на обмен информацией между узлами были предложены различные методы уменьшения количества передаваемых данных [19-20]. В данной работе исследуется применимость техники компрессии к распределенным вариационным неравенствам, сочетая

преимущества сжатия и эффективной обработки данных для минимизации коммуникационных затрат и повышения устойчивости алгоритмов.

2. Обзор литературы

2.1 Методы со сжатием

Для преодоления возникающих проблем с коммуникационными затратами сообщество применяет различные методики. Идея уменьшения количества передаваемой информации в векторах градиентов была исследована в работе [21]. Авторы предлагают модификацию градиентного спуска, при которой на каждом шаге обновляются только некоторые случайные координаты. Такие операторы впоследствии стали называть RAND-K [22]. Через несколько лет метод координатного градиентного спуска был адаптирован для распределенной оптимизации [23]. Альтернативной методикой борьбы с коммуникационными издержками является техника сжатия передаваемой информации, основанная на использовании квантизации градиентов или параметров моделей [24]. Данные методы направлены на уменьшение объема данных, за счет ограничения количества бит, необходимых для представления чисел с плавающей точкой [25-27]. Обобщение техники сжатия (будь то с помощью выбора координат или квантизации) было сделано в работе [28] посредством введения общего определения оператора сжатия. Авторы предлагают добавить оператор в обычный градиентный спуск. Однако этот метод не сходится к истинному решению, а лишь к некоторой окрестности, зависящей от дисперсии квантизованных оценок градиентов [29].

Принимая во внимание данную проблему, сообщество исследователей в области оптимизации и машинного обучения активно разрабатывает распределенные алгоритмы, применяющие технику уменьшения дисперсии. В стандартной нераспределенной задаче стохастической оптимизации она была использована в методах SVRG [30], SAG [31], SAGA [32], SARAH [33-34], а затем адаптирована для распределенной оптимизации в методе DIANA [35] путем сжатия не самого градиента, а разности градиентов. В дальнейшем, идея была развита и использована в более продвинутых алгоритмах: VR-DIANA [36], FEDCOMGATE [37], FEDSTEPH [38]. Одной из наиболее продвинутых методик применения данных техник в невыпуклой распределенной оптимизации является MARINA [39]. В отличие от всех известных подходов, использующих несмещенные операторы сжатия, MARINA базируется на смещенной оценке градиента. Тем не менее, доказано, что данный метод обеспечивает гарантии сходимости, превосходящие все ранее известные техники.

2.2 Методы решения вариационных неравенств

Применение различных методов для решения задач вариационных неравенств и задач седловых точек является предметом обширных исследований [40-48]. Упомянутая выше техника редукции дисперсии была также перенесена и на вариационные неравенства [49-57]. Большинство из этих методов основаны на подходе SVRG, однако некоторые исследования были проведены в анализе более привлекательного с практической точки зрения для задач минимизации метода SARAH [58-59].

Методы борьбы за эффективность процесса общения также исследовались в общности вариационных неравенств [9, 60-65], исследовались и алгоритмы с компрессией. Для липшицевых операторов это было сделано в работах [66-68] на основе редукции дисперсии из работы [52].

Говоря о стандартных предположениях, которые используются в анализе методов решения вариационных неравенств, кокоэрсивный режим является одним из наиболее часто встречающихся, несмотря на то что данное требование является более строгим, нежели липшицевость операторов [69]. В частности, для кокоэрсивных вариационных неравенств существует общий анализ стохастических методов [70], включающий и методы со сжатием.

В разрезе работ, использующих предположение кокоэрсивности, стоит выделить основанный на алгоритме SARAH метод [59]. В то же время метод MARINA, как раз и базирующийся на SARAH, еще не был исследован в контексте вариационных неравенств. Шаг MARINA по факту является шагом SARAH с добавлением сжатия. Как уже отмечалось ранее, метод MARINA является наиболее привлекательным с точки зрения уменьшения коммуникационных издержек в парадигме распределенного обучения. Цель данной статьи – исследовать теоретическую и практическую применимость алгоритма MARINA к кокоэрсивным вариационным неравенствам.

3. Основные результаты

- **Новый метод.** В рамках исследования распределенной постановки представляется новый метод, который использует метод с компрессией MARINA для решения задач вариационных неравенств.
- **Теоретическая значимость.** В работе представлен полный теоретический анализ предложенного метода для кокоэрсивных монотонных вариационных неравенств.
- **Адаптивный анализ.** Данное исследование предлагает несколько возможных расширений. Например, полученные результаты могут быть обобщены для случая произвольного оператора квантования и различных размеров подмножеств данных, используемых клиентами.
- **Экспериментальная валидация.** Проведены численные эксперименты, которые подчеркивают применимость нашего метода с различными техниками сжатия (квантизацией, выбором координат).

4. Постановка

Приведем нотацию, которая используется в работе. Будем обозначать множество векторов размерности d с элементами, являющимися вещественными числами, как R^d, положительные вещественные коэффициенты греческими и латинскими буквами α, γ, μ и ℓ, Евклидову норму вектора $u \in R^d$ как $\| u \| = \sqrt{\langle u, u \rangle}$, математическое ожидание случайной величины ξ как $E[\xi]$. Напомним, что рассматривается задача 1, где оператор F принимает вид (2). Введем следующие предположения на оператор.

Предположение 1 (Кокоэрсивность). *Каждый оператор F_i является ℓ-кокоэрсивным, если для любых $u, v \in R^d$ выполняется*

$$\| F_i(u) - F_i(v) \|^2 \leq \ell \langle F_i(u) - F_i(v), u - v \rangle.$$

Это предположение является более сильным аналогом предположения на липшицевость F_i. Для задач выпуклой минимизации ℓ-липшицевость и ℓ-кокоэрсивность эквивалентны. Сравнение этих предположений для задачи вариационных неравенств приведено в работе [69].

Предположение 2 (Сильная монотонность). *Оператор F является -сильно монотонным, то есть для любых $u, v \in R^d$ выполняется*

$$\langle F(u) - F(v), u - v \rangle \geq \mu \| u - v \|^2.$$

Для задач минимизации это свойство означает сильную выпуклость, а для задач поиска седловой точки – сильную выпуклость-сильную вогнутость.

Предположение 3. *Отображение $C: R^d \to R^d$ является несмещенным компрессором, если для любого $u \in R^d$ существует $\alpha > 0$ такое, что выполняется*

$$E[C(u)] = u,$$
$$E\|C(u) - u\|^2 \leq \alpha \| u \|^2.$$

Это классическое предположение на компрессоры, используемые как в статье MARINA [39], так и в других работах [28, 71].

Предположение 4. *Компрессор* $C(\cdot)$ *оставляет долю информации равную* $\delta \in (0,1]$.

Например, для компрессоров, которые производят выбор только части координат, это предположение означает, что из d координат, они оставят только $d\delta$.

5. Метод и анализ сходимости

Для решения общей постановки задачи вариационных неравенств с липшицевыми операторами обычно используют методы, основанные на подходе EXTRAGRADIENT [40]. Однако в этой работе рассматриваются кокоэрсивные вариационные неравенства, поэтому достаточно использовать классический подход SGD [72-73]. Комбинируя его с продвинутой техникой сжатия MARINA, представляем формальное описание рассматриваемого алгоритма (Алгоритм 1). Отметим, что наш подход несколько отличается от предложенного в оригинальной статье [39]. Там предлагается производить рестарт рекурсивных обновлений аппроксимаций локальных градиентов с некоторой заранее выбранной вероятностью, здесь же это делается раз в фиксированное число итераций, которое называется эпохой (строка 6 в Алгоритме 1). Далее приведен теоретический анализ сходимости Алгоритма 1.

Алгоритм 1 MARINA для кокоэрсивных вариационных неравенств

1: **Параметры:** Шаг обучения $\gamma > 0$, количество итераций обучения K.
2: **Инициализация:** $z^0 \in \mathbb{R}^d$.
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad z^0 = \tilde{z}^{s-1}$
5: $\quad g^0 = F(z^0)$
6: $\quad z^1 = z^0 - \gamma g^0$
7: $\quad$ **for** $k = 1, 2, \ldots, K-1$ **do**
8: $\quad\quad$ Отправить g^{k-1} на каждое устройство
9: $\quad\quad$ **for** $i = 1, 2, \ldots, n$ **do**
10: $\quad\quad\quad g_i^k = g^{k-1} + \mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right)$
11: $\quad\quad\quad$ Отправить g_i^k на сервер
12: $\quad\quad$ **end for**
13: $\quad\quad g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$
14: $\quad\quad z^{k+1} = z^k - \gamma g^k$
15: $\quad$ **end for**
16: $\quad \tilde{z}^s = z^K$
17: **end for**

Лемма 1. *Пусть выполнены Предположения 1, 2, 3. Тогда для Алгоритма 1 с* $\gamma \leq \frac{1}{2\ell\left(1+\frac{\alpha}{n}\right)}$ *верна следующая оценка:*

$$E \parallel g^K \parallel^2 \leq \left(1 - \frac{2\gamma\mu}{3}\right)^K E \parallel F(z^0) \parallel^2.$$

Доказательство. Зафиксируем любое s в Алгоритме 1 и рассмотрим обновление аппроксимаций градиента g^k:

$$\|g^k\|^2 = \left\| \frac{1}{n}\sum_{i=1}^{n} g_i^k \right\|^2$$

$$
\begin{aligned}
&= \left\| g^{k-1} + \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right)\right\|^2 \\
&= \left\| g^{k-1} + \frac{1}{n}\sum_{i=1}^{n} \left(F_i(z^k) - F_i(z^{k-1})\right)\right\|^2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right) - \left(F_i(z^k) - F_i(z^{k-1})\right)\right)\right\|^2 \\
&\quad + 2\Big\langle g^{k-1} + \frac{1}{n}\sum_{i=1}^{n} \left(F_i(z^k) - F_i(z^{k-1})\right), \\
&\qquad\qquad \frac{1}{n}\sum_{i=1}^{n} \left(\mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right) - \left(F_i(z^k) - F_i(z^{k-1})\right)\right)\Big\rangle.
\end{aligned}
$$

Пользуясь Предположением 3, получим, что математическое ожидание правой части скалярного произведения равно нулю. Отсюда следует, что и скалярное произведение становится равным нулю. Более того, используя Предположение 3 для второй нормы, получаем

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\Big(\mathcal{C}\big(F_i(z^k) - F_i(z^{k-1})\big) - \big(F_i(z^k) - F_i(z^{k-1})\big)\Big)\right\|^2 \\
&= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n}\left\|\mathcal{C}\big(F_i(z^k) - F_i(z^{k-1})\big) - \big(F_i(z^k) - F_i(z^{k-1})\big)\right\|^2 \\
&\leq \frac{\alpha}{n^2}\sum_{i=1}^{n}\|F_i(z^k) - F_i(z^{k-1})\|^2,
\end{aligned}
$$

так как все попарные скалярные произведения равны нулю. Таким образом,

$$
\begin{aligned}
\mathbb{E}\|g^k\|^2 &\leqslant \mathbb{E}\|g^{k-1}\|^2 + \left\|F(z^k) - F(z^{k-1})\right\|^2 \\
&\quad + 2\left\langle g^{k-1}, F(z^k) - F(z^{k-1})\right\rangle \\
&\quad + \frac{\alpha}{n^2}\sum_{i=1}^{n}\left\|F_i(z^k) - F_i(z^{k-1})\right\|^2 \\
&= \mathbb{E}\|g^{k-1}\|^2 + \left\|F(z^k) - F(z^{k-1})\right\|^2 \\
&\quad - \frac{2}{\gamma}\left\langle z^k - z^{k-1}, F(z^k) - F(z^{k-1})\right\rangle \\
&\quad + \frac{\alpha}{n^2}\sum_{i=1}^{n}\left\|F_i(z^k) - F_i(z^{k-1})\right\|^2 \\
&= \mathbb{E}\|g^{k-1}\|^2 + \left\|F(z^k) - F(z^{k-1})\right\|^2 \\
&\quad + \frac{\alpha}{n^2}\sum_{i=1}^{n}\left\|F_i(z^k) - F_i(z^{k-1})\right\|^2 \\
&\quad - \frac{2}{3\gamma}\frac{1}{n}\sum_{i=1}^{n}\left\langle z^k - z^{k-1}, F_i(z^k) - F_i(z^{k-1})\right\rangle
\end{aligned}
$$

$$-\frac{2}{3\gamma}\left\langle z^k - z^{k-1}, F(z^k) - F(z^{k-1})\right\rangle$$
$$-\frac{2}{3\gamma}\left\langle z^k - z^{k-1}, F(z^k) - F(z^{k-1})\right\rangle.$$

Для первого и второго скалярного произведений используется Предположение 1, а именно $\langle z^k - z^{k-1}, F_i(z^k) - F_i(z^{k-1})\rangle \geq \frac{1}{\ell}\|F_i(z^k) - F_i(z^{k-1})\|^2$, а для третьего – Предположение 2, а именно $\langle z^k - z^{k-1}, F(z^k) - F(z^{k-1})\rangle \geq \mu\|F(z^k) - F(z^{k-1})\|^2$. Тогда

$$\begin{aligned}\mathbb{E}\|g^k\|^2 \quad &\leqslant \quad \mathbb{E}\|g^{k-1}\|^2 + \left(1 - \frac{2}{3\gamma\ell}\right)\left\|F(z^k) - F(z^{k-1})\right\|^2 \\ &+ \left(\frac{\alpha}{n} - \frac{2}{3\gamma\ell}\right)\frac{1}{n}\sum_{i=1}^{n}\left\|F_i(z^k) - F_i(z^{k-1})\right\|^2 \\ &-\frac{2\mu}{3\gamma}\|z^k - z^{k-1}\|^2.\end{aligned}$$

Выберем $\gamma \leq \frac{1}{2\ell\left(1+\frac{\alpha}{n}\right)}$. С учетом этого, получим

$$E \parallel g^k \parallel^2 \quad \leq \quad \left(1 - \frac{2\gamma\mu}{3}\right)E \parallel g^{k-1} \parallel^2.$$

Запустим рекурсию до первой итерации в эпохе и учтем, что $g^0 = F(z^0)$:

$$E \parallel g^K \parallel^2 \leq \left(1 - \frac{2\gamma\mu}{3}\right)^K E \parallel F(z^0) \parallel^2.$$

□

Следующая лемма дает нам представление о разнице между полным оператором $F(\cdot)$ и его сжатой версией g в течение внутреннего цикла Алгоритма 1.

Лемма 2. *Пусть выполнены Предположения 1, 2, 3. Тогда для Алгоритма 1 верна следующая оценка:* $E\|F(z^K) - g^K\|^2 \leq \frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}E \parallel F(z^0) \parallel^2$.

Доказательство. Рассмотрим следующую цепочку:

$$\begin{aligned}\mathbb{E}\left\|F(z^k) - g^k\right\|^2 \quad &= \quad \mathbb{E}\left\|\left[F(z^{k-1}) - g^{k-1}\right] + \left[F(z^k) - F(z^{k-1})\right] - \left[g^k - g^{k-1}\right]\right\|^2 \\ &= \quad \mathbb{E}\left\|F(z^{k-1}) - g^{k-1}\right\|^2 + \mathbb{E}\left\|F(z^k) - F(z^{k-1})\right\|^2 \\ &\quad +\mathbb{E}\left\|g^k - g^{k-1}\right\|^2 + 2\mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, F(z^k) - F(z^{k-1})\right\rangle \\ &\quad -2\mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, g^k - g^{k-1}\right\rangle \\ &\quad -2\mathbb{E}\left\langle F(z^k) - F(z^{k-1}), g^k - g^{k-1}\right\rangle. \qquad (3)\end{aligned}$$

Теперь отдельно оценим второе скалярное произведение:

$$\begin{aligned}&\mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, g^k - g^{k-1}\right\rangle \\ &\quad = \mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, \frac{1}{n}\sum_{i=1}^{n}\mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right)\right\rangle\end{aligned}$$

$$= \mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, \frac{1}{n}\sum_{i=1}^{n}\left(F_i(z^k) - F_i(z^{k-1})\right)\right\rangle$$
$$+\mathbb{E}\Big\langle F(z^{k-1}) - g^{k-1},$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right) - \left(F_i(z^k) - F_i(z^{k-1})\right)\right)\Big\rangle.$$

Так как второе скалярное произведение равно нулю ввиду Предположения 3, получим

$$\mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, g^k - g^{k-1}\right\rangle = \mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, F(z^k) - F(z^{k-1})\right\rangle. \quad (4)$$

Делая аналогичные выкладки, можно оценить третье скалярное произведение в (3), как

$$\begin{aligned}\mathbb{E}\left\langle F(z^k) - F(z^{k-1}), g^k - g^{k-1}\right\rangle &= \mathbb{E}\left\langle F(z^k) - F(z^{k-1}), F(z^k) - F(z^{k-1})\right\rangle \\ &= \left\|F(z^k) - F(z^{k-1})\right\|^2. \end{aligned} \quad (5)$$

Подставляя (4) и (5) в (3), получим

$$\begin{aligned}\mathbb{E}\left\|F(z^k) - g^k\right\|^2 &= \mathbb{E}\left\|F(z^{k-1}) - g^{k-1}\right\|^2 + \mathbb{E}\left\|F(z^k) - F(z^{k-1})\right\|^2 \\ &\quad +\mathbb{E}\left\|g^k - g^{k-1}\right\|^2 \\ &\quad +2\mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, F(z^k) - F(z^{k-1})\right\rangle \\ &\quad -2\mathbb{E}\left\langle F(z^{k-1}) - g^{k-1}, F(z^k) - F(z^{k-1})\right\rangle \\ &\quad -2\mathbb{E}\left\|F(z^k) - F(z^{k-1})\right\|^2 \\ &\leqslant \mathbb{E}\left\|F(z^{k-1}) - g^{k-1}\right\|^2 + \mathbb{E}\left\|g^k - g^{k-1}\right\|^2.\end{aligned}$$

Запустим рекурсию до первой итерации в эпохе и учтем, что $g^0 = F(z^0)$*:*

$$\mathbb{E}\left\|F(z^K) - g^K\right\|^2 \leqslant \sum_{k=1}^{K}\mathbb{E}\|g^k - g^{k-1}\|^2. \quad (6)$$

Пользуясь Предположением 3,

$$\begin{aligned}\mathbb{E}\|g^k - g^{k-1}\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right)\right\|^2 \\ &\leqslant \frac{1+\frac{\alpha}{n}}{n}\sum_{i=1}^{n}\left\|F_i(z^k) - F_i(z^{k-1})\right\|^2. \end{aligned} \quad (7)$$

Далее, аналогично доказательству Леммы 1:

$$\begin{aligned}\mathbb{E}\|g^k\|^2 &= \mathbb{E}\left\|g^{k-1} + \frac{1}{n}\sum_{i=1}^{n}\mathcal{C}\left(F_i(z^k) - F_i(z^{k-1})\right)\right\|^2 \\ &\leqslant \mathbb{E}\left\|g^{k-1} + \frac{1}{n}\sum_{i=1}^{n}\left(F_i(z^k) - F_i(z^{k-1})\right)\right\|^2\end{aligned}$$

$$
\begin{aligned}
&+\frac{\alpha}{n^2}\sum_{i=1}^{n}\left\|F_i(z^k)-F_i(z^{k-1})\right\|^2\\
\leqslant\ & \mathbb{E}\|g^{k-1}\|^2+\frac{1+\frac{\alpha}{n}}{n}\sum_{i=1}^{n}\left\|F_i(z^k)-F_i(z^{k-1})\right\|^2\\
&+2\left\langle g^{k-1},F(z^k)-F(z^{k-1})\right\rangle\\
=\ & \mathbb{E}\|g^{k-1}\|^2+\frac{1+\frac{\alpha}{n}}{n}\sum_{i=1}^{n}\left\|F_i(z^k)-F_i(z^{k-1})\right\|^2\\
&-\frac{1}{\gamma}\frac{1}{n}\sum_{i=1}^{n}\left\langle z^k-z^{k-1},F_i(z^k)-F_i(z^{k-1})\right\rangle\\
&-\frac{1}{\gamma}\left\langle z^k-z^{k-1},F(z^k)-F(z^{k-1})\right\rangle\\
\leqslant\ & (1-\gamma\mu)\mathbb{E}\|g^{k-1}\|^2\\
&+\left(1-\frac{1}{\gamma\ell\left(1+\frac{\alpha}{n}\right)}\right)\frac{1+\frac{\alpha}{n}}{n}\sum_{i=1}^{n}\left\|F_i(z^k)-F_i(z^{k-1})\right\|^2.
\end{aligned}
$$

Выражая сумму,

$$
\begin{aligned}
&\frac{1+\frac{\alpha}{n}}{n}\sum_{i=1}^{n}\|F_i(z^k)-F_i(z^{k-1})\|^2\\
\leq&\frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}[(1-\gamma\mu)E\parallel g^{k-1}\parallel^2-E\parallel g^{k-1}\parallel^2]\\
\leq&\frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}[E\parallel g^{k-1}\parallel^2-E\parallel g^{k}\parallel^2].
\end{aligned}
$$

Комбинируя полученную оценку с (7) и (6), получаем

$$
E\|F(z^K)-g^K\|^2\leq\frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}E\parallel g^0\parallel^2=\frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}E\parallel F(z^0)\parallel^2.
$$

□

Теперь объединим результаты Леммы 1 и 2 и получим основную теорему данной статьи.

Теорема 1. *Пусть выполнены Предположения 1, 2, 3. Тогда для Алгоритма 1 с* $\gamma=\frac{1}{8\ell\left(1+\frac{\alpha}{n}\right)}$ *и* $K=\frac{30\ell\left(1+\frac{\alpha}{n}\right)}{\mu}$ *верна следующая оценка:*

$$
E\parallel F(z^s)\parallel^2\leq\frac{1}{2}E\parallel F(z^{s-1})\parallel^2.
$$

Доказательство. Начнем с

$$
E\|F(z^K)\|^2\leq 2E\|F(z^K)-g^K\|^2+2E\|g^K\|^2.
$$

Далее применим Леммы 1, 2:

$$\mathbb{E}\left\|F(z^K)\right\|^2 \leqslant \left[2\frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}+2\left(1-\frac{2\gamma\mu}{3}\right)^K\right]\mathbb{E}\left\|F(z^0)\right\|^2$$
$$\leqslant \left[2\frac{\gamma\ell\left(1+\frac{\alpha}{n}\right)}{1-\gamma\ell\left(1+\frac{\alpha}{n}\right)}+2\exp\left(-\frac{2}{3}\gamma\mu K\right)\right]\mathbb{E}\left\|F(z^0)\right\|^2.$$

Здесь использовалось, что $\gamma\mu \in (0;1)$ и $(1-\gamma\mu) \leq exp(-\gamma\mu)$. Подставив $\gamma = \frac{1}{8\ell\left(1+\frac{\alpha}{n}\right)}$ и $K = \frac{30\ell\left(1+\frac{\alpha}{n}\right)}{\mu}$, получаем оценку

$$E\|F(z^K)\|^2 \leq \frac{1}{2}E\|F(z^0)\|^2.$$

Так как $z^0 = \mathbb{z}^{S-1}$ и $z^K = \mathbb{z}^S$, была получена сходимость за одну эпоху

$$E\|F(\mathbb{z}^S)\|^2 \leq \frac{1}{2}E\|F(\mathbb{z}^{S-1})\|^2.$$

□

Следствие 1. *Пусть выполнены Предположения 1, 2, 3, 4. Тогда Алгоритм 1 с* $\gamma = \frac{1}{8\ell\left(1+\frac{\alpha}{n}\right)}$ *и* $K = \frac{30\ell\left(1+\frac{\alpha}{n}\right)}{\mu}$ *для достижения -точности, где* $\varepsilon^2 \sim E\|F(\mathbb{z}^S)\|^2$, *требует*

$$O\left(\left[1+\delta\frac{\ell}{\mu}\left(1+\frac{\alpha}{n}\right)\right]log_2\frac{\|F(z^0)\|^2}{\varepsilon^2}\right)$$

пересылок градиента для каждого устройства.

Доказательство. Из Теоремы 1:

$$E\|F(\mathbb{z}^S)\|^2 \leq \left(\frac{1}{2}\right)^S E\|F(z^0)\|^2.$$

Тогда для достижения точности $\varepsilon^2 \sim E\|F(\mathbb{z}^S)\|^2$ нам нужно следующее число внешних итераций (эпох) S:

$$S = O\left(log_2\frac{\|F(z^0)\|^2}{\varepsilon^2}\right).$$

На каждой внешней итерации полный оператор пересылается только один раз, а в течении остальных $K-1$ внутренних итераций используются сжатые версии размера δ (Предположение 4). Тогда каждое устройство пересылает следующее количество градиентов:

$$S\times\left(1+\delta\times(K-1)\right) = O\left(\left[1+\delta\frac{\ell}{\mu}\left(1+\frac{\alpha}{n}\right)\right]log_2\frac{\|F(z^0)\|^2}{\varepsilon^2}\right).$$

□

Замечание 1. *Выбирая* $\delta \leq \frac{1}{\alpha}$ *и* $\alpha = n$, *Следствие 1 дает следующую оценку сложности коммуникаций:*

$$O\left(\left[1+\frac{\ell}{\mu n}\right]log_2\frac{\|F(z^0)\|^2}{\varepsilon^2}\right).$$

Сравнивая полученный результат с другими методами, отметим, что в работе [70] была получена такая же оценка сходимости метода DIANA для кокоэрсивных вариационных неравенств.

6. Эксперименты

В данном разделе демонстрируется практическое применение предложенного алгоритма. Наша основная цель – оценить эффективность различных методов для решения вариационных неравенств с кокоэрсивными свойствами. В частности, проводится сравнение производительности метода с несмещенной компрессией, метода с квантизацией [74] и метода без компрессии для алгоритма MARINA.

Рассмотрим задачу седловой точки конечной суммы, определяемую следующим образом:

$$g(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left[x^\top A_i y + a_i^\top x + b_i^\top y + \frac{\lambda}{2}\| x \|^2 - \frac{\lambda}{2}\| y \|^2\right],$$

где $A_i \in R^{d\times d}$, $a_i, b_i \in R^d$. Данная задача является -сильно выпуклой по x и λ-сильно вогнутой по y, а также L-гладкой с $L = \| A \|_2$, где $A = \frac{1}{n}\sum_{i=1}^{n} A_i$.

Положим $n = 10$, $d = 100$, и $\lambda = 1$. Матрицы A_i и векторы a_i, b_i генерируются случайным образом. Примечательно, что константа кокоэрсивности для данной задачи определяется как $\ell = \frac{\|A\|_2^2}{\lambda}$. Эти параметры позволяют исследовать поведение алгоритмов в различных условиях.

В качестве критерия возьмем квадрат нормы градиента на k −ой итерации, отнесенного к квадрату нормы функционала на первой итерации: $\left\|\frac{F(z^k)}{F(z^0)}\right\|^2$. Обозначим $\delta = \frac{K}{d}$ – отношение числа координат в компрессии, которые выбираются, к общей размерности задачи. Для экспериментов использовалась квантизация для перевода чисел из формата fp-32 в формат int-8, что позволяет значительно оптимизировать модели за счет уменьшения размера весов модели в 4 раза и того, что многие процессоры эффективнее обрабатывают 8-битные данные [75-76]. Метод квантизации для удобства на графиках обозначаем Q-MARINA. По оси абсцисс отображается объем передаваемой информации в килобитах, что позволяет оценить эффективность алгоритмов с учетом ограничений на передачу данных. Для комплексной оценки методов проектируются три экспериментальных сценария с разными уровнями кокоэрсивности: низким ($\ell \approx 10^2$), средним ($\ell \approx 10^3$) и высоким ($\ell \approx 10^4$).

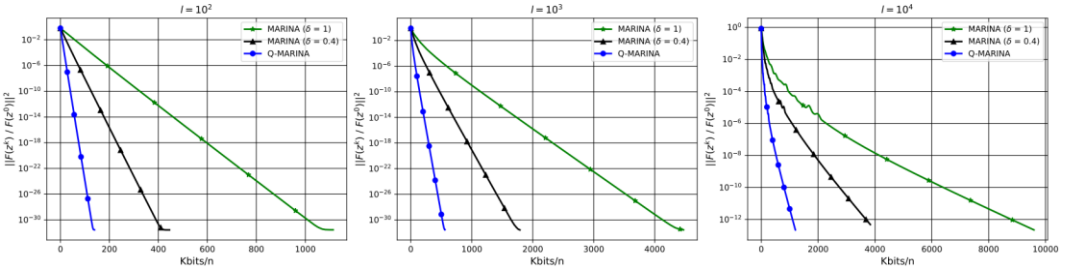

Рис. 1. Сравнение производительности методов на основе метода MARINA для кокоэрсивных вариационных неравенств на билинейной задаче седловой точки.
Fig. 1. Performance comparison of MARINA-based methods for cocoersive variational inequalities on the bilinear saddle point problem.

Результаты экспериментов представлены на рис. 1. Как видно из графиков, методы MARINA с компрессией стабильно превосходит методы без нее во всех сценариях. В то время как сжатие RAND-K демонстрирует приемлемую производительность, оно уступает квантизации. В свою очередь, метод без сжатия показывает значительно более медленную сходимость с точки зрения объемов передаваемой информации, что подчеркивает его ограничения при решении задач больших размерностей.

Таким образом, эксперименты подтверждают эффективность предложенного подхода к решению вариационных неравенств. Метод MARINA с компрессией обеспечивает

значительное сокращение объемов передаваемой информации при сохранении скорости сходимости. Это делает его перспективным инструментом для распределенных систем, где ограниченные ресурсы передачи данных являются критическим фактором.

Список литературы / References

[1]. F. E. Browder, “Nonexpansive nonlinear operators in a banach space,” Proceedings of the National Academy of Sciences, vol. 54, no. 4, pp. 1041–1044, 1965.

[2]. P. Kairouz et al., “Advances and open problems in federated learning,” Foundations and trends in machine learning, vol. 14, no. 1–2, pp. 1–210, 2021.

[3]. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” IEEE signal processing magazine, vol. 37, no. 3, pp. 50–60, 2020.

[4]. J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, “A survey on distributed machine learning,” Acm computing surveys (csur), vol. 53, no. 2, pp. 1–33, 2020.

[5]. J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence,” arXiv preprint arXiv:1610.02527, 2016.

[6]. V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, “Federated multi-task learning,” Advances in neural information processing systems, vol. 30, 2017.

[7]. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.

[8]. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436–444, 2015.

[9]. A. Beznosikov, V. Samokhin, and A. Gasnikov, “Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms,” arXiv preprint arXiv:2010.13112, 2020.

[10]. I. Goodfellow et al., “Generative adversarial networks,” Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[11]. X. Liu et al., “Adversarial training for large neural language models,” arXiv preprint arXiv:2004.08994, 2020.

[12]. C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, “Freelb: Enhanced adversarial training for natural language understanding,” arXiv preprint arXiv:1909.11764, 2019.

[13]. F. Bach, J. Mairal, and J. Ponce, “Convex sparse matrix factorizations,” arXiv preprint arXiv:0812.1869, 2008.

[14]. E. Esser, X. Zhang, and T. F. Chan, “A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science,” SIAM Journal on Imaging Sciences, vol. 3, no. 4, pp. 1015–1046, 2010.

[15]. A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” Journal of mathematical imaging and vision, vol. 40, pp. 120–145, 2011.

[16]. A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, Robust optimization, vol. 28. Princeton university press, 2009.

[17]. J. Von Neumann and O. Morgenstern, Theory of games and economic behavior: By j. Von neumann and o. morgenstern. Princeton university press, 1953.

[18]. F. Facchinei and J.-S. Pang, Finite-dimensional variational inequalities and complementarity problems. Springer, 2003.

[19]. D. Kovalev, A. Beznosikov, E. Borodich, A. Gasnikov, and G. Scutari, “Optimal gradient sliding and its application to optimal distributed optimization under similarity,” Advances in Neural Information Processing Systems, vol. 35, pp. 33494–33507, 2022.

[20]. D. Medyakov, G. Molodtsov, A. Beznosikov, and A. Gasnikov, “Optimal data splitting in distributed optimization for machine learning,” in Doklady mathematics, Pleiades Publishing Moscow, 2023, pp. S465–S475.

[21]. Y. Nesterov, “Efficiency of coordinate descent methods on huge-scale optimization problems,” SIAM Journal on Optimization, vol. 22, no. 2, pp. 341–362, 2012.

[22]. A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan, “On biased compression for distributed learning,” Journal of Machine Learning Research, vol. 24, no. 276, pp. 1–50, 2023.

[23]. P. Richtárik and M. Takáč, “Distributed coordinate descent method for learning with big data,” Journal of Machine Learning Research, vol. 17, no. 75, pp. 1–25, 2016.

[24]. A. T. Suresh, X. Y. Felix, S. Kumar, and H. B. McMahan, “Distributed mean estimation with limited communication,” in International conference on machine learning, PMLR, 2017, pp. 3329–3337.

[25]. J. Wu, W. Huang, J. Huang, and T. Zhang, “Error compensated quantized SGD and its applications to large-scale distributed optimization,” in International conference on machine learning, PMLR, 2018, pp. 5325–5333.
[26]. J. Wangni, J. Wang, J. Liu, and T. Zhang, “Gradient sparsification for communication-efficient distributed optimization,” Advances in Neural Information Processing Systems, vol. 31, 2018.
[27]. W. Wen et al., “Terngrad: Ternary gradients to reduce communication in distributed deep learning,” Advances in neural information processing systems, vol. 30, 2017.
[28]. D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, “QSGD: Communication-efficient SGD via gradient quantization and encoding,” Advances in neural information processing systems, vol. 30, 2017.
[29]. E. Gorbunov, “Distributed and stochastic optimization methods with gradient compression and local steps,” arXiv preprint arXiv:2112.10645, 2021.
[30]. R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” Advances in neural information processing systems, vol. 26, 2013.
[31]. N. Roux, M. Schmidt, and F. Bach, “A stochastic gradient method with an exponential convergence _rate for finite training sets,” Advances in neural information processing systems, vol. 25, 2012.
[32]. A. Defazio, F. Bach, and S. Lacoste-Julien, “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives,” Advances in neural information processing systems, vol. 27, 2014.
[33]. L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, “SARAH: A novel method for machine learning problems using stochastic recursive gradient,” in International conference on machine learning, PMLR, 2017, pp. 2613–2621.
[34]. A. Beznosikov and M. Takáč, “Random-reshuffled SARAH does not need full gradient computations,” Optimization Letters, vol. 18, no. 3, pp. 727–749, 2024.
[35]. K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik, “Distributed learning with compressed gradient differences,” Optimization Methods and Software, pp. 1–16, 2024.
[36]. S. Horvóth, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtárik, “Natural compression for distributed deep learning,” in Mathematical and scientific machine learning, PMLR, 2022, pp. 129–141.
[37]. F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, “Federated learning with compression: Unified analysis and sharp guarantees,” in International conference on artificial intelligence and statistics, PMLR, 2021, pp. 2350–2358.
[38]. R. Das, A. Acharya, A. Hashemi, S. Sanghavi, I. S. Dhillon, and U. Topcu, “Faster non-convex federated learning via global and local momentum,” in Uncertainty in artificial intelligence, PMLR, 2022, pp. 496–506.
[39]. E. Gorbunov, K. P. Burlachenko, Z. Li, and P. Richtárik, “MARINA: Faster non-convex distributed learning with compression,” in International conference on machine learning, PMLR, 2021, pp. 3788–3798.
[40]. A. Juditsky, A. Nemirovski, and C. Tauvel, “Solving variational inequalities with stochastic mirror-prox algorithm,” Stochastic Systems, vol. 1, no. 1, pp. 17–58, 2011.
[41]. G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien, “A variational inequality perspective on generative adversarial networks,” arXiv preprint arXiv:1802.10551, 2018.
[42]. Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, “On the convergence of single-call stochastic extra-gradient methods,” in Advances in neural information processing systems, Curran Associates, Inc., 2019, pp. 6938–6948.
[43]. K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky, “Revisiting stochastic extragradient,” in International conference on artificial intelligence and statistics, PMLR, 2020, pp. 4573–4582.
[44]. Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, “Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling,” in Advances in Neural Information Processing Systems, 2020, pp. 16223–16234.
[45]. E. Gorbunov, H. Berard, G. Gidel, and N. Loizou, “Stochastic extragradient: General analysis and improved rates,” in International conference on artificial intelligence and statistics, PMLR, 2022, pp. 7865–7901.
[46]. A. Beznosikov, B. Polyak, E. Gorbunov, D. Kovalev, and A. Gasnikov, “Smooth monotone stochastic variational inequalities and saddle point problems: A survey,” European Mathematical Society Magazine, no. 127, pp. 15–28, 2023.
[47]. A. Beznosikov, S. Samsonov, M. Sheshukova, A. Gasnikov, A. Naumov, and E. Moulines, “First order methods with markovian noise: From acceleration to variational inequalities,” Advances in Neural Information Processing Systems, vol. 36, 2024.

[48]. V. Solodkin, A. Veprikov, and A. Beznosikov, “Methods for optimization problems with markovian stochasticity and non-euclidean geometry,” arXiv preprint arXiv:2408.01848, 2024.
[49]. B. Palaniappan and F. Bach, “Stochastic variance reduction methods for saddle-point problems,” in Advances in neural information processing systems, 2016, pp. 1416–1424.
[50]. T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien, “Reducing noise in gan training with variance reduced extragradient,” in Advances in Neural Information Processing Systems, 2019, pp. 393–403.
[51]. A. Alacaoglu, Y. Malitsky, and V. Cevher, “Forward-reflected-backward method with variance reduction,” Computational Optimization and Applications, vol. 80, Nov. 2021, doi: 10.1007/s10589-021-00305- 3.
[52]. A. Alacaoglu and Y. Malitsky, “Stochastic variance reduction for variational inequality methods,” arXiv preprint arXiv:2102.08352, 2021.
[53]. D. Kovalev, A. Beznosikov, A. Sadiev, M. Persiianov, P. Richtárik, and A. Gasnikov, “Optimal algorithms for decentralized stochastic variational inequalities,” Advances in Neural Information Processing Systems, vol. 35, pp. 31073–31088, 2022.
[54]. A. Beznosikov, A. Gasnikov, K. Zainulina, A. Maslovskiy, and D. Pasechnyuk, “A unified analysis of variational inequality methods: Variance reduction, sampling, quantization andCoordinate descent,” arXiv preprint arXiv:2201.12206, 2022.
[55]. A. Pichugin, M. Pechin, A. Beznosikov, A. Savchenko, and A. Gasnikov, “Optimal analysis of method with batching for monotone stochastic finite-sum variational inequalities,” in Doklady mathematics, Springer, 2023, pp. S348–S359.
[56]. A. Pichugin, M. Pechin, A. Beznosikov, V. Novitskii, and A. Gasnikov, “Method with batching for stochastic finite-sum variational inequalities in non-euclidean setting,” Chaos, Solitons & Fractals, vol. 187, p. 115396, 2024.
[57]. D. Medyakov, G. Molodtsov, E. Grigoriy, E. Petrov, and A. Beznosikov, “Shuffling heuristic in variational inequalities: Establishing new convergence guarantees,” in International conference on computational optimization,
[58]. L. Chen, B. Yao, and L. Luo, “Faster stochastic algorithms for minimax optimization under polyak-{\l} ojasiewicz condition,” Advances in Neural Information Processing Systems, vol. 35, pp. 13921–13932, 2022.
[59]. A. Beznosikov and A. Gasnikov, “Sarah-based variance-reduced algorithm for stochastic finite-sum cocoercive variational inequalities,” in Data analysis and optimization: In honor of boris mirkin’s 80th birthday, Springer, 2023, pp. 47–57.
[60]. W. Liu, A. Mokhtari, A. Ozdaglar, S. Pattathil, Z. Shen, and N. Zheng, “A decentralized proximal point-type method for saddle point problems,” arXiv preprint arXiv:1910.14380, 2019.
[61]. M. Liu et al., “A decentralized parallel algorithm for training generative adversarial nets,” Advances in Neural Information Processing Systems, vol. 33, pp. 11056–11070, 2020.
[62]. I. Tsaknakis, M. Hong, and S. Liu, “Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack,” in ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2020, pp. 5755–5759.
[63]. Y. Deng and M. Mahdavi, “Local stochastic gradient descent ascent: Convergence analysis and communication efficiency,” in International conference on artificial intelligence and statistics, PMLR, 2021, pp. 1387–1395.
[64]. A. Beznosikov, G. Scutari, A. Rogozin, and A. Gasnikov, “Distributed saddle-point problems under data similarity,” Advances in Neural Information Processing Systems, vol. 34, pp. 8172–8184, 2021.
[65]. A. Beznosikov, P. Dvurechenskii, A. Koloskova, V. Samokhin, S. U. Stich, and A. Gasnikov, “Decentralized local stochastic extra-gradient for variational inequalities,” Advances in Neural Information Processing Systems, vol. 35, pp. 38116–38133, 2022.
[66]. A. Beznosikov, P. Richtárik, M. Diskin, M. Ryabinin, and A. Gasnikov, “Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees,” Advances in Neural Information Processing Systems, vol. 35, pp. 14013–14029, 2022.
[67]. A. Beznosikov and A. Gasnikov, “Compression and data similarity: Combination of two techniques for communication-efficient solving of distributed variational inequalities,” in International conference on optimization and applications, Springer, 2022, pp. 151–162.
[68]. A. Beznosikov, M. Takác, and A. Gasnikov, “Similarity, compression and local steps: Three pillars of efficient communications for distributed variational inequalities,” Advances in Neural Information Processing Systems, vol. 36, 2024.

[69]. N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien, “Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity,” Advances in Neural Information Processing Systems, vol. 34, pp. 19095–19108, 2021.
[70]. A. Beznosikov, E. Gorbunov, H. Berard, and N. Loizou, “Stochastic gradient descent-ascent: Unified theory and new efficient methods,” in International conference on artificial intelligence and statistics, PMLR, 2023, pp. 172–235.
[71]. E. Gorbunov, F. Hanzely, and P. Richtárik, “A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent,” in International conference on artificial intelligence and statistics, PMLR, 2020, pp. 680–690.
[72]. H. Robbins and S. Monro, “A stochastic approximation method,” The annals of mathematical statistics, pp. 400–407, 1951.
[73]. E. Moulines and F. Bach, “Non-asymptotic analysis of stochastic approximation algorithms for machine learning,” Advances in neural information processing systems, vol. 24, 2011.
[74]. B. Jacob et al., “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2704–2713.
[75]. R. Krishnamoorthi, “Quantizing deep convolutional networks for efficient inference: A whitepaper,” arXiv preprint arXiv:1806.08342, 2018.
[76]. H. Wu, P. Judd, X. Zhang, M. Isaev, and P. Micikevicius, “Integer quantization for deep learning inference: Principles and empirical evaluation,” arXiv preprint arXiv:2004.09602, 2020.

Информация об авторах / Information about authors

Даниил Олегович МЕДЯКОВ – магистрант Московского физико-технического института, лаборант лаборатории проблем федеративного обучения, старший лаборант исследовательского центра доверенного искусственного интеллекта Института системного программирования РАН. Сфера научных интересов: стохастическая оптимизация, распределенное и федеративное обучение.

Daniil Olegovich MEDYAKOV – master student at Moscow Institute of Physics and Technology, laboratory assistant at the Laboratory of Federated Learning Problems, senior laboratory assistant at the Research Center of Trusted Artificial Intelligence of the Institute for System Programming of the Russian Academy of Sciences. Research interests: stochastic optimization, distributed and federated learning.

Глеб Львович МОЛОДЦОВ – магистрант Московского физико-технического института, лаборант лаборатории проблем федеративного обучения Института системного программирования РАН. Сфера научных интересов: стохастическая оптимизация, распределенное и федеративное обучение.

Gleb Lvovich MOLODTSOV – master student at Moscow Institute of Physics and Technology, laboratory assistant at the Laboratory of Federated Learning Problems of the Institute for System Programming of the Russian Academy of Sciences. Research interests: stochastic optimization, distributed and federated learning.

Александр Николаевич БЕЗНОСИКОВ – кандидат физико-математических наук, заведующий лабораторией проблем федеративного обучения Института системного программирования РАН, доцент кафедры математических основ управления Московского физико-технического института. Сфера научных интересов: стохастическая оптимизация, распределенное и федеративное обучение.

Aleksandr Nikolaevich BEZNOSIKOV – Cand. Sci. (Phys.-Math.), Head of the Laboratory of Federated Learning Problems of the Institute for System Programming of the Russian Academy of Sciences, Associate Professor of the Department of Mathematical Foundations of Management for the Moscow Institute of Physics and Technology. Research interests: stochastic optimization, distributed and federated learning.