Представление Кашина для распределённого обучения

Егор Шульгин

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

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

Москва 2021

Распределённое обучение: $\min\limits_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

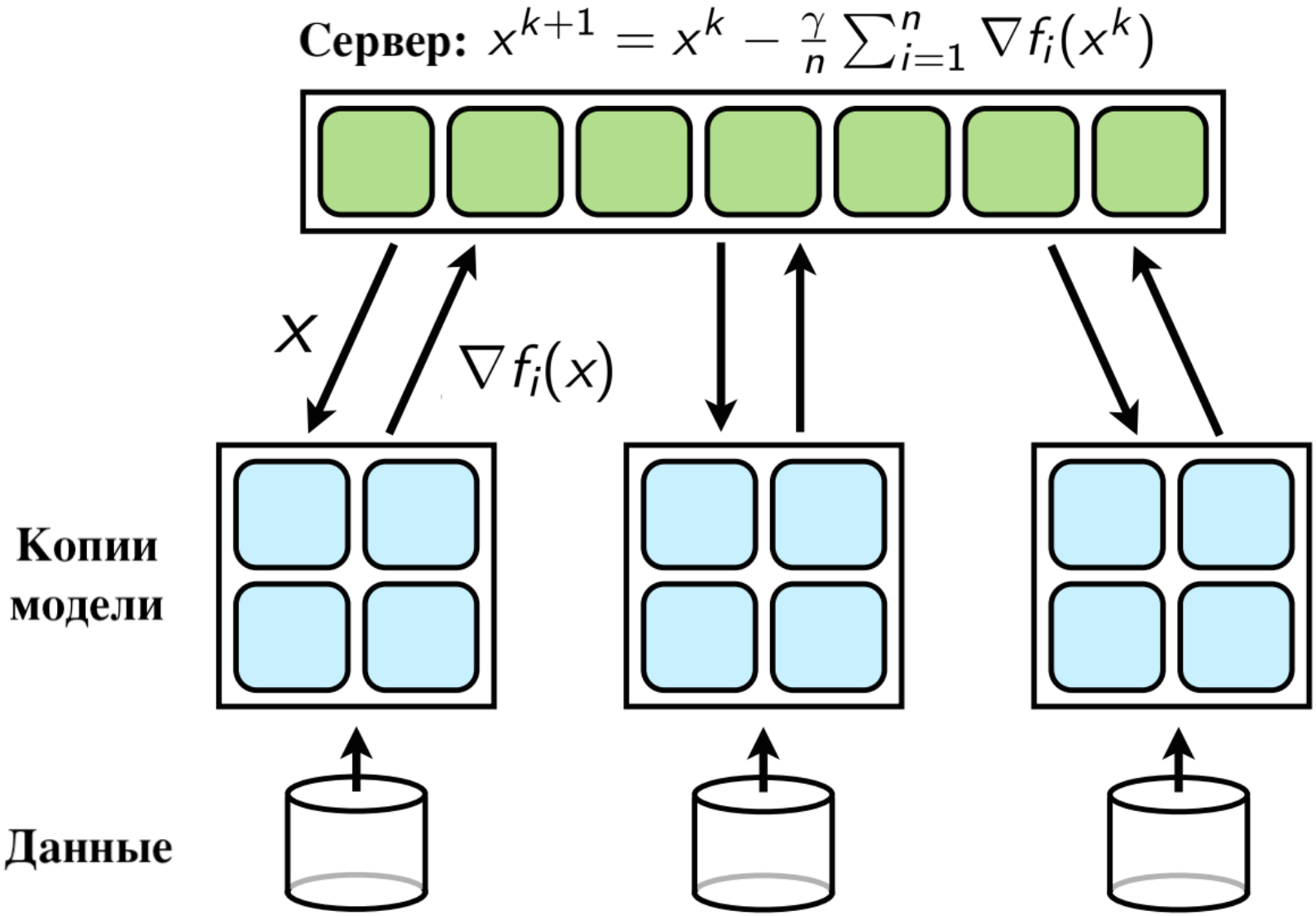


Рис.: Схема распределённого градиентного спуска

Узкое место: отправка $\nabla f_i(x)$ на сервер

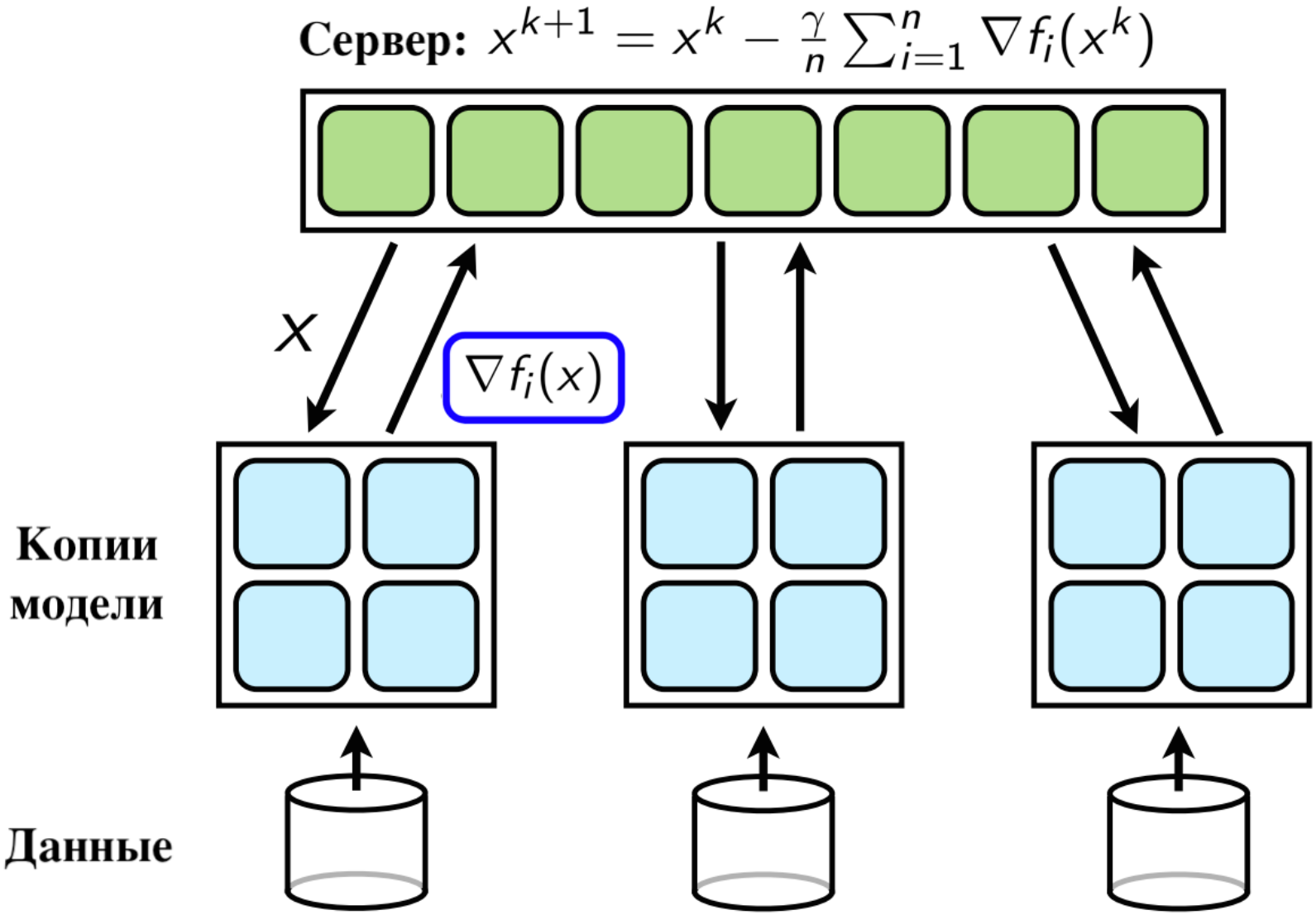


Рис.: Схема распределённого градиентного спуска

Исследуемый подход: Компрессия сообщений

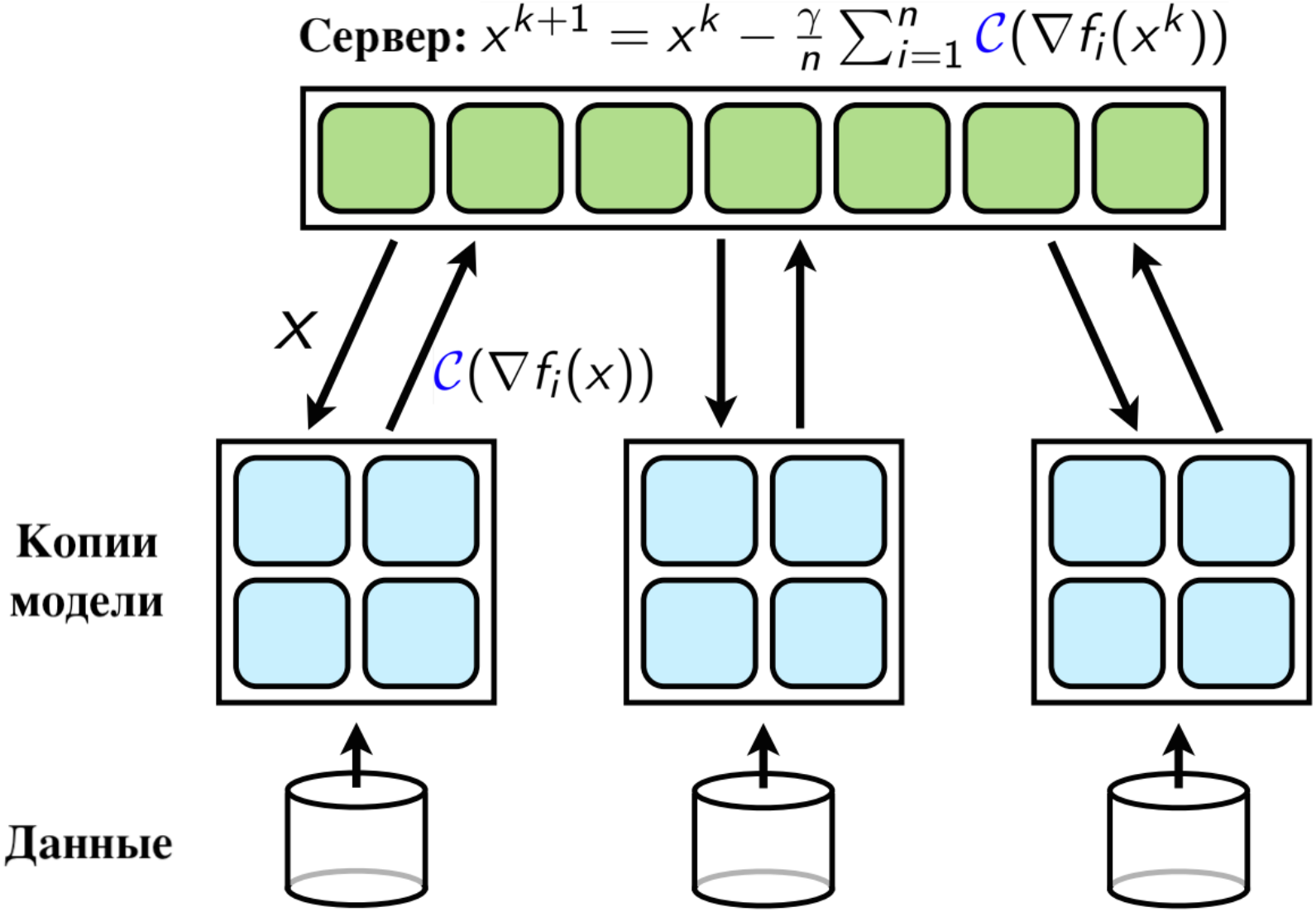


Рис.: Схема распределённого градиентного спуска с компрессией

Основные работы по теме

Boris Kashin
Diameters of some finite-dimensional sets and classes of smooth functions.
Jour. Izv. Akad. Nauk SSSR Ser. Mat. 41 (1977), Nr. 2, S. 334–351.

Yurii Lyubarskii and Roman Vershynin
Uncertainty Principles and Vector Quantization.
IEEE Trans. Inf. Theor. 56 (2010), Juli, Nr. 7, S. 3491–3501.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka
QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding.
Advances in Neural Information Processing Systems 30 (2017), S. 1709–1720.

Mher Safaryan, Egor Shulgin, Peter Richtárik
Uncertainty Principle for Communication Compression in Distributed and Federated Learning and the Search for an Optimal Compressor.
Information and Inference: A Journal of the IMA (2021), iaab006.

Компрессия сообщений

Определение (Несмещённый компрессор)

Стохастическое отображение $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ является *несмещённым компрессором* с "дисперсией" $\omega > 0$ если $\forall x \in \mathbb{R}^d$ выполнено

1. $\mathrm{E}[\mathcal{C}(x)] = x$, [несмещённость]
2. $\mathrm{E}\|\mathcal{C}(x) - x\|^2 \leq \omega \|x\|^2$, [ограниченная "ошибка компрессии"]

Пример: оператор случайного прореживания (`Rand-K`)

Компрессия сообщений

Определение (Несмещённый компрессор)

Стохастическое отображение $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ является *несмещённым компрессором* с "дисперсией" $\omega > 0$ если $\forall x \in \mathbb{R}^d$ выполнено

1. $\mathrm{E}[\mathcal{C}(x)] = x$, [несмещённость]
2. $\mathrm{E}\|\mathcal{C}(x) - x\|^2 \leq \omega\|x\|^2$, [ограниченная "ошибка компрессии"]

Пример: оператор случайного прореживания (`Rand-K`)

Число итераций (сложность) Градиентного Спуска с компрессией зависит от ω линейно $\mathcal{O}(\omega\kappa \log 1/\varepsilon)$ для сильно выпуклых функций.

Компрессия сообщений

Определение (Несмещённый компрессор)

Стохастическое отображение $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ является *несмещённым компрессором* с "дисперсией" $\omega > 0$ если $\forall x \in \mathbb{R}^d$ выполнено

1. $\mathrm{E}[\mathcal{C}(x)] = x$, [несмещённость]
2. $\mathrm{E}\|\mathcal{C}(x) - x\|^2 \leq \omega\|x\|^2$, [ограниченная "ошибка компрессии"]

Пример: оператор случайного прореживания (Rand-K)

Число итераций (сложность) Градиентного Спуска с компрессией зависит от ω линейно $\mathcal{O}(\omega\kappa \log 1/\varepsilon)$ для сильно выпуклых функций.

Метод компрессии	Дисперсия ω	Число бит
Rand-K	$\approx \mathcal{O}(\frac{d}{k})$	$k + \log_2 \binom{d}{k}$
Стохастическая квантизация	$\approx \mathcal{O}(\frac{\sqrt{d}}{s})$	$d \log_2(2s+1)$
Тернарная квантизация	$\approx \mathcal{O}(\sqrt{d})$	$d \log_2 3$
Компрессор Кашина	$\approx \boldsymbol{O(1)}$	$\lambda d \log_2 3$

Системы векторов I

Ортогональная система

Пусть $\{e_1, e_2, \ldots, e_d\}$ – произвольный ортогональный базис в $\mathbb{R}^d$.
Тогда *ортогональное разложение* вектора $x \in \mathbb{R}^d$ определяется как

$$x = \sum_{i=1}^{d} x_i e_i, \qquad x_i = \langle x, e_i \rangle \tag{1}$$

Замечания

- Коэффициенты x_i **независимы**: в случае потери одного из них, вектор x не может быт восстановлен даже приблизительно
- Коэффициенты x_i **не равнозначны**: некоторые из них могут содержать больше информации о векторе x чем другие и быть **более чувствительными к компрессии**.

Системы векторов II

Полная система (Frame Representation)

Система векторов $\{u_1, u_2, \ldots, u_D\} \subset \mathbb{R}^d$ для $D \geq d$ образует *полную систему* если каждый вектор $x \in \mathbb{R}^d$ может быть представлен в виде

$$x = \sum_{i=1}^{D} a_i u_i, \qquad a_i = \langle x, u_i \rangle \tag{2}$$

Замечания

- Если $D > d$, система векторов является линейно зависимой, поэтому представление (2) не единственно.
- Избыточность позволяет выбирать коэффициенты a_i таким образом, чтобы информация была распределена равномерно.
- Обозначим за $U \in \mathbb{R}^{d \times D}$ матрицу, составленную из векторов u_i. Тогда разложение (2) принимает форму $x = Ua$.

Представление Кашина

Представление Кашина

Пусть $\{u_i\}_{i=1}^D$ – полная система в $\mathbb{R}^d$. *Представление Кашина* для вектора $x \in \mathbb{R}^d$ с уровнем K определяется как

$$x = \sum_{i=1}^{D} a_i u_i, \qquad \max_{1 \le i \le D} |a_i| \le \frac{K}{\sqrt{D}} \|x\|_2 \tag{3}$$

- **Оптимальность.** Представление Кашина обладает наименьшим динамическим диапазоном $K/\sqrt{D}$, который в $\sqrt{d}$ раз меньше ортогонального представления (2).

Представление Кашина

Представление Кашина

Пусть $\{u_i\}_{i=1}^D$ – полная система в $\mathbb{R}^d$. *Представление Кашина* для вектора $x \in \mathbb{R}^d$ с уровнем K определяется как

$$x = \sum_{i=1}^{D} a_i u_i, \qquad \max_{1 \le i \le D} |a_i| \le \frac{K}{\sqrt{D}} \|x\|_2 \tag{3}$$

- **Существование.** Не каждая полная система векторов может гарантировать представление Кашина с постоянным уровнем K.

Теорема (Кашин, 1977)

Существуют полные системы векторов $\{u_i\}_{i=1}^D \subset \mathbb{R}^d$ с произвольно малой избыточностью $\lambda = D/d > 1$, и такие что любой вектор $x \in \mathbb{R}^d$ имеет представление Кашина $\{a_i\}_{i=1}^D$ с уровнем $K = K(\lambda)$.

Компрессор Кашина

Идея: Комбинация представления Кашина с квантизацией.

Компрессор Кашина

Пусть коэффициент избыточности $\lambda > 1$ такой что $D = \lambda d \in \mathbb{N}$, а несмещённый компрессор $\mathcal{C} : \mathbb{R}^D \to \mathbb{R}^D$ сохраняет знак и максимальную амплитуду (например, квантизатор):

$$0 \leq \mathcal{C}(a) \times \text{sign}(a) \leq \|a\|_\infty, \qquad a \in \mathbb{R}^D \tag{4}$$

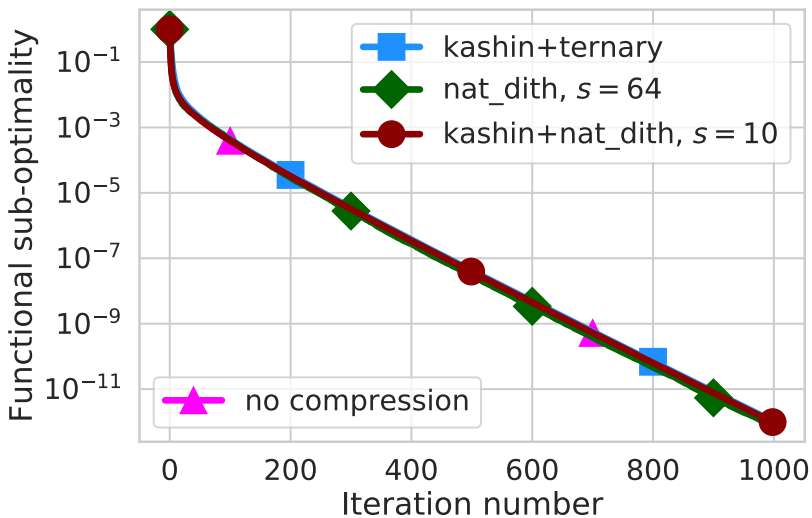
Тогда **компрессия Кашина** вектора $x \in \mathbb{R}^d$ определяется как

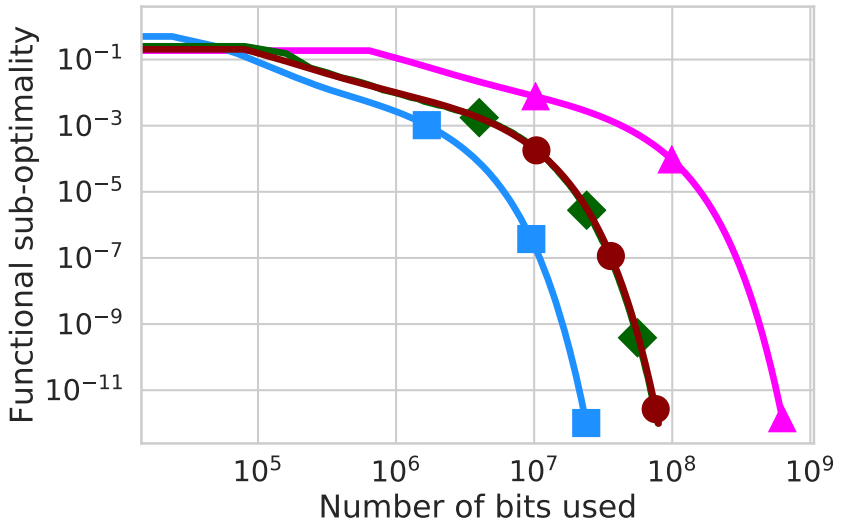
$$\mathcal{C}_\kappa(x) = U \cdot \mathcal{C}(a) \tag{5}$$

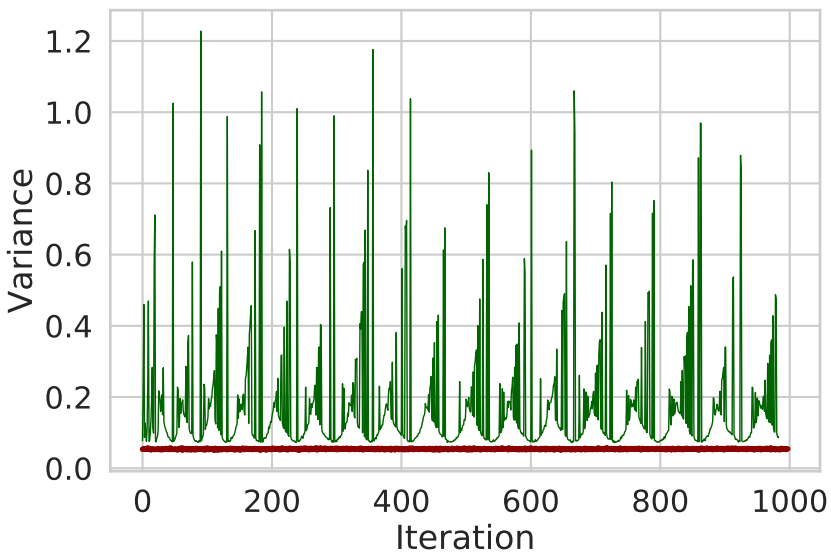
где $a \in \mathbb{R}^D$ - вектор из коэффициентов Кашина для x, что $x = U \cdot a$.

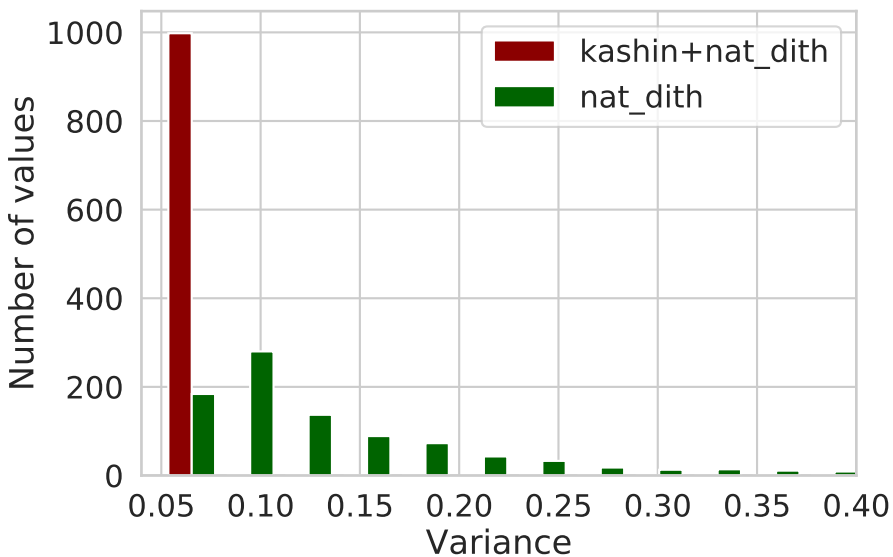
Экспериментальные результаты I: $\min_x\{x^\top Ax - b^\top x\}$

Градиентный спуск с компрессией (CGD): $x^{k+1} = x^k - \gamma\mathcal{C}(\nabla f(x^k))$







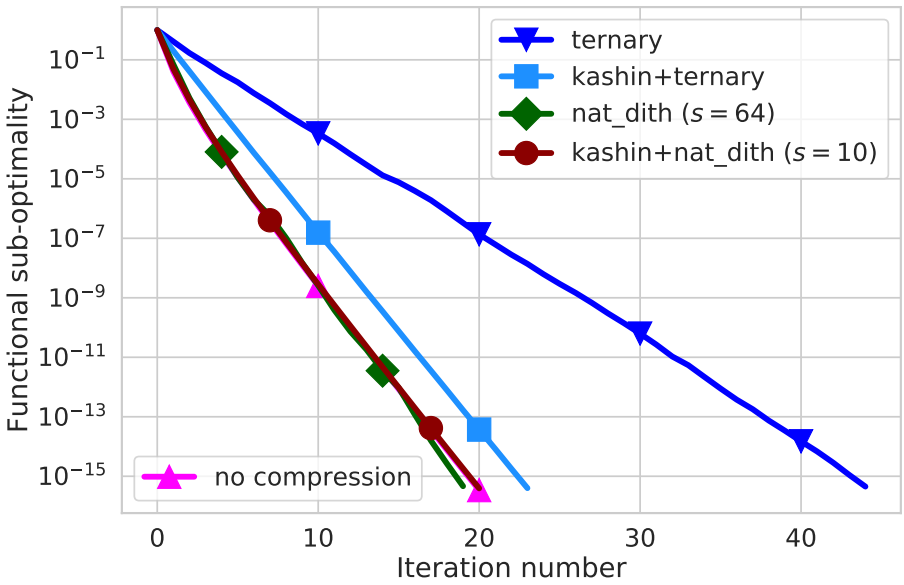


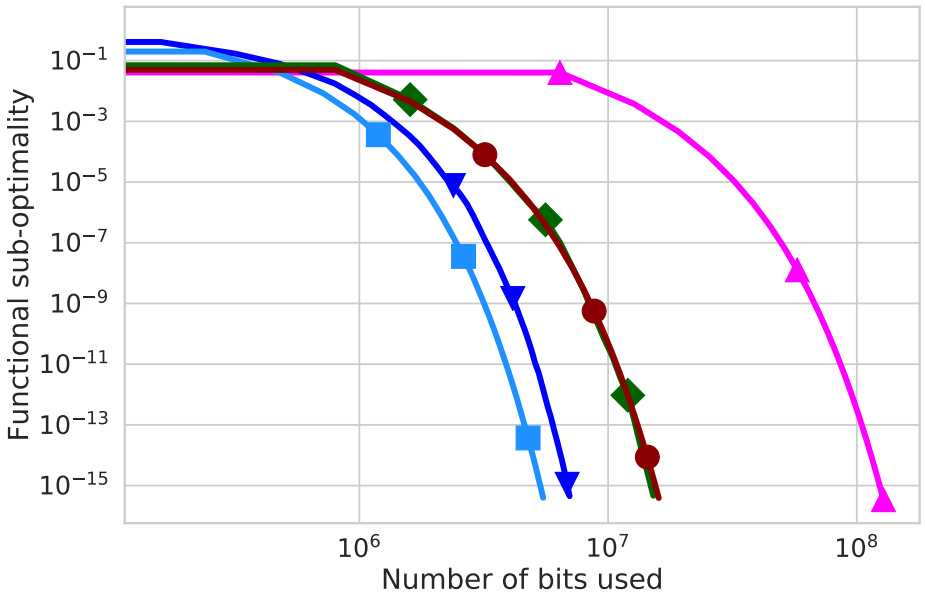
Нормализованная эмпирическая дисперсия: $\omega(x) = \|\mathcal{C}(x) - x\|^2/\|x\|^2$

Экспериментальные результаты II: $\min_x \frac{1}{n}\sum_{i=1}^n x^\top A_i x$

Распределённый Градиентный спуск с компрессией:

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^n \mathcal{C}(\nabla f_i(x^k)) \qquad \text{(DCGD)}$$





Основные выводы

Сочетание представления Кашина с Тернарной квантизацией позволяет минимизировать число переданных бит и практически не замедляет сходимость по итерациям.

Результаты, выносимые на защиту

Разработка

- алгоритма для вычисления представления Кашина,
- методов оптимизации с компрессией,
- различных несмещённых векторных компрессоров.

Экспериментальное сравнение

представления Кашина и ортогонального в сочетании с различными методами квантизации на задачах

- сжатия случайных гауссовских векторов,
- оптимизации квадратичных функций.

Публикации в соавторстве за время обучения

[1] E. Shulgin, P. Richtárik
Shifted Compression Framework: Generalizations and Improvements.
Подано, 2021.

[2] D. Kovalev, E. Shulgin, P. Richtárik, A. Rogozin, A. Gasnikov
ADOM: Accelerated Decentralized Optimization Method for Time-Varying Networks.
Proceedings of the 38th International Conference on Machine Learning, 2021.

[3] M. Safaryan, E. Shulgin, P. Richtárik
Uncertainty Principle for Communication Compression in Distributed and Federated Learning and the Search for an Optimal Compressor.
Information and Inference: A Journal of the IMA (2021), iaab006.

[4] A. Rogozin, V. Lukoshkin, A. Gasnikov, D. Kovalev, E. Shulgin
Towards accelerated rates for distributed optimization over time-varying networks.
Preprint arXiv:1911.11271, 2020.

[5] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, Y. Malitsky
Revisiting Stochastic Extragradient.
Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 2020.

[6] A. Ivanova, D. Pasechnyuk, D. Grishchenko, E. Shulgin, A. Gasnikov, V. Matyukhin
Adaptive Catalyst for smooth convex optimization.
Preprint arXiv:1911.11271, 2019.