# Распределённая оптимизация

## Математика Больших Данных

Александр Безносиков

МФТИ

2 ноября 2022

# План лекции

- Постановка задачи и актуальность
- Организация коммуникаций
- Сжатие: несмещенные и смещенные операторы
- Data similarity

# Постановка задачи и актуальность

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? GD, Accelerated GD, SGD, Adam и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? `GD`, `Accelerated GD`, `SGD`, `Adam` и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? `GD`, `Accelerated GD`, `SGD`, `Adam` и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? `GD`, `Accelerated GD`, `SGD`, `Adam` и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? `GD`, `Accelerated GD`, `SGD`, `Adam` и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? `GD`, `Accelerated GD`, `SGD`, `Adam` и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Задача машинного обучения

- Сформулируем задачу машинного обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i).$$

- **Вопрос:** как решать? `GD`, `Accelerated GD`, `SGD`, `Adam` и так далее.
- Объемы данных растут, одно вычислительное устройство может долго считать даже стохастический градиент.
- **Вопрос:** что делать? Распараллеливать процесс обучения. Использовать несколько вычислительных устройств.
- **Вопрос:** как это сделать? Распределить данные между устройствами/агентами/нодами.

# Распределенное обучение

- Распределенная задача обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) = \frac{1}{M} \sum_{m=1}^{M} \left[ \frac{1}{N_m} \sum_{i=1}^{N} \ell(x, z_i) \right],$$

  где $N = N_1 + \ldots + N_M$.
- Кластерные вычисления: можем разделить данные равномерно ($N_m \approx N_j$), локальные функции потерь будут в некоторой степени похожи между собой.
- Коллаборативные вычисления на открытых данных: можем разделить данные (но скорее всего неравномерно по количеству, но равномерно по природе).
- Федеративное обучение.

# Распределенное обучение

- Распределенная задача обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) = \frac{1}{M} \sum_{m=1}^{M} \left[ \frac{1}{N_m} \sum_{i=1}^{N} \ell(x, z_i) \right],$$

  где $N = N_1 + \ldots + N_M$.
- Кластерные вычисления: можем разделить данные равномерно ($N_m \approx N_j$), локальные функции потерь будут в некоторой степени похожи между собой.
- Коллаборативные вычисления на открытых данных: можем разделить данные (но скорее всего неравномерно по количеству, но равномерно по природе).
- Федеративное обучение.

# Распределенное обучение

- Распределенная задача обучения:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) = \frac{1}{M} \sum_{m=1}^{M} \left[ \frac{1}{N_m} \sum_{i=1}^{N} \ell(x, z_i) \right],$$

  где $N = N_1 + \ldots + N_M$.
- Кластерные вычисления: можем разделить данные равномерно ($N_m \approx N_j$), локальные функции потерь будут в некоторой степени похожи между собой.
- Коллаборативные вычисления на открытых данных: можем разделить данные (но скорее всего неравномерно по количеству, но равномерно по природе).
- Федеративное обучение.

# Федеративное обучение

- Вычислительные устройства – пользовательские устройства: ноутбуки, планшеты, телефоны. Неравномерность вычислительных мощностей.
- Данные часто сильно разнородны, как по размеру так и по природе. Дополнительные проблемы приватности.
- Желание пользователей часто отличается друг от друга и от желания владельца сервиса.

# Федеративное обучение

- Вычислительные устройства – пользовательские устройства: ноутбуки, планшеты, телефоны. Неравномерность вычислительных мощностей.
- Данные часто сильно разнородны, как по размеру так и по природе. Дополнительные проблемы приватности.
- Желание пользователей часто отличается друг от друга и от желания владельца сервиса.

# Федеративное обучение

- Вычислительные устройства – пользовательские устройства: ноутбуки, планшеты, телефоны. Неравномерность вычислительных мощностей.
- Данные часто сильно разнородны, как по размеру так и по природе. Дополнительные проблемы приватности.
- Желание пользователей часто отличается друг от друга и от желания владельца сервиса.

# Коммуникации – главное узкое место

- Выигрываем в параллельности, но платим за это тратами на коммуникации. Хотелось бы плату эту уменьшить.
- Проблема присутствует во всех подходах от кластерного (с подключением по кабелю) до федеравтиного обучения (с неустойчивым интернетом).

# Коммуникации – главное узкое место

- Выигрываем в параллельности, но платим за это тратами на коммуникации. Хотелось бы плату эту уменьшить.
- Проблема присутствует во всех подходах от кластерного (с подключением по кабелю) до федеравтиного обучения (с неустойчивым интернетом).

# Организация коммуникаций

# Типы коммуникационных архитектур: теория

- Централизованная: можем получить **точное** усреднение по всем устройствам (возможно, с задержками, сбоями и так далее)
- Децентрализованная: **точное** усреднение не предусмотрено

# Типы коммуникационных архитектур: теория

- Централизованная: можем получить **точное** усреднение по всем устройствам (возможно, с задержками, сбоями и так далее)
- Децентрализованная: **точное** усреднение не предусмотрено

# Типы коммуникационных архитектур: централизованная

- Посмотрим на примере, как обычный неопределенный GD становится централизованным.

**Алгоритм 1** Централизованный GD

**Вход:** Размер шага $\gamma > 0$, стартовая точка $x_0 \in \mathbb{R}^d$, количество итераций $K$

**for** $k = 0, 1, \dots, K-1$ **do**
    Отправить $x_k$ всем рабочим ▷ выполняется сервером
    **for** $i = 1, \dots, n$ параллельно **do**
        Принять $x_k$ от мастера ▷ выполняется рабочими
        Вычислить градиент $\nabla f_m(x_k)$ в точке $x_k$ ▷ выполняется рабочими
        Отправить $\nabla f_m(x_k)$ мастеру ▷ выполняется рабочими
    **end for**
    Принять $\nabla f_m(x_k)$ от всех рабочих ▷ выполняется сервером
    Вычислить $\nabla f(x_k) = \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(x_k)$ ▷ выполняется сервером
    $x_{k+1} = x_k - \gamma \nabla f(x_k)$ ▷ выполняется сервером
**end for**

**Выход:** $x^K$

# Типы коммуникационных архитектур: практика

- Централизованная: есть некоторая машина-сервер (мастер), с которой соединены все остальные устройства (рабочие).
- Архитектура с AllReduce процедурой (в теории это централизованная): задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, в том чсиле можно организовать усреднение (в теории это также централизованная архитектура).
- Децентрализованная: задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, точное усреднение не используется.

## Типы коммуникационных архитектур: практика

- Централизованная: есть некоторая машина-сервер (мастер), с которой соединены все остальные устройства (рабочие).
- Архитектура с AllReduce процедурой (в теории это централизованная): задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, в том чсиле можно организовать усреднение (в теории это также централизованная архитектура).
- Децентрализованная: задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, точное усреднение не используется.

# Типы коммуникационных архитектур: практика

- Централизованная: есть некоторая машина-сервер (мастер), с которой соединены все остальные устройства (рабочие).
- Архитектура с AllReduce процедурой (в теории это централизованная): задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, в том чсиле можно организовать усреднение (в теории это также централизованная архитектура).
- Децентрализованная: задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, точное усреднение не используется.

# Базовый вариант: Butterfly AllReduce

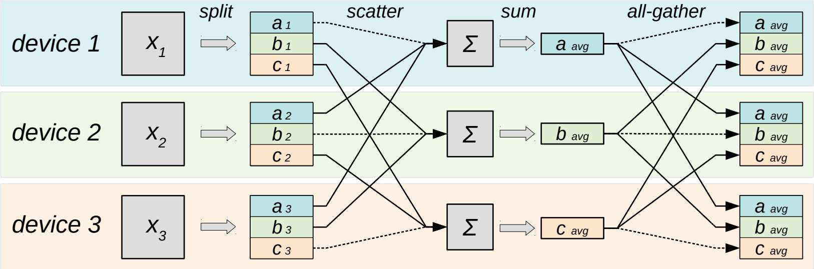


Figure: Картинка из Ryabinin M. Moshpit SGD: Communication-Efficient Decentralized Training on Heterogeneous Unreliable Devices

# Более практичный вариант: Ring AllReduce

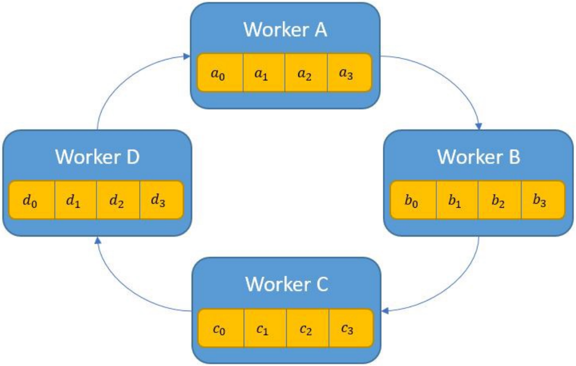


Figure: Картинка отсюда

# Ring AllReduce: первый шаг суммирования

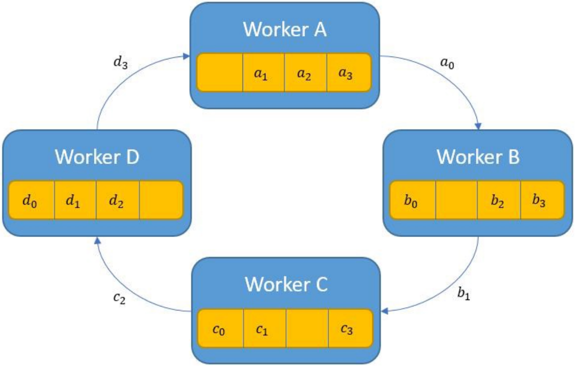


Figure: Картинка отсюда

# Ring AllReduce: второй шаг суммирования

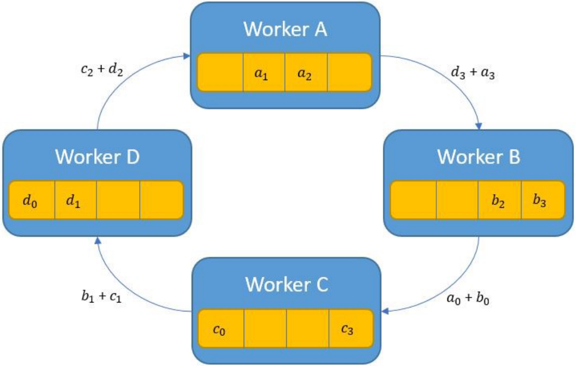


Figure: Картинка отсюда

# Ring AllReduce: первый шаг распространения

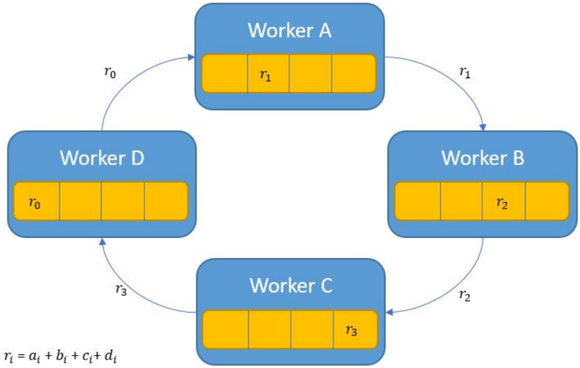


Figure: Картинка отсюда

# Ring AllReduce: итог

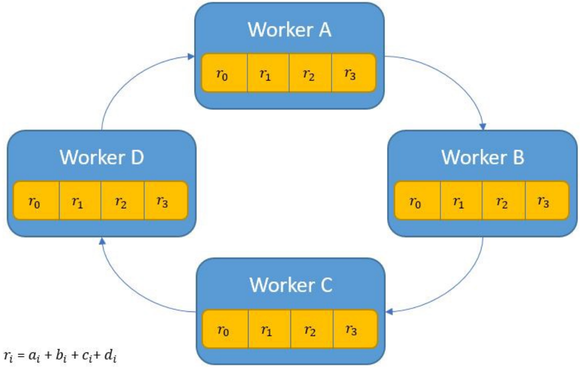


Figure: Картинка отсюда

# Сжатие: несмещенные и смещенные операторы

# Несмещённая компрессия (квантизация)

## Несмещённая компрессия (квантизация)

Будем называть стохастический оператор $\mathcal{Q}(x)$ оператором несмещённой компрессии (квантизации), если для любого $x \in \mathbb{R}^d$ выполняется:

$$\mathbb{E}[\mathcal{Q}(x)] = x, \quad \mathbb{E}[\|\mathcal{Q}(x)\|_2^2] \leq \omega \|x\|_2^2,$$

где $\omega \geq 1$.

# Несмещённая компрессия: примеры

## Случайная спарсификация (выбор случайных компонент)

Рассмотрим стохастический оператор

$$\mathrm{Rand}_k(x) = \frac{d}{k} \sum_{i \in S} x_i e_i,$$

где $k$ — некоторое фиксированное число из множества $\{1, \ldots, d\}$ (количество компонент вектора $x$, которые мы передаём; например, можно выбрать $k = 1$), $S$ — случайное подмножество множества $\{1, \ldots, d\}$ размера $k$ (подмножество $S$ выбирается случайно и равновероятно среди всех возможных подмножеств размера $d$), $(e_1, \ldots, e_d)$ — стандартный базис в $\mathbb{R}^d$. Можно показать, что данный оператор является несмещённой компрессией с константой $\omega = \frac{d}{k}$.

Richtárik P. and Takáč M. Parallel coordinate descent methods for big data optimization

- **Вопрос:** зачем нужен множитель $\frac{d}{k}$? Для несмещенности

# Несмещённая компрессия: примеры

## Случайная спарсификация (выбор случайных компонент)

Рассмотрим стохастический оператор

$$\mathrm{Rand}_k(x) = \frac{d}{k} \sum_{i \in S} x_i e_i,$$

где $k$ — некоторое фиксированное число из множества $\{1, \ldots, d\}$ (количество компонент вектора $x$, которые мы передаём; например, можно выбрать $k = 1$), $S$ — случайное подмножество множества $\{1, \ldots, d\}$ размера $k$ (подмножество $S$ выбирается случайно и равновероятно среди всех возможных подмножеств размера $d$), $(e_1, \ldots, e_d)$ — стандартный базис в $\mathbb{R}^d$. Можно показать, что данный оператор является несмещённой компрессией с константой $\omega = \frac{d}{k}$.

PDF Richtárik P. and Takáč M. Parallel coordinate descent methods for big data optimization

- **Вопрос:** зачем нужен множитель $\frac{d}{k}$? Для несмещенности

# Несмещённая компрессия: примеры

## Случайная спарсификация (выбор случайных компонент)

Рассмотрим стохастический оператор

$$\mathrm{Rand}_k(x) = \frac{d}{k} \sum_{i \in S} x_i e_i,$$

где $k$ — некоторое фиксированное число из множества $\{1, \ldots, d\}$ (количество компонент вектора $x$, которые мы передаём; например, можно выбрать $k = 1$), $S$ — случайное подмножество множества $\{1, \ldots, d\}$ размера $k$ (подмножество $S$ выбирается случайно и равновероятно среди всех возможных подмножеств размера $d$), $(e_1, \ldots, e_d)$ — стандартный базис в $\mathbb{R}^d$. Можно показать, что данный оператор является несмещённой компрессией с константой $\omega = \frac{d}{k}$.

PDF Richtárik P. and Takáč M. Parallel coordinate descent methods for big data optimization

- **Вопрос:** зачем нужен множитель $\frac{d}{k}$? Для несмещенности.

# Несмещённая компрессия: примеры

## Трёхуровневая $\ell_2$-квантизация

Рассмотрим следующий оператор: $[\mathcal{Q}(x)]_i = \|x\|_2 \text{sign}(x_i)\xi_i, \quad i = 1, \ldots, d$, где $[\mathcal{Q}(x)]_i$ — $i$-я компонента вектора $\mathcal{Q}(x)$ и $\xi_i$ — случайная величина, имеющая распределение Бернулли с параметром $\frac{|x_i|}{\|x\|_2}$, то есть

$$\xi_i = \begin{cases} 1 & \text{с вероятностью } \frac{|x_i|}{\|x\|_2}, \\ 0 & \text{с вероятностью } 1 - \frac{|x_i|}{\|x\|_2}. \end{cases}$$

Таким образом, если мы хотим передать вектор $\mathcal{Q}(x)$, то нам нужно передать вектор, состоящий из нулей и $\pm 1$, и вещественное число $\|x\|_2$, причём вероятность обнуления компоненты тем больше, чем компонента меньше по модулю по сравнению с остальными компонентами. Можно показать, что данный оператор является несмещённой компрессией с константой $\omega = \sqrt{d}$.

PDF Alistarh D. et al. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding

# Несмещенная компрессия: примеры

- Еще примеры несмещенных компрессией:

  Beznosikov A. et al. On Biased Compression for Distributed Learning

  Horváth S. et al. Natural compression for distributed deep learning

  Szlendak R. et al. Permutation Compressors for Provably Faster Distributed Nonconvex Optimization

# Несмещенная компрессия: идея

- Самая простая идея, которая приходит в голову, состоит в том, чтобы использовать параллельный GD, но к градиентам, пересылаемым от рабочих на сервер, применять несмещённую компрессию.

# Несмещенная компрессия: идея

- Самая простая идея, которая приходит в голову, состоит в том, чтобы использовать параллельный GD, но к градиентам, пересылаемым от рабочих на сервер, применять несмещённую компрессию.

# Квантизированный GD (QGD)

**Алгоритм 1** `QGD`

**Вход:** размер шага $\gamma > 0$, стартовая точка $x_0 \in \mathbb{R}^d$, количество итераций $K$

**for** $k = 0, 1, \ldots, K-1$ **do**
  Отправить $x_k$ всем рабочим ▷ выполняется сервером
  **for** $i = 1, \ldots, n$ параллельно **do**
    Принять $x_k$ от мастера ▷ выполняется рабочими
    Вычислить градиент $\nabla f_m(x_k)$ в точке $x_k$ ▷ выполняется рабочими
    Независимо сгенерировать $g_{k,m} = \mathcal{Q}(\nabla f_m(x_k))$ ▷ выполняется рабочими
    Отправить $g_{k,m}$ мастеру ▷ выполняется рабочими
  **end for**
  Принять $g_{k,m}$ от всех рабочих ▷ выполняется сервером
  Вычислить $g_k = \frac{1}{M}\sum_{m=1}^{M} g_{k,m}$ ▷ выполняется сервером
  $x_{k+1} = x_k - \gamma g_k$ ▷ выполняется сервером
**end for**

**Выход:** $x^K$

# Несмещенная компрессия: доказательство

- Будем доказывать в случае, когда все $f_m$ являются $L$-гладкими и $\mu$-сильно выпуклыми.
- Рассмотрим одну итерацию метода:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\gamma\langle g_k, x_k - x^*\rangle + \|g_k\|^2.$$

- Берем условное мат.ожидание по случайности только на итерации $k$:

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid x_k\right] = \|x_k - x^*\|^2 - 2\gamma\langle \mathbb{E}\left[g_k \mid x_k\right], x_k - x^*\rangle + \gamma^2\mathbb{E}\left[\|g_k\|^2 \mid x_k\right].$$

# Несмещенная компрессия: доказательство

- Будем доказывать в случае, когда все $f_m$ являются $L$-гладкими и $\mu$-сильно выпуклыми.
- Рассмотрим одну итерацию метода:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\gamma\langle g_k, x_k - x^*\rangle + \|g_k\|^2.$$

- Берем условное мат.ожидание по случайности только на итерации $k$:

$$\begin{aligned} \mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid x_k\right] =& \|x_k - x^*\|^2 - 2\gamma\langle \mathbb{E}\left[g_k \mid x_k\right], x_k - x^*\rangle \\ &+ \gamma^2\mathbb{E}\left[\|g_k\|^2 \mid x_k\right]. \end{aligned}$$

# Несмещенная компрессия: доказательство

- Работаем с $\mathbb{E}\left[g_k \mid x_k\right]$:

$$\mathbb{E}\left[g_k \mid x_k\right] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[g_{k,m} \mid x_k\right]$$

$$= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\mathbb{E}\left[\mathcal{Q}(\nabla f_m(x_k)) \mid \nabla f_m(x_k)\right] \mid x_k\right]$$

$$= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\nabla f_m(x_k) \mid x_k\right] = \frac{1}{M}\sum_{m=1}^{M}\nabla f_m(x_k) = \nabla f(x_k).$$

- Работаем с $\mathbb{E}\left[\|g_k\|^2 \mid x_k\right]$:

$$\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] = \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x_k\right] = \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x^k\right].$$

## Несмещенная компрессия: доказательство

- Работаем с $\mathbb{E}\left[g_k \mid x_k\right]$:

$$\begin{aligned}\mathbb{E}\left[g_k \mid x_k\right] &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[g_{k,m} \mid x_k\right]\\ &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\mathbb{E}\left[\mathcal{Q}(\nabla f_m(x_k)) \mid \nabla f_m(x_k)\right] \mid x_k\right]\\ &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\nabla f_m(x_k) \mid x_k\right] = \frac{1}{M}\sum_{m=1}^{M}\nabla f_m(x_k) = \nabla f(x_k).\end{aligned}$$

- Работаем с $\mathbb{E}\left[\|g_k\|^2 \mid x_k\right]$:

$$\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] = \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x_k\right] = \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x^k\right].$$

# Несмещенная компрессия: доказательство

- Работаем с $\mathbb{E}\left[g_k \mid x_k\right]$:

$$
\begin{aligned}
\mathbb{E}\left[g_k \mid x_k\right] &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[g_{k,m} \mid x_k\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\mathbb{E}\left[\mathcal{Q}(\nabla f_m(x_k)) \mid \nabla f_m(x_k)\right] \mid x_k\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\nabla f_m(x_k) \mid x_k\right] = \frac{1}{M}\sum_{m=1}^{M}\nabla f_m(x_k) = \nabla f(x_k).
\end{aligned}
$$

- Работаем с $\mathbb{E}\left[\|g_k\|^2 \mid x_k\right]$:

$$
\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] = \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x_k\right] = \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x^k\right].
$$

# Несмещенная компрессия: доказательство

- Работаем с $\mathbb{E}\left[g_k \mid x_k\right]$:

$$
\begin{aligned}
\mathbb{E}\left[g_k \mid x_k\right] &= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[g_{k,m} \mid x_k\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\mathbb{E}\left[\mathcal{Q}(\nabla f_m(x_k)) \mid \nabla f_m(x_k)\right] \mid x_k\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\nabla f_m(x_k) \mid x_k\right] = \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(x_k) = \nabla f(x_k).
\end{aligned}
$$

- Работаем с $\mathbb{E}\left[\|g_k\|^2 \mid x_k\right]$:

$$
\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] = \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x_k\right] = \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x^k\right].
$$

# Несмещенная компрессия: доказательство

- Продолжаем и применяем первое свойство (несмещенность) в определении компрессии:

$$
\begin{aligned}
\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] &= \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x^k\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|g_{k,m}\|^2 \mid x^k\right] \\
&\quad + \frac{2}{M^2}\sum_{m\neq l}\mathbb{E}\left[\langle g_{k,m}, g_{k,l}\rangle \mid x^k\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|g_{k,m}\|^2 \mid x^k\right] \\
&\quad + \frac{1}{M^2}\sum_{m\neq l}\mathbb{E}\left[\langle \mathbb{E}\left[g_{k,m} \mid \nabla f_m(x_k)\right], \mathbb{E}\left[g_{k,l} \mid \nabla f_l(x_k)\right]\rangle \mid x^k\right].
\end{aligned}
$$

# Несмещенная компрессия: доказательство

- Продолжаем и применяем первое свойство (несмещенность) в определении компрессии:

$$
\begin{aligned}
\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] &= \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M} g_{k,m}\right\|^2 \mid x^k\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|g_{k,m}\|^2 \mid x^k\right] \\
&\quad + \frac{2}{M^2}\sum_{m\neq l}\mathbb{E}\left[\langle g_{k,m}, g_{k,l}\rangle \mid x^k\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|g_{k,m}\|^2 \mid x^k\right] \\
&\quad + \frac{1}{M^2}\sum_{m\neq l}\mathbb{E}\left[\langle \mathbb{E}\left[g_{k,m} \mid \nabla f_m(x_k)\right], \mathbb{E}\left[g_{k,l} \mid \nabla f_l(x_k)\right]\rangle \mid x^k\right].
\end{aligned}
$$

# Несмещенная компрессия: доказательство

- Продолжаем и применяем второе свойство в определении компрессии:

$$
\begin{aligned}
\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] = & \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|\mathcal{Q}(\nabla f_m(x_k))\|^2 \mid x^k\right] \\
& + \frac{1}{M^2}\sum_{m\neq l}\langle \nabla f_m(x_k), \nabla f_l(x_k)\rangle \\
\leq & \frac{\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x_k)\|^2 \\
& + \|\nabla f(x_k)\|^2 \\
\leq & \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x_k) - \nabla f_m(x^*)\|^2 \\
& + \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 + \|\nabla f(x_k) - \nabla f(x^*)\|^2.
\end{aligned}
$$

# Несмещенная компрессия: доказательство

- Продолжаем и применяем второе свойство в определении компрессии:

$$
\begin{aligned}
\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] \leq & \frac{4\omega L}{M^2}\sum_{m=1}^{M}(f_m(x_k) - f_m(x^*) - \langle \nabla f_m(x^*), x_k - x^* \rangle) \\
& + \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 + 2L(f(x_k) - f(x^*)) \\
= & \frac{4\omega L}{M}(f(x_k) - f(x^*)) \\
& + \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 + 2L(f(x_k) - f(x^*)).
\end{aligned}
$$

# Несмещенная компрессия: доказательство

- Все, что получили:

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid x_k\right] = \|x_k - x^*\|^2 - 2\gamma\langle \mathbb{E}\left[g_k \mid x_k\right], x_k - x^*\rangle + \gamma^2 \mathbb{E}\left[\|g_k\|^2 \mid x_k\right].$$

$$\mathbb{E}\left[g_k \mid x_k\right] = \nabla f(x_k).$$

$$\mathbb{E}\left[\|g_k\|^2 \mid x_k\right] \leq \frac{4\omega L}{M}(f(x_k) - f(x^*)) + \frac{2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 + 2L(f(x_k) - f(x^*)).$$

# Несмещенная компрессия: доказательство

- Объединяем:

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid x_k\right] \leq \|x_k - x^*\|^2 - 2\gamma\langle \nabla f(x_k), x_k - x^*\rangle$$
$$+ 2\gamma^2 L \left(\frac{2\omega}{M} + 1\right)(f(x_k) - f(x^*))$$
$$+ \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 .$$

- Пользуемся сильной выпуклостью:

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid x_k\right] \leq \|x_k - x^*\|^2 - 2\gamma\left(\frac{\mu}{2}\|x_k - x^*\|^2 + f(x_k) - f(x^*)\right)$$
$$+ 2\gamma^2 L \left(\frac{2\omega}{M} + 1\right)(f(x_k) - f(x^*))$$
$$+ \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 .$$

# Несмещенная компрессия: доказательство

- Если взять полное математическое ожидание

$$
\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|x_k - x^*\|^2\right] \\
& - 2\gamma\left[1 - \gamma L\left(\frac{2\omega}{M} + 1\right)\right]\mathbb{E}\left[(f(x_k) - f(x^*))\right] \\
& + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 .
\end{aligned}
$$

- Если $\gamma \leq L^{-1}\left(\frac{2\omega}{M} + 1\right)^{-1}$, то

$$
\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|x_k - x^*\|^2\right] \\
& + \frac{2\gamma^2\omega}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2 .
\end{aligned}
$$

# QGD: сходимость

## Теорема (QGD)

Пусть все локальные функции $f_m$ являются $\mu$-сильно выпуклыми и имеют $L$-Липшицев градиент, тогда если $\eta \leq L^{-1}\left(\frac{2\omega}{M}+1\right)^{-1}$, то

$$\mathbb{E}\left[\|x_K - x^*\|^2\right] = \mathcal{O}\left((1-\gamma\mu)^K\|x_0 - x^*\|^2 + \frac{2\omega}{\mu M^2 K}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2\right).$$

При получении данного результата так же использовался подбор $\gamma$ из работы:

PDF Stich S. Unified Optimal Analysis of the (Stochastic) Gradient Method

- **Вопрос:** какие проблемы есть в этой оценке? Сублинейная сходимость (зависит от гетерогенности данных).

# QGD: сходимость

## Теорема (QGD)

Пусть все локальные функции $f_m$ являются $\mu$-сильно выпуклыми и имеют $L$-Липшицев градиент, тогда если $\eta \leq L^{-1}\left(\frac{2\omega}{M}+1\right)^{-1}$, то

$$\mathbb{E}\left[\|x_K - x^*\|^2\right] = \mathcal{O}\left((1-\gamma\mu)^K\|x_0 - x^*\|^2 + \frac{2\omega}{\mu M^2 K}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2\right).$$

При получении данного результата так же использовался подбор $\gamma$ из работы:

PDF Stich S. Unified Optimal Analysis of the (Stochastic) Gradient Method

- **Вопрос:** какие проблемы есть в этой оценке? Сублинейная сходимость (зависит от гетерогенности данных).

# QGD: сходимость

## Теорема (QGD)

Пусть все локальные функции $f_m$ являются $\mu$-сильно выпуклыми и имеют $L$-Липшицев градиент, тогда если $\eta \leq L^{-1}\left(\frac{2\omega}{M}+1\right)^{-1}$, то

$$\mathbb{E}\left[\|x_K - x^*\|^2\right] = \mathcal{O}\left((1-\gamma\mu)^K\|x_0 - x^*\|^2 + \frac{2\omega}{\mu M^2 K}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|^2\right).$$

При получении данного результата так же использовался подбор $\gamma$ из работы:

Stich S. Unified Optimal Analysis of the (Stochastic) Gradient Method

- **Вопрос:** какие проблемы есть в этой оценке? Сублинейная сходимость (зависит от гетерогенности данных).

# Несмещенная компрессия: больше

- Решена проблема с гетерогенностью за счет "памяти":
  - Mishchenko K. et al. Distributed Learning with Compressed Gradient Differences
- Ускоренная версия:
  - Li Z. et al. Acceleration for compressed gradient descent in distributed and federated optimization

# Несмещенная компрессия: больше

- Решена проблема с гетерогенностью за счет "памяти":
  - Mishchenko K. et al. Distributed Learning with Compressed Gradient Differences
- Ускоренная версия:
  - Li Z. et al. Acceleration for compressed gradient descent in distributed and federated optimization

# Смещенная компрессия

## Смещённая компрессия

Будем называть (стохастический) оператор $(x)$ оператором смещённой компрессии, если для любого $x \in \mathbb{R}^d$ выполняется:

$$\mathbb{E}[\|C(x) - x\|_2^2] \leq \left(1 - \frac{1}{\delta}\right)\|x\|_2^2,$$

где $\delta \geq 0$.

# Смещенная компрессия: примеры

## "Жадная" спарсификация (выбор наибольших по модулю компонент)

Рассмотрим стохастический оператор

$$\mathrm{Top}_k(x) = \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)},$$

где $k$ — некоторое фиксированное число из множества $\{1, \ldots, d\}$ (количество компонент вектора $x$, которые мы передаём; например, можно выбрать $k = 1$), при этом координаты отсортированы по модулю: $|x_{(1)}| \leq |x_{(2)}| \leq \ldots \leq |x_{(d)}|$, $(e_1, \ldots, e_d)$ — стандартный базис в $\mathbb{R}^d$. Можно показать, что данный оператор является смещённой компрессией с константой $\delta = \frac{d}{k}$.

PDF Alistarh D. et al. The convergence of sparsified gradient methods

# Смещенная компрессия: примеры

- Еще примеры смещенных компрессией:

PDF Beznosikov A. et al. On Biased Compression for Distributed Learning

PDF Vogels T. et al. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization

# Смещенная компрессия: идея и доказательство в случае 1 ноды

- Использовать тот же подход, что и в несмещенном случае (QGD).
- Докажем в случае одной ноды:

$$x_{k+1} = x_k - \gamma C(\nabla f(x_k)).$$

Пусть $f$ имеет $L$-Липшицев градиент и является $\mu$-сильно выпуклой.

- Начнем с того, что воспользуемся Липшицевостью градиента:

$$\begin{aligned} f(x_{k+1}) &= f(x_k - \gamma C(\nabla f(x_k))) \\ &\leq f(x_k) + \langle \nabla f(x_k), -\gamma C(\nabla f(x_k)) \rangle + \frac{L}{2} \| - \gamma C(\nabla f(x_k)) \|^2 \\ &= f(x_k) - \gamma \langle C(\nabla f(x_k)), \nabla f(x_k) \rangle + \frac{\gamma^2 L}{2} \| C(\nabla f(x_k)) \|^2. \end{aligned}$$

# Смещенная компрессия: идея и доказательство в случае 1 ноды

- Использовать тот же подход, что и в несмещенном случае (QGD).
- Докажем в случае одной ноды:

$$x_{k+1} = x_k - \gamma C(\nabla f(x_k)).$$

Пусть $f$ имеет $L$-Липшицев градиент и является $\mu$-сильно выпуклой.

- Начнем с того, что воспользуемся Липшицевостью градиента:

$$\begin{aligned} f(x_{k+1}) &= f(x_k - \gamma C(\nabla f(x_k))) \\ &\leq f(x_k) + \langle \nabla f(x_k), -\gamma C(\nabla f(x_k)) \rangle + \frac{L}{2} \| - \gamma C(\nabla f(x_k)) \|^2 \\ &= f(x_k) - \gamma \langle C(\nabla f(x_k)), \nabla f(x_k) \rangle + \frac{\gamma^2 L}{2} \| C(\nabla f(x_k)) \|^2. \end{aligned}$$

# Смещенная компрессия: идея и доказательство в случае 1 ноды

- Использовать тот же подход, что и в несмещенном случае (QGD).
- Докажем в случае одной ноды:

$$x_{k+1} = x_k - \gamma C(\nabla f(x_k)).$$

Пусть $f$ имеет $L$-Липшицев градиент и является $\mu$-сильно выпуклой.

- Начнем с того, что воспользуемся Липшицевостью градиента:

$$\begin{aligned} f(x_{k+1}) &= f(x_k - \gamma C(\nabla f(x_k))) \\ &\leq f(x_k) + \langle \nabla f(x_k), -\gamma C(\nabla f(x_k)) \rangle + \frac{L}{2}\| - \gamma C(\nabla f(x_k))\|^2 \\ &= f(x_k) - \gamma \langle C(\nabla f(x_k)), \nabla f(x_k) \rangle + \frac{\gamma^2 L}{2}\|C(\nabla f(x_k))\|^2. \end{aligned}$$

# Смещенная компрессия: доказательство в случае 1 ноды

- Определение компрессора:

$$\|\nabla f(x_k)\|^2 - 2\mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right]$$
$$= \mathbb{E}_C\left[\|C(\nabla f(x_k)) - \nabla f(x_k)\|^2\right] \leq \left(1 - \frac{1}{\delta}\right)\|\nabla f(x_k)\|^2.$$

- Откуда:

$$-\gamma\mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \frac{\gamma}{2}\mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right] \leq -\frac{\gamma}{2\delta}\|\nabla f(x_k)\|^2.$$

# Смещенная компрессия: доказательство в случае 1 ноды

- Определение компрессора:

$$\|\nabla f(x_k)\|^2 - 2\mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right] = \mathbb{E}_C\left[\|C(\nabla f(x_k)) - \nabla f(x_k)\|^2\right] \leq \left(1 - \frac{1}{\delta}\right)\|\nabla f(x_k)\|^2.$$

- Откуда:

$$-\gamma\mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \frac{\gamma}{2}\mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right] \leq -\frac{\gamma}{2\delta}\|\nabla f(x_k)\|^2.$$

# Смещенная компрессия: доказательство в случае 1 ноды

- С двух предыдущих слайдов:

$$f(x_{k+1}) - \leq f(x_k) - \gamma\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle + \frac{\gamma^2 L}{2}\|C(\nabla f(x_k))\|^2.$$

$$-\gamma\mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \frac{\gamma}{2}\mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right] \leq -\frac{\gamma}{2\delta}\|\nabla f(x_k)\|^2.$$

- Сложим, вычтем из обоих частей $f(x^*)$ и возьмем полное мат. ожидание:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2}(1-\gamma L)\,\mathbb{E}\left[\|C(\nabla f(x_k))\|^2\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

- Возьмем $\gamma \leq \frac{1}{L}$:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

# Смещенная компрессия: доказательство в случае 1 ноды

- С двух предыдущих слайдов:

$$f(x_{k+1}) - \leq f(x_k) - \gamma\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle + \frac{\gamma^2 L}{2}\|C(\nabla f(x_k))\|^2.$$

$$-\gamma \mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \frac{\gamma}{2}\mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right] \leq -\frac{\gamma}{2\delta}\|\nabla f(x_k)\|^2.$$

- Сложим, вычтем из обоих частей $f(x^*)$ и возьмем полное мат. ожидание:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2}(1 - \gamma L)\,\mathbb{E}\left[\|C(\nabla f(x_k))\|^2\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

- Возьмем $\gamma \leq \frac{1}{L}$:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

# Смещенная компрессия: доказательство в случае 1 ноды

- С двух предыдущих слайдов:

$$f(x_{k+1}) - \leq f(x_k) - \gamma\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle + \frac{\gamma^2 L}{2}\|C(\nabla f(x_k))\|^2.$$

$$-\gamma\mathbb{E}_C\left[\langle C(\nabla f(x_k)), \nabla f(x_k)\rangle\right] + \frac{\gamma}{2}\mathbb{E}_C\left[\|C(\nabla f(x_k))\|^2\right] \leq -\frac{\gamma}{2\delta}\|\nabla f(x_k)\|^2.$$

- Сложим, вычтем из обоих частей $f(x^*)$ и возьмем полное мат. ожидание:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2}(1 - \gamma L)\mathbb{E}\left[\|C(\nabla f(x_k))\|^2\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

- Возьмем $\gamma \leq \frac{1}{L}$:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

# Смещенная компрессия: доказательство в случае 1 ноды

- С предыдущего слайда:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

- Сильная выпуклость (или даже более слабое условие PL):

$$2\mu(f(x_k) - f(x^*)) \leq \|\nabla f(x_k)\|^2.$$

- Соединим два предыдущих:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{\delta}\right)\mathbb{E}\left[f(x_k) - f(x^*)\right].$$

# Смещенная компрессия: доказательство в случае 1 ноды

- С предыдущего слайда:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

- Сильная выпуклость (или даже более слабое условие PL):

$$2\mu(f(x_k) - f(x^*)) \leq \|\nabla f(x_k)\|^2.$$

- Соединим два предыдущих:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{\delta}\right)\mathbb{E}\left[f(x_k) - f(x^*)\right].$$

# Смещенная компрессия: доказательство в случае 1 ноды

- С предыдущего слайда:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma}{2\delta}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right].$$

- Сильная выпуклость (или даже более слабое условие PL):

$$2\mu(f(x_k) - f(x^*)) \leq \|\nabla f(x_k)\|^2.$$

- Соединим два предыдущих:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{\delta}\right)\mathbb{E}\left[f(x_k) - f(x^*)\right].$$

# Смещенная компрессия: теорема в случае 1 ноды

**Теорема (сходимость QGD со смещенной компрессией в случае 1 ноды)**

Пусть $f$ $\mu$-сильно выпуклая (или PL) и имеет $L$-Липшицев градиент, тогда QGD для одной ноды с шагом $\gamma \leq 1/L$ и со смещенным компрессором с параметром $\delta$ сходится и выполнено:

$$f(x_K) - f(x^*) \leq \left(1 - \frac{\gamma\mu}{\delta}\right)^K (f(x_0) - f(x^*)).$$

# Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(x) = \langle a, x \rangle^2 + \frac{1}{4}\|x\|^2, \ f_2(x) = \langle b, x \rangle^2 + \frac{1}{4}\|x\|^2, \ f_3(x) = \langle c, x \rangle^2 + \frac{1}{4}\|x\|^2,$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $x_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(x_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать $Top_1$ компрессию?

$$x_1 = (t, t, t) + \eta \cdot \frac{11}{6}(t, t, t) = \left(1 + \frac{11\eta}{6}\right) x_0.$$

- Мы удаляемся от решения геометрически для любого $\eta \geq 0$.

# Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(x) = \langle a, x\rangle^2 + \frac{1}{4}\|x\|^2, \; f_2(x) = \langle b, x\rangle^2 + \frac{1}{4}\|x\|^2, \; f_3(x) = \langle c, x\rangle^2 + \frac{1}{4}\|x\|^2,$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $x_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(x_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать $Top_1$ компрессию?

$$x_1 = (t, t, t) + \eta \cdot \frac{11}{6}(t, t, t) = \left(1 + \frac{11\eta}{6}\right) x_0.$$

- Мы удаляемся от решения геометрически для любого $\eta \geq 0$.

# Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(x) = \langle a, x\rangle^2 + \frac{1}{4}\|x\|^2,\ f_2(x) = \langle b, x\rangle^2 + \frac{1}{4}\|x\|^2,\ f_3(x) = \langle c, x\rangle^2 + \frac{1}{4}\|x\|^2,$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $x_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(x_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать $Top_1$ компрессию?

$$x_1 = (t, t, t) + \eta \cdot \frac{11}{6}(t, t, t) = \left(1 + \frac{11\eta}{6}\right) x_0.$$

- Мы удаляемся от решения геометрически для любого $\eta \geq 0$.

# Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(x) = \langle a, x\rangle^2 + \frac{1}{4}\|x\|^2,\ f_2(x) = \langle b, x\rangle^2 + \frac{1}{4}\|x\|^2,\ f_3(x) = \langle c, x\rangle^2 + \frac{1}{4}\|x\|^2,$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $x_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(x_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать $Top_1$ компрессию?

$$x_1 = (t, t, t) + \eta \cdot \frac{11}{6}(t, t, t) = \left(1 + \frac{11\eta}{6}\right) x_0.$$

- Мы удаляемся от решения геометрически для любого $\eta \geq 0$.

# Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(x) = \langle a, x\rangle^2 + \frac{1}{4}\|x\|^2, \; f_2(x) = \langle b, x\rangle^2 + \frac{1}{4}\|x\|^2, \; f_3(x) = \langle c, x\rangle^2 + \frac{1}{4}\|x\|^2,$$

  где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.
- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $x_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(x_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(x_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(x_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать $Top_1$ компрессию?

$$x_1 = (t, t, t) + \eta \cdot \frac{11}{6}(t, t, t) = \left(1 + \frac{11\eta}{6}\right) x_0.$$

- Мы удаляемся от решения геометрически для любого $\eta \geq 0$.

# Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(x_0) - C(0 + \gamma \nabla f_m(x_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(x_1))$$

- На произвольной итерации это записывается так:

$$\text{Посылка:} C(e_{k,m} + \gamma \nabla f_m(x_k)),$$
$$e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(x_k) - C(e_{k,m} + \gamma \nabla f_m(x_k))$$

- Это техника называется компенсация ошибка (error feedback).

PDF Stich S. et al. Sparsified SGD with memory

# Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(x_0) - C(0 + \gamma \nabla f_m(x_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(x_1))$$

- На произвольной итерации это записывается так:

$$\text{Посылка:} C(e_{k,m} + \gamma \nabla f_m(x_k)),$$
$$e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(x_k) - C(e_{k,m} + \gamma \nabla f_m(x_k))$$

- Это техника называется компенсация ошибка (error feedback).

PDF Stich S. et al. Sparsified SGD with memory

# Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(x_0) - C(0 + \gamma \nabla f_m(x_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(x_1))$$

- На произвольной итерации это записывается так:

$$\text{Посылка:} C(e_{k,m} + \gamma \nabla f_m(x_k)),$$
$$e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(x_k) - C(e_{k,m} + \gamma \nabla f_m(x_k))$$

- Это техника называется компенсация ошибка (error feedback).

PDF Stich S. et al. Sparsified SGD with memory

# Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(x_0) - C(0 + \gamma \nabla f_m(x_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(x_1))$$

- На произвольной итерации это записывается так:

$$\text{Посылка:} C(e_{k,m} + \gamma \nabla f_m(x_k)),$$
$$e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(x_k) - C(e_{k,m} + \gamma \nabla f_m(x_k))$$

- Это техника называется компенсация ошибка (error feedback).

PDF Stich S. et al. Sparsified SGD with memory

# GD с error feedback

**Алгоритм 1** GD с error feedback

**Вход:** Размер шага $\gamma > 0$, стартовая точка $x_0 \in \mathbb{R}^d$, стартовые ошибки $e_{0,m} = 0$ для всех $m$ от 1 до $M$, количество итераций $K$

```
for k = 0, 1, ..., K − 1 do
    Отправить x_k всем рабочим                              ▷ выполняется сервером
    for m = 1, ..., M параллельно do
        Принять x_k от мастера                              ▷ выполняется рабочими
        Вычислить градиент ∇f(x_k) в точке x_k              ▷ выполняется рабочими
        Сгенерировать g_{k,m} = C(e_{k,m} + γ∇f(x_k))       ▷ выполняется рабочими
        Вычислить e_{k+1,m} = e_{k,m} + γ∇f_m(x_k) − g_{k,m} ▷ выполняется рабочими
        Отправить g_{k,m} мастеру                           ▷ выполняется рабочими
    end for
    Принять g_{k,m} от всех рабочих                         ▷ выполняется сервером
    Вычислить g_k = (1/M) Σ_{m=1}^{M} g_{k,m}               ▷ выполняется сервером
    x_{k+1} = x_k − g_k                                     ▷ выполняется сервером
end for
```

**Выход:** $x_K$

# GD c error feedback: сходимость

## Теорема GD с error feedback

Пусть все локальные функции $f_m$ являются $\mu$-сильно выпуклыми и имеют $L$-Липшицев градиент, тогда если $\eta \leq \frac{1}{28\delta L}$, то

$$\mathbb{E}\left[f(\tilde{x}_K) - f(x^*)\right] \leq \mathcal{O}\left(\delta L \|x_0 - x^*\|^2 \exp\left(-\frac{\gamma\mu K}{2}\right) + \frac{\delta}{\mu K} \cdot \frac{1}{M}\sum_{m=1}^{M} \|\nabla f_m(x^*)\|^2\right).$$

PDF Stich S. and Karimireddy S. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication

PDF Beznosikov A. et al. On Biased Compression for Distributed Learning

# Смещенная компрессия: больше

- Решена проблема с гетерогенностью за счет "памяти":
  Richtarik P. et al. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback
- Ускоренная версия:
  Qian X. Error Compensated Distributed SGD Can Be Accelerated

# Смещенная компрессия: больше

- Решена проблема с гетерогенностью за счет "памяти":
  - Richtarik P. et al. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback
- Ускоренная версия:
  - Qian X. Error Compensated Distributed SGD Can Be Accelerated

# Несмещенная против смещенной

- Лучшая оценка на число коммуникаций для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[1+\frac{\omega}{M}\right]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- Лучшая оценка на число коммуникаций для неускоренного метода со смещенной компрессией (EF-21):

$$\mathcal{O}\left([1+\delta]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- **Вопрос:** что можно о них сказать? Как они соотносятся с несжатыми методами? Они хуже. Но важно не число коммуникаций, а количество передаваемой информации.

# Несмещенная против смещенной

- Лучшая оценка на число коммуникаций для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[1+\frac{\omega}{M}\right]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- Лучшая оценка на число коммуникаций для неускоренного метода со смещенной компрессией (EF-21):

$$\mathcal{O}\left([1+\delta]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- **Вопрос:** что можно о них сказать? Как они соотносятся с несжатыми методами? Они хуже. Но важно не число коммуникаций, а количество передаваемой информации.

# Несмещенная против смещенной

- Лучшая оценка на число коммуникаций для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[1 + \frac{\omega}{M}\right]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- Лучшая оценка на число коммуникаций для неускоренного метода со смещенной компрессией (EF-21):

$$\mathcal{O}\left([1 + \delta]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- **Вопрос:** что можно о них сказать? Как они соотносятся с несжатыми методами? Они хуже. Но важно не число коммуникаций, а количество передаваемой информации.

# Несмещенная против смещенной

- Компрессоры сжимают информацию в $\beta$ раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[\frac{1}{\beta} + \frac{1}{M}\right] \frac{L}{\mu} \log \frac{1}{\varepsilon}\right).$$

- Несмещенный компрессор доказуемо улучшает число передаваемой информации, фактор улучшения: $\left[\frac{1}{\beta} + \frac{1}{M}\right]$.
- Смещенный компрессор не улучшает число передаваемой информации в общем случае.

# Несмещенная против смещенной

- Компрессоры сжимают информацию в $\beta$ раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[\frac{1}{\beta} + \frac{1}{M}\right]\frac{L}{\mu}\log\frac{1}{\varepsilon}\right).$$

- Несмещенный компрессор доказуемо улучшает число передаваемой информации, фактор улучшения: $\left[\frac{1}{\beta} + \frac{1}{M}\right]$.
- Смещенный компрессор не улучшает число передаваемой информации в общем случае.

# Несмещенная против смещенной

- Компрессоры сжимают информацию в $\beta$ раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[\frac{1}{\beta} + \frac{1}{M}\right] \frac{L}{\mu} \log \frac{1}{\varepsilon}\right).$$

- Несмещенный компрессор доказуемо улучшает число передаваемой информации, фактор улучшения: $\left[\frac{1}{\beta} + \frac{1}{M}\right]$.
- Смещенный компрессор не улучшает число передаваемой информации в общем случае.

# Несмещенная против смещенной

- Компрессоры сжимают информацию в $\beta$ раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O}\left(\left[\frac{1}{\beta} + \frac{1}{M}\right] \frac{L}{\mu} \log \frac{1}{\varepsilon}\right).$$

- Несмещенный компрессор доказуемо улучшает число передаваемой информации, фактор улучшения: $\left[\frac{1}{\beta} + \frac{1}{M}\right]$.
- Смещенный компрессор не улучшает число передаваемой информации в общем случае.

# Data similarity

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}\ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}\ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1} \ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \le \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}\ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1} \ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.
- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}\ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Data similarity

- И снова распределенная задача обучения:

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}\ell(x, z_i)\right],$$

где $z_i$ – элемент выборки, $\ell$ – функция потерь.
- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Что это может дать? Похожесть локальных функций потерь.
- Утверждается, что для любого $x$

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta.$$

- **Вопрос:** если $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция, то в общем случае (если не предполагать равномерность распределения данных), что можно сказать о $\delta$? $\delta \sim L$.

# Матричное неравенство Хёфдинга

## Теорема (Матричное неравенство Хёфдинга)

Рассмотрим конечную последовательность случайных квадратных матриц $\{X_i\}_{i=1}^N$. Пусть в этой последовательности матрицы независимы, эрмитовы и имеют размерность $d$. Предположим так же, что $\mathbb{E}[X_i] = 0$, и $X_i^2 \preceq A^2$ почти наверное, где $A$ – неслучайная эрмитова матрица. Тогда с вероятностью $1 - p$ выполнено, что

$$\left\| \sum_{i=1}^N X_i \right\| \leq \sqrt{8N\|A^2\| \cdot \ln(d/p)}.$$

PDF Tropp J. An introduction to matrix concentration inequalities

PDF Tropp J. User-friendly tail bounds for sums of random matrices

# Параметр схожести

- Локальная функция потерь:

$$f_m(x) = \frac{1}{N}\sum_{i=1}^{N} l(x, z_i).$$

- $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция (например, квадратичная или логрегрессия). Тогда имеем $\nabla^2 l(x, z_i) \preceq LI$ для любого $x$ и $z_i$ (здесь $I$ – единичная матрица.).
- Распределим все данные равномерно по всем нодам.
- **Вопрос:** что нужно взять в качестве $X_i$ в неравенстве Хёфдинга? $X_i = \frac{1}{N}\left[\nabla l(x, z_i) - \nabla f(x)\right]$. Легко проверить, что все условия матричного неравенства Хёфдинга для нее выполнены, в частности, $A^2 = \frac{4L^2}{N^2} I$.

# Параметр схожести

- Локальная функция потерь:

$$f_m(x) = \frac{1}{N} \sum_{i=1}^{N} l(x, z_i).$$

- $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция (например, квадратичная или логрегрессия). Тогда имеем $\nabla^2 l(x, z_i) \preceq LI$ для любого $x$ и $z_i$ (здесь $I$ – единичная матрица.).
- Распределим все данные равномерно по всем нодам.
- **Вопрос:** что нужно взять в качестве $X_i$ в неравенстве Хёфдинга? $X_i = \frac{1}{N}\left[\nabla l(x, z_i) - \nabla f(x)\right]$. Легко проверить, что все условия матричного неравенства Хёфдинга для нее выполнены, в частности, $A^2 = \frac{4L^2}{N^2} I$.

# Параметр схожести

- Локальная функция потерь:

$$f_m(x) = \frac{1}{N}\sum_{i=1}^{N} l(x, z_i).$$

- $\ell$ – $L$-гладкая ($L$-Липшицев градиент), выпуклая, дважды дифференцируемая функция (например, квадратичная или логрегрессия). Тогда имеем $\nabla^2 l(x, z_i) \preceq LI$ для любого $x$ и $z_i$ (здесь $I$ – единичная матрица.).
- Распределим все данные равномерно по всем нодам.
- **Вопрос:** что нужно взять в качестве $X_i$ в неравенстве Хёфдинга? $X_i = \frac{1}{N}\left[\nabla l(x, z_i) - \nabla f(x)\right]$. Легко проверить, что все условия матричного неравенства Хёфдинга для нее выполнены, в частности, $A^2 = \frac{4L^2}{N^2} I$.

# Параметр схожести: итог

- В итоге имеем
$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta \sim \frac{L}{\sqrt{N}}.$$

- Для квадратичных задач можно получить оценку вида:
$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta \sim \frac{L}{N}.$$

PDF Hendrikx H. et al. Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization

- В любом случае следует вывод: чем больше размер локальной выборки, тем меньше параметр схожести (похожи между собой гессианы).

# Параметр схожести: итог

- В итоге имеем
$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta \sim \frac{L}{\sqrt{N}}.$$
- Для квадратичных задач можно получить оценку вида:
$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta \sim \frac{L}{N}.$$

PDF Hendrikx H. et al. Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization

- В любом случае следует вывод: чем больше размер локальной выборки, тем меньше параметр схожести (похожи между собой гессианы).

# Параметр схожести: итог

- В итоге имеем

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta \sim \frac{L}{\sqrt{N}}.$$

- Для квадратичных задач можно получить оценку вида:

$$\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta \sim \frac{L}{N}.$$

PDF Hendrikx H. et al. Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization

- В любом случае следует вывод: чем больше размер локальной выборки, тем меньше параметр схожести (похожи между собой гессианы).

# Метод в общем виде

- Рассмотрим зеркальный спуск:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(\gamma \langle \nabla f(x_k), x \rangle + V(x, x_k)\right),$$

где $V(x,y)$ – дивергенция Брегмана, порожденная функцией строго-выпуклой функцией $\varphi(x)$:

$$V(x,y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y); x - y \rangle.$$

- **Вопрос:** Какой метод получится, если $\varphi(x) = \frac{1}{2}\|x\|^2$? Градиентный спуск.

# Метод в общем виде

- Рассмотрим зеркальный спуск:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(\gamma\langle \nabla f(x_k), x\rangle + V(x, x_k)\right),$$

где $V(x, y)$ – дивергенция Брегмана, порожденная функцией строго-выпуклой функцией $\varphi(x)$:

$$V(x, y) = \varphi(x) - \varphi(y) - \langle \nabla\varphi(y); x - y\rangle.$$

- **Вопрос:** Какой метод получится, если $\varphi(x) = \frac{1}{2}\|x\|^2$? Градиентный спуск.

# Метод в общем виде

- Рассмотрим зеркальный спуск:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(\gamma \langle \nabla f(x_k), x \rangle + V(x, x_k)\right),$$

где $V(x,y)$ – дивергенция Брегмана, порожденная функцией строго-выпуклой функцией $\varphi(x)$:

$$V(x,y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y); x - y \rangle.$$

- **Вопрос:** Какой метод получится, если $\varphi(x) = \frac{1}{2}\|x\|^2$? Градиентный спуск.

# Сходимость в общем виде

## Определение (относительная гладкость и сильная выпуклость)

Пусть $\varphi : \mathbb{R}^d \to \mathbb{R}$ является выпуклой и дважды дифференцируемой. Будем говорить, что функция $f$ является $L_\varphi$-гладкой и $\mu_\varphi$-сильно выпуклой относительно $\varphi$, если для любого $x \in \mathbb{R}^d$ выполнено

$$\mu_\varphi \nabla^2 \varphi(x) \preceq \nabla^2 f(x) \preceq L_\varphi \nabla^2 \varphi(x),$$

или эквивалентно для любых $x, y \in \mathbb{R}^d$

$$\mu_\varphi V(x, y) \leq f(x) - f(y) - \langle \nabla f(y); x - y \rangle \leq L_\varphi V(x, y).$$

PDF Lu H. et al. Relatively-Smooth Convex Optimization by First-Order Methods, and Applications

## Сходимость в общем виде: доказательство

- Первое условие оптимальности для шага зеркального спуска:

$$\gamma \nabla f(x_k) + \nabla \varphi(x_{k+1}) - \nabla \varphi(x_k) = 0.$$

- Из него (здесь $x^*$ – оптимум):

$$\langle \gamma \nabla f(x_k) + \nabla \varphi(x_{k+1}) - \nabla \varphi(x_k), x_{k+1} - x^* \rangle = 0.$$

$$\langle \gamma \nabla f(x_k), x^{k+1} - x^* \rangle = \langle \nabla \varphi(x_k) - \nabla \varphi(x_{k+1}), x^{k+1} - x^* \rangle$$
$$= V(x^*, x_k) - V(x^*, x_{k+1}) - V(x_{k+1}, x_k).$$

  (последнее утверждение называется теоремой Пифагора для дивергенций Брегмана и проверяется по определению)

- Небольшие перестановки дадут:

$$\langle \gamma \nabla f(x_k), x_{k+1} - x_k \rangle + V(x_{k+1}, x_k)$$
$$= V(x^*, x_k) - V(x^*, x_{k+1}) - \langle \gamma \nabla f(x_k), x_k - x^* \rangle.$$

# Сходимость в общем виде: доказательство

- Первое условие оптимальности для шага зеркального спуска:

$$\gamma \nabla f(x_k) + \nabla \varphi(x_{k+1}) - \nabla \varphi(x_k) = 0.$$

- Из него (здесь $x^*$ – оптимум):

$$\langle \gamma \nabla f(x_k) + \nabla \varphi(x_{k+1}) - \nabla \varphi(x_k), x_{k+1} - x^* \rangle = 0.$$

$$\begin{aligned} \langle \gamma \nabla f(x_k), x^{k+1} - x^* \rangle &= \langle \nabla \varphi(x_k) - \nabla \varphi(x_{k+1}), x^{k+1} - x^* \rangle \\ &= V(x^*, x_k) - V(x^*, x_{k+1}) - V(x_{k+1}, x_k). \end{aligned}$$

(последнее утверждение называется теоремой Пифагора для дивергенций Брегмана и проверяется по определению)

- Небольшие перестановки дадут:

$$\begin{aligned} &\langle \gamma \nabla f(x_k), x_{k+1} - x_k \rangle + V(x_{k+1}, x_k) \\ &\qquad = V(x^*, x_k) - V(x^*, x_{k+1}) - \langle \gamma \nabla f(x_k), x_k - x^* \rangle. \end{aligned}$$

# Сходимость в общем виде: доказательство

- Первое условие оптимальности для шага зеркального спуска:

$$\gamma \nabla f(x_k) + \nabla \varphi(x_{k+1}) - \nabla \varphi(x_k) = 0.$$

- Из него (здесь $x^*$ – оптимум):

$$\langle \gamma \nabla f(x_k) + \nabla \varphi(x_{k+1}) - \nabla \varphi(x_k), x_{k+1} - x^* \rangle = 0.$$

$$\begin{aligned}\langle \gamma \nabla f(x_k), x^{k+1} - x^* \rangle &= \langle \nabla \varphi(x_k) - \nabla \varphi(x_{k+1}), x^{k+1} - x^* \rangle \\ &= V(x^*, x_k) - V(x^*, x_{k+1}) - V(x_{k+1}, x_k).\end{aligned}$$

(последнее утверждение называется теоремой Пифагора для дивергенций Брегмана и проверяется по определению)

- Небольшие перестановки дадут:

$$\begin{aligned}&\langle \gamma \nabla f(x_k), x_{k+1} - x_k \rangle + V(x_{k+1}, x_k) \\ &\qquad = V(x^*, x_k) - V(x^*, x_{k+1}) - \langle \gamma \nabla f(x_k), x_k - x^* \rangle.\end{aligned}$$

# Сходимость в общем виде: доказательство

- Подставим $\gamma = \frac{1}{L_\varphi}$:

$$\langle \nabla f(x_k), x^{k+1} - x^k \rangle + L_\varphi V(x_{k+1}, x_k) = L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Воспользуемся определением гладкость относительно $\varphi$ с $x = x_{k+1}$, $y = x_k$:

$$f(x_{k+1}) - f(x_k) \le \langle \nabla f(x_k); x_{k+1} - x_k \rangle + L_\varphi V(x_{k+1}, x_k).$$

- Соединим два предыдущих:

$$f(x_{k+1}) - f(x_k) \le L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

# Сходимость в общем виде: доказательство

- Подставим $\gamma = \frac{1}{L_\varphi}$:

$$\langle \nabla f(x_k), x^{k+1} - x^k \rangle + L_\varphi V(x_{k+1}, x_k) = L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Воспользуемся определением гладкость относительно $\varphi$ с $x = x_{k+1}$, $y = x_k$:

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k); x_{k+1} - x_k \rangle + L_\varphi V(x_{k+1}, x_k).$$

- Соединим два предыдущих:

$$f(x_{k+1}) - f(x_k) \leq L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

# Сходимость в общем виде: доказательство

- Подставим $\gamma = \frac{1}{L_\varphi}$:

$$\langle \nabla f(x_k), x^{k+1} - x^k \rangle + L_\varphi V(x_{k+1}, x_k) \\ = L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Воспользуемся определением гладкость относительно $\varphi$ с $x = x_{k+1}$, $y = x_k$:

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k); x_{k+1} - x_k \rangle + L_\varphi V(x_{k+1}, x_k).$$

- Соединим два предыдущих:

$$f(x_{k+1}) - f(x_k) \leq L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

# Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(x_{k+1}) - f(x_k) \leq L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(x^*, x_k) \leq f(x^*) - f(x_k) - \langle \nabla f(x_k); x^* - x_k \rangle$$

- Сложим два предыдущих и немного поперемещаем:

$$f(x_{k+1}) - f(x^*) \leq (L_\varphi - \mu_\varphi) V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}).$$

- В силу того, что $x^*$ – оптимум:

$$V(x^*, x_{k+1}) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right) V(x^*, x_k).$$

# Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(x_{k+1}) - f(x_k) \leq L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(x^*, x_k) \leq f(x^*) - f(x_k) - \langle \nabla f(x_k); x^* - x_k \rangle$$

- Сложим два предыдущих и немного поперемещаем:

$$f(x_{k+1}) - f(x^*) \leq (L_\varphi - \mu_\varphi) V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}).$$

- В силу того, что $x^*$ – оптимум:

$$V(x^*, x_{k+1}) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right) V(x^*, x_k).$$

# Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(x_{k+1}) - f(x_k) \leq L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(x^*, x_k) \leq f(x^*) - f(x_k) - \langle \nabla f(x_k); x^* - x_k \rangle$$

- Сложим два предыдущих и немного поперемещаем:

$$f(x_{k+1}) - f(x^*) \leq (L_\varphi - \mu_\varphi) V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}).$$

- В силу того, что $x^*$ – оптимум:

$$V(x^*, x_{k+1}) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right) V(x^*, x_k).$$

# Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(x_{k+1}) - f(x_k) \leq L_\varphi V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}) - \langle \nabla f(x_k), x_k - x^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(x^*, x_k) \leq f(x^*) - f(x_k) - \langle \nabla f(x_k); x^* - x_k \rangle$$

- Сложим два предыдущих и немного поперемещаем:

$$f(x_{k+1}) - f(x^*) \leq (L_\varphi - \mu_\varphi) V(x^*, x_k) - L_\varphi V(x^*, x_{k+1}).$$

- В силу того, что $x^*$ – оптимум:

$$V(x^*, x_{k+1}) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right) V(x^*, x_k).$$

# Сходимость в общем виде: теорема

**Теорема (сходимость зеркального спуска)**

Пусть $\varphi$ и $f$ удовлетворяют определению выше, тогда зеркальный спуск с шагом $\gamma = \frac{1}{L_\varphi}$ сходится и выполнено:

$$V(x^*, x_K) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right)^K V(x^*, x_0).$$

# Метод для задачи data similarity

- Зеркальный спуск:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left( \gamma \langle \nabla f(x_k), x \rangle + V(x, x_k) \right),$$

где с дивергенция Брегмана $V(x, y)$, порожденной функцией $\varphi(x)$ (тут нужно потребовать, чтобы $f_1$ была выпуклой):

$$\varphi(x) = f_1(x) + \frac{\delta}{2}\|x\|^2.$$

Функция $f_1$ хранится на сервере.

- **Вопрос:** Какое число коммуникаций происходит $K$ итераций такого зеркального спуска? $K$ коммуникаций (количество подсчетов градиента $\nabla f$), вычисления $\arg\min$ требуют только вычислений на сервере.

# Метод для задачи data similarity

- Зеркальный спуск:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left( \gamma \langle \nabla f(x_k), x \rangle + V(x, x_k) \right),$$

где с дивергенция Брегмана $V(x, y)$, порожденной функцией $\varphi(x)$ (тут нужно потребовать, чтобы $f_1$ была выпуклой):

$$\varphi(x) = f_1(x) + \frac{\delta}{2}\|x\|^2.$$

Функция $f_1$ хранится на сервере.

- **Вопрос:** Какое число коммуникаций происходит $K$ итераций такого зеркального спуска? $K$ коммуникаций (количество подсчетов градиента $\nabla f$), вычисления $\arg\min$ требуют только вычислений на сервере.

# Метод для задачи data similarity

- Зеркальный спуск:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(\gamma \langle \nabla f(x_k), x \rangle + V(x, x_k)\right),$$

где с дивергенция Брегмана $V(x, y)$, порожденной функцией $\varphi(x)$ (тут нужно потребовать, чтобы $f_1$ была выпуклой):

$$\varphi(x) = f_1(x) + \frac{\delta}{2}\|x\|^2.$$

Функция $f_1$ хранится на сервере.

- **Вопрос:** Какое число коммуникаций происходит $K$ итераций такого зеркального спуска? $K$ коммуникаций (количество подсчетов градиента $\nabla f$), вычисления $\arg\min$ требуют только вычислений на сервере.

# Метод для задачи data similarity

**Алгоритм 1** Зеркальный спуск для задачи data similarity

**Вход:** Размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций $K$

**for** $k = 0, 1, \ldots, K-1$ **do**
  Отправить $x_k$ всем рабочим ▷ выполняется сервером
  **for** $m = 1, \ldots, M$ параллельно **do**
    Принять $x_k$ от мастера ▷ выполняется рабочими
    Вычислить градиент $\nabla f_m(x_k)$ в точке $x_k$ ▷ выполняется рабочими
    Отправить $\nabla f_m(x_k)$ мастеру ▷ выполняется рабочими
  **end for**
  Принять $\nabla f_m(x_k)$ от всех рабочих ▷ выполняется сервером
  Вычислить $\nabla f(x_k) = \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(x_k)$ ▷ выполняется сервером
  $x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(\gamma \langle \nabla f(x_k), x\rangle + V(x, x_k)\right)$ ▷ выполняется сервером
**end for**

**Выход:** $x_K$

# Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_\varphi \nabla^2 \varphi(x) \preceq \nabla^2 f(x) \preceq L_\varphi \nabla^2 \varphi(x),$$

- В нашем случае:

$$\mu_\varphi \left(\delta I + \nabla^2 f_1(x)\right) \preceq \nabla^2 f(x) \preceq L_\varphi \left(\delta I + \nabla^2 f_1(x)\right)$$

- Найдем $L_\varphi$:

$$\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta \Rightarrow \nabla^2 f(x) - \nabla^2 f_1(x) \preceq \delta I$$
$$\Rightarrow \nabla^2 f(x) \preceq \delta I + \nabla^2 f_1(x) \Rightarrow L_\varphi = 1.$$

# Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_\varphi \nabla^2 \varphi(x) \preceq \nabla^2 f(x) \preceq L_\varphi \nabla^2 \varphi(x),$$

- В нашем случае:

$$\mu_\varphi \left(\delta I + \nabla^2 f_1(x)\right) \preceq \nabla^2 f(x) \preceq L_\varphi \left(\delta I + \nabla^2 f_1(x)\right)$$

- Найдем $L_\varphi$:

$$\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta \Rightarrow \nabla^2 f(x) - \nabla^2 f_1(x) \preceq \delta I$$
$$\Rightarrow \nabla^2 f(x) \preceq \delta I + \nabla^2 f_1(x) \Rightarrow L_\varphi = 1.$$

# Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_\varphi \nabla^2 \varphi(x) \preceq \nabla^2 f(x) \preceq L_\varphi \nabla^2 \varphi(x),$$

- В нашем случае:

$$\mu_\varphi \left(\delta I + \nabla^2 f_1(x)\right) \preceq \nabla^2 f(x) \preceq L_\varphi \left(\delta I + \nabla^2 f_1(x)\right)$$

- Найдем $L_\varphi$:

$$\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta \Rightarrow \nabla^2 f(x) - \nabla^2 f_1(x) \preceq \delta I$$
$$\Rightarrow \nabla^2 f(x) \preceq \delta I + \nabla^2 f_1(x) \Rightarrow L_\varphi = 1.$$

# Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:
$$\mu_\varphi \nabla^2 \varphi(x) \preceq \nabla^2 f(x) \preceq L_\varphi \nabla^2 \varphi(x),$$
- В нашем случае:
$$\mu_\varphi \left(\delta I + \nabla^2 f_1(x)\right) \preceq \nabla^2 f(x) \preceq L_\varphi \left(\delta I + \nabla^2 f_1(x)\right)$$
- Найдем $L_\varphi$:
$$\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta \Rightarrow \nabla^2 f(x) - \nabla^2 f_1(x) \preceq \delta I$$
$$\Rightarrow \nabla^2 f(x) \preceq \delta I + \nabla^2 f_1(x) \Rightarrow L_\varphi = 1.$$

# Сходимость для задачи data similarity: доказательство

- Найдем $\mu_\varphi$. Из сильно выпуклости функции $f$:

$$\mu I \preceq \nabla^2 f(x) \Rightarrow \delta I \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Из $\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta$ имеем:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \delta I.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Откуда:

$$\nabla^2 f_1(x) + \delta I \preceq \frac{2\delta + \mu}{\mu}\nabla^2 f(x) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

# Сходимость для задачи data similarity: доказательство

- Найдем $\mu_\varphi$. Из сильно выпуклости функции $f$:

$$\mu I \preceq \nabla^2 f(x) \Rightarrow \delta I \preceq \frac{2\delta}{\mu} \nabla^2 f(x) - \delta I.$$

- Из $\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta$ имеем:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \delta I.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \frac{2\delta}{\mu} \nabla^2 f(x) - \delta I.$$

- Откуда:

$$\nabla^2 f_1(x) + \delta I \preceq \frac{2\delta + \mu}{\mu} \nabla^2 f(x) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

# Сходимость для задачи data similarity: доказательство

- Найдем $\mu_\varphi$. Из сильно выпуклости функции $f$:

$$\mu I \preceq \nabla^2 f(x) \Rightarrow \delta I \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Из $\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta$ имеем:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \delta I.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Откуда:

$$\nabla^2 f_1(x) + \delta I \preceq \frac{2\delta + \mu}{\mu}\nabla^2 f(x) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

# Сходимость для задачи data similarity: доказательство

- Найдем $\mu_\varphi$. Из сильно выпуклости функции $f$:

$$\mu I \preceq \nabla^2 f(x) \Rightarrow \delta I \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Из $\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta$ имеем:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \delta I.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Откуда:

$$\nabla^2 f_1(x) + \delta I \preceq \frac{2\delta + \mu}{\mu}\nabla^2 f(x) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

# Сходимость для задачи data similarity: доказательство

- Найдем $\mu_\varphi$. Из сильно выпуклости функции $f$:

$$\mu I \preceq \nabla^2 f(x) \Rightarrow \delta I \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Из $\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta$ имеем:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \delta I.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Откуда:

$$\nabla^2 f_1(x) + \delta I \preceq \frac{2\delta + \mu}{\mu}\nabla^2 f(x) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

# Сходимость для задачи data similarity: доказательство

- Найдем $\mu_\varphi$. Из сильно выпуклости функции $f$:

$$\mu I \preceq \nabla^2 f(x) \Rightarrow \delta I \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Из $\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta$ имеем:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \delta I.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(x) - \nabla^2 f(x) \preceq \frac{2\delta}{\mu}\nabla^2 f(x) - \delta I.$$

- Откуда:

$$\nabla^2 f_1(x) + \delta I \preceq \frac{2\delta + \mu}{\mu}\nabla^2 f(x) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

# Сходимость для задачи data similarity: теорема

**Теорема (сходимость для задачи data similarity)**

Пусть $f$ сильно выпуклая, $f_i$ выпуклые, а $\ell$ - гладкие, а $\varphi(x) = f_1(x) + \delta\|x\|^2$, тогда зеркальный спуск с шагом $\gamma = 1$ сходится и выполнено:

$$V(x^*, x_K) \leq \left(1 - \frac{\mu}{\mu + 2\delta}\right)^K V(x^*, x_0).$$

- Это означает, что если нам необходимо достигнуть точности $\varepsilon$ ($V(x^*, x_K) \sim \varepsilon$), то нам необходимо

$$K = \mathcal{O}\left(\frac{\mu + 2\delta}{\mu} \log \frac{V(x^*, x_0)}{\varepsilon}\right) = \left(\left[1 + \frac{\delta}{\mu}\right] \log \frac{V(x^*, x_0)}{\varepsilon}\right) \text{ коммуникаций.}$$

# Сходимость для задачи data similarity: теорема

## Теорема (сходимость для задачи data similarity)

Пусть $f$ сильно выпуклая, $f_i$ выпуклые, а $\ell$ - гладкие, а $\varphi(x) = f_1(x) + \delta\|x\|^2$, тогда зеркальный спуск с шагом $\gamma = 1$ сходится и выполнено:

$$V(x^*, x_K) \leq \left(1 - \frac{\mu}{\mu + 2\delta}\right)^K V(x^*, x_0).$$

- Это означает, что если нам необходимо достигнуть точности $\varepsilon$ ($V(x^*, x_K) \sim \varepsilon$), то нам необходимо

$$K = \mathcal{O}\left(\frac{\mu + 2\delta}{\mu} \log \frac{V(x^*, x_0)}{\varepsilon}\right) = \left(\left[1 + \frac{\delta}{\mu}\right] \log \frac{V(x^*, x_0)}{\varepsilon}\right) \text{ коммуникаций.}$$

# Лучше?

- Оценка на число коммуникаций в условиях data similarity:

$$K = \mathcal{O}\left(\left[1 + \frac{\delta}{\mu}\right] \log \frac{1}{\varepsilon}\right).$$

- Оценка на число коммуникаций для обычного распределенного градиентного спуска:

$$K = \mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\varepsilon}\right).$$

- Напомним, что $\delta \sim \frac{L}{\sqrt{N}}$, т.е. может быть значительное улучшение.

# Лучше?

- Оценка на число коммуникаций в условиях data similarity:
$$K = \mathcal{O}\left(\left[1 + \frac{\delta}{\mu}\right] \log \frac{1}{\varepsilon}\right).$$
- Оценка на число коммуникаций для обычного распределенного градиентного спуска:
$$K = \mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\varepsilon}\right).$$
- Напомним, что $\delta \sim \frac{L}{\sqrt{N}}$, т.е. может быть значительное улучшение.

# Лучше?

- Но есть ведь и ускоренный градиентный метод, который дает оценки:

$$K = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right).$$

Более того, эти оценки улучшаемые в общем случае.

PDF Scaman K. et al. Optimal Convergence Rates for Convex Distributed Optimization in Networks

**Вопрос:** как в общем случае доказывается какого-то алгоритма (необязательно оптимизационного) для класса задач?

- Для задачи data similarity так же имеются нижние оценки (в 2015 году):

$$K = \Omega\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right),$$

т.е. предполагается возможно ускорение.

PDF Arjevani Y. and Shamir O. Communication complexity of distributed convex learning and optimization

# Лучше?

- Но есть ведь и ускоренный градиентный метод, который дает оценки:

$$K = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right).$$

Более того, эти оценки улучшаемые в общем случае.

Scaman K. et al. Optimal Convergence Rates for Convex Distributed Optimization in Networks

**Вопрос:** как в общем случае доказывается какого-то алгоритма (необязательно оптимизационного) для класса задач?

- Для задачи data similarity так же имеются нижние оценки (в 2015 году):

$$K = \Omega\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right),$$

т.е. предполагается возможно ускорение.

Arjevani Y. and Shamir O. Communication complexity of distributed convex learning and optimization

# Лучше?

- Но есть ведь и ускоренный градиентный метод, который дает оценки:

$$K = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right).$$

Более того, эти оценки улучшаемые в общем случае.

PDF Scaman K. et al. Optimal Convergence Rates for Convex Distributed Optimization in Networks

**Вопрос:** как в общем случае доказывается какого-то алгоритма (необязательно оптимизационного) для класса задач?

- Для задачи data similarity так же имеются нижние оценки (в 2015 году):

$$K = \Omega\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right),$$

т.е. предполагается возможно ускорение.

PDF Arjevani Y. and Shamir O. Communication complexity of distributed convex learning and optimization

# Оптимальный алгоритм

- У данной проблемы довольно большая история:

| Reference | Communication complexity | Local gradient complexity | Order | Limitations |
|---|---|---|---|---|
| DANE [42] | $\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$ | — (2) | 1st | quadratic |
| DiSCO [51] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon}+C^2\Delta F_0)\log\frac{L}{\mu}\right)$ | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon}+C^2\Delta F_0)\log\frac{L}{\mu}\right)$ | 2nd | $C$ - self-concordant (3) |
| AIDE [40] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4) | 1st | quadratic |
| DANE-LS [50] | $\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5) | 1st/2nd | quadratic (6) |
| DANE-HB [50] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5) | 1st/2nd | quadratic (6) |
| SONATA [45] | $\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ | — (2) | 1st | decentralized |
| SPAG [21] | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1) | — (2) | 1st | $M$ - Lipshitz hessian |
| DiRegINA [12] | $\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}+\sqrt{\frac{M\delta R_0}{\mu}}\right)$ | — (2) | 2nd | $M$ -Lipshitz hessian |
| ACN [1] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}+\sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$ | — (2) | 2nd | $M$ -Lipshitz hessian |
| Acc SONATA [46] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ | — (2) | 1st | decentralized |
| This paper | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ | 1st | |

Figure: Таблица из статьи Kovalev D. et al. Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity

В частности, подход зеркального спуска с необычной дивергенцией называется DANE.

- Оптимальный алгоритм был предложен в 2022 году:

  PDF Kovalev D. et al. Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity

# Оптимальный алгоритм

- У данной проблемы довольно большая история:

| Reference | Communication complexity | Local gradient complexity | Order | Limitations |
|---|---|---|---|---|
| DANE [42] | $\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$ | — (2) | 1st | quadratic |
| DiSCO [51] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon}+C^2\Delta F_0)\log\frac{L}{\mu}\right)$ | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon}+C^2\Delta F_0)\log\frac{L}{\mu}\right)$ | 2nd | $C$ - self-concordant (3) |
| AIDE [40] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4) | 1st | quadratic |
| DANE-LS [50] | $\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5) | 1st/2nd | quadratic (6) |
| DANE-HB [50] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5) | 1st/2nd | quadratic (6) |
| SONATA [45] | $\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ | — (2) | 1st | decentralized |
| SPAG [21] | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1) | — (2) | 1st | $M$ - Lipshitz hessian |
| DiRegINA [12] | $\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}+\sqrt{\frac{M\delta R_0}{\mu}}\right)$ | — (2) | 2nd | $M$ -Lipshitz hessian |
| ACN [1] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}+\sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$ | — (2) | 2nd | $M$ -Lipshitz hessian |
| Acc SONATA [46] | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ | — (2) | 1st | decentralized |
| This paper | $\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ | 1st | |

Figure: Таблица из статьи Kovalev D. et al. Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity

В частности, подход зеркального спуска с необычной дивергенцией называется DANE.

- Оптимальный алгоритм был предложен в 2022 году:

  PDF Kovalev D. et al. Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity

Все! Спасибо за внимание!