Федеративное обучение. SCAFFOLD

Методы оптимизации

Федоренко Екатерина

Московский физико-технический институт

9 декабря 2023

Федеративное обучение

- Проблема:
  Мобильные устройства сейчас имеют доступ к обширным данным, пригодным для обучения моделей, что в свою очередь может значительно улучшить пользовательский опыт на устройстве. Однако эти богатые данные часто являются конфиденциальными, имеют большой объем или и то, и другое, что может исключить возможность регистрации в центре обработки данных и обучения там с использованием традиционных методов.

Федеративное обучение

- Основная идея:
  Оставление распределенных данных для обучения на мобильных устройствах и обучение общей модели путем агрегирования локальных вычисленных обновлений.
- Свойства задач для федеративного обучения:
  1)Обучение на реальных данных с мобильных устройств предоставляет явное преимущество перед обучением данных, которые обычно доступны в центре обработки данных.
  2)Данные являются конфиденциальными или имеют большой размер
  3)Метки на данных могут выводиться из взаимодействия с пользователем.

Постановка задачи

- Рассматривали такую задачу:

$$\min_{x \in \mathbb{R}^d} f(x).$$

- Теперь сформулируем задачу следующим образом:
  Задача: минимизация суммы стохастических функций, имея доступ только к стохастическим выборкам:

$$\min_{x \in R^d} \{f(x) := \frac{1}{N} \sum_{i=1}^{N} (f_i(x) := \mathbb{E}_{\zeta_i}[f_i(x, \zeta_i)])\}$$

  Функции f_i представляют функцию потерь на клиенте

Постановка задачи

- Предполагаем что:
  1) функция f ограничена снизу значением f^*, а функция f_i является β-гладкой.
  2)$g_i(x) := \nabla f_i(x; \zeta_i)$ является несмещенным стохастическим градиентом f_i с дисперсией, ограниченной σ^2.
  3) Для некоторых результатов мы предполагаем, что $\mu \geq 0$ (сильная) выпуклость. σ ограничивает дисперсию внутри клиентов.
- Вводим два термина, нестандартные для данного контекста:

Постановка задачи

- Новые термины:
  Условие 1:
  (G, B)-BGD, или ограниченная диссимиларность градиента: существуют константы $G \geq 0$ и $B \geq 1$ такие, что
  $$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 \leq G^2 + B^2\|\nabla f(x)\|^2, \quad \forall x.$$
  Если f_i являются выпуклыми, мы можем усилить предположение до
  $$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 \leq G^2 + 2\beta B^2(f(x) - f^*), \quad \forall x.$$

Постановка задачи

- Новые термины:
  Условие 2:
  δ-BHD, или ограниченная диссимиларность гессиана:
  $$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta, \quad \forall x.$$
  Кроме того, f_i является δ-слабо выпуклой, то есть
  $$\nabla^2 f_i(x) \succeq -\delta I$$
  .
  Предположения из условия 1 и 2 ортогональны — возможно иметь $G = 0$ и $\delta = 2\beta$, или $\delta = 0$, но $G > 1$.

FedAvg

Алгоритм 1 FedAvg

Вход: стартовое значение сервера x, размер шага для сервера - η_g, размер шага для клиентов - η_g, количество раундов - R, количество батчей - K

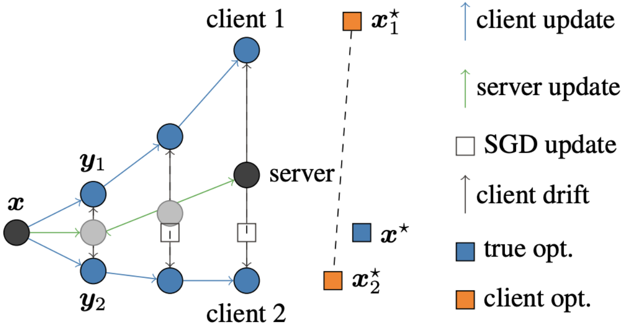
```
for r = 1, ..., R do
    Выбираем подмножество клиентов S
    Каждому клиенту S передаем значение x, хранящиеся на сервере
    for i ∈ S do
        y_i ← x
        for k = 0, ..., K-1 do
            Посчитать g_i(y_i) (градиент мини-батча в y_i)
            y_i ← y_i − η_l g_i(y_i)
        Δy_i ← y_i − x
    Δx ← (1/|S|) ∑ Δy_i
    x ← x + η_g Δx
```

- **for** r = 1, ..., R **do**
  - Выбираем подмножество клиентов S
  - Каждому клиенту S передаем значение x, хранящиеся на сервере
  - **for** $i \in S$ **do**
    - $y_i \leftarrow x$
    - **for** k = 0, ..., K-1 **do**
      - Посчитать $g_i(y_i)$ (градиент мини-батча в y_i)
      - $y_i \leftarrow y_i - \eta_l g_i(y_i)$
    - $\Delta y_i \leftarrow y_i - x$
  - $\Delta x \leftarrow \frac{1}{|S|} \sum \Delta y_i$
  - $x \leftarrow x + \eta_g \Delta x$

Выход: обновленное значение на сервере - x

FedAvg

- Недостаток FedAvg:
  При различных функциях f_i локальные обновления FEDAVG на каждом клиенте подвергаются дрейфу, что замедляет сходимость.



FedAvg

- Пусть x^* будет глобальным оптимумом $f(x)$, а x_i^* - оптимумом функции потерь каждого клиента $f_i(x)$.
- В случае гетерогенных данных каждый x_i^* далек от других и от глобального оптимума x^*. Даже если все клиенты начинают с одной и той же точки x, каждый y_i будет двигаться к своему оптимуму x_i^*.
- Среднее обновление клиентов (которое представляет собой обновление сервера) движется к $\frac{1}{N}\sum_{i=1}^{N} x_i^*$.
- Разница между $\frac{1}{N}\sum_{i=1}^{N} x_i^*$ и истинным оптимумом x^* вызывает дрейф на клиенте.

FedAvg

- Предполагаем, что:
  1) *FEDAVG* запускается с $\eta_g = 1$, $K > 1$, и произвольными, возможно адаптивными положительными значениями шагов $\{\eta_1, \ldots, \eta_R\}$, используемыми с $\eta_r \leq \frac{1}{\mu}$ и фиксированными в течение раунда для всех клиентов.
  2) Обновление сервера является выпуклой комбинацией обновлений клиентов с неадаптивными весами.
  Теорема. Для любых положительных констант G, μ существуют μ-сильно выпуклые функции, удовлетворяющие условию 1, для которых выход *FEDAVG* имеет ошибку для любого $r \geq 1$:

$$f(x^r) - f(x^*) \geq \Omega\left(\min f(x^0) - f(x^*), \frac{G^2}{\mu R^2}\right)$$

Доказательство нижней оценки

Доказательство.

- Рассмотрим следующие простые одномерные функции для любых данных μ и G:

$$f_1(x) := \mu x^2 + Gx, \quad f_2(x) := -Gx,$$

- где $f(x) = \frac{1}{2}(f_1(x) + f_2(x)) = \frac{\mu}{2}x^2$ с оптимумом в точке $x = 0$
- f - μ-сильно выпуклая, и f_1 и f_2 удовлетворяют условию 1 с $B = 3$
- Начнем FEDAVG с $x_0 > 0$. Одиночное локальное обновление для f_1 и f_2 в раунде $r \geq 1$ соответственно:

$$y_1 = y_1 - \eta_r(2\mu x + G), \quad y_2 = y_2 + \eta_r G.$$

Доказательство нижней оценки

- Запишем, как выглядит шаг нашего метода:

$$x^r = x^{r-1} + \frac{1}{|S|}\sum_{i \in S_r}(y_{ri,K} - x^{r-1})$$

- Введем коэффициент усреднения $\alpha = \frac{1}{|S|}, \alpha \in [0, 1]$

$$x^r = x^{r-1} + \alpha(y_{r1,K} - x^{r-1} + y_{r2,K} - x^{r-1})$$

- Найдем $y_{r1,K}$ и $y_{r2,K}$

$$y_{ri,k} = y_{ri,k-1} - \eta_r g_i(y_{ri,k-1})$$

$$y_{r2,K} = x^{r-1} - \eta_r GK$$

Доказательство нижней оценки

- Теперь запишем реккуренту для $y_{r1,K}$

$$y_{r1,k} = y_{r1,k-1} - \eta_r 2\mu y_{r1,k-1} - \eta_r G = y_{r1,k-1}(1 - 2\eta_r\mu) - \eta_r G$$

$$y_{r1,0} = x^{r-1}$$

- Получаем:

$$y_{r1,k} = x^{r-1}(1 - 2\eta_r\mu)^K - \sum_{t=0}^{K-1} \eta_r G(1 - 2\eta_r\mu)^t$$

Доказательство нижней оценки

- Подставим $y_{r1,K}$ и $y_{r2,K}$ в нашу формулу и выразим x^r

$$x^r = x^{r-1}((1-\alpha)(1-2\mu\eta_r)^K + \alpha) + \eta_r G \sum_{t=0}^{K-1} (\alpha - (1-\alpha)(1-2\mu\eta_r)^t)$$

- Поскольку α было выбрано непредвзято, мы можем предположить, что $\alpha \leq 0.5$. В обратном случае, поменяем определения f1 и f2 и знак x0.

$$x^r \geq x^{r-1} \left(\left(\frac{1-2\mu\eta_r)^K + 1}{2} + \frac{\eta_r G}{2} \sum_{t=0}^{K-1} (1 - (1-2\mu\eta_r)^t) \right)$$

Доказательство нижней оценки

- Немного преобразуем

$$x^r \geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2}\sum_{t=0}^{K-1}(1-(1-2\mu\eta_r)^t)$$

- В выражении выше правая сторона возрастает с увеличением η_r — это представляет собой влияние дрейфа клиента и увеличивает ошибку при увеличении размера шага. Левая сторона убывает с η_r — это сходимость при выполнении градиентных шагов. Покажем, что даже с тщательным балансированием двух переменных воздействия G не может быть устранено.

Доказательство нижней оценки

- Пусть $\gamma_r = \mu\eta_r R(K-1)$
  Такое значение γ_r существует и положительно, так как $K \geq 2$.
  Тогда верно, что:

$$(1-(2\mu\eta_r))^{\frac{K-1}{2}} = (1 - \frac{2\gamma_r}{R(K-1)})^{\frac{K-1}{2}} \leq \exp\left(-\frac{\gamma_r}{R}\right).$$

- Проведем некоторые преобразования с неравенством

$$x^r \geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2}\sum_{t=0}^{K-1}(1-(1-2\mu\eta_r)^t)$$

$$\geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2}\sum_{t=(K-1)/2}^{K-1}(1-(1-2\mu\eta_r)^t)$$

Доказательство нижней оценки

- Теперь подставим оценку на $(1-(2\mu\eta_r))^{\frac{K-1}{2}}$

$$x^r \geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G(K-1)}{4}(1-(1-2\mu\eta_r)^{\frac{K-1}{2}})$$

$$\geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\gamma_r G}{4\mu}(1-\exp(-\gamma_r/R))$$

- Рассмотрим 2 варианта:
  1) $\gamma_r \geq R/8$ или $8\mu\eta_r(K-1) \geq 1$
  2) $\gamma_r \leq R/8$ или $8\mu\eta_r(K-1) \leq 1$
- Начнем с первого пункта:

$$2\mu\eta_r \geq \frac{1}{4(K-1)} \geq \frac{1}{4}$$

Доказательство нижней оценки

$(1 - 2\mu\eta_r) \geq \frac{1}{4} > 0$. Значит:

$$x^r \geq \frac{\gamma_r G}{4\mu}(1 - \exp(-\gamma_r/R))$$

У нас есть константа $c_1 \in (0, \frac{1}{32})$ такая, что:

$$x^r \geq \frac{c_1 G}{\mu}$$

Доказательство нижней оценки

- Теперь разберемся со вторым пунктом:

$$\gamma_r < \frac{R}{8}$$

Получаем более строгое неравенство(из ряда тейлора с остаточным членом в интегральной форме):

$$(1-(2\mu\eta_r))^{\frac{K-1}{2}} = (1-\frac{2\gamma_r}{R(K-1)})^{\frac{K-1}{2}} \leq 1-\frac{\gamma_r}{R}$$

- Подставим эту оценку в неравенство для x^r

$$x^r \geq x^{r-1}(1-\frac{2\gamma_r}{R(K-1)})^K + \frac{\eta_r G(K-1)}{4}(1-(1-\frac{2\gamma_r}{R(K-1)})^{\frac{K-1}{2}})$$

$$\geq x^{r-1}(1-\frac{2\gamma_r}{R(K-1)})^K + \frac{G\gamma_r^2}{4\mu R}$$

Доказательство нижней оценки

- Воспользуемся неравенством Бернулли:

$$(1 - \frac{2\gamma_r}{R(K-1)})^K \geq (1 - \frac{2\gamma_r K}{R(K-1)})$$

- $K - 1 \geq \frac{K}{2}, K \geq 2$

$$(1 - \frac{2\gamma_r}{R(K-1)})^K \geq (1 - \frac{2\gamma_r K}{R(K-1)}) \geq (1 - \frac{4\gamma_r}{R})$$

- Подставим в наше неравенство:

$$x^r \geq x^{r-1}(1 - \frac{4\gamma_r}{R}) + \frac{G\gamma_r^2}{4\mu R}$$

Доказательство нижней оценки

- В выражении выше правая сторона возрастает с γ_r — это представляет собой влияние дрейфа клиента и увеличивает ошибку при увеличении размера шага.
- Левая сторона убывает с γ_r — это обычная сходимость, наблюдаемая при выполнении градиентных шагов.
- Остальная часть доказательства направлена на показ того, что даже с тщательным балансированием двух терминов, воздействие G не может быть устранено.

Доказательство нижней оценки

- Предположим, что все раунды после $r_0 \geq 0$ имеют малый размер шага, т.е. $\gamma_r \leq \frac{R}{8}$ для всех $r > r_0$:

$$x^r \geq x^{r-1}(1 - \frac{4\gamma_r}{R}) + \frac{G\gamma_r^2}{4\mu R}$$

- Докажем по индукции, что для этого случая верно:

$$x^r \geq \min(c_r x^{r_0}, \frac{G}{256\mu R}),$$

где константы $c_r := (1 - \frac{1}{2R})^{r-r_0}$

- База. Для $r = r_0$ утверждение тривиальное

Доказательство нижней оценки

- Переход. Для $r > r_0$
  Рассмотрим 2 случая:
  1) $\gamma_r \geq \frac{1}{8}$
  $$(1 - \frac{4\gamma_r}{R}) \geq (1 - \frac{1}{2R})$$
  2) $\gamma_r \geq \frac{1}{8}$
  $$\frac{\gamma_r^2 G}{4\mu R} \geq \frac{G}{256\mu R}$$
  Получили:
  $$\begin{aligned} x^r &\geq x^{r-1}(1 - \frac{4\gamma_r}{R}) + \frac{\gamma_r^2 G}{4\mu R} \\ &\geq \min\left(x^{r-1}(1 - \frac{1}{2R}), \frac{G}{256\mu R}\right) \end{aligned}$$

Доказательство нижней оценки

- Воспользуемся предположением индукции:

$$x^r \geq \min\left(c_r x^{r_0}, \frac{G}{256\mu R}\right).$$

- По неравенству Бернулли: $c_R \geq \frac{1}{2}$

$$x^R \geq \min\left(\frac{1}{2} x^{r_0}, \frac{G}{256\mu R}\right).$$

- Теперь предположим, что $\gamma_{r_0} > \frac{R}{8}$.

$$x^R \geq c\frac{G}{\mu R}$$

Если такого $r_0 \geq 1$ не существует, то мы можем установить $r_0 = 0$. Предыдущее доказательство не делало никаких предположений о R, и неравенство справедливо для всех $r \geq 1$.

Доказательство нижней оценки

Получили, что:

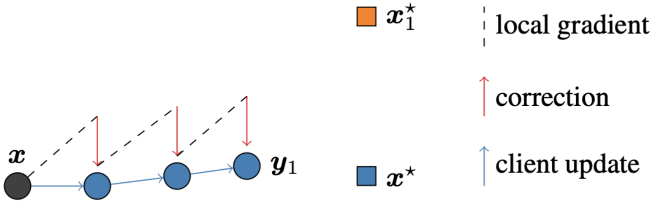
$$x^r \geq c \min(x^0, \frac{G}{\mu R})$$

Завершаем доказательство, отмечая, что $f(x^r) = \frac{\mu}{2}(x^r)^2$
Получили нижнюю оценку для FedAvg:

$$f(x^r) - f(x^*) \geq \Omega\left(\min f(x^0) - f(x^*), \frac{G^2}{\mu R^2}\right)$$

SCAFFOLD

- Решение проблемы на гетерогенных данных - SCAFFOLD
  SCAFFOLD состоит из трех основных шагов:
  1) локальное обновление модели клиента
  2)локальное обновление управляющей переменной клиента
  3)агрегирование обновлений



SCAFFOLD

Вместе с моделью сервера x, SCAFFOLD поддерживает состояние для каждого клиента (c_i) и для сервера (c). Они инициализируются так, чтобы $c = \frac{1}{N}\sum_i c_i$, и могут быть безопасно инициализированы значением 0. На каждом этапе обмена данными параметры сервера (x, c) передаются участвующим клиентам $S \subseteq [N]$.

SCAFFOLD

Алгоритм 2 SCAFFOLD

Вход: стартовое значение сервера x, размер шага сервера - η_g, размер шага клиентов - η_g, количество раундов - R, количество батчей - K, стартовое значение у.п. - c, стартовое значение у.п. клиентов - c_i

```
for r = 1, ..., R do
    Выбираем подмножество клиентов S
    Каждому клиенту S передаем значения (x, c), хранящиеся на сервере
    for i ∈ S do
        y_i ← x
        for k = 0, ..., K-1 do
            Посчитать g_i(y_i) (градиент мини-батча в y_i)
            y_i ← y_i − η_l(g_i(y_i) − c_i + c)
        c_i^+ ← c_i − c + 1/(K η_l) (x − y_i)
        (Δy_i, Δc_i) ← (y_i − x, c_i^+ − c_i)
    (Δx, Δc) ← 1/|S| ∑ (Δy_i, Δc_i)
    x ← x + η_g Δx
    c ← c + |S|/N Δc
```

Формулы:

$$y_i \leftarrow x$$

$$y_i \leftarrow y_i - \eta_l(g_i(y_i) - c_i + c)$$

$$c_i^+ \leftarrow c_i - c + \frac{1}{K\eta_l}(x - y_i)$$

$$(\Delta y_i, \Delta c_i) \leftarrow (y_i - x, c_i^+ - c_i)$$

$$(\Delta x, \Delta c) \leftarrow \frac{1}{|S|}\sum(\Delta y_i, \Delta c_i)$$

$$x \leftarrow x + \eta_g \Delta x$$

$$c \leftarrow c + \frac{|S|}{N}\Delta c$$

Выход: обновленное значение на сервере - x

Управляющие переменные

Если стоимость коммуникации не учитывается, то идеальное обновление на клиенте i было бы

$$y_i \leftarrow y_i + \frac{1}{N}\sum_j g_j(y_i).$$

Такое обновление фактически вычисляет несмещенный градиент f и, следовательно, становится эквивалентным выполнению FEDAVG в случае $|S| = N$. К сожалению, такое обновление требует связи со всеми клиентами на каждом этапе обновления. Вместо этого SCAFFOLD использует управляющие переменные такие, что

$$c_j \approx g_j(y_i) \quad \text{и} \quad c \approx \frac{1}{N}\sum_j g_j(y_i).$$

Еще немного оценок

Для любых функций $\{fi\}$, являющихся β-гладкими, вывод SCAFFOLD имеет ожидаемую ошибку, меньшую чем ε, в каждом из следующих трех случаев для некоторых значений η_l и η_g, с учетом следующего ограничения на R:

- Сильно выпуклая:

$$R = \tilde{O}\left(\frac{\sigma^2}{\mu K S \varepsilon} + \frac{\beta}{\mu} + \frac{N}{S}\right),$$

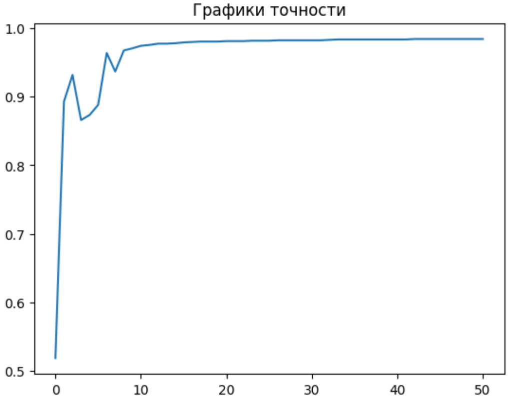
- Обобщенная выпуклая:

$$R = \tilde{O}\left(\frac{\sigma^2 D^2}{K S \varepsilon^2} + \frac{\beta D^2}{\varepsilon} + \frac{NF}{S}\right),$$

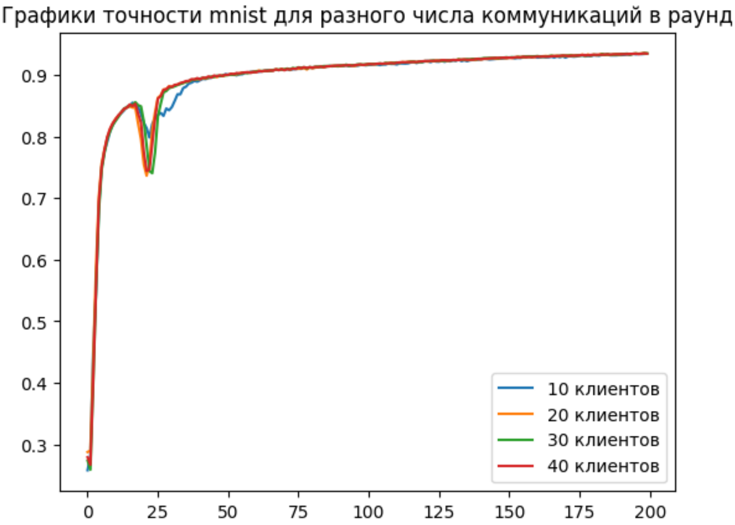
где $D := \|x^0 - x^*\|_2$, а $F := f(x^0) - f^*$

Немного практических результатов

Смоделировали федеративное обучение путем разделения данных на несколько частей(на клиентов)
Пример работы на задаче из домашнего задания(mushrooms.txt):



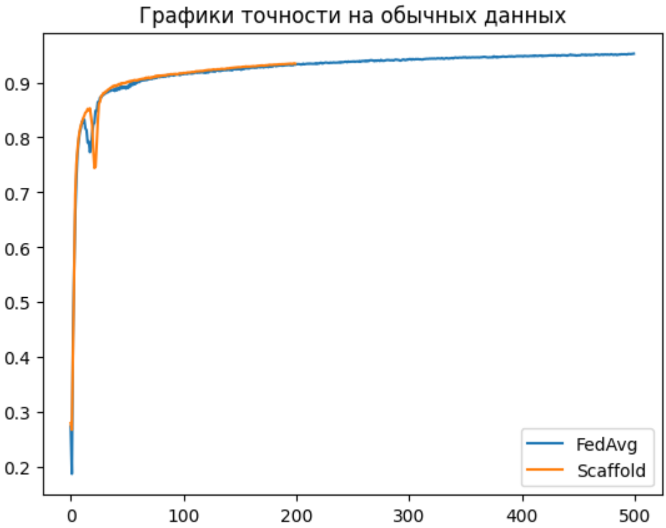
Немного практических результатов. MNIST



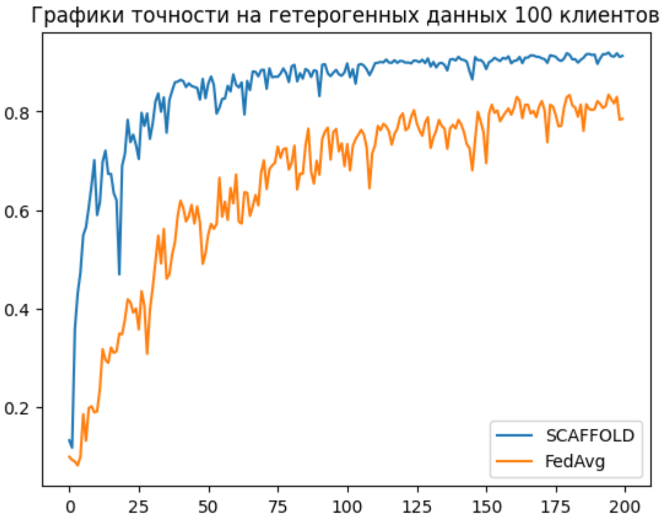
Гетерогенные данные

Пусть у каждого клиента хранятся данные только по конкретной цифре. Разделим датасет на 100 клиентов(на каждую цифру - 10 клиентов)

Сравнение с FedAvg



Сравнение с FedAvg



Репозиторий проекта

Тут можно посмотреть реализацию оптимизаторов и эксперименты.