Империя наносит ответный удар: Византийские атаки в мире распределенного обучения и как с ними бороться

Садчиков Андрей, Никитин Артемий

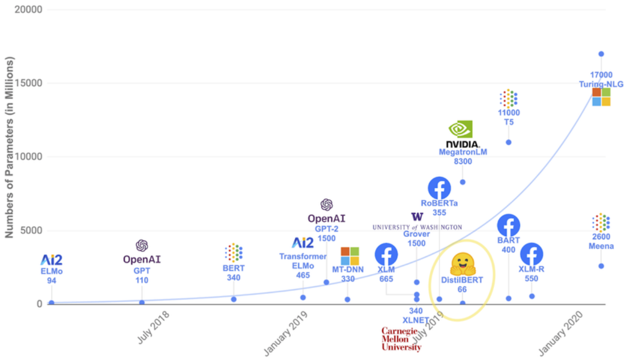
Московский физико-технический институт

6 декабря 2024

ФПМИ

Зачем нужно распределенное обучение

- Делает возможным обучение больших моделей



- Позволяет сохранить приватность данных

Реализация

Классический метод распределенного обучения - Federated Averaging

- Идея - усреднять посчитанные локально на устройствах градиенты, чтобы уменьшить дисперсию

Алгоритм 1 Federated Averaging

1: **Input:** Stepsize $\eta \geq 0$, initial vector x_0.
2: **for** $t = 0, 1, \dots$ **do**
3:     **for** $i = 1, \dots, n$ in parallel **do**
4:         Sample $z_i \sim D_i \equiv D$.
5:         Compute $g_t^i = \nabla F(x_t^i, z_i)$.
6:     **end for**
7:     $x_{t+1} = \frac{1}{n} \sum_{i=1}^{n} (x_t - \gamma g_t^i)$
8: **end for**

Проблема с федеративным обучением

Устройства могут обманывать

- могут присылать шум вместо стох.градиентов, а свои вычислительные мощности использовать в личных целях
- особо опасные могут намеренно портить процесс обучения
- назовём множество обманщиков $\mathcal{B}$ **византийцами**, а их долю как $\delta = \frac{|\mathcal{B}|}{n}$. Также обозначим честных игроков как $\mathcal{G} = [n] \setminus \mathcal{B}$.

Пример:

- Пусть у нас 3 устройства: $Device_1, Device_2, Device_3$
- $Device_3$ - злонамеренный игрок
- Если $g_t^3 := -g_t^1 - g_t^2$, то среднее градиентов будет нулём
- **FedAvg не будет сходиться вообще!**

Как бороться с атаками?

Идея: давайте измением наш ***аггрегатор***

$x_{t+1} = x_t - \frac{\gamma}{n}\sum_{i=1}^n g_t^i \Longrightarrow x_{t+1} = x_t - \gamma \cdot \mathrm{AGG}(g_t^1, ..., g_t^n)$

- Покоординатная медиана:
  $[\mathrm{CM}(x_1, \ldots, x_n)]_j = \mathrm{median}([x_1]_j, \ldots, [x_n]_j)$
- Урезанное среднее (для каждой координаты выкинули δn слева и справа):
  $[\mathrm{TM}(x_1, \ldots, x_n)]_j = \frac{1}{n-2\delta n}\sum_{i=\delta n}^{n-\delta n}[x_{\Pi_j(i)}]_j.$
- Геометрическая медиана:
  $\mathrm{RFA}(x_1, \ldots, x_n) = \arg\min_v \sum_{i=1}^n \|v - x_i\|_2.$

На практике все эти аггрегаторы ведут себя плохо!

Centered Clipping

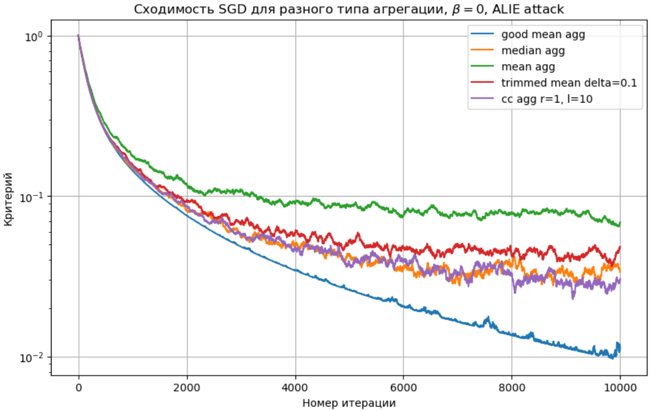
Авторы предлагают более продвинутый итеративный алгоритм аггрегации.

Алгоритм 2 AGG - Centered Clipping (CC)

1: **Input:** $(m_1, \dots, m_n)$, τ, v, L
2: **Default:** $L = 1$ and $v = \hat{m}$ (previous round aggregation)
3: **for** each iteration $l = 1, \dots, L$ **do**
4: $\quad c_i \leftarrow (m_i - v) \min \left(1, \frac{\tau}{\|m_i - v\|}\right)$
5: $\quad v \leftarrow v + \frac{1}{n} \sum_{i \in [n]} c_i$
6: **end for**
7: **Output:** v

Эксперимент 1 (FedAvg + разные алгоритмы аггрегации)

- $\delta = 0.2$, то есть 2 из 10 устройств - византийцы
- Ни один из методов аггрегации, даже продвинутый, не даёт сходимости.



Почему нет сходимости?
Комбинация FedAvg с представленными выше аггрегаторами не использует исторические данные, поэтому плохо работает на практике.
Давайте обозначим этот результат формально.

Формальная постановка

Введём формальные предположения и обозначим решаемую задачу.

Предположение

Градиент $\nabla f(x)$ - Липшицев с константой L.

Предположение

Все игроки имеют доступ к стох. градиенту $\nabla f(x, z)$, где $\mathbb{E}_z[\nabla f(x, z)] = \nabla f(x)$ и $\mathbb{E}_z[\nabla f(x, z) - \nabla f(x)] = \sigma^2$.

Предположение

На любом шаге t Византийцы имеют доступ к стохастическим градиентам g_t^j честных игроков. Номера Византийцев не меняются от шага к шагу.

Теоретические результаты

Решаем такую задачу:

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left[f(x) := \mathbb{E}_{z \sim \mathcal{D}}[f(x, z)] \right].$$

Под предположениями с предыдущего слайда, получаем:

Теорема

Если алгоритм оптимизации ALG ***не использует исторические данные****, а также функция f яв-ся $\mu > 0$ сильно выпуклой, то после $t = o(e^{\delta n})$ итераций любой такой ALG имеет ошибку:*

$$\mathbb{E}[f(\hat{x}_t)] - f(x^*) \geq \Omega(\frac{\delta \sigma^2}{\mu})$$

Устойчивый к атакам алгоритм

Значит, чтобы получить устойчивый алгоритм, нам необходимо учитывать данные с прошлых итераций. Таким свойством накопления информации обладает **моментум**

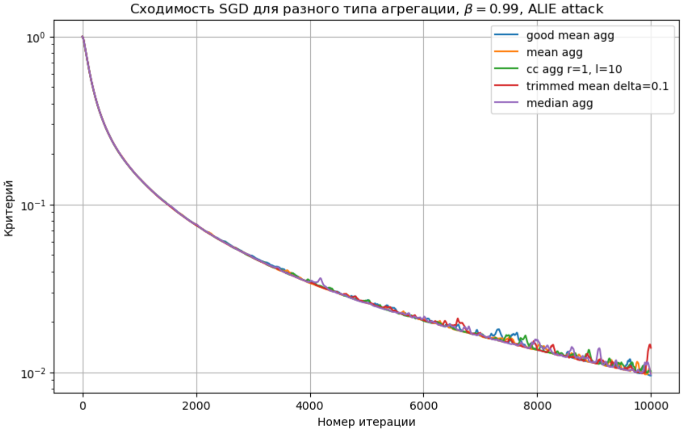
SGDm

Давайте улучшим FedAvg, добавив к нему моментум.

Алгоритм 3 SGDm

Input: x_0, η, β, AGG
Initialize: $m_i \leftarrow 0 \ \forall i \in [n]$
for each round $t = 1, \dots$ **do**
    Server communicates x_t to workers
    for worker $i \in \mathcal{G}$ **in parallel do**
        Compute stochastic gradient $g_t^i(x_t)$
        Compute $m_t^i \leftarrow (1-\beta) g_t^i(x) + \beta m_t^i$
        Communicate m_i to server
    end for
    Aggregate $\hat{m} \leftarrow \mathrm{AGG}(m_t^1, \dots, m_t^n)$
    Update $x_{t+1} \leftarrow x_t - \eta \hat{m}_t$
end for

Эксперимент 3



Добавление моментума сильно улучшило ситуацию

Теперь получим оценки сходимости.
Начнём с аггрегатора.

Сходимость Centered Clipping

Наша цель - приблизить среднее градиентов честных игроков, т.е. $\bar{g} := \frac{1}{n-\delta n} \sum_{i \in [n] \setminus \mathcal{G}} g_i$.

Теорема

При $\delta \leq 0.1$ и при $\rho^2 = \max_{i,j \in \mathcal{G}} \|g_i - g_j\|^2$, то спустя l итераций алгоритм CC с начальной точкой v_0 будет иметь следующую ошибку:

$$\mathbb{E}[\|v_l - \bar{g}\|] \leq (9.7\delta)^l \cdot \|v_0 - \bar{g}\|^2 + 4000\delta\rho^2$$

Сходимость линейная. На практике достаточно $l = 1$ для приемлемого качества.

Финальная теорема

Логическое завершение - скомбинировать моментум с продвинутым алгоритмом аггрегации.

Теорема

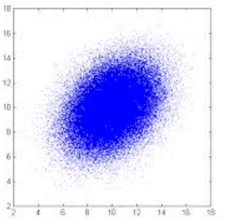
Существует послед-ть шагов η_t и послед-ть β_t, при которой по модулю предыдущих предположений, SGDm в паре с CC при $l = 1$ и с начальной точкой за T итераций сходится следующим образом:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_{t-1})\|^2] = \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\cdot\left(\frac{1}{n}+\delta\right)}\right)$$

Атаки

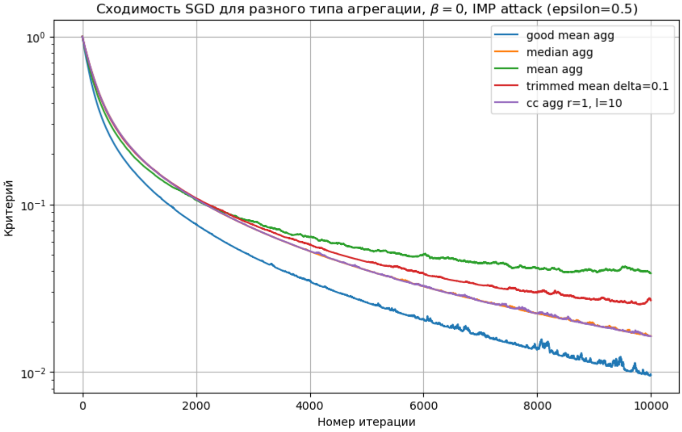
Аггрегатор - метод защиты от атак. Но какие атаки вообще бывают?
Будем считать, что атака - когда все Византийцы выбирают свои градиенты следующим образом:

- Random Noise (RN)
  $\forall j \in \mathcal{B} \quad (g_j := \nabla f(x, z) + \xi$, где $\xi \sim \mathcal{N}(0, \Sigma)$
- A Little Is Enough (ALIE)
  $\forall j \in \mathcal{B} \quad (g_j := \mu - \sigma \cdot z_{max})$
- Inner Product Manipulation (IPM)
  $\forall j \in \mathcal{B} \quad (g_j := -\epsilon \cdot \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} g_i)$



Эксперимент 3

Inner-product manipulation attack (IPM)

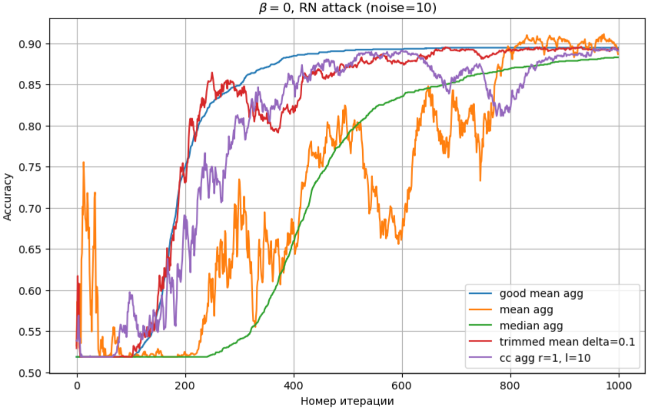


Эксперимент 5

В данном случае CC обыгрывает все остальные аггрегаторы.

Эксперимент 6

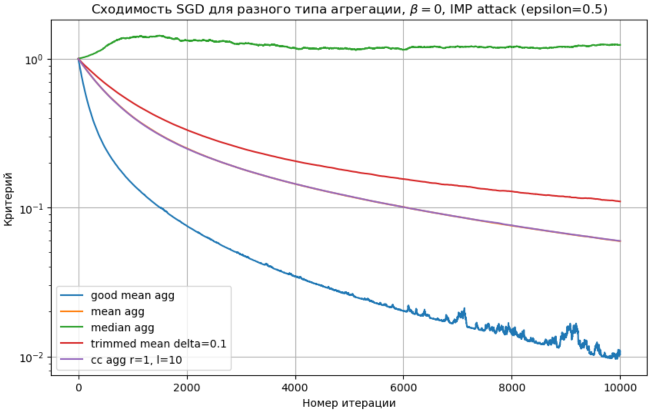
До этого считали, что распределение данных на всех устройствах одинаковое (т.н. homogenous settpmng). Что произойдёт, когда на разных устройствах будет разное распределение данных?



В не-гомогенном случае не видим тотальной доминации CC над другими методами аггрегации.

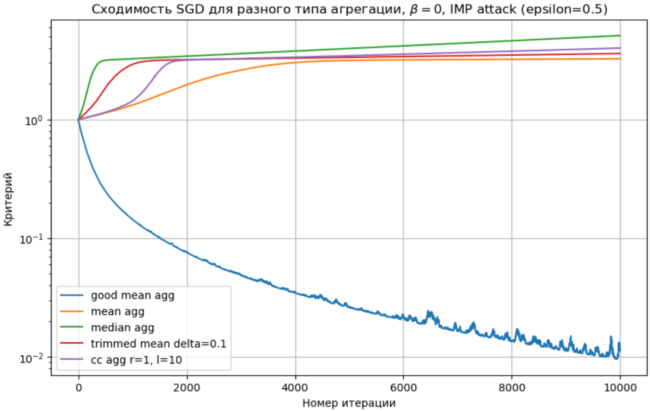
Эксперименты

Сделаем половину (5/10) устройств Византийцами



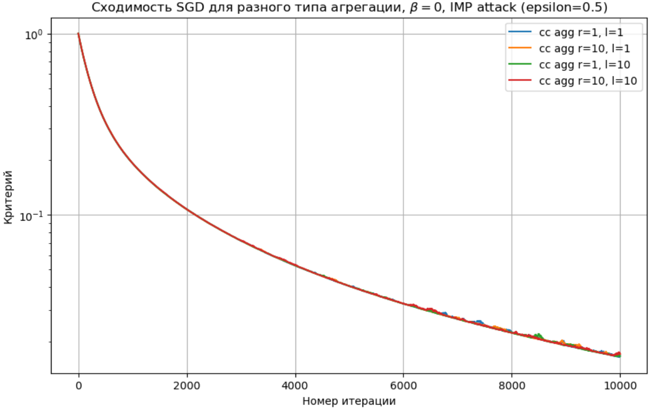
Эксперименты

Сделаем больше половины (7/10) устройств Византийцами



Эксперименты

Проверим устойчивость Centered Clipping к выбору гиперпараметров



Выводы

- Авторы рассматривают невыпуклый случай, что интересно
- Другие предположение (равномерно ограниченная дисперсия) редко имеют место на практике
- Также авторы предполагают $\delta \leq 0.1$, что довольно мало
- Статья довольно старая, но основополагающая. Последующие работы накладывали менее общие предположения и добавляли новые атаки к экспериментам.