Vertical federative learning for distributed networks

Pyotr Lisov
Ivan Toropin

Table of contents

Problem statement

Задача распределенной оптимизации в общем виде:

$$\min_{\mathbf{\Theta}} \ell(\mathbf{\Theta}; \mathcal{D}) \triangleq \frac{1}{N} \sum_{i=1}^{N} f\left(\mathbf{\Theta}; \mathbf{x}_i, y_i\right)$$

x_i , y_i - датасет. Theta - некая обобщенная модель

Problem statement

Что такое Gossip fast mix?

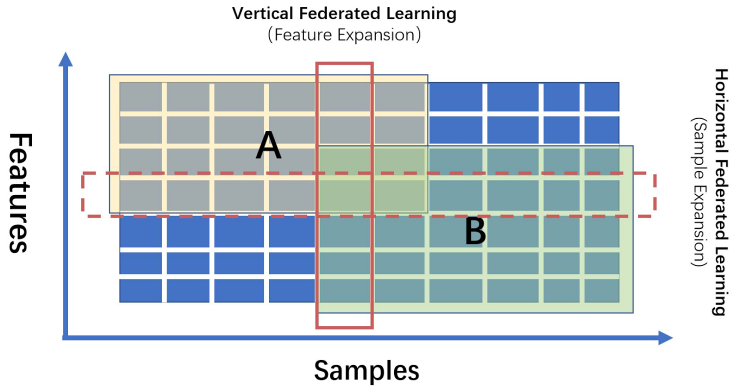
Основная идея - имея n штук “псевдоградиентов” в идеале сойтись к единому “усредненному” значению градиента

Algorithm 4 FastMix

Parameters: Vectors $z_1, ..., z_M$, communic. rounds P.
Initialization: Construct matrix $\mathbf{z}$ with rows $z_1^T, ..., z_M^T$, choose $\mathbf{z}^{-1} = \mathbf{z}$, $\mathbf{z}^0 = \mathbf{z}$, $\eta = \frac{1-\sqrt{1-\lambda_2^2(\tilde{W})}}{1+\sqrt{1-\lambda_2^2(\tilde{W})}}$.
for $h = 0, 1, 2, \ldots, P-1$ **do**
$\quad \mathbf{z}^{h+1} = (1+\eta)\tilde{W}\mathbf{z}^h - \eta\mathbf{z}^{h-1}$,
end for
Output: rows $z_1, ..., z_M$ of $\mathbf{z}^P$.

Motivation

Вертикальное обучение - подход к обучению моделей, при котором мы делим исходные данные на признаки. Это позволяет эффективно использовать данные, находящиеся в различных узлах сети, к тому же, такой подход является более безопасным.



Related works

В [2] задача ставится так:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{s} \left\| \sum_{i=1}^{n} A_i x_i - b \right\|^2 \right]$$

Каждая нода имеет свой собственный набор фич A_i, x_i и вычисляет градиент независимо с помощью L-Katyusha, а затем передает его с помощью функции AllReduce[3].

Related works

Работаем в 3 предположениях

1) μ сильная выпуклость
2) L гладкость
3) Собственные значения $AA^T \in [\mu, L]$

Related works

В данных предположениях достигается такая оценка на число итераций.

$$\mathcal{O}\left(\frac{s}{b_s}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{1}{\beta}\left(\frac{1}{s}+(w-1)\frac{sL^2}{\mu^2}\right)}+1\right)\log\frac{1}{\varepsilon}\right)$$

Эта статья примечательна прежде всего тем, что она является одной из первых, которая дала гарантии теоретической сходимости в VFL, а не только в численных тестах

Related works - Pseudocode

for $k = 0, 1, 2, \dots K$ **do**
  for $i = 1 \dots n$ in parallel **do**
    Select uniformly batch $J^k,\ |J^k| = b_s$
    $x_i^k \leftarrow \theta_1 z_i^k + \theta_2 \omega_i^k + (1 - \theta_1 - \theta_2) y_i^k$
    for $j \in J^k$ **do**
      Compute $\langle A_{ji}^T, x_i^k - w_i^k \rangle$
      Using communications
        broadcast $Q_i \left(\langle A_{ji}^T, x_i^k - w_i^k \rangle \right)$
    end for

$$g_i^k \leftarrow \frac{2}{b_s} \sum_{j \in J^k} A_{ji}^T \sum_{i=1}^{n} Q_i \left(\langle A_{ji}^T, x_i^k - w_i^k \rangle \right) + \frac{2}{s} \left(A^T A w^k - A^T b \right)_i$$

$$z_i^{k+1} \leftarrow \frac{1}{1+\eta\sigma} (\eta\sigma x_i^k + z_i^k - \frac{\eta}{L} g_i^k)$$

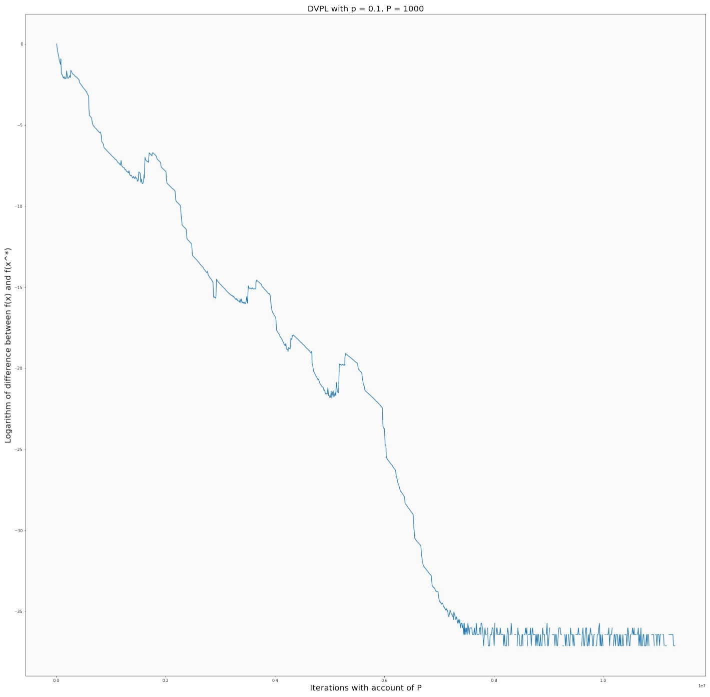
$$y_i^{k+1} \leftarrow x_i^k + \theta_1 (z_i^{k+1} - z_i^k)$$

$$\omega_i^{k+1} \leftarrow \begin{cases} y_i^k, & \text{with probability } p \\ \omega_i^k, & \text{with probability } 1 - p \end{cases}$$

    if $w_i^{k+1} = y_i^k$ **then**
      for $j = 1 \dots s$ **do**
        Compute $\langle A_{j_i}^T, x_i^k - w_i^k \rangle$
        Using communications
          broadcast $\langle A_{ji}^T, x_i^k - w_i^k \rangle$
      end for
      Compute $\frac{2}{s} \left(A^T A w^k - A^T b \right)_i$
    end if
  end for
end for

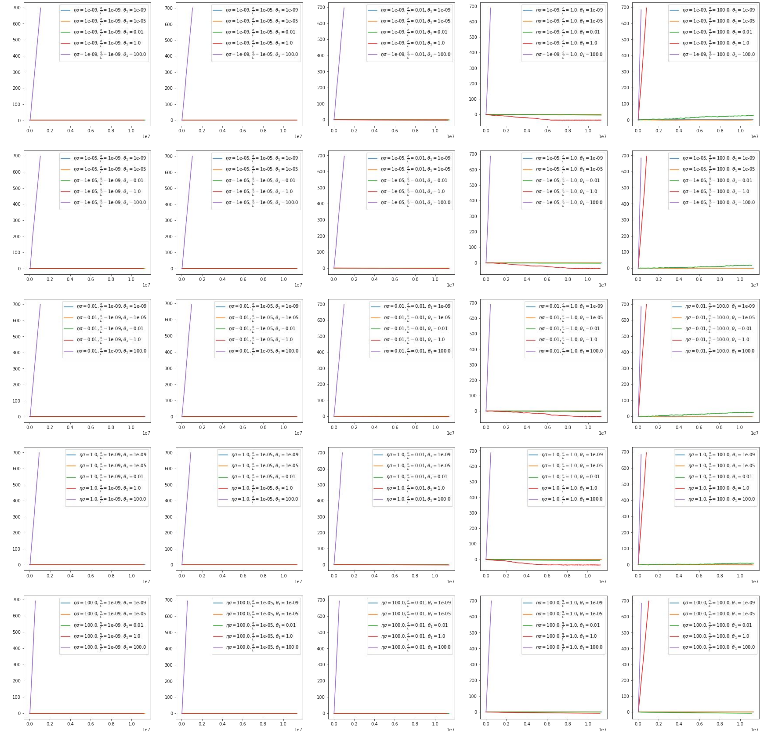
New results

Написан код для процедуры Gossip fast mix с более конкретной децентрализацией, показана сходимость.



New results

В предыдущей статье[2], эксперимент показал, что параметры сходимости отличаются порядком в 10^-8 раз. Сейчас получено обратное. Сходимость с параметрами par = max(par_i)



Github link

github.com/Lhesnor/DVPL

Plans and Expectations

Что делать дальше?

- Оценить время работы fast-mix (Просто)
- Доказать сходимость для процедуры Gossip уже аналитически.
- Доказать сходимость для нелинейной функции.

В условиях общей задачи распределенной оптимизации авторами [2], в предположении несмещенности градиента не была доказана сходимость для нелинейного случая

Подробнее можно посмотреть в доказательстве леммы C.1 в [2]

Bibliography

[1] Allen-Zhu, Z. (2016). Katyusha: The first direct acceleration of stochastic gradient methods

[2] Anonymous Authors (2024). Accelerated Methods with Compression for Horizontal and Vertical Federated Learning

[3] Kovalev, D., Horvath, S., and Richtarik, P. (2019). Don’t Jump Through Hoops and Remove Those Loops: SVRG and Katyusha are Better Without the Outer Loop

[4] Ernie Chan, Marcel Heimlich, Avi Purkayastha and Robert van de Geijn (2006). Collective communication: theory, practice, and experience. Wiley InterScience.

[5] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, Sebastian U. Stich (2020). A Unified Theory of Decentralized SGD with Changing Topology and Local Updates