Обобщенная теория децентрализованного стохастического градиентного спуска с изменяющейся топологией и локальными шагами

Общероссийский семинар по оптимизации

1 июля 2020

Machine Learning and Optimization Lab

EPFL

https://www.epfl.ch/labs/mlo/

Обобщенная теория децентрализованного стохастического градиентного спуска с изменяющейся топологией и локальными шагами

A Unified Theory of Decentralized SGD with Changing Topology and Local Updates

Anastasia Koloskova *[1] Nicolas Loizou[2] Sadra Boreiri[1] Martin Jaggi[1] Sebastian U. Stich *[1]

Abstract

Decentralized stochastic optimization methods have gained a lot of attention recently, mainly because of their cheap per iteration cost, data locality, and their communication-efficiency. In this

recently—though yet still at a smaller scale than federated learning (Lian et al., 2017; Assran et al., 2019; Koloskova et al., 2020). However, the community has identified a host of challenges that come along with decentralized training: notably, high communication cost (Tang et al., 2018a; Wang

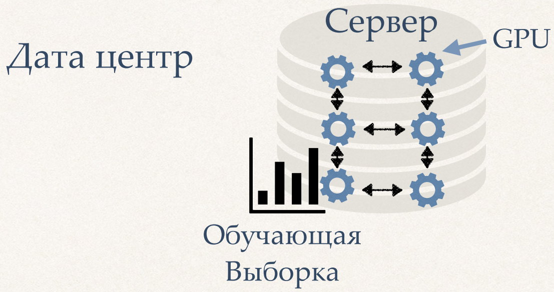
ICML 2020 https://arxiv.org/pdf/2003.10422.pdf

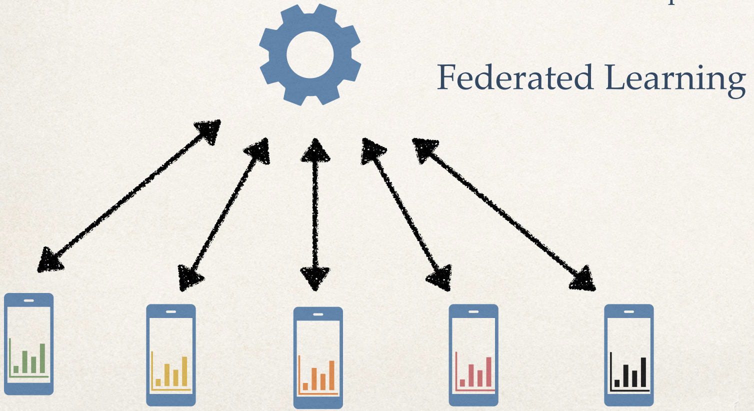
Параллельные методы оптимизации

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})$$

Централизованные

Децентрализованные



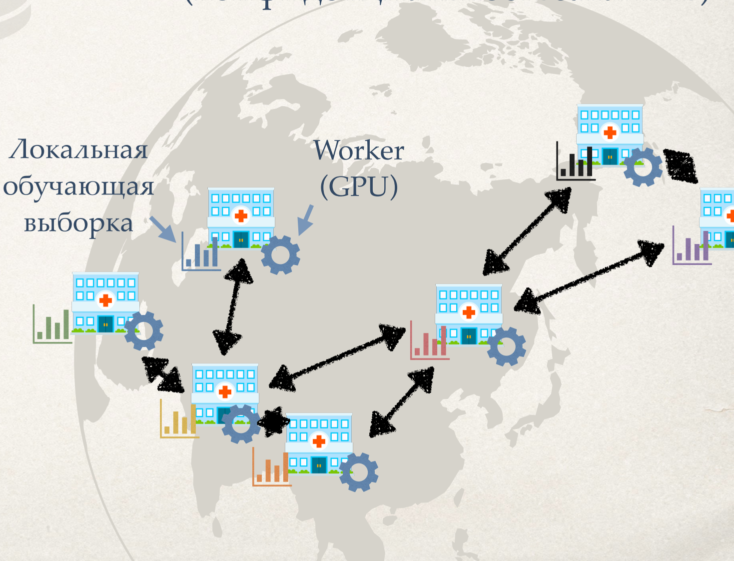


Нет центрального устройства

Локальная передача данных

Распределенная выборка

(конфиденциальность&память)



Вариации

Разные статьи используют разную технику доказательства, не все поддурживают non-iid распределенную выборку

Local SGD

(Li et al. 2020), (Khaled et al. 2020), (Wang&Joshi 2018), (Yu et al. 2019), (Basu et al. 2019), (Patel&Dieuleveut 2019), (Stich & Karimireddy 2019), (Li et al. 2019)

Decentralized Local SGD

(Wang&Joshi 2018), (Li et al. 2019)

Decentralized SGD

(Lian et al. 2017), (Wang&Joshi 2018), (Olshevsky et al. 2019), (Koloskova et al. 2019), (Li et al. 2019)

Decentralized SGD с изменяющейся топологией

(Nedic&Olshevsky 2014), (Wang et al. 2019)

and others

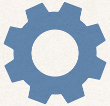
Основные результаты

- Мы анализируем все частные случаи в одном фреймворке
- Покрываем non-iid data
- Скорость сходимости либо совпадает, либо улучшает предыдущие результаты
- Предоставляем нижнюю оценку скорости сходимости

- Может помочь понять недостатки алгоритмов и для дизайна новых методов

Decentralized SGD

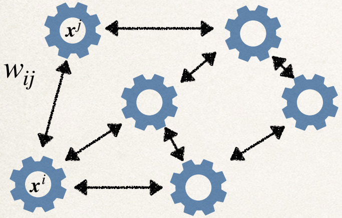
устройство i



$$\boldsymbol{x}_i^{(t+\frac{1}{2})} := \boldsymbol{x}_i^{(t)} - \gamma_t \nabla F_i(\boldsymbol{x}_i^{(t)}, \xi_{j_t})$$

шаг SGD

стохастический градиент



$$\boldsymbol{x}_i^{(t+1)} := \sum_j w_{ij}\, \boldsymbol{x}_j^{(t+\frac{1}{2})}$$

шаг усреднения

Например, для централизованной системы:

$$\boldsymbol{x}^{(t+1)} := \frac{1}{n}\sum_j \boldsymbol{x}_j^{(t+\frac{1}{2})} = \boldsymbol{x}^{(t)} - \gamma_t \frac{1}{n}\sum_j \nabla F_i(\boldsymbol{x}^{(t)}, \xi_{j_t})$$

Decentralized SGD

$$\boldsymbol{x}_i^{(t+\frac{1}{2})} := \boldsymbol{x}_i^{(t)} - \gamma_t \nabla F_i(\boldsymbol{x}_i^{(t)}, \xi_{j_t}^i)$$

$$\boldsymbol{x}_i^{(t+1)} := \sum_j w_{ij}\, \boldsymbol{x}_j^{(t+\frac{1}{2})}$$

Mixing matrix:

$$W = \{w_{ij}\}$$

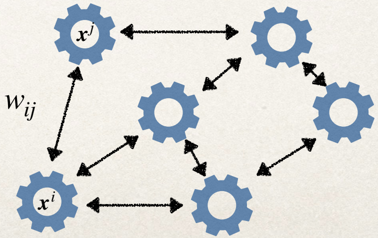
$$X^{(t)} = \left[\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_n^{(t)}\right]$$

Шаг усреднения также можно записать как:

$$X^{(t+1)} := X^{(t+\frac{1}{2})} W$$

$$W = \{w_{ij}\}$$

- Дважды стохастическая
- Симметричная



Decentralized SGD

$$x_i^{(t+\frac{1}{2})} := x_i^{(t)} - \gamma_t \nabla F_i(x_i^{(t)}, \xi_{j_t})$$

$$x_i^{(t+1)} := \sum_j w_{ij}\, x_j^{(t+\frac{1}{2})}$$

- Шаг усреднения действительно усредняет

$$\mathbb{E}\|XW - \bar{X}\|^2 \leq (1 - \underline{p})\|X - \bar{X}\|^2$$

$$p \in [0,1]$$

$$X^{(t)} = \left[x_1^{(t)}, \ldots, x_n^{(t)}\right]$$

$$\bar{X}^{(t)} = \left[\bar{x}^{(t)}, \ldots, \bar{x}^{(t)}\right]$$

$$\boxed{1 = |\lambda_1(W)|} > |\lambda_2(W)| \geq \ldots$$

$\delta = \underbrace{1 - |\lambda_2(W)|}$ — spectral gap

Decentralized SGD

$$\mathbb{E}\|XW - \bar{X}\|^2 \leq (1-p)\|X - \bar{X}\|^2$$

$$\boxed{X^{(t)}} = \left[\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_n^{(t)}\right]$$

$$\bar{X}^{(t)} = \left[\bar{\boldsymbol{x}}^{(t)}, \ldots, \bar{\boldsymbol{x}}^{(t)}\right]$$

$$\|X\underline{W} - \bar{X}\|_2^2 = \|(\cancel{X} - \bar{X})W - \underbrace{(X-\bar{X})\frac{11^T}{n}}_{0}\|_2^2 =$$

$$\bar{X} = X\frac{11^T}{n}$$

$$\bar{X} = \bar{X}\cdot W$$

$$= \|(X-\bar{X})(W - \frac{11^T}{n})\|_2^2 \leq$$

$$\leq \|X-\bar{X}\|_2^2 \underbrace{\|W - \frac{11^T}{n}\|_2^2}_{\lambda_2(W)} = \lambda_2^2(W)\|X-\bar{X}\|^2$$

$$\lambda_2^2(W) = (1-\underline{\delta})^2 = (1-p)$$

$$\boxed{p} = 1-(1-\delta)^2 = 1-1+2\delta-\delta^2 = 2\delta-\delta^2 \underset{\delta\to 0}{\approx} 2\delta = \Theta(\delta)$$

Decentralized SGD

$$x_i^{(t+\frac{1}{2})} := x_i^{(t)} - \gamma_t \nabla F_i(x_i^{(t)}, \xi_{j_t})$$

$$x_i^{(t+1)} := \sum_j w_{ij}\, x_j^{(t+\frac{1}{2})}$$

На практкике часто выбирают:

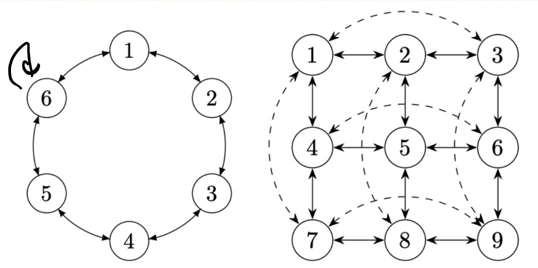
- Если все степени вершин одинаковые,

$$w_{ij} = \frac{1}{deg_i + 1}$$

- Если разные, Metropolis-Hastings правило

$$w_{ij} = \frac{1}{\max(deg_i, deg_j) + 1}$$

$$w_{ii} = 1 - \sum w_{ij}$$



graph/topology	δ^{-1}	node degree
ring	$\mathcal{O}(n^2)$	2
2d-torus	$\mathcal{O}(n)$	4
fully connected	$\mathcal{O}(1)$	$n-1$

Обобщенный decentralized SGD

$$\boldsymbol{x}^i_{t+\frac{1}{2}} := \boldsymbol{x}^i_t - \gamma_t \nabla F_i(\boldsymbol{x}^i_t, \xi_{j_t})$$

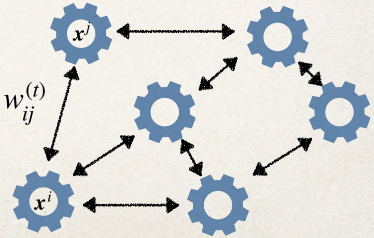
шаг SGD

стохастический градиент

$$\boldsymbol{x}^i_{t+1} := \sum_j w^{(t)}_{ij} \, \boldsymbol{x}^j_{t+\frac{1}{2}}$$

шаг усреднения с весами $\{w^{(t)}_{ij}\} \sim \mathscr{W}^{(t)}$

- Возможно менять топологию сети на каждой итерации
- Разрешено случайным образом выбирать веса из какого-то распределения



Предположение

$$x^i_{t+\frac{1}{2}} := x^i_t - \gamma_t \nabla F_i(x^i_t, \xi_{j_t})$$

$$x^i_{t+1} := \sum_j w^{(t)}_{ij}\, x^j_{t+\frac{1}{2}}$$

$$W^{(t)} = \{w^{(t)}_{ij}\}$$

- Дважды стохастическая

$$\underbrace{W^{(l+1)\tau-1}\dots W^{l\tau}}_{W_{l,\tau}}\dots\underbrace{W^{(2\tau-1)}\dots W^{(\tau)}}_{W_{1,\tau}}\underbrace{W^{(\tau-1)}\dots W^{(0)}}_{W_{0,\tau}}$$

- Нужно только что комбинация каждых τ шагов усредняет, а не каждый отдельный шаг

$$\mathbb{E}\|XW_{l,\tau} - \bar{X}\|^2 \leq (1-p)\|X - \bar{X}\|^2 \qquad \forall l$$

Пример: Decentralized Local SGD

$$\underbrace{WI\cdots I}_{W_{l,\tau}} \cdot \cdots \cdot \underbrace{WI\cdots I}_{W_{1,\tau}} \cdot \underbrace{WI\cdots I}_{W_{0,\tau}}$$

I — единичная матрица, без передачи данных

Сходимость

$f_i(x) = \mathbb{E}_\xi F_i(x,\xi)$

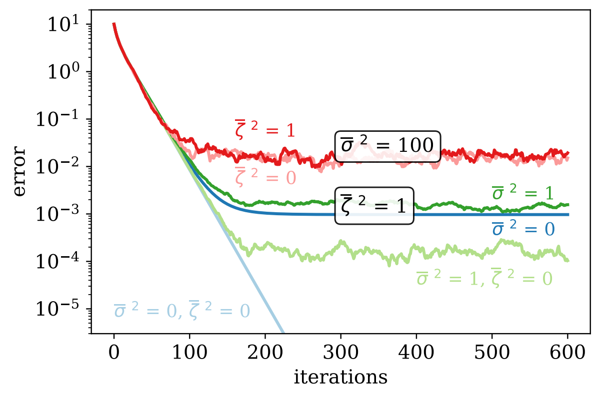
$$f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \le \langle \nabla f_i(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle$$

$$\|\nabla F_i(\boldsymbol{y},\xi) - \nabla F_i(\boldsymbol{x},\xi)\| \le L\|\boldsymbol{y}-\boldsymbol{x}\|$$

Сильно выпуклый случай

$$\tilde{\mathcal{O}}\left(\frac{\bar\sigma^2}{nT} + \frac{\tau^2\bar\zeta^2 + \tau p\bar\sigma^2}{p^2T^2} + \frac{\tau}{p}\exp\left[-\frac{Tp}{\tau}\right]\right)$$

стохастические члены   оптимизационные члены



$$\bar\sigma^2 = \frac{1}{n}\sum_i \sigma_i^2, \quad \sigma_i^2 := \mathbb{E}_{\xi_i}\|\nabla F_i(\boldsymbol{x}^\star,\xi_i) - \nabla f_i(\boldsymbol{x}^\star)\|_2^2 \longleftarrow$$ дисперсия градиентов

$$\bar\zeta^2 = \frac{1}{n}\sum_i \zeta_i^2, \qquad \zeta_i^2 := \|\nabla f_i(\boldsymbol{x}^\star)\|_2^2 \longleftarrow$$ разница функций

Сходимость: идея доказательства

$\eta_t = \gamma_t$

Лемма 1

$$\mathbb{E}\|\bar{\boldsymbol{x}}^{(t+1)} - \boldsymbol{x}^\star\|^2 \le \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}\|\bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^\star\|^2 - \eta_t(f(\bar{\boldsymbol{x}}^{(t)}) - f(\bar{\boldsymbol{x}}^\star)) + \frac{\bar{\sigma}^2}{n}\eta_t^2 + \eta_t 3L\Xi_t$$

$$\Xi_t = \frac{1}{n}\mathbb{E}_t\sum_{i=1}^{n}\|\boldsymbol{x}_i^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|^2$$

$$\tilde{W} \sim \begin{cases} W & \text{w.p. } \frac{1}{\tau} \\ I & \text{w.p. } 1-\frac{1}{\tau}. \end{cases}$$

Лемма 2

$$\Xi_t \le \left(1 - \frac{p}{2}\right)\Xi_{m\tau} + \frac{p}{16\tau}\sum_{j=m\tau}^{t-1}\Xi_j + 36L\frac{\tau}{p}\sum_{j=m\tau}^{t-1}\eta_j^2(f(\bar{\boldsymbol{x}}^{(t)}) - f(\bar{\boldsymbol{x}}^\star)) + \left(\bar{\sigma}^2 + \frac{18\tau}{p}\bar{\zeta}^2\right)\sum_{j=m\tau}^{t-1}\eta_j^2$$

Нижняя оценка

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{nT} + \frac{\tau^2\bar{\zeta}^2 + \tau p\bar{\sigma}^2}{p^2T^2} + \frac{1}{p}\exp\left[-\frac{Tp}{\tau}\right]\right)$$

Сильно выпуклый случай

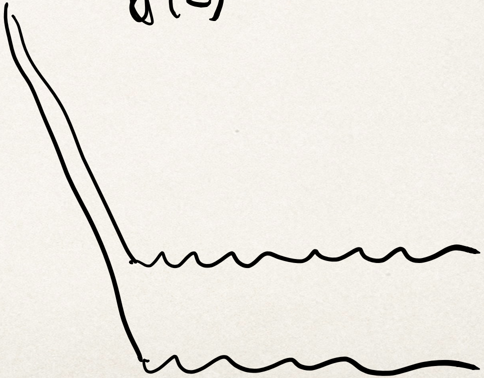
$$\bar{\sigma}^2 = 0 \qquad \tau = 1$$

$$\tilde{\Omega}\left(\frac{\bar{\zeta}^2(1-p)}{p^2T^2}\right)$$

Сублинейный член необходим

$\boxed{\varepsilon}$

$\gamma(\varepsilon)$



Нижняя оценка: идея

$$x^i_{t+\frac{1}{2}} := x^i_t - \gamma_t \nabla F_i(x^i_t, \xi_{j_t})$$
$$x^i_{t+1} := \sum_j w^{(t)}_{ij}\, x^j_{t+\frac{1}{2}}$$

Decentralized SGD с матрицей W

$$f_i(x) = \frac{1}{2}(x - y_i)^2 \qquad x \in \mathbb{R}^1$$

$\boldsymbol{x}^{(0)} = (x^{(0)}_1, \ldots, x^{(0)}_n)^\top$ — собственный вектор W, отвечающий второму собственному числу p

$$\boldsymbol{y} = \boldsymbol{x}^{(0)} + \mathbf{1}$$

Итерация равносильна:

$$\bar{\sigma}^2 = 0 \qquad \tau = 1$$

$$\boldsymbol{x}^{(t+1)} = W\boldsymbol{x}^{(t+\frac{1}{2})} = W((1-\gamma)\boldsymbol{x}^{(t)} + \gamma \boldsymbol{y}) =$$

$$\underline{W^t}(1-\gamma)^t \underline{x^{(0)}} + \gamma \sum_\tau \underline{(1-\gamma)^{t-\tau}} W^{t-\tau} y$$

$$\tilde{\Omega}\left(\frac{\bar{\zeta}^2(1-p)}{p^2T^2}\right)$$

Частный случай: Local SGD

Наилучшая скорость сходимости ранее:

$$\mathcal{O}\left(\frac{\bar{\sigma}^2}{nT} + \frac{\tau^2\bar{\zeta}^2}{T}\right)$$ (Li et al. 2020)

Эта работа:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{nT} + \frac{\tau^2\bar{\zeta}^2 + \tau\bar{\sigma}^2}{T^2} + \tau \exp\left[-\frac{T}{\tau}\right]\right)$$

Скорость сходимости совпадает с ранее известным результатом для невыпуклых функций

вычисление градиентов, без общения

$W^{(t)} = I$

$W^{(t)} = I$

$\dots$

τ

2τ

time

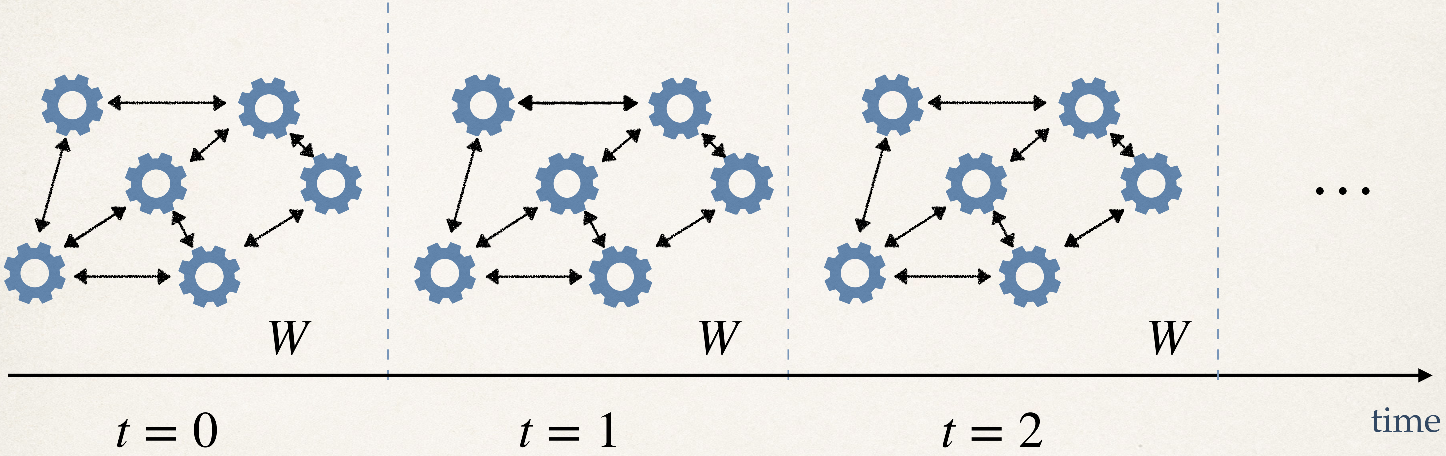
$$W^{(t)} = \frac{11^\top}{n}$$ (полный граф)

централизованная all-to-all агрегация

Улучшает ранее известный результат для выпуклых и сильно выпуклых функций

Также в литературе называют FedAvg, популярен в Federated Learning

Частный случай: Decentralized SGD



Наилучшая скорость сходимости ранее:

$$\mathcal{O}\left(\frac{\bar{\sigma}^2}{nT} + \frac{G^2}{p^2T^2} + \frac{G^2}{p^3T^3}\right)$$

(Koloskova et al., 2019)
(Olshevsky et al., 2019)

Эта работа:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{nT} + \frac{\bar{\zeta}^2 + p\bar{\sigma}^2}{p^2T^2} + \frac{1}{p}\exp\left[-Tp\right]\right)$$

$$G^2 \geq \bar{\zeta}^2 + \bar{\sigma}^2$$

Совпадает с ранее известным для невыпуклых функций

Улучшает ранее известный результат для выпуклых и сильно выпуклых функций

Частный случай: Decentralized Local SGD

вычисление градиентов, без общения

$W^{(t)} = I$

$W^{(t)} = I$

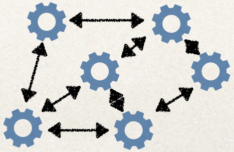
$\dots$

Эта работа:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{nT} + \frac{\tau^2\bar{\zeta}^2 + \tau p\bar{\sigma}^2}{p^2T^2} + \frac{\tau}{p}\exp\left[-\frac{Tp}{\tau}\right]\right)$$

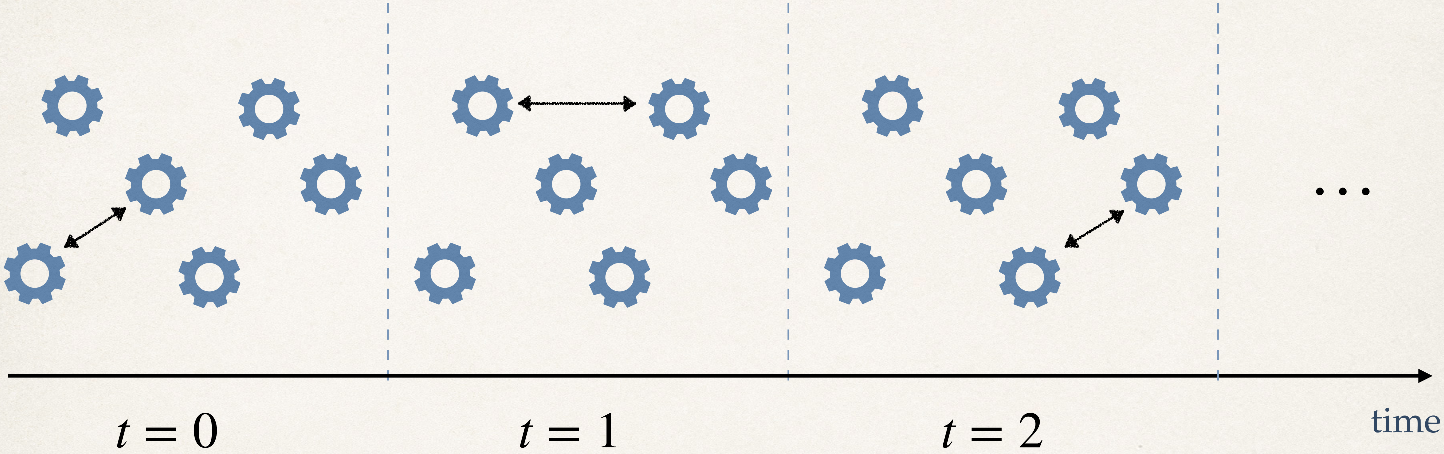
τ $\quad 2\tau \quad$ time

$W^{(t)} = W$



Улучшает известные результаты во всех случаях

Частный случай: Pairwise randomised gossip



$$\frac{1}{T+1}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^{(t)})\|$$

Эта работа:

$$\frac{1}{T}\sum_{t=1}^{T}\underbrace{\mathbb{E} f(\bar{x}^t) - f(x^\star)} \leq \tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{nT} + \frac{\bar{\zeta}^2 + p\bar{\sigma}^2}{p^2T^2} + \frac{1}{p}\exp\left[-Tp\right]\right)$$

Частный случай: Overparametrized regime

В оптимуме, функция потерь равна нулю для каждого из обучающего примера

$$\nabla F_i(\boldsymbol{x}^\star, \xi) = 0 \quad \forall i,\ \xi$$

$$\bar{\sigma} = 0 \qquad\qquad \bar{\zeta} = 0$$

Линейная сходимость

$$\tilde{\mathcal{O}}\left(\frac{L\tau R_0^2}{p}\exp\left[-\frac{\mu T p}{\tau L}\right]\right)$$

Заключение

- Предложен обобщенный фреймворк для параллельной оптимизации
- Вывели теоретическую скорость сходимости с non-iid data
- Скорость сходимости совпадает и улучшает предыдущие результаты
- Вывели нижнюю оценку для decentralized SGD