Comparison to Other Methods (in strongly convex case)

Defining $\zeta$ as function heterogeneity $\sum_{i=1}^{n} \|\nabla f_i(x_i) - \nabla f_j(x_j)\|_2^2 \leq \zeta^i$.

[Li, +, 17], [Koloskova+, 20] $\tilde{O}\left(\frac{\sigma^2}{\mu n} + \frac{\sqrt{\sigma^2}{\sqrt{\mu}}}{\mu n} + \frac{p L}{\mu n} \log \frac{1}{\epsilon}\right)$ (D-SGD)

[Tang+, 18], [Yuan & Alghunaimi, 21] $O\left(\frac{\sigma^2}{\mu n} + \frac{\sqrt{\sigma^2}{\sqrt{\mu}}}{\mu n} + \frac{p L}{\mu n} \log \frac{1}{\epsilon}\right)$ $\geq \frac{c}{\zeta^i \sqrt{d}}$ ($D^2$)

Important Advances for GT (in strongly convex case)

Reference | rate of convergence to $\epsilon$-accuracy | considered stochastic noise
--- | --- | ---
[Lojdani+, 16] | $O\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$ | $\times$
[Alghunaimi+, 19] | $O\left(\frac{\log L}{\sqrt{\log L}}\right)$ | $\times$
[Qu & Li, 17] | $O\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$ | $\times$
[Lu & Nedic, 20] | $O\left(\frac{\sigma^2}{\mu n} + \frac{\sqrt{\sigma^2}{\sqrt{\mu}}}{\mu n} + \frac{p L}{\mu n} \log \frac{1}{\epsilon}\right)$ | $\checkmark$

Important Advances for GT (in non-convex case)

Reference | rate of convergence to $\epsilon$-accuracy | considered stochastic noise
--- | --- | ---
[Lorenzo & Scutari, 16] | asymptotic convergence guarantees | $\times$
[Zhao & You, 20] | $O\left(\frac{\log L}{\sqrt{\log L}}\right)$ | $\checkmark$
[Lu+, 19] | $O\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$ | $\checkmark$

| [we] | $O\left(\frac{\sigma^2}{\mu n} + \frac{\sqrt{\sigma^2}{\sqrt{\mu}}}{\mu n} + \frac{p L}{\mu n} \log \frac{1}{\epsilon}\right)$ | $\checkmark$

Proof Idea

In matrix notation, GT is equal to

$$\begin{align*}
\Delta x^{(t+1)} &= (I - \Omega L) \hat{\Delta} x^{(t)} \\
\hat{\Delta} x^{(t+1)} &= (I - \Omega L) \hat{\Delta} x^{(t)}
\end{align*}$$

To get convergence we need contraction properties on $\Omega$, how ever it is not a conservative operator, i.e. $\|\tilde{J}\| > 1$. But we can prove

$$\tilde{J} \geq \frac{1}{\tilde{J}} \text{ it holds that } \|\tilde{J}\| \leq 1$$

Thus we measure progress only after every $\tau$ steps

$$\Psi_{t+\tau} = \Psi_{t+\tau} \frac{1}{\tau} \sum_{i=1}^{\tau} E_{\tau_{i+\tau}}$$

- Parameter $c$ comes from careful estimation of the gradient term $\sum_{j=1}^{\tau} E_{\tau_{i+\tau}}$.
- Another technical difficulty arises from possibility of divergence during intermediate steps, due to $\|\tilde{J}\| > 1$. But we can prove

Discussion

- Derived improved complexity bounds for the GT method, that improve over all previous results.
- The smallest eigenvalue of the mixing matrix has a strong impact on the performance of GT.
- The smallest eigenvalue can often be controlled in practice by choosing large enough self-weights $w_i$ of the mixing matrix $W_i$. 

**Problem Setup**

Decentralized Optimization Problem on $n$ nodes:

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

$f_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ can be stochastic: $f_i(x) = E[Z_i(x, \xi_i)]$

Assumptions & Notation:
- nodes can only communicate with neighbors in graph $G$
- $G = (\mathbb{N}, E)$, averaging weights $W_{ij} \geq 0 \iff (i, j) \in E$, $W$ is doubly stochastic ($W^2 = W$) and symmetric ($W^{T} = W$).
- $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ are $L$-smooth $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|_2$

For some of the results we assume convexity $f$ / strong-convexity $f(x_i) - f(y_i) + \frac{\gamma}{2} \|x_i - y_i\|_2^2 \leq (\nabla f_i(x), x_i - y_i)$

- access to gradient oracles $g_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$, s.t. $\forall x \in \mathbb{R}^d$

$\nabla g_i(x) = \nabla f_i(x)$, $\forall x \in \mathbb{R}^d$,

- $\sigma$ - Sometimes we will use matrix notation, for vectors $x_i \in \mathbb{R}^d$ defined for $i \in [n]$

$Z = [x_1, \ldots, x_n] \in \mathbb{R}^{n \times d}$, $\Delta Z = Z - \frac{1}{n}ż$.

**Algorithm: Gradient Tracking** ([Lorenzo & Scutari, 16], [Nedic+, 16])

**Important Advances for GT (in strongly convex case)**

**Assumptions on Mixing Matrix $W$**

$$\begin{align*}
1 = \lambda_1(W) > \lambda_2(W) \geq \cdots \geq \lambda_n(W) \geq -1
\end{align*}$$

We assume that $p > 0$ (and consequently $c > 0$).

- $c \geq p$ for all graphs.
- If $w_i \geq p > 0$ (self-weight), then $c \geq \min \{2p, 1\}$ (corollary of Gershgorin’s circle theorem).
- Thus $c$ can be controlled in practice by choosing large enough self-weights $w_i$.
- For common Metropolis-Hastings rule, $w_i = w_j = \min \{\frac{1}{\sum_{l=1}^{n} w_l}, \frac{1}{\sum_{l=1}^{n} w_l}\}$. 

PEEDL

An Improved Analysis of Gradient Tracking for Decentralized Machine Learning

Anastasia Koloskova, Tao Lin, Sebastian Stich