NeurIPS, Dec 6–14, online

An Improved Analysis of Gradient Tracking for Decentralized Machine Learning

Anastasia Koloskova, Tao Lin, Sebastian Stich

Problem Setup

Decentralized Optimization Problem on n nodes:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})\right]$$

$f_i\colon \mathbb{R}^d \to \mathbb{R}$ can be stochastic: $f_i(\mathbf{x}) := \mathbb{E}_{\xi_i} F_i(\mathbf{x}, \xi_i)$

Assumptions & Notation:

- nodes can only communicate with *neighbors* in network G
- $G = ([n], E)$, averaging weights $W_{ij} \geq 0 \Leftrightarrow \{i,j\} \in E$, W is doubly stochastic ($W\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top W = \mathbf{1}^\top$) and symmetric ($W^\top = W$).
- $f_i\colon \mathbb{R}^n \to \mathbb{R}$ are L-smooth $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$.
  For some of the results we assume convexity / μ-strong convexity
  $$f_i(\mathbf{x}) - f_i(\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \langle \nabla f_i(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle\,.$$
- access to gradient oracles, $\mathbf{g}_i\colon \mathbb{R}^d \to \mathbb{R}^d$, s.t. $\forall \mathbf{x}\in\mathbb{R}^n$:
  $$\mathbb{E}\,\mathbf{g}_i(\mathbf{x}) = \nabla f_i(\mathbf{x})\,, \qquad \operatorname{Var} \mathbf{g}_i \leq \sigma^2$$
- Sometimes we will use matrix notation, for vectors $\mathbf{z}_i \in \mathbb{R}^d$ defined for $i \in [n]$
  $$Z := [\mathbf{z}_1, \dots, \mathbf{z}_n] \in \mathbb{R}^{d\times n}\,, \qquad \bar{Z} := [\bar{\mathbf{z}}, \dots, \bar{\mathbf{z}}] \equiv Z\tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\,, \qquad \Delta Z := Z - \bar{Z}\,.$$

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

input: Initial values $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$ on each node $i \in [n]$, communication graph $G = ([n], E)$ and mixing matrix W, stepsize γ, initialize $\mathbf{y}_i^{(0)} = \nabla F_i(\mathbf{x}_i^{(0)}, \xi_i^{(0)})$, $\mathbf{g}_i^{(0)} = \mathbf{y}_i^{(0)}$ in parallel for $i \in [n]$.

1: **for** $t = 0, \dots, T-1$ **do**
2: each node i sends $\left(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}\right)$ to is neighbors
3: $\mathbf{x}_i^{(t+1)} = \sum_{j:\{i,j\}\in E} w_{ij}\left(\mathbf{x}_j^{(t)} - \gamma \mathbf{y}_j^{(t)}\right)$ ▷ update model parameters
4: Sample $\xi_i^{(t+1)}$, compute gradient $\mathbf{g}_i^{(t+1)} = \nabla F_i\left(\mathbf{x}_i^{(t+1)}, \xi_i^{(t+1)}\right)$
5: $\mathbf{y}_i^{(t+1)} = \sum_{j:\{i,j\}\in E} w_{ij}\mathbf{y}_j^{(t)} + \left(\mathbf{g}_i^{(t+1)} - \mathbf{g}_i^{(t)}\right)$ ▷ update tracking variable
6: **end for**

- GT algorithm is not affected by functions / data heterogeniety due to tracking of gradients in variables $\mathbf{y}_i$.
- We develop a new, and improved, analysis of the GT algorithm.
- Our analysis improves over all existing results that analyze the GT algorithm.
- We use a novel proof technique, that might be of independent interest.

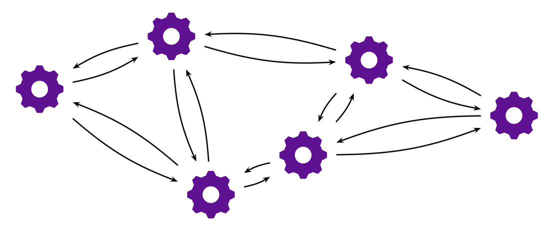
Assumptions on Mixing Matrix W

$$1 = \lambda_1(W) > \lambda_2(W) \geq \dots \geq \lambda_n(W) > -1 \qquad \text{eigenvalues of } W.$$

$$p = 1 - \max\{|\lambda_2(W)|, |\lambda_n(W)|\}^2\,, \qquad c = 1 - \min\{\lambda_n(W), 0\}^2\,.$$

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

- $c \geq p$ for all graphs.
- If $w_{ii} \geq \rho > 0$ (self-weight), then $c \geq \min\{2\rho, 1\}$ (corollary of Gershgorin's circle theorem).
- Thus c can be controlled in practice by choosing large enough self-weights w_{ii}.
- For common Metropolis-Hastings rule, $w_{ij} = w_{ji} = \min\left\{\frac{1}{\deg(i)+1}, \frac{1}{\deg(j)+1}\right\}$, $w_{ii} = 1 - \sum_{j=1}^n w_{ij} \geq \frac{1}{\max_{j\in[n]}\deg(j)}$



Details

Theorem (GT convergence in the general case): There exists a stepsize γ such that if $T > \frac{2}{p}\log\left(\frac{50}{p}(1 + \log\frac{1}{p})\right)$, GT converges at the rate:

Non-convex:
$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{n\varepsilon} + \frac{\sigma}{\sqrt{p}c\varepsilon^{3/2}} + \frac{1+L}{pc\varepsilon}\right)\cdot L$$

Strongly-convex:
$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\varepsilon} + \frac{\sqrt{L}\sigma}{\mu\sqrt{p c}\sqrt{\varepsilon}} + \frac{L}{\mu p c}\log\frac{1}{\varepsilon}\right)$$

General convex:
$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{n\varepsilon^2} + \frac{\sqrt{L}\sigma}{\sqrt{p}c\varepsilon^{3/2}} + \frac{L}{pc\varepsilon}\right).$$

Comparison to Other Methods (in strongly convex case)

Defining ζ as function heterogeneity $\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(\mathbf{x}^\star) - \nabla f(\mathbf{x}^\star)\|_2^2 \leq \zeta^2$,

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

[Tang+, 18], [Yuan & Alghunaim, 21] $\quad \mathcal{O}\left(\frac{\sigma^2}{\mu n\varepsilon} + \frac{\sqrt{L}\sigma}{\mu\sqrt{p}\sqrt{\varepsilon}} + \frac{L}{\mu p}\log\frac{1}{\varepsilon}\right)$ $c \geq \frac{8}{9}$ (D^2)

Important Advances for GT (in strongly convex case)

Reference	rate of convergence to ϵ-accuracy	considered stochastic noise
[Nedic+, 16]	$\mathcal{O}\left(\frac{L^3}{\mu^3 p^2}\log\frac{1}{\varepsilon}\right)$	✗
[Alghunaim+, 19]	$\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon} + \frac{1}{p^2}\log\frac{1}{\varepsilon}\right)$	✗
[Qu & Li, 17]	$\mathcal{O}\left(\frac{L^2}{\mu^2 p^2}\log\frac{1}{\varepsilon}\right)$	✗
[Pu & Nedic, 20]	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\varepsilon} + \frac{\sqrt{L}\sigma}{\mu\sqrt{p}p\sqrt{\varepsilon}} + \frac{C_1}{\sqrt{\varepsilon}}\right)$ [a]	✓
[we]	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\varepsilon} + \frac{\sqrt{L}\sigma}{\mu\sqrt{p}c\sqrt{\varepsilon}} + \frac{L}{\mu p c}\log\frac{1}{\varepsilon}\right)$	✓

[a] C_1 is a constant that is independent of ε, but can depend on other parameters, such as σ, μ, L, p

Important Advances for GT (in non-convex case)

Reference	rate of convergence to ϵ-accuracy	considered stochastic noise
[Lorenzo & Scutari, 16]	asymptotic convergence guarantees	✗
[Zhang & You, 20]	$\mathcal{O}\left(\frac{Ln\sigma^2}{\varepsilon^2} + \frac{Ln}{p^3\varepsilon}\right)$	✓
[Lu+, 19]	$\mathcal{O}\left(\frac{C_1 + C_2\sigma}{\varepsilon^2}\right)$ [a]	✓
[we]	$\tilde{\mathcal{O}}\left(\frac{L\sigma^2}{n\varepsilon^2} + \frac{L\sigma}{\sqrt{p}c\varepsilon^{3/2}} + \frac{L}{pc\varepsilon}\right)$	✓

[a] C_1 and C_2 are constants that are independent of ε, but can depend on other parameters, such as σ, n, L, p.

Proof Idea

In matrix notation, GT is equal to

$$\begin{pmatrix}\Delta X^{(t+1)}\\ \gamma\Delta Y^{(t+1)}\end{pmatrix}^\top = \underbrace{\begin{pmatrix}\Delta X^{(t)}\\ \gamma\Delta Y^{(t)}\end{pmatrix}^\top}_{=:\Psi_t}\underbrace{\begin{pmatrix}\tilde{W} & 0\\ -\tilde{W} & \tilde{W}\end{pmatrix}}_{=:J} + \gamma\underbrace{\begin{pmatrix}0\\ \left(\nabla F(X^{t+1}, \xi^{t+1}) - \nabla F(X^t, \xi^t)\right)\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\end{pmatrix}^\top}_{=:E_t},$$

To get convergence we need contraction properties on J, however it is not a contractive operator, i.e. $\|J\| > 1$. But we can prove

For $\tau \geq \tilde{\mathcal{O}}\left(\frac{1}{p}\right)$ it holds that $\|J^\tau\| \leq \frac{1}{2}$.

Thus we measure progress only after every τ steps

$$\Psi_{t+\tau} = \Psi_t J^\tau + \gamma\sum_{j=1}^{\tau-1} E_{t+j-1}J^{\tau-j}\,.$$

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

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_{ii} of the mixing matrix W.