Incremental Consensus based Collaborative Deep Learning

A short report on [Jiang et al., 2018]

Nazarov Ivan

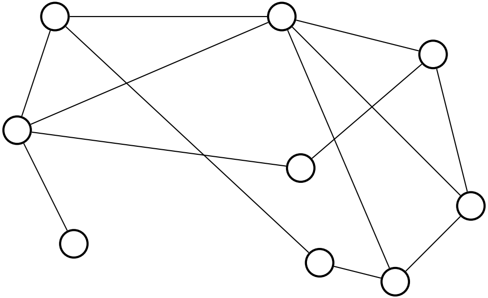
Skoltech

July 27, 2018

General Setting

Agents connected within a communication graph

- want to collaboratively solve an optimization problem



Generic Optimisation in ML

Agents are nodes in a undirected connected graph $G = (V, E)$

Each node $v \in G$ has a set of objectives $S_v = (l_{vi} : \mathbb{R}^d \to \mathbb{R})_{i=1}^{m_v}$

- $l_{vi}(\theta) = \text{Loss}(x_{vi}, y_{vi}; \theta)$ – supervised learning
- $l_{vi}(\theta) = \text{Loss}(z_{vi}; \theta)$ – unsupervised learning

Goal – to solve the global problem

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \sum_{v \in G} \sum_{i \in v} l_{vi}(\theta) \,. \qquad \text{(Glbl-P)}$$

Distributed Optimisation in ML

Assumptions on G

- two-way communication: if $uv \in E$ then $vu \in E$
- for any $v \in G$ the neighbours $G_v = \{u \in G \colon vu \in E\}$
- single connected component

Assumptions on l_{vi}: γ_{vi}-smooth, proper and coercive

- $l_{iv}(y) \leq l_{iv}(x) + \langle \nabla l_{iv}(x), y - x \rangle + \frac{\gamma_{vi}}{2} \|y - x\|^2$ for all y
- $dom\, l_{vi} = \{\theta \colon l_{vi}(\theta)\} \neq \emptyset$
- coercive $l_{vi}(\theta) \to \infty$ as $\|\theta\| \to \infty$

Distriubuted and Decentralized Optimisation

Consensus problem

$$\begin{aligned} \underset{\theta_v \in \Theta}{\text{minimize}} \quad & \sum_{v \in G} \sum_{i \in v} l_{vi}(\theta_v)\,. \\ \text{subject to} \quad & \theta_v = \theta_u\,, \forall v \in G \;\forall u \in G_v\,. \end{aligned} \tag{Cons-P}$$

Equivalent problem: if A – symmetric adjacency matrix of G then

$$\begin{aligned} \underset{\boldsymbol{\theta} \in \Theta^G}{\text{minimize}} \quad & \sum_{v \in G} \sum_{i \in v} l_{vi}(\boldsymbol{\theta}_v)\,. \\ \text{subject to} \quad & A\boldsymbol{\theta} = \boldsymbol{\theta}\,. \end{aligned}$$

The Optimization Updates in Distributed Optimization

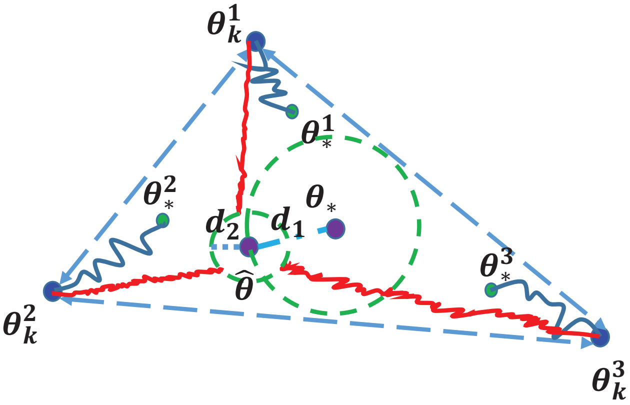


Figure 1: **Blue** *dots* ($\boldsymbol{\theta}_k^i \sim \theta_v^t$) *represent the current states;* **Green** *dots represent the private local optima* ($\boldsymbol{\theta}_*^i \sim \theta_v^*$); **Purple** *dot* ($\boldsymbol{\theta}_* \sim \theta^*$) *represents the ideal global optimal solution; another* **purple** *dot* $\hat{\boldsymbol{\theta}} \sim \hat{\theta}^t$ *represents a possible consensus point.*

Consensus Distributed SGD

CD-SGD update looks like this

$$\boldsymbol{\theta}_v^{t+1} \leftarrow \underbrace{\left\{ \frac{1}{|G_v|} \sum_{u \in G_v} \boldsymbol{\theta}_u^t \right\}}_{\text{synchronisation step}} - \underbrace{g_v^t(\boldsymbol{\theta}_v^t)\eta}_{\text{gradient step}} \;,$$

where $g_v^t(\boldsymbol{\theta}_v^t)$ is the stochastic gradient of l_v at $\boldsymbol{\theta}_v^t$

$$g_v^t(\boldsymbol{\theta}_v^t) = \nabla_\theta \hat{\mathbb{E}}_{l \sim S_v} l(\theta) \Big|_{\theta = \boldsymbol{\theta}_v^t} \;.$$

Proposal of the paper

Instead of this ... (CDSGD)

$$\boldsymbol{\theta}^{t+1} \leftarrow \left(\Pi \otimes I_d\right)\boldsymbol{\theta}^t - \boldsymbol{g}(\boldsymbol{\theta}^t)\eta\,.$$

... why not do this

$$\boldsymbol{\theta}^{t+1} \leftarrow \left(\Pi^\tau \otimes I_d\right)\boldsymbol{\theta}^t - \boldsymbol{g}(\boldsymbol{\theta}^t)\eta\,,$$

thereby increasing the communication complexity by $\tau \geq 1$.

... and get

- the *incremental*-CDSGD (i-CDSGD)

Results: consensus bound

Consensus with fixed step size, i-CDSGD

Let Π be the double stochastic version of A with $\lambda_2 < 1$:

- $\lambda_j = \lambda_j(\Pi)$ – the j-th largest eigenvalue of Π

If assumptions hold and $0 < \alpha \leq \frac{r_1 - (1-\lambda_N^\tau)B_m}{\gamma_m B_m}$, then iterates of i-CDSGD satisfy the following inequality: $\forall t \geq 1$

$$\mathbb{E}\left\|\theta_v^t - \frac{1}{|G|}\sum_{v\in G}\theta_v^t\right\| \leq \frac{\alpha\sqrt{B + B_m W^2}}{1-\lambda_2^\tau}, \qquad (1)$$

where $s^t = \frac{1}{N}\sum_{v\in G}\theta_v^t$, and $\boldsymbol{P} = \boldsymbol{\Pi}\otimes I_d$, $\gamma_m = \max_{v\in G, i\in v}\gamma_{vi}$, and B_m, B, and W are constants determined by the Lyapunov analysis.

Results

Provide vanilla and momentum variants of Gradient Descent iterations

Prove convergence for $\tau \geq 1$ using Lyapunov stability analysis

Conduct experimental validation of the modification on CIFAR-10 with a deep CNN:

- sparse network topology with 5 nodes
- both balanced and imbalanced data

References

Jiang, Z., Balu, A., Hegde, C., and Sarkar, S. (2018).
On consensus-optimality trade-offs in collaborative deep learning.