Decentralized optimization with non-identical sampling in presence of stragglers

Tharindu Adikari, Stark Draper

University of Toronto

ICASSP, May 2020
Setup:
- Decentralized data/computation
- $Q_i$: data distribution of $i$th worker

\[ F_i(w) = \mathbb{E}_{X \sim Q_i}[f(w, X)] \]

- Want $n$ workers to collectively minimize

\[ F(w) = \frac{1}{n} \sum_{i=1}^{n} F_i(w) \]
Setup:

- Decentralized data/computation
- $Q_i$: data distribution of $i$th worker
  
  $F_i(w) = \mathbb{E}_{X \sim Q_i}[f(w, X)]$

- Want $n$ workers to collectively minimize
  
  $F(w) = \frac{1}{n} \sum_{i=1}^{n} F_i(w)$

Assumption 1:

- Non-identical data distributions$^1$
  
  e.g.: MNIST with 10 workers, worker $i$ only has images of digit $i - 1$.

Assumption 2:

- Variable amount of work$^2$
  
  e.g.: Mini-batch size 10 for stragglers (slow workers), 100 for fast workers

---


$^2$ Nuwan Ferdinand et al. “Anytime minibatch: Exploiting stragglers in online distributed optimization”. In: ICLR. New Orleans, 2019
Consensus optimization through random-walk

$W_k, G_k$: $n$-column matrices

\begin{align*}
W_{k+1} &= W_k - \eta G_k \quad \text{(decoupled update)} \\
W_{k+1} &= (W_k - \eta G_k) P \quad \text{(consensus update)}
\end{align*}

\begin{itemize}
  \item $n$ columns for $n$ workers
  \item store \textit{weights} and \textit{gradients} $\nabla F_i$
\end{itemize}
Consensus optimization through random-walk

\[ W_k, G_k: n\text{-column matrices} \]
\[ n \text{ columns for } n \text{ workers} \]
\[ \text{store weights and gradients } \nabla F_i \]

\[ W_{k+1} = W_k - \eta G_k \quad \text{(decoupled update)} \]
\[ W_{k+1} = (W_k - \eta G_k) P \quad \text{(consensus update)} \]

\( j, l, m: \) neighbours of worker \( i \)

\[ \tilde{w}^i \leftarrow \tilde{w}^i P_{ii} + \tilde{w}^j P_{ji} + \tilde{w}^l P_{li} + \tilde{w}^m P_{mi} \]
Consensus optimization through random-walk

$W_k, G_k$: $n$-column matrices

\[
W_{k+1} = W_k - \eta G_k \quad \text{(decoupled update)}
\]

\[
W_{k+1} = (W_k - \eta G_k) P \quad \text{(consensus update)}
\]

$j, l, m$: neighbours of worker $i$  
\[
\tilde{w}^i \leftarrow \tilde{w}^i P_{ii} + \tilde{w}^j P_{ji} + \tilde{w}^l P_{li} + \tilde{w}^m P_{mi}
\]

$P_{i,j} > 0$ only if workers $i, j$ connected  
$P$ - doubly stochastic matrix  
Entries in $[P]^m$ converge to $\frac{1}{n}$ for large $m$  

$W_T = W_0 [P]^{T} - \eta \sum_{k=0}^{T-1} G_k \quad \left[ P \right]^{T-k}$

averaging effect on gradients
Assumption 2: Variable amount of work

- $\bar{g}_i$: $i$th column of $G = \text{avg. gradient}$ of a size $b_i (\geq 1)$ mini-batch
- $Q_i$: data distribution of $i$th worker

\[
\bar{g}_i = \frac{1}{b_i} \sum_{l=1}^{b_i} \nabla_w f(w, X_l); \quad X_l \sim Q_i
\]

In slides, assume all distributions are equally important ($\implies n\gamma_i = 1$ for the $\gamma_i$ discussed in paper).
Assumption 2: Variable amount of work

- $\bar{g}_i$: $i$th column of $G = \text{avg. gradient}$ of a size $b_i \ (\geq 1)$ mini-batch
- $Q_i$: data distribution of $i$th worker

\[
\bar{g}_i = \frac{1}{b_i} \sum_{l=1}^{b_i} \nabla_w f(w, X_l); \quad X_l \sim Q_i
\]

Assumption 2: Workers complete different amounts of work

- $b_i$ i.i.d. across workers and iterations
- $b_i \neq b_j$ in general $\implies$ confidence of $\bar{g}_i$ vary across $i$

\[
W_{k+1} = (W_k - \eta G_k)P \quad \text{(consensus update)}
\]

- Columns of $G_k$ treated equally, irrespective of $b_i$ $\implies$ Equal weighting
- How should we account for the variability in confidences?

In slides, assume all distributions are equally important ( $\implies n\gamma_i = 1$ for the $\gamma_i$ discussed in paper).
Our proposal: Treat confident workers better!

- Give a **higher weight** to confident gradients
- \( V \): diagonal matrix, \( V_{i,i} \propto b_i \)

\[
W_{k+1} = (W_k - \eta V G_k)P
\]

**Concerns:**
- Columns of \( W_{k+1} \) pulled towards confident workers
- Will the oscillatory effect hurt convergence?
Confirming numerically

- Fashion-MNIST dataset: 10 classes
- Multinomial logistic regression
- 1-hidden layer neural network
- 10 workers for each class

Simulate stragglers by sampling $b_i$

$$b_i = \begin{cases} 
60 & \text{with probability 0.8} \\
1 & \text{with probability 0.2}
\end{cases}$$

Simulation results

Cost function:
- Convex: no activation in the hidden layer
- Non-convex: ReLU in the hidden layer

Consensus:
- Approximate: 10 consensus rounds
- Perfect: All entries in $P$ set to $\frac{1}{n}$

Experiments:
- Top: Convex, Perfect consensus
- Middle: Convex, Approx. consensus
- Bottom: Non-convex, Approx. consensus
Theoretical guarantees: Perfect consensus

- $\text{Var}(\nabla_w f(w, X)) \leq \sigma^2$: measures local variance within one worker

- $\nabla F_i = \mathbb{E}_{X \sim Q_i} [\nabla_w f(w, X)]$ and $\nabla F = \frac{1}{n} \sum_{i=1}^{n} F_i(w)$

- $\sum_{i=0}^{n} ||\nabla F_i - \nabla F||^2 \leq n^2 D$: measures global variation among all workers
Theoretical guarantees: Perfect consensus

- $\text{Var}(\nabla_w f(w, X)) \leq \sigma^2$: measures local variance within one worker

- $\nabla F_i = \mathbb{E}_{X \sim Q_i} [\nabla_w f(w, X)]$ and $\nabla F = \frac{1}{n} \sum_{i=1}^{n} F_i(w)$

- $\sum_{i=0}^{n} \| \nabla F_i - \nabla F \|^2 \leq n^2 D$: measures global variation among all workers

Main results:

- Proportional weighting converges!
- Faster than Equal weighting if:

\[
\frac{D}{\sigma^2} \leq \frac{(\mu_2 - n^2 \mu_3)}{(n^4 s^2)}
\]

- $\mu_2 = \mathbb{E}[1/b_i]$
- $\mu_3 = \mathbb{E}[b_i/(\sum_{i=1}^{n} b_i)^2]$
- $s^2 = \text{Var}(b_i/b)$

\[7/9\]
Visualizing the condition

\( g_i = \nabla_w f(w, X) \) for \( X \sim Q_i \)

\( \text{For small } \sigma, \text{ even } b_i = 1 \text{ enough to accurately estimate } \nabla F_i. \)
Conclusions/Next steps

▶ Account for the variability in confidences
▶ Proposed proportional method
▶ Sufficient conditions for faster convergence

Planned work
▶ Proof for approximate consensus.
▶ Generalize to include $b_i = 0$ case.

Thank you.