Communication-Efficient Gradient Descent

Sai Ganesh Nagarajan, Assistant Professor @ IMADA

AI801 Lecture on 11.05.2026

University of
Southern Denmark

Special Topics Outline

- Last week: How to optimize in large-scale distributed systems?
  - Computational/Statistical and Communication trade-offs.
  - Parallel SGD
  - Synchronous SGD
- Three main topics:
  - Online Gradient Descent (April 27th)
  - Distributed Gradient Descent (May 4th)
  - **Communication-Efficient Gradient Descent (May 11th)**
- Today: How to optimize in large-scale distributed systems and communicate less?

Stochastic Gradient Descent in ML

Suppose $x_i \in \mathcal{X}$ is a feature vector and $y_i \in \{-1, 1\}$ is the label such that $\{(x_i, y_i)\}$ are **i.i.d samples** from a fixed (but unknown) distribution. Let there be N data points.

Empirical Loss: $h_N(w) := \frac{1}{N} \sum_{i=0}^{N-1} h(w; (x_i, y_i))$

Noisy Gradient at w_t on sample $j \in [N]$: $\nabla h(w_t; (x_j, y_j))$

Note: $\nabla h(w_t; (x_j, y_j)) \neq \nabla h_N(w_t)$

Parallelized SGD

Initialize with $w_0^{(k)} = w_0$

For $k = 1$ to M in parallel do,

$$w_T^{(k)} = \text{SGD}(h^{(k)}, w_0; \eta^{(k)})$$

Return $\bar{w}_T = \frac{1}{M} \sum_{k=1}^{M} w_T^{(k)}$

Note: Only one round of communication at the end.

Synchronous SGD/Federated SGD

$$w_{t+1} = w_t - \eta_t \sum_{k=1}^{M} \mu_k g_t^{(k)}$$

Gradients here are local and can be from a sample or a mini-batch.

$$h(w) = \sum_{k=1}^{M} \mu_k h^{(k)}(w)$$

Note: Every round involves communication from nodes to a server or between all nodes to synchronize their weights at every step and do the gradient updates.

Convergence Rates of Synchronised SGD

	General (Return average-iterate)	Strongly convex (return last-iterate)
Synch SGD	$\frac{1}{\sqrt{MT}}$	$\frac{1}{\alpha MT}$

Synch SGD converges to a minimizer in the expectation, i.e., $\mathbb{E}[h(\bar{x}_T)] - h(x^*) \leq O(r(T))$, under an appropriate choice of the sequence of step-sizes $\{\eta_t\}$.

Convergence Rates of Synchronised SGD

	General (Return average-iterate)	Strongly convex (return last-iterate)
Synch SGD/FedSGD	$\frac{1}{\sqrt{MT}}$	$\frac{1}{\alpha MT}$

Average gradients from M workers, helps reduces the variance, by the factor leading to improved convergence rates. But there is a clear trade-off between the communication overhead!

SGD converges to a minimizer in the expectation, i.e., $\mathbb{E}[h(\bar{x}_T)] - h(x^*) \leq O(r(T))$, under an appropriate choice of the sequence of step-sizes $\{\eta_t\}$.

Synchronous SGD/Federated SGD

$$w_{t+1} = w_t - \eta_t \sum_{k=1}^{M} \mu_k g_t^{(k)}$$

Gradients here are local and can be from a sample or a mini-batch.

$$h(w) = \sum_{k=1}^{M} \mu_k h^{(k)}(w)$$

Note: How to reduce communication, but still get the same guarantees?

Local SGD

$$w_{t+1}^{(m)} := \begin{cases} w_t^{(m)} - \eta_t \nabla h_{i_t^m}(w_t^{(m)}), & \text{if } t+1 \notin \{H, 2H, \ldots, T\}, \\ \dfrac{1}{M}\displaystyle\sum_{j=1}^{M}\left(w_t^{(j)} - \eta_t \nabla h_{i_t^j}(w_t^{(j)})\right), & \text{if } t+1 \in \{H, 2H, \ldots, T\}. \end{cases}$$

h h h h

.....

$w_t^{(1)}, g_t^{(1)}$ $w_t^{(2)}, g_t^{(2)}$ $w_t^{(3)}, g_t^{(3)}$ $w_t^{(M)}, g_t^{(M)}$

$$h(w) = \frac{1}{N}\sum_{i=1}^{N} h(w; (x_i, y_i)) = \frac{1}{N}\sum_{i=1}^{N} h_i(w)$$

Local SGD Features

$$w_{t+1}^{(m)} := \begin{cases} w_t^{(m)} - \eta_t \nabla h_{i_t^m}(w_t^{(m)}), & \text{if } t+1 \notin \{H, 2H, \ldots, T\}, \\ \frac{1}{M} \sum_{j=1}^{M} \left(w_t^{(j)} - \eta_t \nabla h_{i_t^j}(w_t^{(j)}) \right), & \text{if } t+1 \in \{H, 2H, \ldots, T\}. \end{cases}$$

H=1 — $H = O\left(\sqrt{T/(Mb)}\right)$ — H=T

Synch SGD | Local SGD | Parallel SGD

$$\mathbb{E}[h(\bar{x}_T)] - h(x^*) \leq \text{OptError} + \text{NoiseError} + \text{DriftError}$$

Note: H is set so that no one term dominates the others and balances all 3 sources of errors

Local SGD Features

$$w_{t+1}^{(m)} := \begin{cases} w_t^{(m)} - \eta_t \nabla h_{i_t^m}(w_t^{(m)}), & \text{if } t+1 \notin \{H, 2H, \ldots, T\}, \\ \dfrac{1}{M}\displaystyle\sum_{j=1}^{M}\left(w_t^{(j)} - \eta_t \nabla h_{i_t^j}(w_t^{(j)})\right), & \text{if } t+1 \in \{H, 2H, \ldots, T\}. \end{cases}$$

H=1 — $H = O\left(\sqrt{T/(Mb)}\right)$ — H=T

Synch SGD | Local SGD | Parallel SGD

Note: b is the batch-size for SGD. b=1 implies standard SGD with one sample.

The number of rounds of communication is T/H and this increases with increase in b or M!!

Why?

Convergence Rates of Local SGD

	General (Return average-iterate)	Strongly convex (return last-iterate)
Local SGD	$\frac{1}{\sqrt{MbT}}$	$\frac{1}{\alpha MbT}$

Local-SGD converges to a minimizer in the expectation, i.e., $\mathbb{E}[h(\bar{x}_T)] - h(x^*) \leq O(r(T))$, under an appropriate choice of the sequence of step-sizes $\{\eta_t\}$, with reduced communication!

FedAvg and Fed Prox

$h^{(1)}$ $h^{(2)}$ $h^{(3)}$ ..... $h^{(M)}$

$w_t^{(1)}, g_t^{(1)}$ $w_t^{(2)}, g_t^{(2)}$ $w_t^{(3)}, g_t^{(3)}$ $w_t^{(M)}, g_t^{(M)}$

$$h(w) = \sum_{m=1}^{M} p_m h^{(m)}(w)$$

FedAvg and FedProx

$h^{(1)}$ $h^{(2)}$ $h^{(3)}$ $h^{(M)}$

Key Issue: The clients/workers here can drift away, due to heterogeneity and does not necessarily satisfy i.i.d assumptions.

$w_t^{(1)}, g_t^{(1)}$ $w_t^{(2)}, g_t^{(2)}$ $w_t^{(3)}, g_t^{(3)}$ $w_t^{(M)}, g_t^{(M)}$

$$h(w) = \sum_{m=1}^{M} p_m h^{(m)}(w)$$

Algorithm

FedProx. $h(w) = \sum_{m=1}^{M} p_m h^m(w)\,.$

At round r, the server selects $S_r \subseteq \{1,\dots,M\}$ and sends w^r to $m \in S_r$.

Each selected client approximately solves

$$w_m^{r+1} \approx \arg\min_w \left\{ h^m(w) + \frac{\mu}{2}\|w - w^r\|^2 \right\}.$$

This can be implemented by H local SGD steps:

$$w_m^{r,0} = w^r,$$

$$w_m^{r,s+1} = w_m^{r,s} - \eta\left(g_m^{r,s} + \mu(w_m^{r,s} - w^r)\right), \quad s = 0,\dots,H-1.$$

The server aggregates

$$w^{r+1} = \sum_{m\in S_r} \frac{p_m}{\sum_{j\in S_r} p_j} w_m^{r,H}\,.$$

FedProx and FedAvg Features

- When $\mu = 0$, FedProx reduces to FedAvg.
- FedAvg converges under bounded-heterogeneity to the same rates as SGD for convex and strongly convex functions.
- FedProx converges to a stationary point, under a similar bounded gradient dissimilarity condition and converges in roughly $O(1/R)$ steps.

$$\sum_{m=1}^{M} p_m \left\| \nabla h^{(m)}(w) - \nabla h(w) \right\|^2 \leq \zeta^2, \qquad \forall w \,.$$

Summary

- We saw new algorithms for reducing communication.
- We looked at cases with homogeneous and heterogeneous clients and problems associated with that.
- More complex trade-offs between stochasticity and drift due to other reasons (heterogeneity, non-IID ness etc.)
- After the break: Assignment discussion.