SLowcalSGD: Slow Query Points Improve Local-SGD for Stochastic Convex Optimization

Tehila Dahan and Kfir Y. Levy
Technion - Israel Institute of Technology

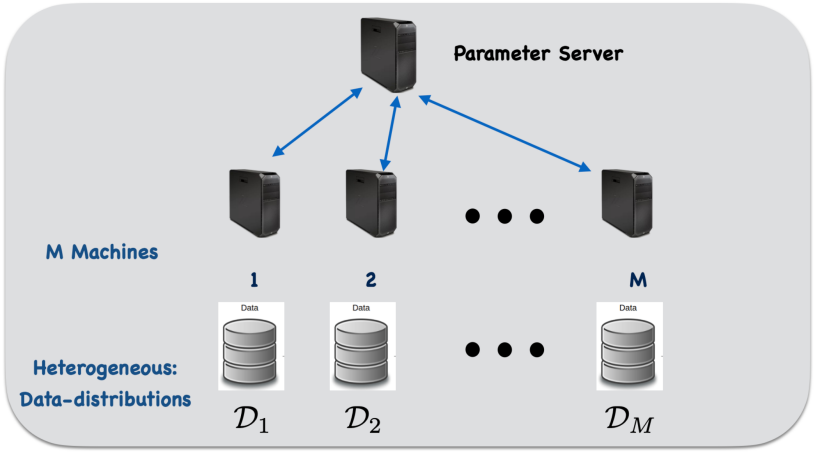
Objective

Goal: find solution **x** that minimizes the following

$$f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_{z^i\sim\mathcal{D}_i} f_i(x, z^i)$$

Distributed Training

- R := #Communication rounds with Parameter Server
- K := #Stochastic Gradient computations per-machine per round
- M := #Machines
- G* := (heterogeneity) dissimilarity measure



Our Approach

Local-SGD degrades <- **Bias** between different machines,

$$\|w_t^{(i)} - w_t^{(j)}\|^2$$

Our idea: rely on **AnyTime-SGD** – a baseline Opt. Approach that uses **slowly changing query points.**

We develop a **local update variant** of AnyTime-SGD:
- Each Machine (i), maintains **slowly changing** weights: $x_t^{(i)}$
- This ensures **low bias,**

$$\|x_t^{(i)} - x_t^{(j)}\|^2 \ll \|w_t^{(i)} - w_t^{(j)}\|^2$$

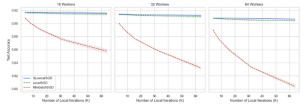
At Round $r \in [R]$

01: WORKER

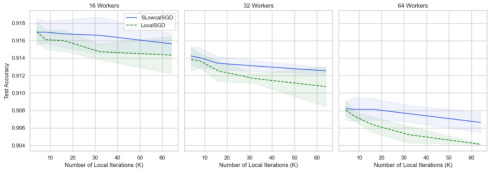
for $k = 0, \ldots, K-1$ **do**
  Set $t = rK + k$
  for each machine $i \in [M]$ **do**
    Draw a fresh sample $z_t^i \sim \mathcal{D}_i$
    Compute the gradient $g_t^i = \nabla f_i(x_t^i, z_t^i)$
  end for
  for each machine $i \in [M]$ **do**
    Update parameters:
    $w_{t+1}^i = w_t^i - \eta\alpha_t g_t^i$
    $x_{t+1}^i = \left(1 - \frac{\alpha_{t+1}}{\alpha_{0:t+1}}\right) x_t^i + \frac{\alpha_{t+1}}{\alpha_{0:t+1}} w_{t+1}^i$
  end for
end for

02: SERVER

After K **Local Steps:**
$t := r(K+1) - 1$
Update global parameters:
$w_{t+1} = \frac{1}{M}\sum_{i\in[M]} w_{t+1}^i$
$x_{t+1} = \frac{1}{M}\sum_{i\in[M]} x_{t+1}^i$



MNIST



Guarantees: Minimal # Communication Rounds

Method	Rate	$R_{\min}$ $(\sigma = 1)$
MiniBatch SGD [10]	$\frac{1}{R} + \frac{\sigma}{\sqrt{MKR}}$	MK
Accelerated MiniBatch SGD [10, 24]	$\frac{1}{R^2} + \frac{\sigma}{\sqrt{MKR}}$	$(MK)^{1/3}$
Local SGD [36]	$\frac{G^{2/3}}{R^{2/3}} + \frac{\sigma^{2/3}}{(\sqrt{KR})^{2/3}} + \frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}}$	$G^4 \cdot (MK)^3 + M^3K$
SCAFFOLD [20]	$\frac{1}{R} + \frac{\sigma}{\sqrt{MKR}}$	MK
SLowcal-SGD (This paper)	$\frac{\sigma^{1/2}+G_*^{1/2}}{K^{1/4}R} + \frac{1}{KR} + \frac{1}{K^{1/3}R^{4/3}} + \frac{1}{R^2} + \frac{\sigma}{\sqrt{MKR}}$	$G_* \cdot MK^{1/2} + MK^{1/2}$
Lower Bound: Local-SGD [38]	$\frac{G_*^{2/3}}{R^{2/3}} + \frac{\sigma_*^{2/3}}{(\sqrt{KR})^{2/3}} + \frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}}$	$G_*^4 \cdot (MK)^3 + M^3K$

Rmin to allow linear Speedup (lower is better)

sensitive to data heterogeneity

✓ **LOW COMMUNICATION COST**

✓ **ENHANCES PERFORMANCE UNDER DATA HETEROGENEITY**