LOCK-FREE PARALLEL SGD ON DENSE DATA

{ HUAYU ZHANG, FAN GAO }   UNIVERSITY OF WISCONSIN-MADISON

INTRODUCTION

- We conduct an initial study on lock-free algorithms for dense data.
- Data is partitioned randomly and based on correlation.
- Implementation of the hogwild algorithm and a basic evaluation platform.
- Theoretical proof of the convergence

HOGWILD

Algorithm 1: Hogwild

Input: $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$: data, P: number of cores, T: number of iterations, γ: learning rate
Output: $\boldsymbol{w}$: model parameters (shared)
Data: $(D_1, D_2, \ldots, D_P)$: partitioned data

Initialize $\boldsymbol{w}$;
$(D_1, D_2, \ldots, D_P) \leftarrow$ `partition` (D);
for $i = 1$ **to** P **do**
  `create_thread` (`serialSGD`, $\boldsymbol{w}$, $D_i, \gamma, T/P$);
end
`wait all threads`;
return $\boldsymbol{w}$;

Algorithm 2: Serial SGD

Input: $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$: data, T: number of iterations, γ: learning rate, $\boldsymbol{w}$: model parameters
Data: $\boldsymbol{g}$: gradient, s_i: uniformly sampled from $[n]$

for $i = 1$ **to** n **do**
  $s_i \leftarrow$ `uniformly_sample` $([n])$;
  $\boldsymbol{g} \leftarrow$ `compute_grad` $(\boldsymbol{w}, \boldsymbol{x}_{s_i}, y_{s_i})$;
  $\boldsymbol{w} \leftarrow \boldsymbol{w} - \gamma \boldsymbol{g}$;
end

DATA PARTITION

Partition $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ to k balanced subgroups $\{D_1, D_2, \ldots, D_k\}$.

- **Random partition**
  Randomly shuffle the indexes and assign $i \in [n]$ to group $D_j, j = \mod(i, k)$.
- **Correlation-based partition**
  - Correlation graph $G = (V, E), V = [n], E = \{e_{ij} =< \boldsymbol{x}_i, \boldsymbol{x}_j >: i, j \in [n], i < j\}$
  - Choose the partition by maximizing the intra-group correlation and minimizing the inter-group correlation.
  - Greedy algorithm: pick the vertex i and group p with minimum $\sum_{j=0, j\neq p}^{k} \sum_{l \in D_j} e_{il} - \sum_{l \in D_p} e_{il}$. Add i to D_p. Complexity: $O(k \mid E \mid + \mid V \mid^2)$.

EXPERIMENTS

- **Simulation** Model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$. $\boldsymbol{x}_i \in \mathcal{N}(0, \boldsymbol{I})$. Given $(\boldsymbol{X}, \boldsymbol{y})$, estimate $\hat{\boldsymbol{w}}$. $\boldsymbol{X} \in \mathbb{R}^{5000 \times 200}$

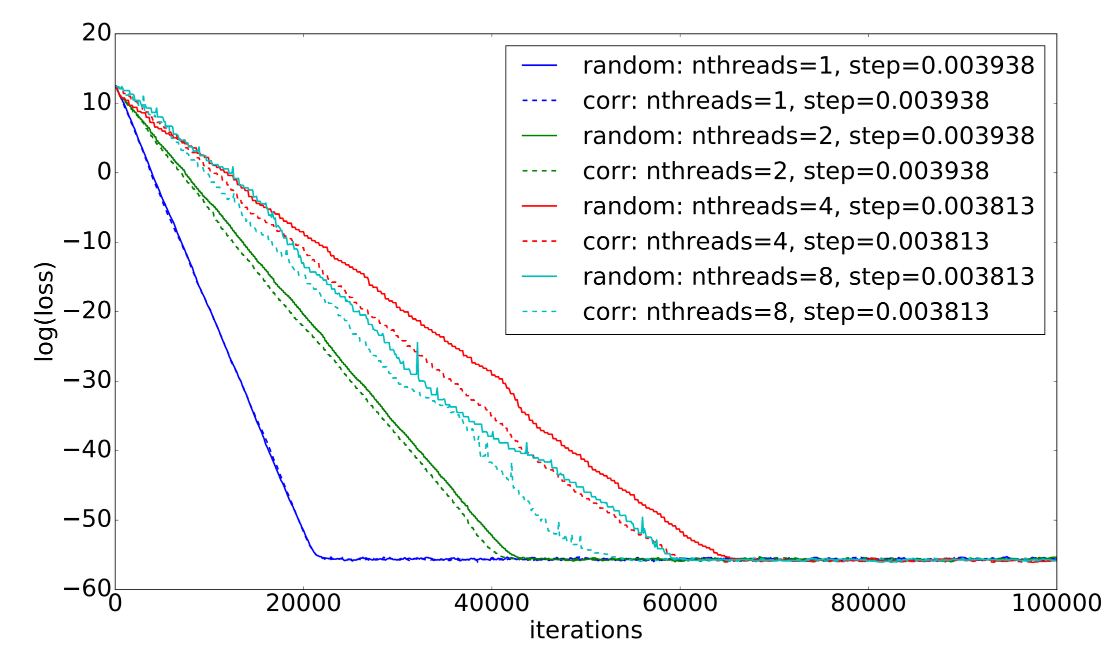

Figure 1: Loss vs. iterations

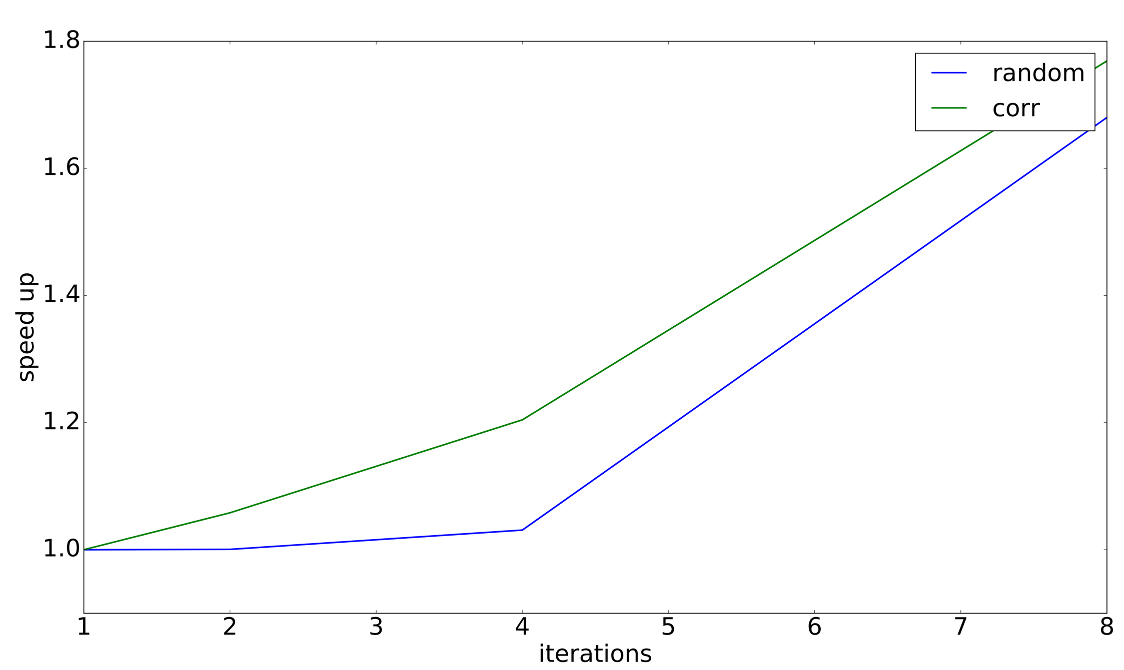

Figure 2: Speed up. The time elapsed when reaching $\frac{\epsilon_i}{\epsilon_0} \leq 10^{-20}$

- **MNIST** Hand write digit classification

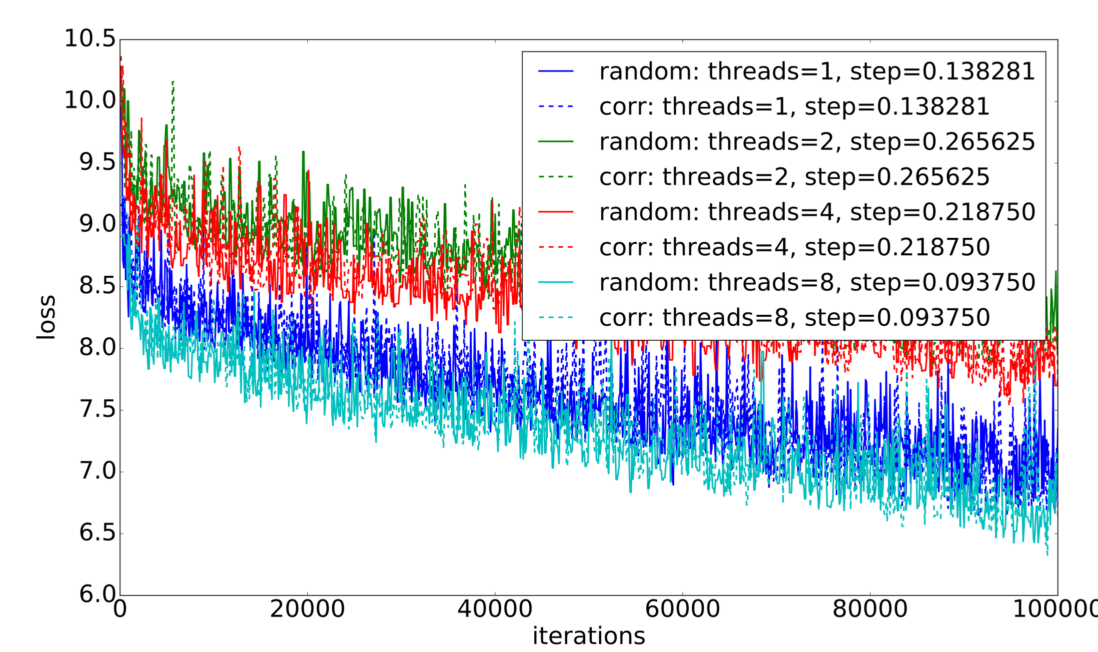

Figure 3: Loss vs. iterations

Threads	2	4	8
intra	38.01	41.32	44.20
inter	31.57	32.62	33.45

Table 1: Average inter and intra correlation after partition

Threads	1	2	4	8
random	0.858	0.872	0.878	0.891
correlation	0.873	0.860	0.842	0.865

Table 2: Classification accuracy

CONVERGENCE

Assumptions in the main theorem,

- f is bounded by C.
- f is m strongly convex.
- Uniform bound on stochastic gradient assumption:

$$\mathbb{E}\|\nabla f_s(w)\|^2 \leq M^2$$

- Low inter-group correlation assumption: in each partitioned data, $\langle x_i, x_j \rangle \leq \delta$.

Theorem. If the number of samples that overlap in time with a single sample during the execution of our algorithm is bounded as

$$\tau = O\left(M^2 \cdot \min\left\{\frac{1}{\epsilon m^2}, \frac{1}{\delta C^2}\right\}\right),$$

our algorithm with the step size $\gamma = \epsilon m / M^2$, after

$$T = O\left(\frac{M^2 \log(\Delta_1/\epsilon)}{\epsilon m^2}\right)$$

iterations, obtains $\mathbb{E}(\Delta_{T+1}) \leq \epsilon$, where Δ_k denotes the distance between the k-th iterate and the optimum i.e. $\Delta_k = \|w_k - w^*\|^2$.

CONCLUSION

- Hogwild converges even on dense data.
- The convergence rate of parallel SGD is slower than that of serial SGD, but a slight gain in speedup is achievable.
- The partition algorithm based on correlation does not accelerate the convergence significantly.