Beyond spectral gap:
The role of the topology in decentralized learning

Sergei Ovcharov, Anton Korznikov, Georgii Budnik

October 2023

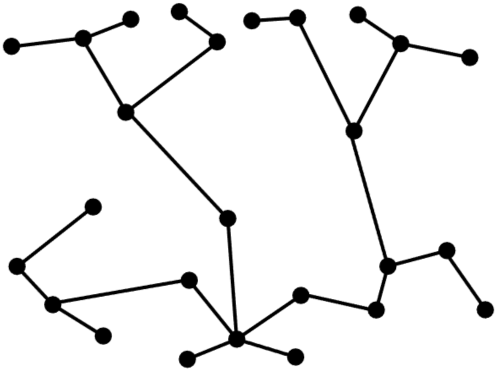
Decentralized learning

The group of nodes is not coordinated by any centralized server. Each node locally holds f_i and exchanges information only with its immediate neighbors.

$$\min_{x_1,\dots,x_m \in \mathbb{R}^d} \sum_{i=1}^{m} f_i(x_i)$$
$$\text{s.t. } x_1 = \dots = x_m.$$

The optimal point in the decentralized sense should be consensual and optimal, i.e.

$$x_1 = \dots = x_m = x^* = \arg\min_{x \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$



Decentralized Stochastic Gradient Descent (D-SGD)

Algorithm:

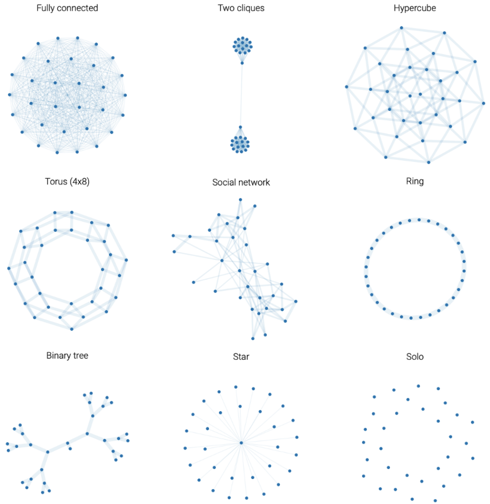
Choose step-size $\alpha > 0$ and pick any $x_i^{(0)} \in \mathbb{R}^n$;

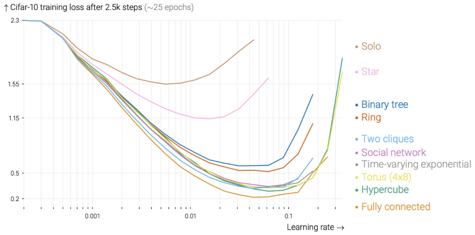
for $k = 0, 1, \ldots$ **do**

$$x_i^{(k+1)} = \sum_{j=1}^n w_{ij} x_j^{(k)} - \alpha \nabla f_i(x_i^{(k)}), \quad i = 1, 2, \ldots, m;$$

end

Article overview





Beyond spectral gap:
The role of the topology in decentralized learning

Thijs Vogels* EPFL

Hadrien Hendrikx* EPFL

Martin Jaggi EPFL

Abstract

In data-parallel optimization of machine learning models, workers collaborate

Spectral Gap vs Effective number of neighbors

The effective number of neighbors measures the ratio of the asymptotic variance of the processes:

$$n_{W}(\gamma) = lim_{t\to\infty} \frac{\sum_{i=1}^{n} Var[\mathbf{y}_i^{(t)}]}{\sum_{i=1}^{n} Var[\mathbf{z}_i^{(t)}]}$$

, where

$$\mathbf{y}^{(t+1)} = \sqrt{\gamma} * y^{(t)} + \xi^{(t)},\ y^{(t)} \in \mathbb{R},\ \xi^{(t)} \sim \mathcal{N}^n(0,1)$$
$$\mathbf{z}^{(t+1)} = W\left(\sqrt{\gamma} * z^{(t)} + \xi^{(t)}\right),\ z^{(t)} \in \mathbb{R},\ \xi^{(t)} \sim \mathcal{N}^n(0,1)$$

We call **y** and **z** **random walks** because workers repeatedly add noise to their state, somewhat like SGD's parameter updates. This should not be confused with a 'random walk' over nodes in the graph.

Paper's results

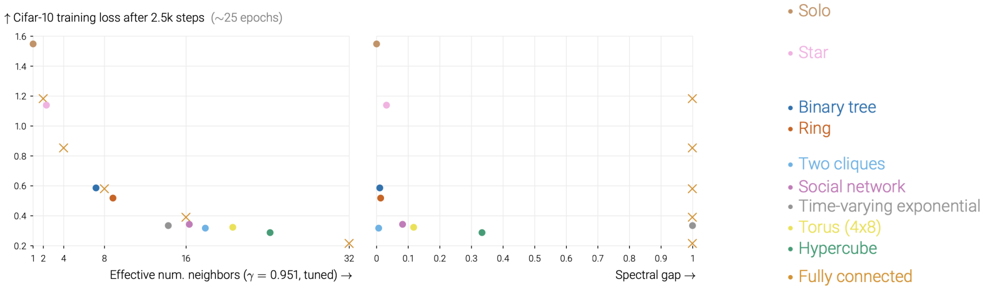


Figure: Cifar-10 training loss after 2.5k steps for all studied topologies with their optimal learning rates.

Our Reproduction of the Experiment

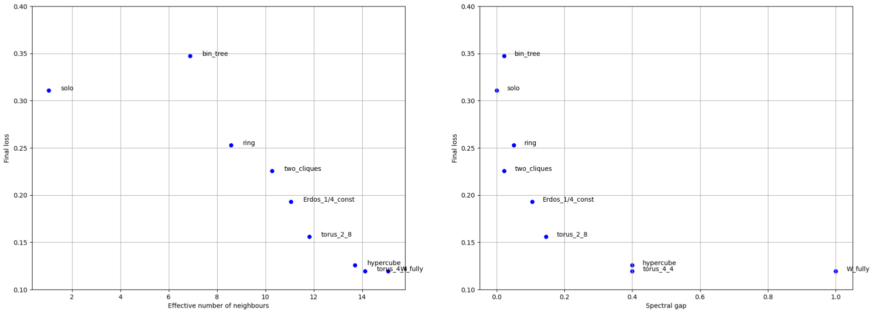


Figure: MNIST training loss after 200 steps for studied **static** topologies with their optimal learning rates.

Time Varying Topologies

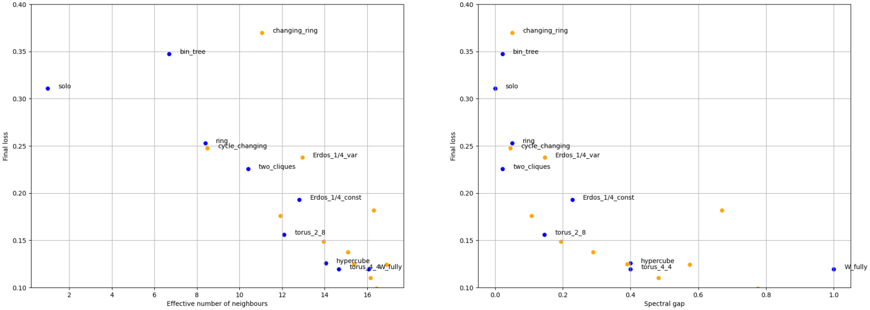


Figure: MNIST training loss after 200 steps for **all** studied topologies with their optimal learning rates.

Regular and Irregular Graphs

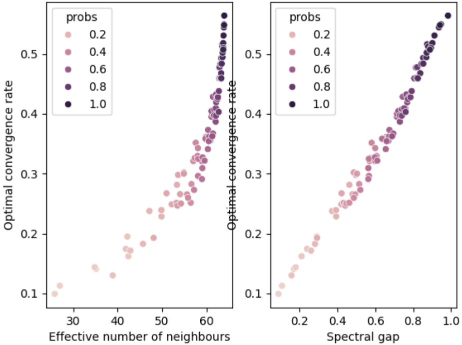


Figure: Optimal convergence rate dependence from ENN and SG for **Erdös-Rényi** random graphs.

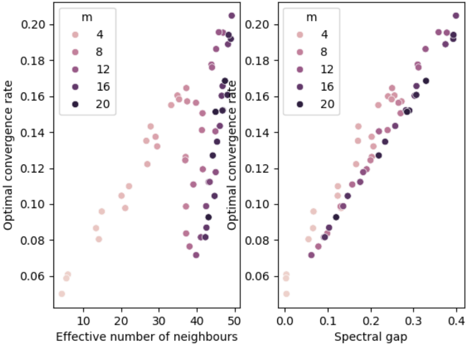
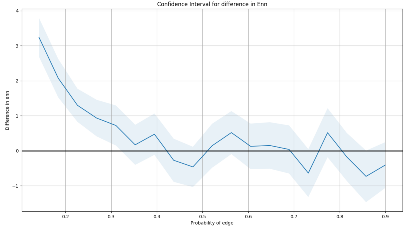
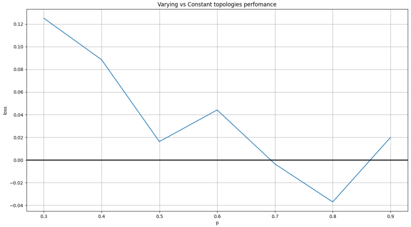


Figure: Optimal convergence rate dependence from ENN and SG for **Barabási-Albert** random graphs.

ENN for random graphs





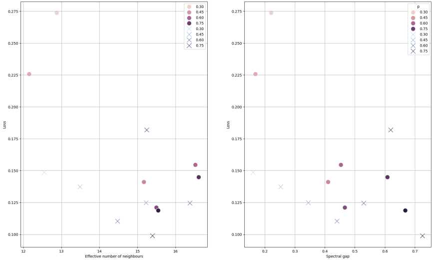


Figure: Time Varying vs Constant random topologies.

Maximization of ENN for a Fixed Graph

G is a fixed graph, W is its weights.

$$\max_W n_W(\gamma) := \frac{\frac{1}{1-\gamma}}{\sum_i \frac{\lambda_i^2}{1-\gamma\lambda_i^2}}$$

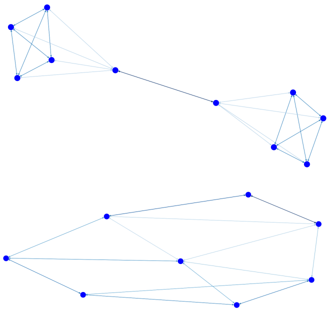
$$\max_W n_W(\gamma) \textit{ is equiv to } \min_W Tr\left(I - \gamma W^2\right)^{-1}$$

$$\text{s.t.}\, W^T = W,\ W\mathbb{1} = \mathbb{1},\ 0 \le W \le G \qquad c$$

$$\min_{W,X,Y} Tr\ X$$

$$\text{s.t.}\, W^T = W,\ W\mathbb{1} = \mathbb{1},\ 0 \le W \le G$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \succeq 0, \quad \begin{bmatrix} I - Y & W \\ W & \frac{1}{\gamma} I \end{bmatrix} \succeq 0$$



DIGing Algorithm

Algorithm:

Choose step-size $\alpha > 0$ and pick any $x^{(0)} \in \mathbb{R}^{n \times p}$;

Initialize $y^{(0)} = \nabla f(x^{(0)})$;

for $k = 0, 1, \ldots$ **do**

$$x^{(k+1)} = W^{(k)} x^{(k)} - \alpha y^{(k)};$$

$$y^{(k+1)} = W^{(k)} y^{(k)} + \nabla f(x^{(k+1)}) - \nabla f(x^{(k)});$$

end

DIGing with static topologies

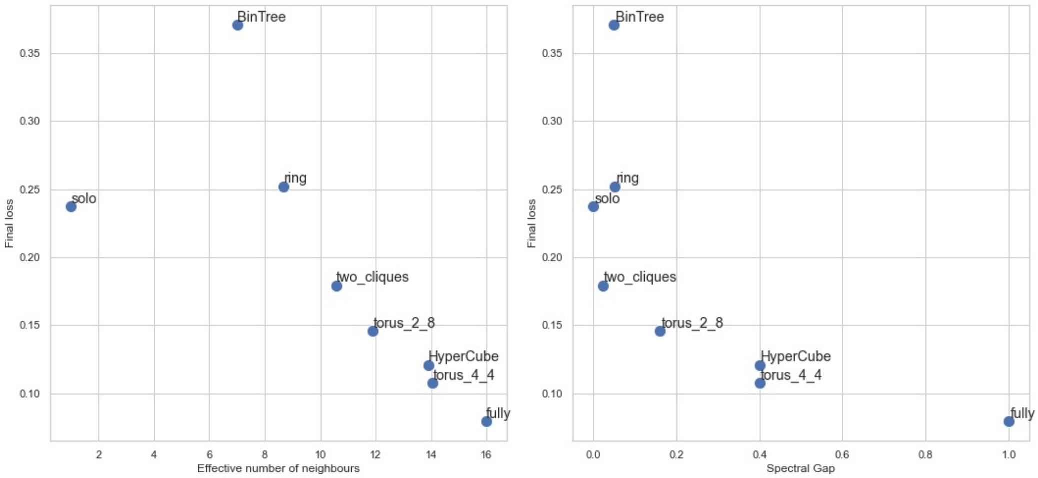


Figure: MNIST training loss after 200 steps for **all** studied topologies with their optimal learning rates.

DIGing with static and varying topologies

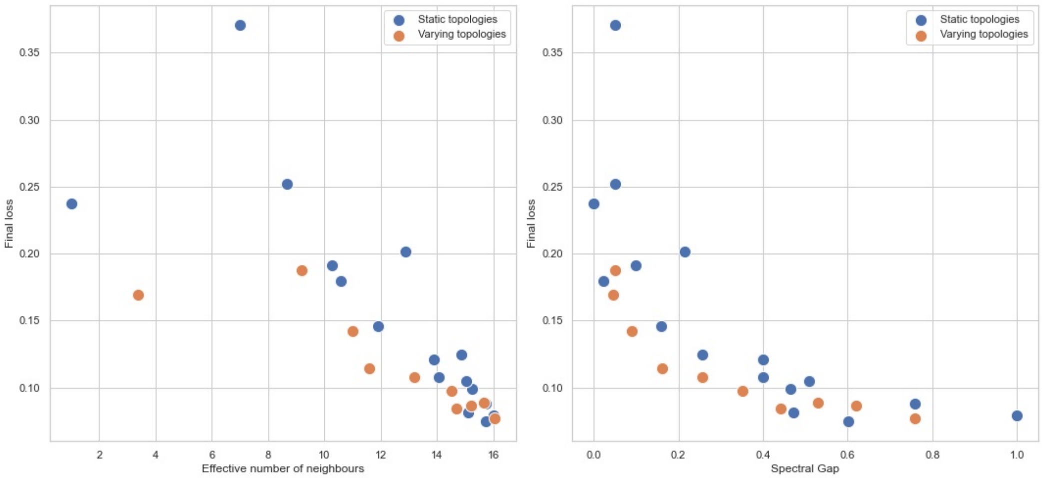


Figure: MNIST training loss after 200 steps for **all** studied topologies with their optimal learning rates.

DIGing with Erdos topologies

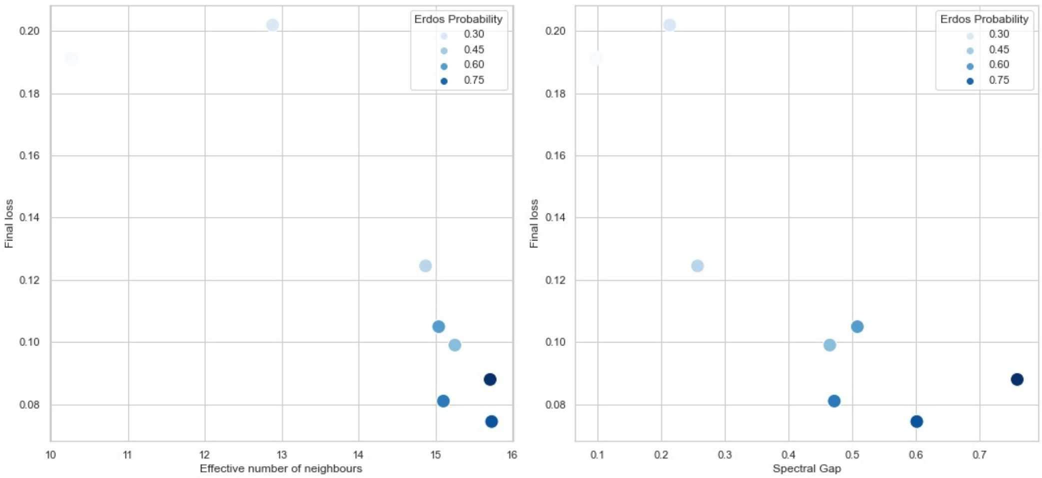


Figure: MNIST training loss after 200 steps for all studied topologies with their optimal learning rates.

DIGing with Varying Erdos topologies

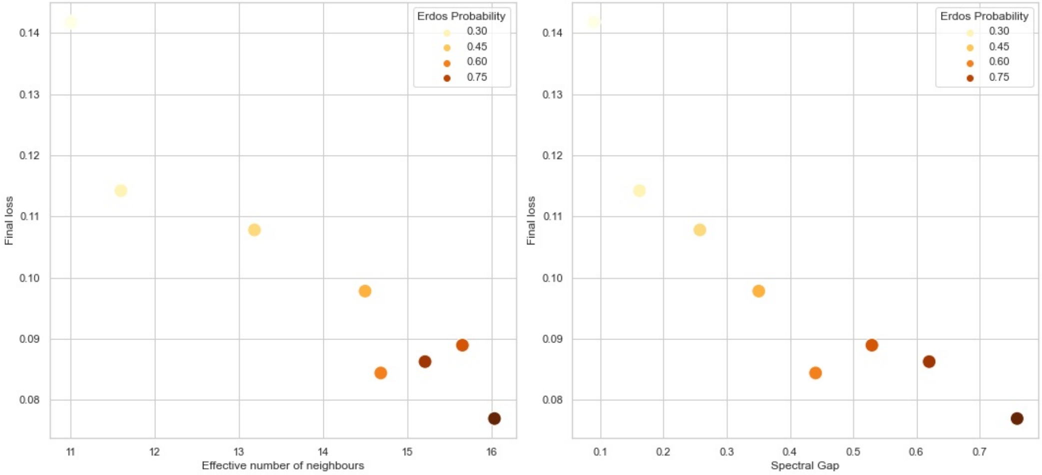


Figure: MNIST training loss after 200 steps for all studied topologies with their optimal learning rates.

Comparison of DSGD and DIGing algorithms

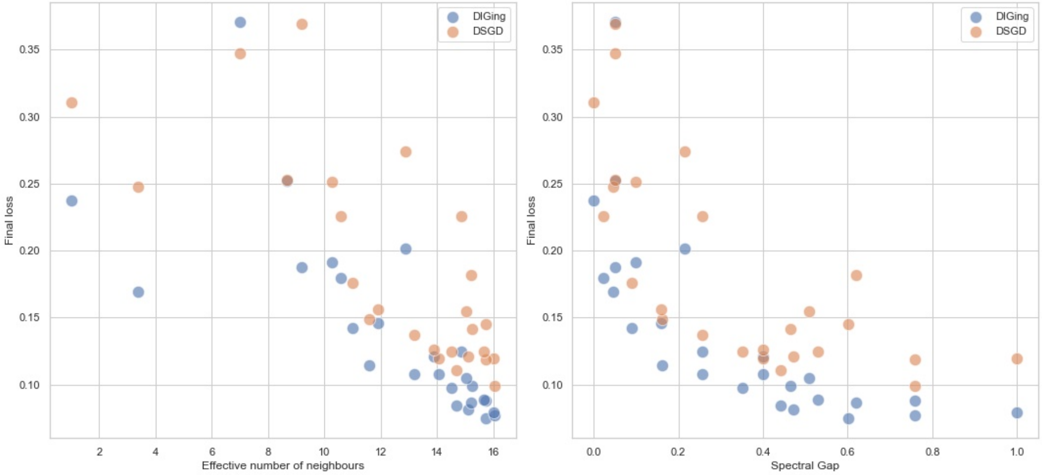


Figure: MNIST training loss after 200 steps for **all** studied topologies with their optimal learning rates.

Thank you for your attention!